Chapter 1 Introduction to Bayesian Analyses

1.1 Introduction

Objectives

- Introduce the basic concepts of Bayesian analysis.
- Illustrate the differences between Bayesian analysis and classical statistics.
- Illustrate the Metropolis sampling algorithm.
- Introduce the Markov chain diagnostic statistics.
- Explain the advantages and disadvantages of Bayesian analysis.

3

What Is Bayesian Analysis?

- *Bayesian analysis* is a field of statistics that is based on the notion of conditional probability.
- It can be viewed as the formalization of the process of incorporating scientific knowledge using probabilistic tools.
- It provides uncertainty quantification of parameters by its conditional distribution in the light of available data.

4

Statistical inference is an inductive process for making inferences about parameters of interest. Bayesian analysis uses conditional probability to quantify the uncertainty of parameters of interest. In general, Bayesian statistical methods start with a prior distribution for all unknown parameters, updates this prior distribution in the light of the data (for example, using likelihood) to construct the posterior distribution, and then uses the posterior distribution for inferential decisions.

Bayes' Theorem

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

- P(A) is the prior probability of event A. It is called the *prior* because it does not take into account any information about event B.
- P(B|A) is the conditional probability of event B given event A.
- P(B) is the prior or marginal probability of event B.
- P(A|B) is the conditional probability of event A given event B. It is called the posterior probability because it is derived from the specified value of event B.

5

The term *Bayesian* comes from the prevalent usage of Bayes' theorem, which was named after the Reverend Thomas Bayes, an eighteenth-century Presbyterian minister. Bayes was interested in solving the question of inverse probability: after observing a collection of events, what is the probability of one event? Therefore, the theorem relates the conditional and marginal probabilities of two random events and it is often used to compute posterior probabilities given observations. In Bayesian analysis, the theorem describes the way in which one's beliefs about observing A is updated by having observed B.

Bayesian Analysis

- The Bayesian approach to statistical inference treats parameters as random variables.
- It includes the incorporation of prior knowledge and its uncertainty in making inferences on unknown quantities (model parameters, missing data, and so on).
- It expresses the uncertainty concerning the parameter through probability statements and distributions.

6

Bayesian methods offer alternatives to classical statistical inference. Instead of treating parameters as fixed constants, Bayesian methods treat parameters as random variables. These parameters cannot be determined exactly, and uncertainty about the parameter is expressed through probability statements and distributions. Bayesian inference about the parameters is based on the probability distribution for the parameter.

Bayes' theorem provides a tool to update uncertainty in the light of observed data. For example, suppose you want to estimate the incidence rate of a disease using a statistical model. The Bayes approach would suggest that you start with prior information (collected from experts, clinicians, Centers for Disease Control database, and so on) about that disease incidence rate and after gathering the data (binomial counts) you update the incidence rate using the posterior distribution.

Frequentist Approach to Statistics

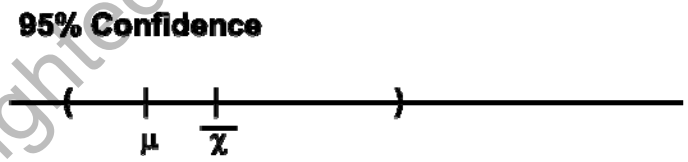
- *Classical methods* consider the parameters to be fixed but unknown.
- They do not enable you to make probability statements about parameters because they are fixed.
- They are based on probabilities that are only for observations given the unknown parameters.
- They are judged by how they perform in an infinite number of hypothetical repetitions of the experiments.

8

sas

In classical statistical inference, the probability of an event is defined as the proportion of times the event will occur in an infinitely long series of repeated identical situations. This is known as the *frequentist* perspective, as it rests on the frequency with which specific events occur. In the frequentist approach, the parameters are fixed, unknown constants and you cannot make any probabilistic statements about them.

Confidence Intervals – Classical Approach

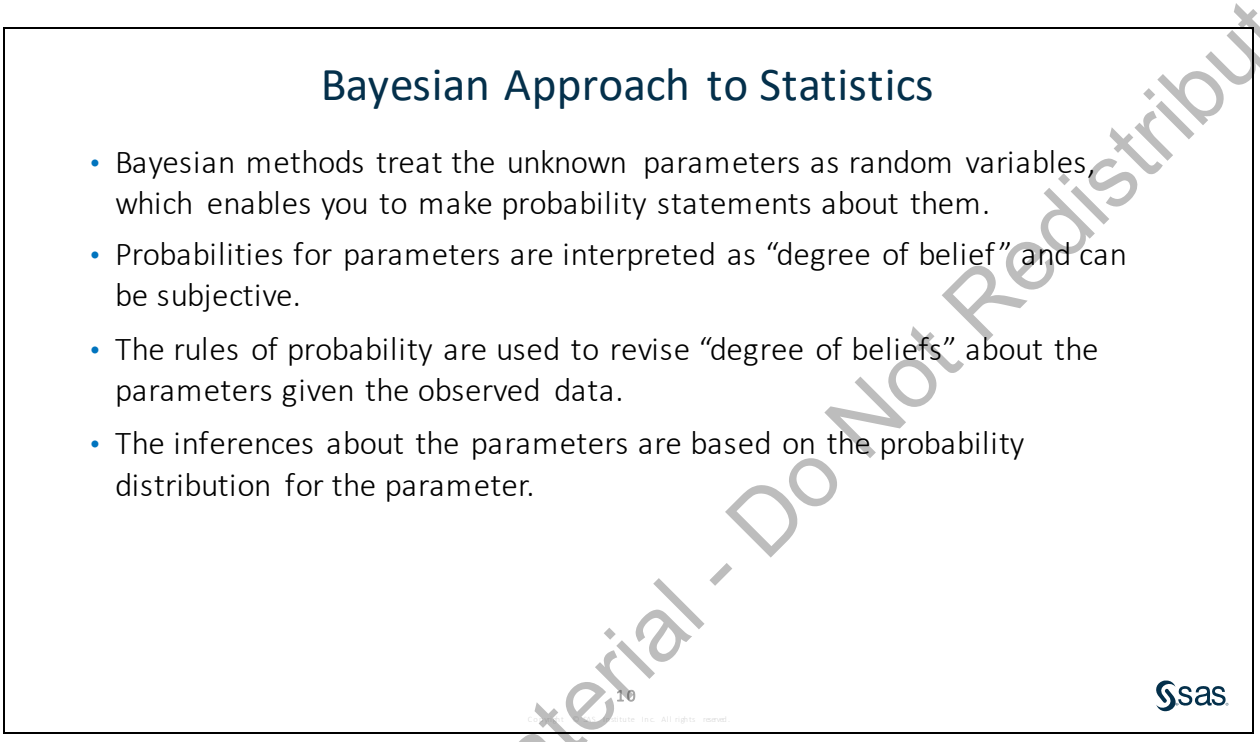


- A 95% confidence interval states that you are 95% confident that random interval contains the true mean.
  - In other words, if 100 different samples were drawn from the same population and 100 intervals were calculated, approximately 95 of them would contain the population mean.

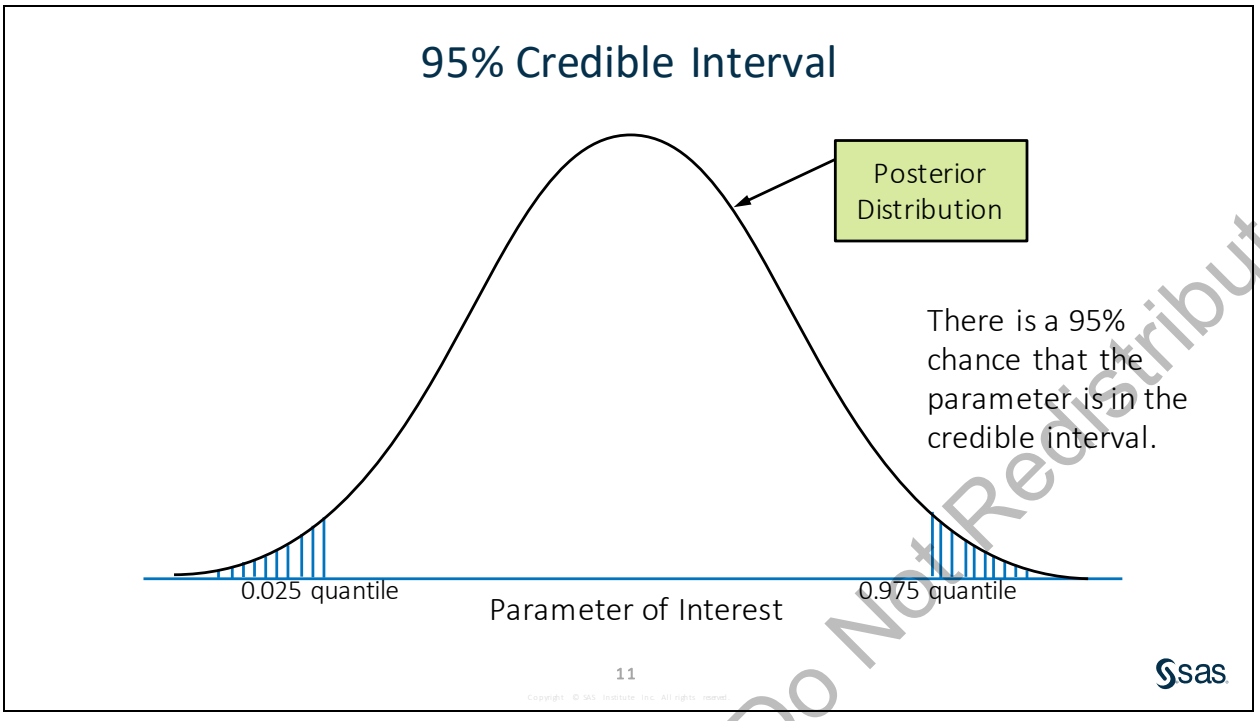
9

sas

An example of the frequentist approach is the confidence interval. A 95% confidence interval in classical statistical inference states that in the long run 95% of the realized confidence intervals cover the true parameter. You cannot say “The true parameter is in the confidence interval with a 95% probability.” The true parameter is either inside or outside the confidence interval, not with any measurable probability. This classical interpretation reflects the uncertainty in the sampling procedure because the parameter is fixed but the interval is random.



In Bayesian inference, the parameters are random variables and you can make probability statements about them.



Interval estimates in Bayesian statistics are called *credible intervals*. Because the parameter is random, you can make the claim that the parameter is inside the credible intervals with measurable probability.

In other words, you can state that there is a 95% chance that the parameter is in this credible interval (SAS Institute, Inc. 2010).

Steps Involved in Bayesian Inference

1. The probability distribution of the parameter, known as the *prior distribution*, is formulated.
2. Given the observed data, you choose a statistical model (referred to as the likelihood) that describes the distribution of the data given the parameters.
3. You update your beliefs about the parameter by combining information from the prior distribution and the data through the calculation of the posterior distribution. This is carried out by using Bayes' theorem; hence the term Bayesian analysis.

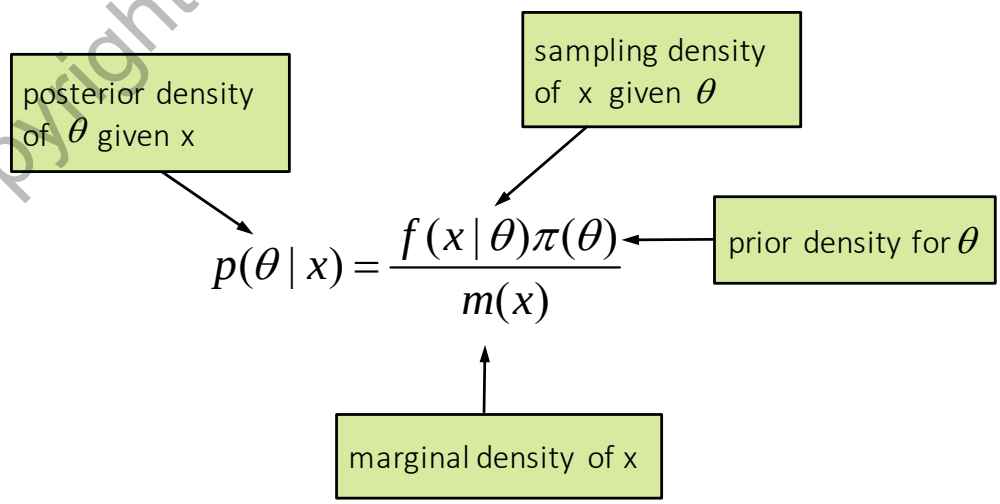
12

sas

The prior distribution expresses your current beliefs, for example, on the location, scale, and shape of the parameter distribution. It is the probability distribution that is simply intended to summarize reasonable uncertainty given evidence external to the study in question.

You can think of step 2 as formulating the likelihood function. In step 3, appropriately combining the prior distribution and the likelihood function using Bayes Rule leads to the posterior distribution of the parameter. You use the posterior distribution to carry out all inferences.

The Bayes' Rule



13

sas

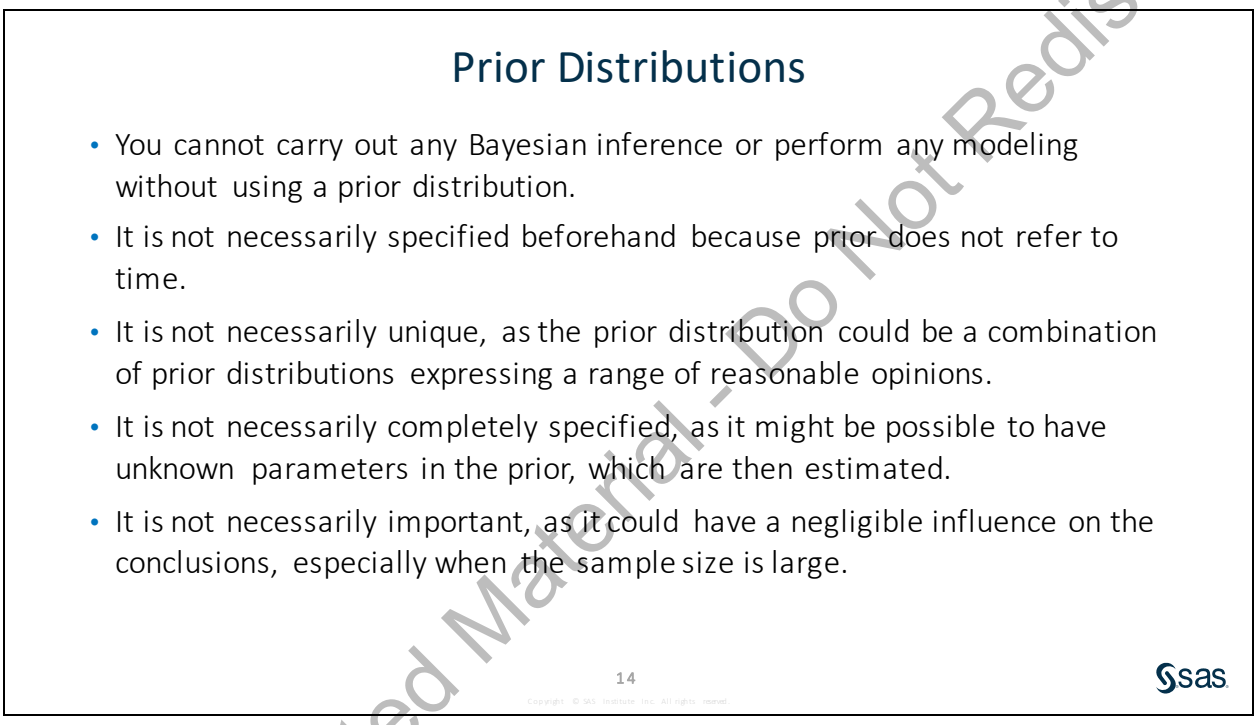
The above equation is often verbalized as

posterior density=(likelihood*prior)/marginal likelihood

The marginal density of x is an integral defined as

$$\int f(x \mid \theta)\pi(\theta)d\theta$$

The posterior density or distribution describes the distribution of the parameter of interest with respect to the data and prior. The posterior distribution is necessary for probabilistic prediction and for sequential updating.



Although the name *prior* suggests a temporal relationship, it is feasible for a prior distribution to be decided after seeing the results of the study (for example, empirical Bayes methods). Cox (1999) points out that the prior distribution refers to a situation where you assess what the evidence would have been if you had no data. This assessment can be made after seeing the data, but there are issues in this.

There is no such thing as the 'correct' prior. In fact, researchers such as Kass and Greenhouse (1989) have suggested using a 'community of priors' to describe the range of reasonable opinions.

Even though Bayesian analysis is driven by the prior distribution, it is sometimes not important in the analysis. As the sample size increases, the prior will usually be overwhelmed by the likelihood and will exert a negligible influence on the conclusions. However, Bayesian analysis is not based on this assumption.

Noninformative Priors

- A prior distribution is *noninformative* if it is flat, relative to the likelihood function.
- It will have a minimal impact on the posterior distribution.
- In the SAS Bayesian procedures, noninformative priors are flat priors, which assign equal likelihood to all possible values of the parameter.

15

sas

A *common prior distribution* is the uniform distribution, which is a flat prior. A *flat prior* is a prior distribution that assigns equal likelihood on all possible values of the parameter. A flat prior is usually a noninformative prior, which means it has minimal impact on the posterior distribution of the parameter. Many statisticians favor noninformative priors because they appear to be more objective, but noninformative priors can lead to improper posteriors (it is no longer a probability distribution). 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.

Informative Priors

- An *informative prior* is a prior that is not strongly dominated by the likelihood and might have an impact on the posterior distribution.
- The information can be obtained from the elicitation of expert opinion or the derivation from historical data.
- It is recommended that you conduct a sensitivity analysis to assess the impact of a particular prior distribution on the conclusions of the analysis.

16

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. The informative prior is reasonable to use if one has real prior information from a previous similar study. However, informative priors must be specified with care in actual practice because you can get misleading results (SAS Institute, Inc. 2010).

1.01 Multiple Choice Poll

Which of the following statements is true regarding Bayesian analysis?

a. Bayesian methods treat parameters as fixed but unknown.

b. Credible intervals enable you to state that there is a 95% chance that the parameter is in this interval.

c. You can carry out any Bayesian inference or perform any modeling without using a posterior distribution.

d. You can carry out any Bayesian inference or perform any modeling without using a prior distribution.

17

Computational Issues

- The posterior distribution or any of its summary measures can be obtained only in closed form for a restricted set of relatively simple models.
- For many models, including generalized linear models, nonlinear models, random-effects models, and survival models, the posterior distribution does not have a closed form.
- In these situations, exact inference is not possible.

19

sas

The development of the posterior distribution might be difficult. The specific problem is carrying out the integrations necessary to obtain the posterior distributions of quantities of interest in situations where nonstandard prior distributions are used. These problems in integration, for many years, restricted Bayesian applications to rather simple examples involving conjugate priors.

Table of Conjugate Distributions

Likelihood	Model Parameters	Conjugate Prior Distribution
Normal with known variance σ^2	μ (mean)	Normal
Normal with known mean μ	σ^2 (variance)	Inverse Gamma Distribution
Binomial	p (probability)	Beta
Poisson	λ (rate)	Gamma
Negative Binomial	p (probability)	Beta

20

sas

If the posterior distributions are in the same family as the prior probability distribution, the prior and posterior are then called *conjugate distributions*, and the prior is called a *conjugate prior* for the likelihood. For example, if the likelihood function is Gaussian with a known variance, choosing a Gaussian prior over the mean will ensure that the posterior distribution is also Gaussian. In simple conjugate cases where the prior and the posterior belong to the same distributional family, it is possible to obtain closed-form solutions for the posterior distribution.

Monte Carlo Methods

- *Monte Carlo methods* involve the use of random sampling techniques based on computer simulation to obtain approximate solutions to integration problems.
- They have the aim of evaluating integrals or sums by simulation rather than exact or approximate analytic methods.
- These methods are useful for Bayesian analysis to obtain posterior summaries from nonstandard distributions.

Most Bayesian analyses require sophisticated computations, including the use of simulation methods such as the Monte Carlo methods, to generate samples from the posterior distribution. The basic idea of Monte Carlo is to simulate the sampling process from a defined population repeatedly by using a computer instead of actually drawing multiple samples to estimate the population summaries of the events of interest.

Markov Chain Monte Carlo Methods

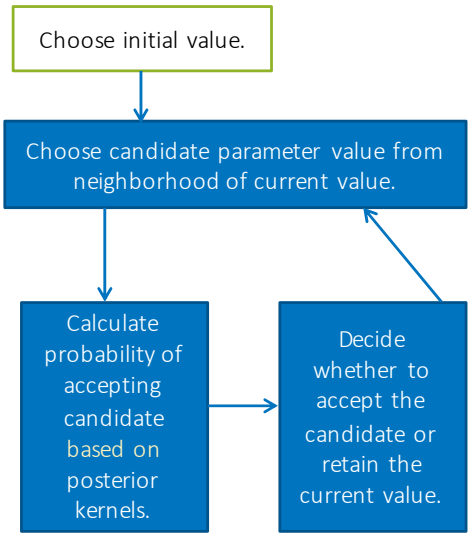
- *Markov Chain Monte Carlo (MCMC)* methods are an effective means of sampling from the posterior distribution of interest even when the posterior has no known closed algebraic form.
- The basic idea behind MCMC is to generate samples from the posterior distribution and to use these samples to approximate expectations of quantities of interest.
- Any inferences that you want to make about the parameters are derived from the sampled values.
- MCMC replaces analytic integration with empirical summaries of sampled values.

22

Markov Chain Monte Carlo methods (MCMC) enable researchers to directly sample sequences of values from the posterior distribution of interest, foregoing the need for closed-form analytic solutions. With MCMC, you use these samples to estimate the posterior distribution's quantities of interest. MCMC methods sample successively from a target distribution. Each sample depends on the previous one, hence the notion of the Markov chain. 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 given the immediate past value. Where it moves next is entirely dependent on where it is now.

The Markov chain method has been quite successful in modern Bayesian computing. One reason is that if the simulation algorithm is implemented correctly, the Markov chain is guaranteed to converge to the target distribution under rather broad conditions, regardless of the initial values of the parameters. Therefore, the Markov chain is able to improve its approximation to the true distribution at each step in the simulation. Furthermore, the simulation algorithm is easily extensible to models with a large number of parameters or high complexity (SAS Institute, Inc. 2010).

Metropolis Algorithm



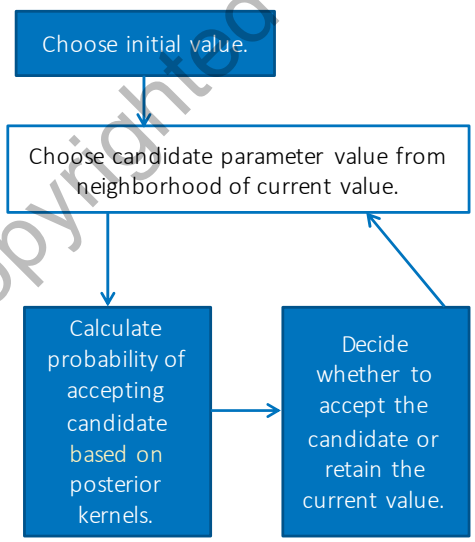
- Several ways to select initial values:
  - Randomly from prior distribution.
  - Using statistics like MLE estimators.
  - Specified directly by user.
- Initial value acts as current value for iteration 1.

23

sas

The Metropolis algorithm is simple and hence 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. For example, suppose you wanted to obtain samples from a univariate distribution with probability density function $p(\theta \mid x)$. To use the Metropolis algorithm, you need to have an initial value for the parameter and a symmetric proposal or candidate density.

Metropolis Algorithm



- Candidate is chosen randomly from a normal distribution whose mean is the current parameter value and a given variance.
- Changing this variance aids in tuning process to achieve desired acceptance rate.

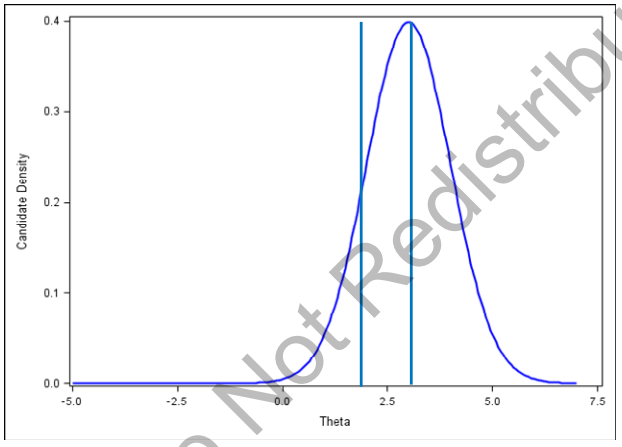
24

sas

At each iteration, the algorithm generates a sample from the proposal density based on the current sample.

Metropolis Algorithm

- Parameter current value is $\theta_{cur}=3$.
- Candidate distribution is normal centered at θ_{cur} with pre-set variance.
- Candidate value θ_{new} is selected $\theta_{new}=2$.

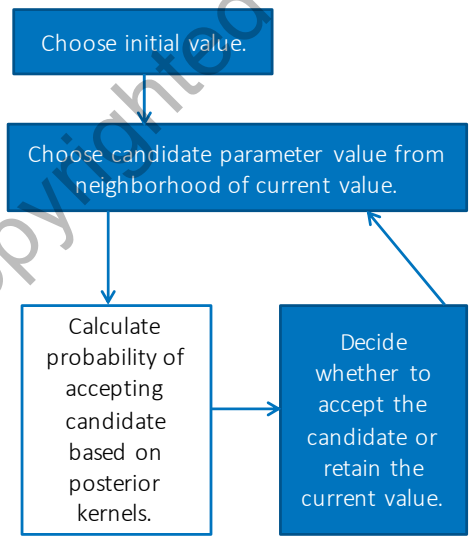


25

sas

The above slide illustrates the first steps of the Metropolis algorithm. The initial value was 3 and a new sample from the proposal density had a value of 2.

Metropolis Algorithm



- calculate r:
- $r=\min\left(1, \frac{p(\theta_{new} \mid x)}{p(\theta_{cur} \mid x)}\right)$,
- the probability of transitioning to the candidate value,
- based on the ratio of posterior kernels at the candidate and current parameter values

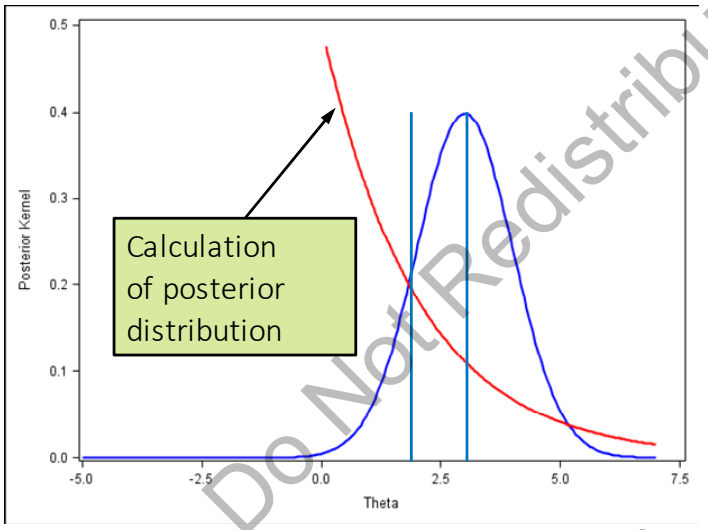
26

sas

In the next step, the statistic *r* is computed, which is the ratio of the posterior kernels at the candidate and current parameter values, or 1 whichever is less.

Metropolis Algorithm

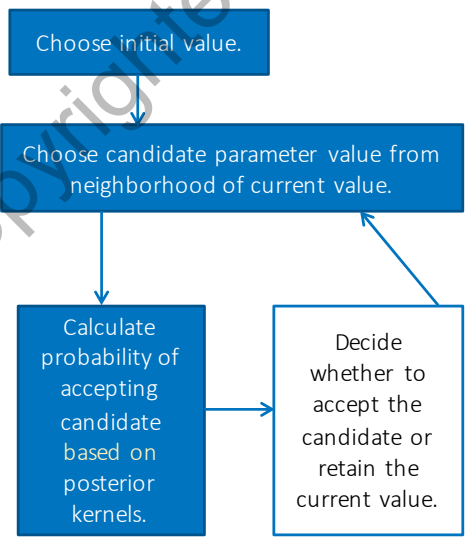
- Value of Posterior Kernel is calculated at each of θ_{cur} and θ_{new}.
- Candidate evaluation is larger than current evaluation. Thus, θ_{new} is accepted with probability 1.



27

The values $p(\theta_{new} \mid x)$ and $p(\theta_{cur} \mid x)$ are computed using Bayes Theorem. In this example, the ratio of posterior kernels at the candidate and current parameter values is greater than 1. This makes the *r* statistic 1 and the new parameter value is accepted with probability 1.

Metropolis Algorithm



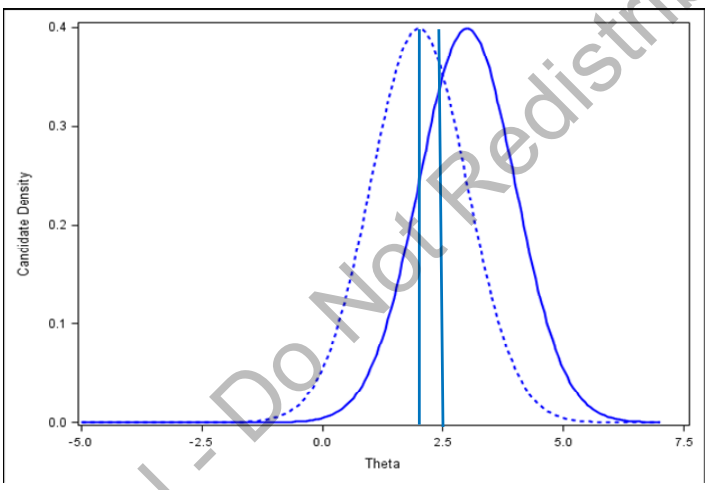
- Randomly select *u* from Uniform (0,1).
- If ($u<r$), then the candidate value replaces the current parameter value. Otherwise, retain the current value for usage in the next iteration.
- $\theta_{next.cur} = \begin{cases} \theta_{new}, & u < r \\ \theta_{cur}, & u \geq r \end{cases}$

28

If the ratio of posterior kernels is less than 1, then a value from the uniform distribution is selected and compared with the statistic *r*. If the *r* statistic is greater than the value selected from the uniform distribution, then the candidate or new parameter value replaces the current parameter value. If the *r* statistic is less than the value selected from the uniform distribution, the current parameter value is retained for usage in the next iteration.

Metropolis Algorithm

- Parameter current value is $\theta_{cur}=2$.
- Candidate distribution is normal centered at θ_{cur} with same pre-set variance as earlier.
- New Candidate value is selected $\theta_{new}=2.5$.



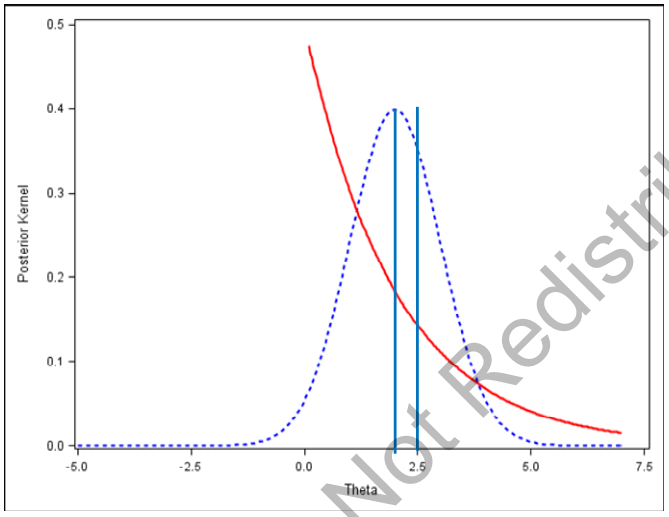
29

To illustrate how a new sample is rejected, suppose the current parameter value is 2 and the new sampled parameter value is 2.5.

Metropolis Algorithm

- Posterior Kernel evaluated at θ_{new} is smaller than at θ_{cur}.
- Candidate is accepted with probability *r*.

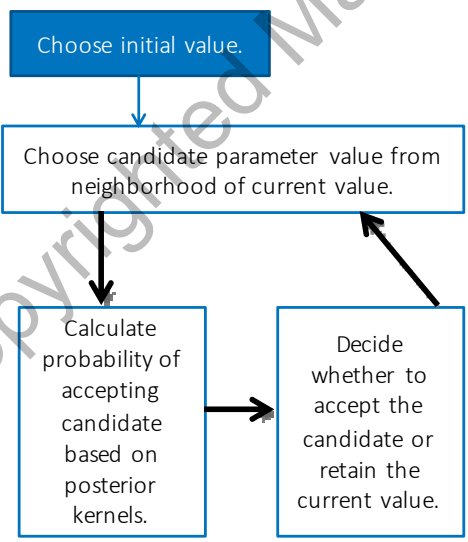
$$r = \frac{p(\theta_{new} \mid x)}{p(\theta_{cur} \mid x)}$$



30

In the slide above, the *r* statistic is the ratio of the posterior kernels because the ratio is less than 1. If the *r* statistic is less than the value selected from the uniform distribution, the new sample is rejected.

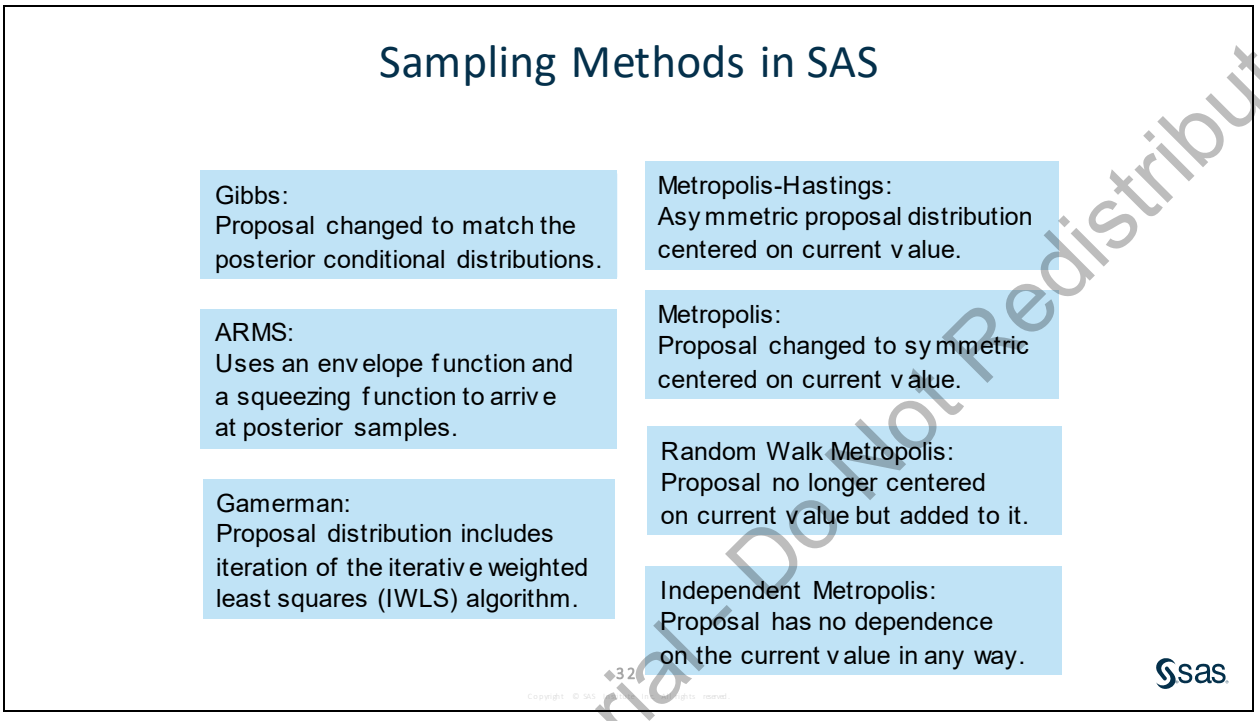
Metropolis Algorithm



- Repeat the process of choosing a candidate and deciding whether to accept the candidate or retain the current for the pre-specified number of iterations.

31

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 the total number of samples needed in advance.



Other sampling methods available in SAS include the Metropolis-Hastings algorithm. The difference between the Metropolis-Hastings and the Metropolis algorithms is that a different proposal distribution is used. Another type of Metropolis algorithm is the independence sampler. It is called the independence sampler because the proposal distribution in the algorithm does not depend on the current point as it does with the random-walk Metropolis algorithm.

Another method is the Gibbs sampler, which is an algorithm that sequentially samples from a joint distribution of two or more random variables. It is a special case of the Metropolis and Metropolis-Hastings Algorithms in which the proposal distributions exactly match the posterior conditional distributions and proposals are accepted 100% of the time. This is the case when the prior and posterior are conjugate distributions. Gibbs sampling requires you to decompose the joint posterior distribution into full conditional distributions for each parameter in the model and then 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 researchers favor this algorithm because it does not require an instrumental proposal distribution as Metropolis methods do. However, although deriving the conditional distributions can be relatively easy, it is not always possible to find an efficient way to sample from these conditional distributions.

Another sampling method available in SAS is the adaptive rejection sampling (ARS) algorithm, which is a rejection algorithm that samples parameters sequentially from their univariate full conditional distributions. Given that the log of the density is concave, the algorithm constructs an envelope to the density by using linear segments. The algorithm then uses the linear segment envelope as a proposal density in the rejection sampling.

Other sampling methods include the Gamerman algorithm, which is used only for generalized linear models. The algorithm uses iterative weighted least squares to generate the proposal distribution.

Markov Chain Convergence

- *Convergence* means that a Markov chain has reached its stationary (target) distribution.
- Assessing the Markov chain convergence is very important, as no valid inferences can be drawn if the chain is not converged.
- It is important to check the convergence for all the parameters and not just the ones of interest.
- Assessing convergence is a difficult task, as the chain converges to a distribution and not to a fixed point.

33

An important aspect of any Bayesian analysis is assessing the convergence of the Markov chains. Inferences based on non-converged Markov chains can be both inaccurate and misleading. First, you have to decide whether the Markov chain has reached its stationary, or its target, posterior distribution. Second, you have to determine the number of iterations to keep after the Markov chain has reached stationarity. Convergence diagnostics help 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. Furthermore, 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 slow convergence, you cannot get accurate posterior inference for parameters that appear to have good convergence (SAS Institute, Inc. 2010).

Burn-In and Thinning

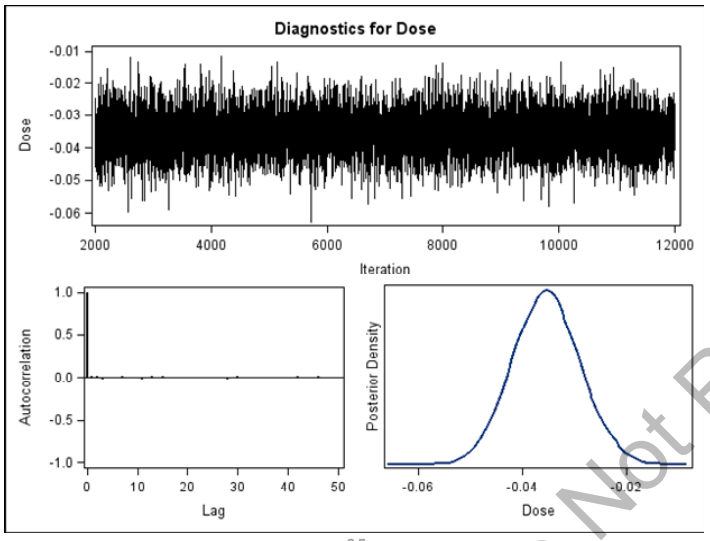
- *Burn-in* refers to the practice of discarding an initial portion of a Markov chain sample so that the effect of the initial values on the posterior inference is minimized.
- *Thinning* refers to the practice of keeping every k^{th} simulated draw from each sequence in order to reduce sample autocorrelations.
- Autocorrelations do not lead to biased Monte Carlo estimates, but rather it is an indicator of poor sampling efficiency.

34

In theory, if the Markov chain runs for an infinite amount of time, the effect of the initial values vanishes. In practice, you do not have the luxury of infinite samples. Therefore, you assume that after a specific number of iterations, the chain has reached its target distribution and you can throw away the early portion and use the remaining samples for posterior inference.

SAS Bayesian procedures compute for each variable the posterior autocorrelations of lags 1, 5, 10, and 50. If the sample autocorrelations are high, a common strategy is to ***thin*** the Markov chain in order to reduce sample autocorrelations. You thin a chain by keeping every k^{th} simulated draw from each sequence. It is important to note that thinning a Markov chain can be wasteful because you are throwing away a (k-1)/k fraction of all the posterior samples generated. MacEachern and Berliner (1994) show that sub-sampling loses information and actually increases the variance of sample mean estimators. Therefore, you always get more precise posterior estimates if the entire Markov chain is used.

Diagnostic Plots – Good Mixing

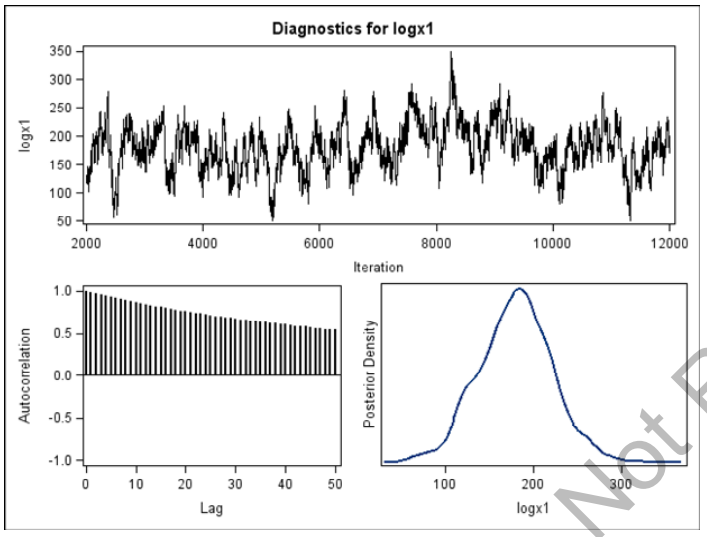


35

Trace plots of samples versus the simulation index (shown in the above slide) 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. The trace plot in the above slide shows a chain that mixes well.

You can also assess the convergence of the generated Markov chain by examining the autocorrelation function plot. You want the autocorrelations to be quite small. The above slide shows plots that exemplify the behavior of a converged Markov chain (SAS Institute, Inc. 2010).

Diagnostic Plots – Poor Mixing



36

The above slide shows a trace plot with poor mixing. *Poor mixing* (or slow convergence) of the Markov chain can happen when parameters are highly correlated with each other. Poor mixing means that the Markov chain slowly traverses the parameter space and the chain has high dependence. The above slide also shows a problematic autocorrelation function plot. High sample autocorrelation can result in high Monte Carlo standard errors.

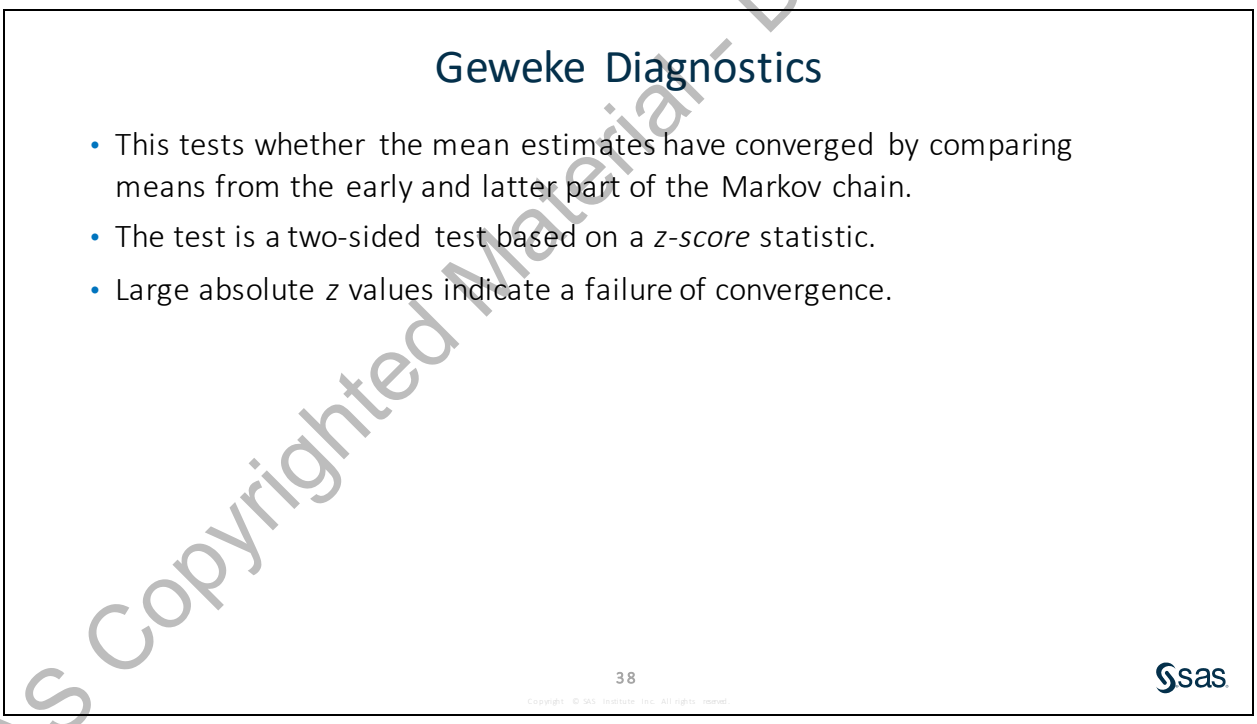
Gelman and Rubin Diagnostics

- This test uses multiple simulated MCMC chains with dispersed initial values and compares the variances within each chain and the variance between the chains.
- Large deviations between these two variances indicate non-convergence.
- A one-sided test based on a variance ratio test statistic is reported where large values indicate a failure to converge.

37

The Gelman and Rubin diagnostics (Gelman and Rubin 1992; Brooks and Gelman 1997) are based on the use of parallel chains with dispersed initial values to test whether they all converge to the same target distribution. For example, suppose you have ***M*** parallel MCMC chains that were initialized from various parts of the target distribution. If all ***M*** chains have reached the target distribution, then the posterior variance estimate (a weighted average of the between-chain variance and the within-chain variance) should be very close to the within-chain variance. Therefore, you would expect to see the ratio of the posterior variance estimate to the within-chain estimate to 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, which could be the result of the presence of a multi-mode posterior distribution (different chains converge to different local modes) or the need to run a longer chain. If the PSRF is close to 1, then you can conclude that each of the ***M*** chains has stabilized, and they are likely to have reached a common target distribution. A refined version of the PSRF is reported in the SAS Bayesian procedures, and because you are concerned only if the ratio is large, only the upper confidence bound is reported.

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, which is something the convergence test cannot detect. If you do not supply initial values for all the different chains, the procedures that support Bayesian methods (PHREG, GENMOD, and LIFEREG) generate them for you (SAS Institute, Inc. 2010).



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. This is a two-sided test, and large absolute *z*-scores indicate convergence problems.

Heidelberger and Welch Diagnostics

- These tests consist of the following two parts:
  - a stationary portion test, which assesses the stationarity of a Markov chain by testing the hypothesis that the chain comes from a covariance stationary process
  - a half-width test, which checks whether the Markov chain sample size is adequate to estimate the mean values accurately.
- The stationary test is a one-sided test based on a Cramér-von Mises statistic.
- The half-width test indicates non-convergence if the relative half-width statistic is greater than a predetermined accuracy measure.

39

sas

The stationarity test is one-sided; rejection occurs when the *p*-value is less than alpha. To perform the half-width test, you need to select an alpha level (the default of which is 0.05) and a predetermined accuracy value (the default is 0.1). If the calculated relative half width of the confidence interval is greater than the accuracy value, you conclude that there are not enough data to accurately estimate the mean with 1-alpha confidence under that specific accuracy value. (Heidelberger and Welch 1981 and 1983).

Raftery and Lewis Diagnostics

- The test evaluates the accuracy of the estimated percentiles by reporting the number of samples needed to reach the desired accuracy of the percentiles.
- If the total number of samples needed are greater than the Markov chain sample, then the desired precision was not obtained.
- The test is specifically designed for the percentile of interest and does not provide information about convergence of the chain as a whole.

40

sas

If your interest lies in posterior percentiles, you would want to examine the Raftery-Lewis diagnostic test (Raftery and Lewis 1992 and 1996), which evaluates the accuracy of the estimated percentiles. If the test indicates failure, then a longer Markov chain might be needed.

Effective Sample Size

- *Effective sample size* is a measure of how well a Markov chain is mixing.
- It takes autocorrelation into account.
- It shows good mixing when it is close to the total sample size.

41

sas

Although you can use autocorrelation and trace plots to examine the mixing of a Markov chain, the effective sample size is also a useful measure that examines mixing. When the effective sample size is much lower than the actual sample size, slower mixing of the Markov chain can be evident.

Summary of Convergence Diagnostics

- There are no definitive tests of convergence.
- Visual inspection of the trace plots is often the most useful approach.
- Geweke and Heidelberger-Welch tests sometimes are statistically significant even when the trace plots look good.
- Oversensitivity to minor departures from stationarity does not impact inferences. Different convergence diagnostics are designed to protect you against different potential pitfalls.

42

sas

Different convergence diagnostic statistics check different aspects of convergence. Even if some of the statistics are statistically significant, the inferences should not be biased if the trace plots look good.

1.02 Multiple Choice Poll

Which of the following statements is true regarding Markov chain convergence?

a. It is important to check only the convergence for the parameters of interest.

b. Geweke diagnostics tests whether the mean estimates have converged by comparing variances from the early and latter part of the Markov chain.

c. A trace plot with a constant mean and variance indicates a non-converged Markov chain.

d. Gelman and Rubin diagnostics uses multiple simulated MCMC chains with dispersed initial values and compares the variances within each chain and the variance between the chains.

43

sas

Deviance Information Criterion (DIC)

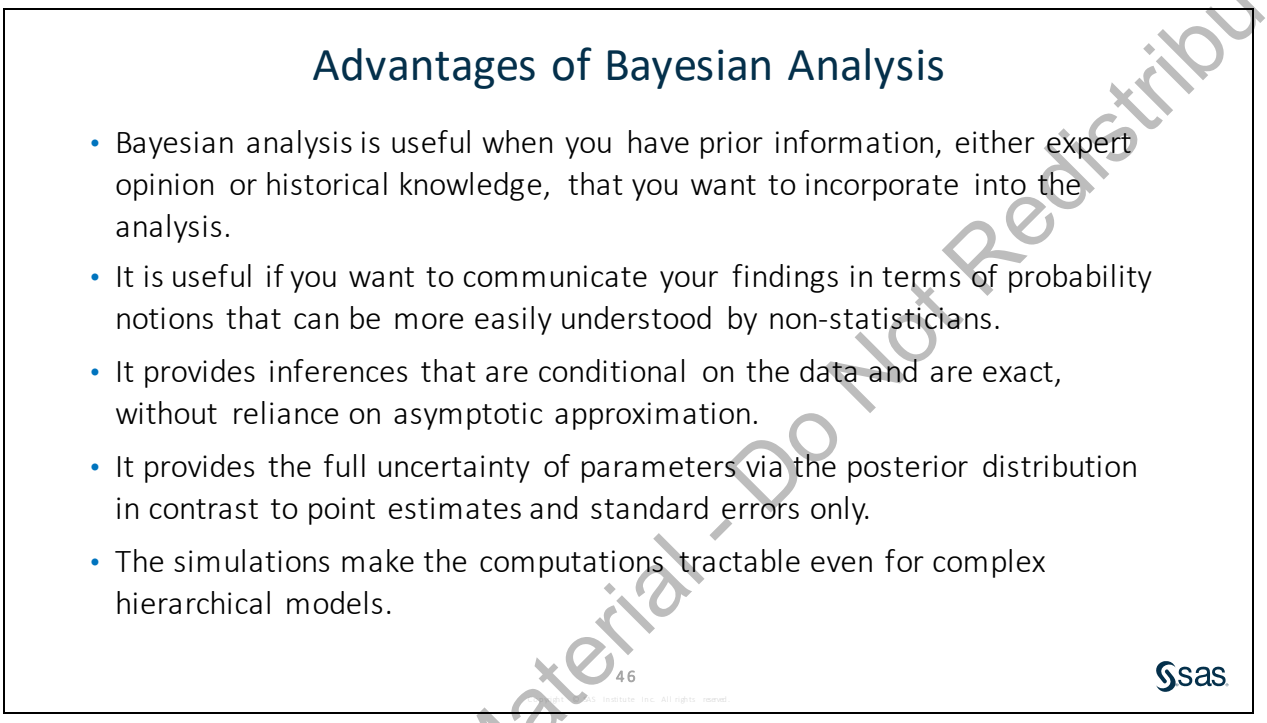
- *Deviance Information Criterion* (DIC) is a Bayesian alternative to AIC and BIC.
- It is a statistic where the smaller value indicates a better fit to the data set.
- DIC can be applied to non-nested models and models that have random effects.

45

sas

The Deviance Information Criterion (DIC) is a model assessment tool, which uses the posterior densities, which means that it takes the prior information into account. Calculation of the DIC does not require maximization over the parameter space, like the Akaike's information criterion (AIC) and Bayesian information criterion (BIC). A smaller DIC indicates a better fit to the data set.

Note: DIC also provides the effective number of parameters, which is useful in the presence of random effects.



The strength of Bayesian analysis is that it 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 are derived from the posterior distribution.

Bayesian analysis provides interpretable answers, such as "the true parameter has a probability of 0.95 of falling in a 95% credible interval." Bayesian analysis can also answer specific scientific questions directly. For example, you can compare the posterior probability of competing hypotheses directly, instead of just using *p*-values.

It should be noted that when the sample size is large, Bayesian inference often provides results for parametric models that are very similar to the results produced by classical inferential methods (SAS Institute, Inc. 2010).

Disadvantages of Bayesian Analysis

- It does not tell you how to select a prior and there is no one 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.
- It often comes with a high computational cost, especially in models with a large number of parameters.

47

The disadvantage of Bayesian analysis is in the selection of the prior distribution, which is also true about the choice of sampling distributions. If the posterior distribution is heavily influenced by the prior distribution, then it might sometimes be difficult to convince subject matter experts who do not agree with the validity of the chosen prior.

1.2 Fitting a Bayesian Model Using SAS Procedures

Objectives

- Illustrate the capabilities of the Bayesian procedures in SAS.
- Fit a Bayesian logistic regression model in the GENMOD procedure.
- Fit a Bayesian survival model in the PHREG procedure.

50

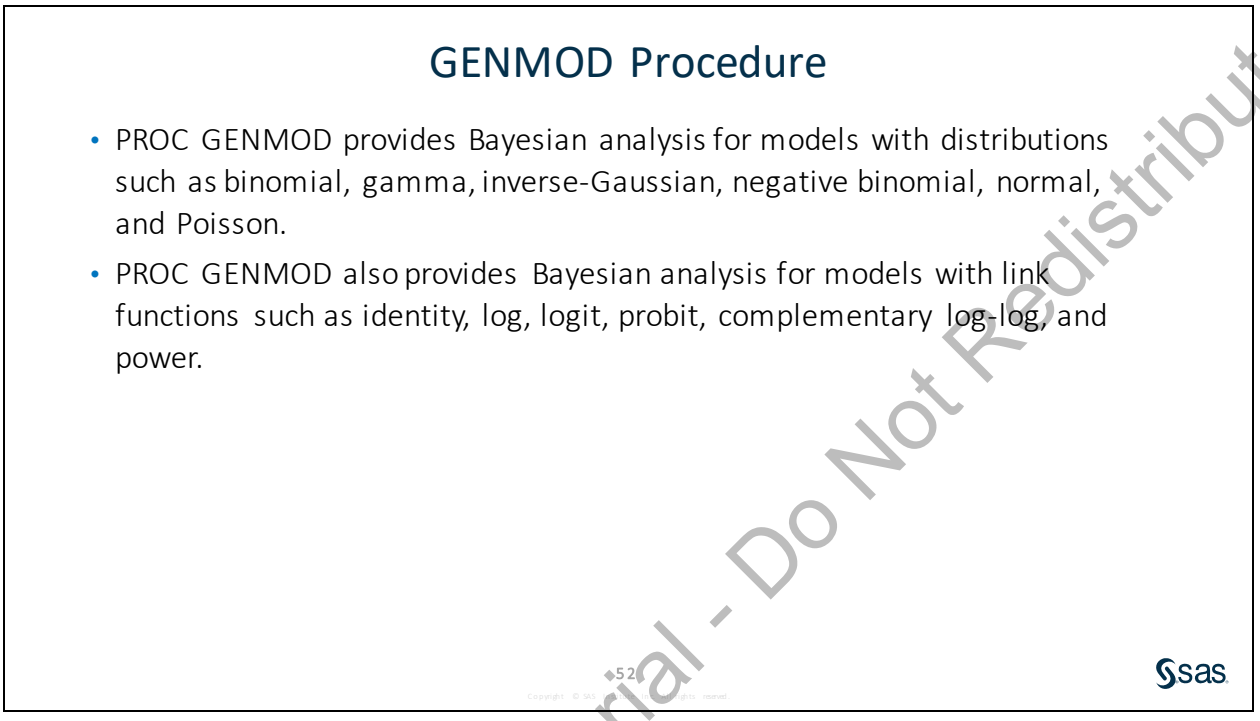
Bayesian Analysis in SAS

Bayesian methods in SAS/STAT 14.2 are found in the following procedures:

- the GENMOD procedure, which fits generalized linear models
- the PHREG procedure, which performs regression analysis of survival data based on the Cox proportional hazards model
- the LIFEREG procedure, which fits parametric models to survival data
- the MCMC procedure, which is a general purpose Markov chain Monte Carlo simulation procedure that is designed to fit Bayesian models.

51

The GENMOD, LIFEREG, and PHREG procedures provide Bayesian analysis in addition to the standard frequentist analyses that they have always performed. PROC PHREG also supports Bayesian analysis for piecewise exponential models. The MCMC procedure is a general procedure that fits Bayesian models with arbitrary priors and likelihood functions.



The ASSESS, CONTRAST, ESTIMATE, OUTPUT, and REPEATED statements, if specified, are ignored when you specify a Bayesian analysis. Furthermore, the PLOTS= option in the PROC GENMOD statement is ignored and the following options are ignored in the MODEL statement: ALPHA=, CORRB, COVB, TYPE1, TYPE3, SCALE=DEVIANCE (DSCALE), SCALE=PEARSON (PSCALE), OBSTATS, RESIDUALS, XVARS, PREDICTED, DIAGNOSTICS, and SCALE= for Poisson and binomial distributions. The multinomial and zero-inflated Poisson distributions are not available for Bayesian analysis in PROC GENMOD.

PHREG Procedure

- PROC PHREG provides Bayesian analysis for Cox regression models with time-independent and time- dependent predictor variables and accommodates all the methods handling ties.
- PROC PHREG also provides Bayesian analysis for piecewise exponential models where you can divide the time axis into sections having its own hazard rate.

53

The ASSESS, CONTRAST, ID, OUTPUT, and TEST statements, if specified, are ignored when you specify a Bayesian analysis. Furthermore, the COVM and COVS options in the PROC PHREG statement are ignored and the following options are ignored in the MODEL statement: BEST=, CORRB, COVB, DETAILS, HIERARCHY=, INCLUDE=, MAXSTEP=, NOFIT, PLCONV=, SELECTION=, SEQUENTIAL, SLENTRY=, and SLSTAY=.

LIFEREG Procedure

PROC LIFEREG provides Bayesian analysis for parametric location-scale survival models with distributions such as exponential, generalized gamma, log-logistic, lognormal, logistic, normal, and Weibull.

54

The OUTPUT and PROBPLOT statements, if specified, are ignored when you specify a Bayesian analysis. The PLOTS=PROBPLOT option in the PROC LIFEREG statement and the CORRB and COVB options in the MODEL statement are also ignored.

BAYES Statement

- The BAYES statement requests a Bayesian analysis of the regression model.
- The Bayesian posterior samples (also known as the chain) for the regression parameters can be written to a SAS data set.

55

sas

For all three procedures, the BAYES statement requests a Bayesian analysis.

BAYES Statement Options

The following options appear in all three procedures:

- 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.
- COEFFPRIOR= specifies the prior of the regression coefficients.
- DIAGNOSTICS= displays convergence diagnostics.
- PLOTS= displays diagnostic plots.
- STATISTICS= displays summary statistics.
- OUTPOST= names a SAS data set for the posterior samples.

56

sas

The default number of burn-in iterations before the chains are saved is 2000. The default number of iterations after the burn-in is 10000. 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 plots, the plots are displayed by parameters unless the global plot option GROUPBY (GROUPBY=TYPE in PROC PHREG) is specified.

BAYES Statement Options – PROC PHREG

- PIECEWISE= specifies details of the piecewise exponential model.
- SAMPLING= specifies the sampling algorithm.

57

sas

The piecewise exponential model is an extension of the exponential hazard model, where instead of a single hazard, you can divide the time axis into sections with each section having its own hazard rate. There are also two sampling algorithms available in PROC PHREG: adaptive rejection Metropolis sampling (ARMS) algorithm and the random walk Metropolis (RWM) algorithm. The default sampling algorithm is ARMS.

BAYES Statement Options – PROC GENMOD

- INITIALMLE specifies that maximum likelihood estimates be used as initial values of the chain.
- METROPOLIS= specifies the use of a Metropolis step in the ARMS algorithm.
- SAMPLING= specifies the sampling algorithm.
- DISPERSIONPRIOR= specifies the prior of the dispersion parameter.
- PRECISIONPRIOR= specifies the prior of the precision parameter.
- SCALEPRIOR= specifies the prior of the scale parameter.

58

There are three sampling algorithms available in PROC GENMOD: ARMS, the Gamerman algorithm (default method except for the normal distribution with a conjugate prior), and the independent Metropolis algorithm. The default prior distribution for the regression coefficients is the uniform or flat prior.

BAYES Statement Options – PROC LIFEREG

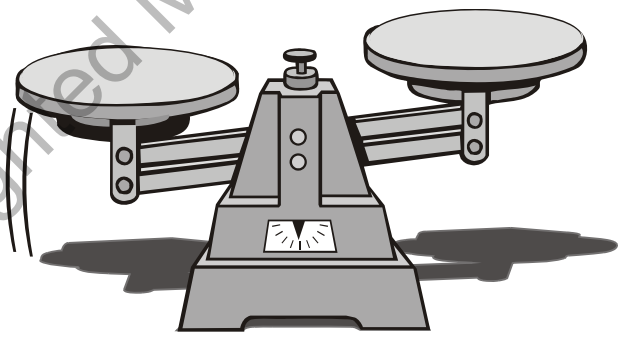
- INITIALMLE specifies that maximum likelihood estimates be used as initialvalue of the chain.
- METROPOLIS= specifies the use of a Metropolis step in the ARMS.
- EXPONENTIALSCALEPRIOR= specifies the prior of the exponential scale parameter.
- GAMMASHAPEPRIOR= specifies the prior of the three-parameter gamma shape 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.

59

sas

In PROC LIFEREG, you can specify that the Metropolis step be used to generate Gibbs samples for posterior distributions that are not log concave.

Low Birth Weight Data Set



61

sas

Example: Babies with low birth weights (defined to be less than 2500 grams) are a concern because of their potential medical problems. Health researchers want to identify possible contributing factors to low birth weight and recommend strategies to reduce the number of low birth weight babies.

These are the variables in the data set:

low	low birth weight (1=yes, 0=no)
mother_wt	mother's weight at last menstrual period
alcohol	drinking status during pregnancy (1=yes, 0=no)
prev_pretrm	history of preterm labor (0=none, 1=one or more)
hist_hyp	history of hypertension (1=yes, 0=no)

The data are stored in a SAS data set named **sasuser.birth**.

Note: The data were modified from an example in Hosmer and Lemeshow (2000).

Bayesian Analysis in PROC GENMOD

Example: Generate a Bayesian analysis of the low birth weight data set. Fit a logistic regression model and specify **low** as the response variable (use the DESC option in the PROC GENMOD statement to model the probability of low birth weight) and specify **alcohol**, **hist_hyp**, **mother_wt**, and **prev_pretrm** as the predictor variables. In the MODEL statement, use the DIST= binomial option to specify a binomial distribution and a LINK=LOGIT option to specify the logit link function. Use the BAYES statement with the default settings except specify a seed of 27513.

```
/* stbay01d01.sas */
proc genmod data=sasuser.birth desc;
   model low=alcohol hist_hyp mother_wt prev_pretrm
            / dist=binomial link=logit;
   bayes seed=27513;
   title 'Bayesian Analysis of Low Birth Weight Model';
run;
```

Selected PROC GENMOD statement option:

DESC specifies that the levels of the response variable be sorted in the reverse of the default order (sorted highest to lowest).

Selected MODEL statement options:

DIST= specifies the built-in probability distribution to use in the model. If you specify no distributionand no link function, then the GENMOD procedure defaults to the normal distribution with the identity link function.

LINK= specifies the link function to use in the model.

Selected BAYES statement option:

SEED= 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.

```
                Bayesian Analysis of Low Birth Weight Model

                           The GENMOD Procedure

                             Bayesian Analysis

                            Model Information

          Data Set              SASUSER.BIRTH
          Burn-In Size                  2000
          MC Sample Size               10000
          Thinning                         1
          Sampling Algorithm         Gamerman
          Distribution               Binomial
          Link Function                 Logit
          Dependent Variable              low    Indicator for Birth Weight
```

The THINNING= option controls the thinning of the Markov chain. Every k^{th} sample is used when THINNING=*k*. The default is THINNING=1, which means all the samples are used.

```
                          Number of Observations Read         189
                          Number of Observations Used         189
                          Number of Events                     59
                          Number of Trials                    189
                                Response Profile

                               Ordered                 Total
                                 Value    low       Frequency

                                     1    1              59
                                     2    0             130

PROC GENMOD is modeling the probability that low='1'.

          Algorithm converged.

                     Analysis Of Maximum Likelihood Parameter Estimates

                                              Standard     Wald 95% Confidence
          Parameter      DF    Estimate         Error           Limits

          Intercept       1      0.7843      0.8670      -0.9150      2.4835
          alcohol         1      0.5129      0.3450      -0.1633      1.1892
          hist_hyp        1      1.8458      0.7044       0.4651      3.2265
          mother_wt       1     -0.0170      0.0068      -0.0302     -0.0037
          prev_pretrm     1      1.2876      0.4366       0.4319      2.1433
          Scale           0      1.0000      0.0000       1.0000      1.0000

NOTE: The scale parameter was held fixed.

                                  Parameter      Prior

                                  Intercept      Constant
                                  alcohol        Constant
                                  hist_hyp       Constant
                                  mother_wt      Constant
                                  prev_pretrm    Constant

          Algorithm converged.

                                Initial Values of the Chain

                                                                                 prev_
 Chain          Seed    Intercept     alcohol    hist_hyp    mother_wt       pretrm

     1         27513      0.78428    0.512933    1.845811     -0.01697     1.287597

                                       Fit Statistics

                        DIC (smaller is better)                  218.069
                        pD (effective number of parameters)        5.088
```

The initial values of the Markov chain are estimates of the mode of the posterior distribution obtained by optimization. You can specify that maximum likelihood estimates of the model parameters be used as initial values of the Markov chain by using the INITIALMLE option.

Posterior Summaries

Parameter	N	Mean	Standard Deviation	Percentiles 25%	50%	75%
Intercept	10000	0.9221	0.8813	0.3293	0.9156	1.5089
alcohol	10000	0.5148	0.3520	0.2760	0.5171	0.7513
hist_hyp	10000	1.9270	0.7448	1.4224	1.9070	2.4022
mother_wt	10000	-0.0183	0.00697	-0.0228	-0.0181	-0.0135
prev_pretrm	10000	1.3302	0.4477	1.0246	1.3278	1.6307

Posterior Intervals

Parameter	Alpha	Equal-Tail Interval		HPD Interval	
Intercept	0.050	-0.7432	2.7326	-0.7764	2.6862
alcohol	0.050	-0.1642	1.2034	-0.1785	1.1815
hist_hyp	0.050	0.5461	3.4611	0.4740	3.3617
mother_wt	0.050	-0.0323	-0.00533	-0.0311	-0.00461
prev_pretrm	0.050	0.4771	2.2179	0.4451	2.1835

The 95% equal-tail interval corresponds to the 2.5th and 97.5th percentiles of the posterior distribution. Some statisticians prefer this interval because it is invariant under transformations.

A 100(1-alpha)% highest posterior density 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 interval is an interval in which most of the distribution lies. Some statisticians prefer this interval because it is the smallest interval.

Posterior Correlation Matrix

Parameter	Intercept	alcohol	hist_hyp	mother_wt	prev_ pretrm
Intercept	1.000	-0.188	0.276	-0.961	-0.080
alcohol	-0.188	1.000	-0.007	0.033	-0.165
hist_hyp	0.276	-0.007	1.000	-0.346	0.048
mother_wt	-0.961	0.033	-0.346	1.000	0.007
prev_pretrm	-0.080	-0.165	0.048	0.007	1.000

Posterior Autocorrelations

Parameter	Lag 1	Lag 5	Lag 10	Lag 50
Intercept	0.3605	0.0460	0.0059	-0.0104
alcohol	0.3037	0.0161	0.0076	0.0069
hist_hyp	0.3338	0.0380	0.0014	0.0141
mother_wt	0.4058	0.0527	0.0065	-0.0156

prev_pretrm 0.2682 0.0250 0.0232 0.0099

Geweke Diagnostics

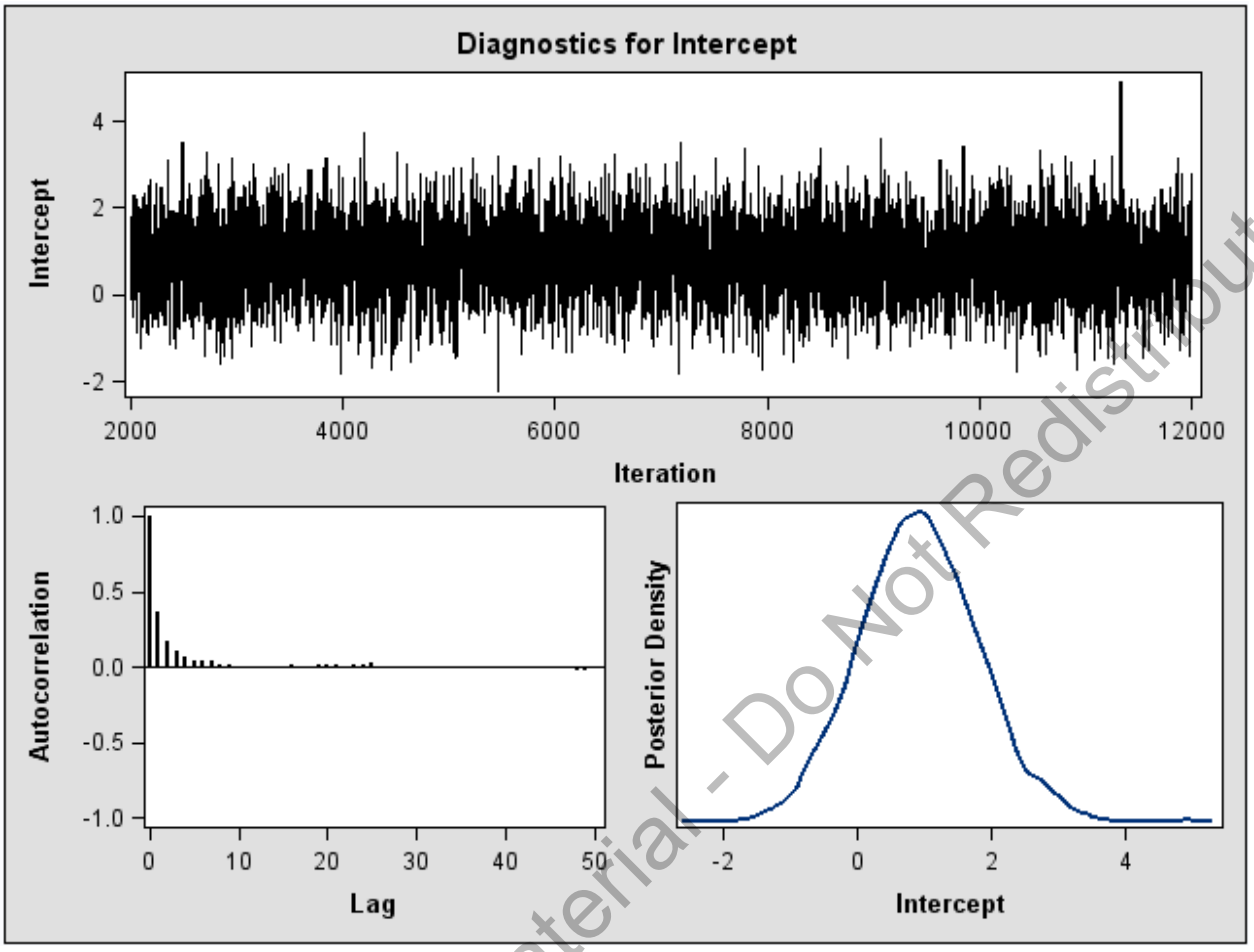
Parameter	z	Pr > |z|
Intercept	-0.2258	0.8214
alcohol	1.5482	0.1216
hist_hyp	-1.3412	0.1799
mother_wt	0.4302	0.6670
prev_pretrm	0.2263	0.8210

Effective Sample Sizes

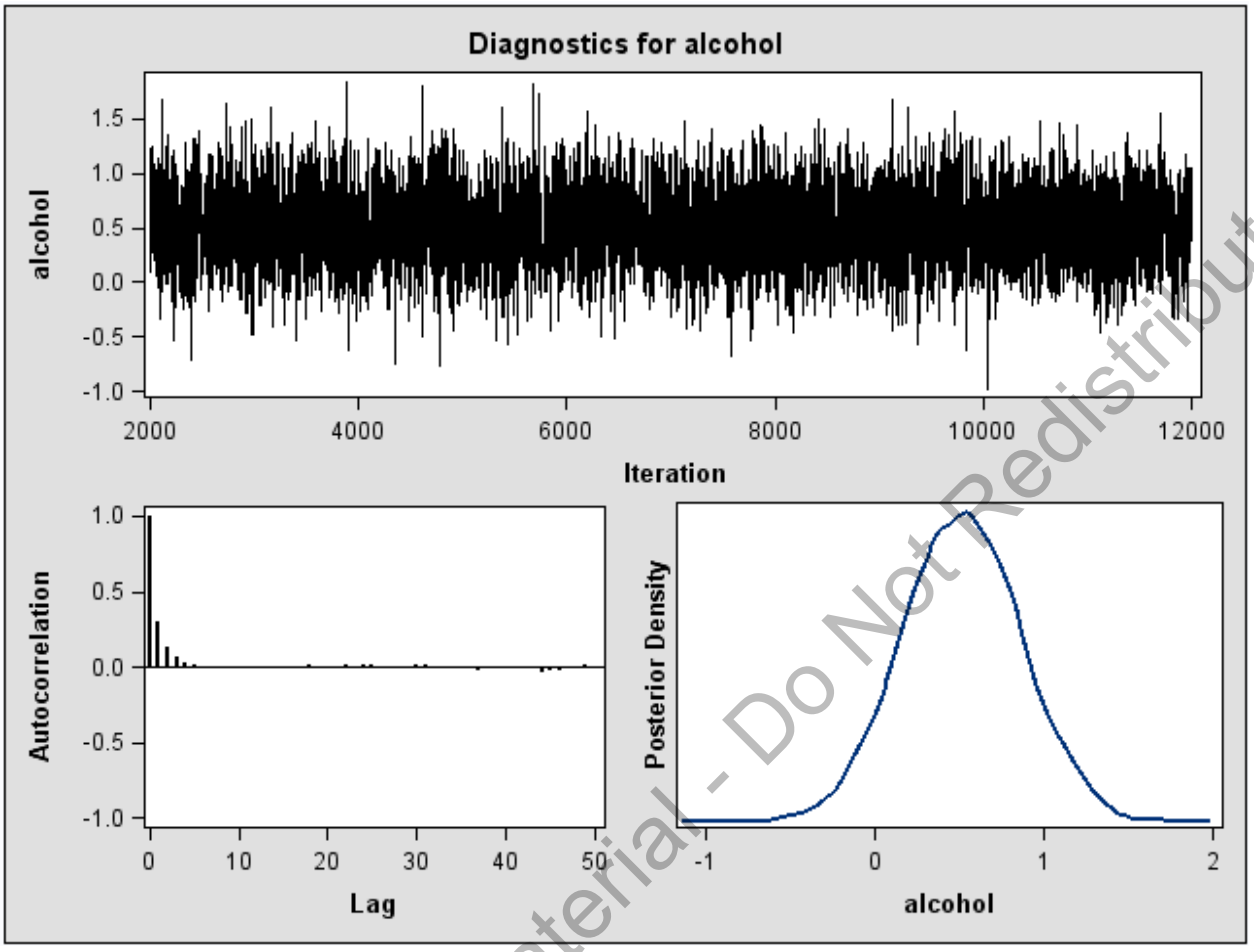
Parameter	ESS	Autocorrelation Time	Efficiency
Intercept	3680.2	2.7173	0.3680
alcohol	4828.8	2.0709	0.4829
hist_hyp	3899.4	2.5645	0.3899
mother_wt	3349.4	2.9856	0.3349
prev_pretrm	4405.0	2.2701	0.4405

The results of the diagnostic statistics show very little evidence that the Markov chain has not converged. The posterior autocorrelations are small after lag 1, the Geweke diagnostics are not significant, and the autocorrelation time is relatively close to 1.

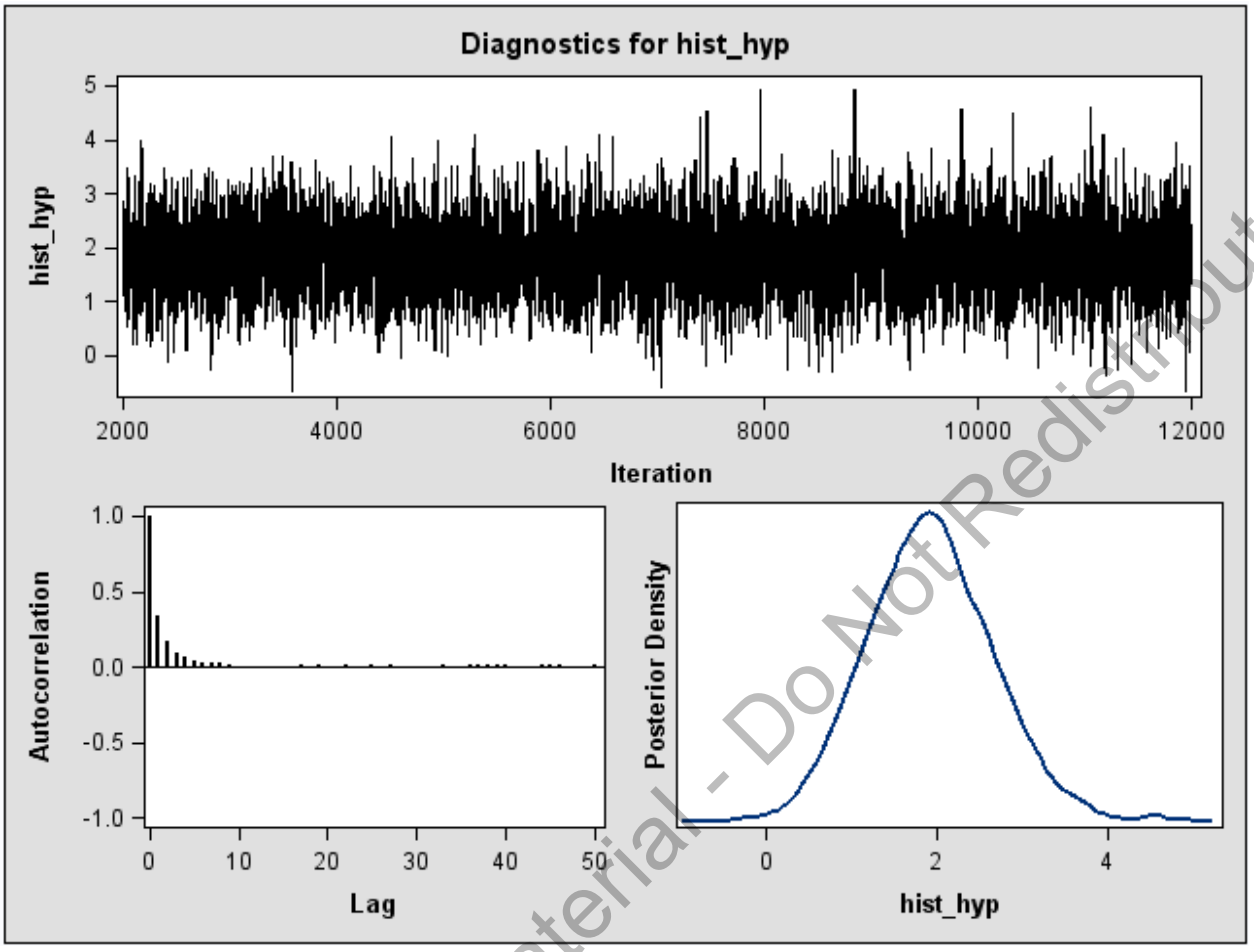
MCMC chains do not produce independent samples. Each sample point depends on the point before it. In this case, the autocorrelation time estimate, read from the effective sample sizes table, is between two to three. This means that it takes two to three observations from the MCMC output to make inferences about the parameters with the same precision that you would get from using an independent sample. A large autocorrelation time estimate indicates poor mixing.



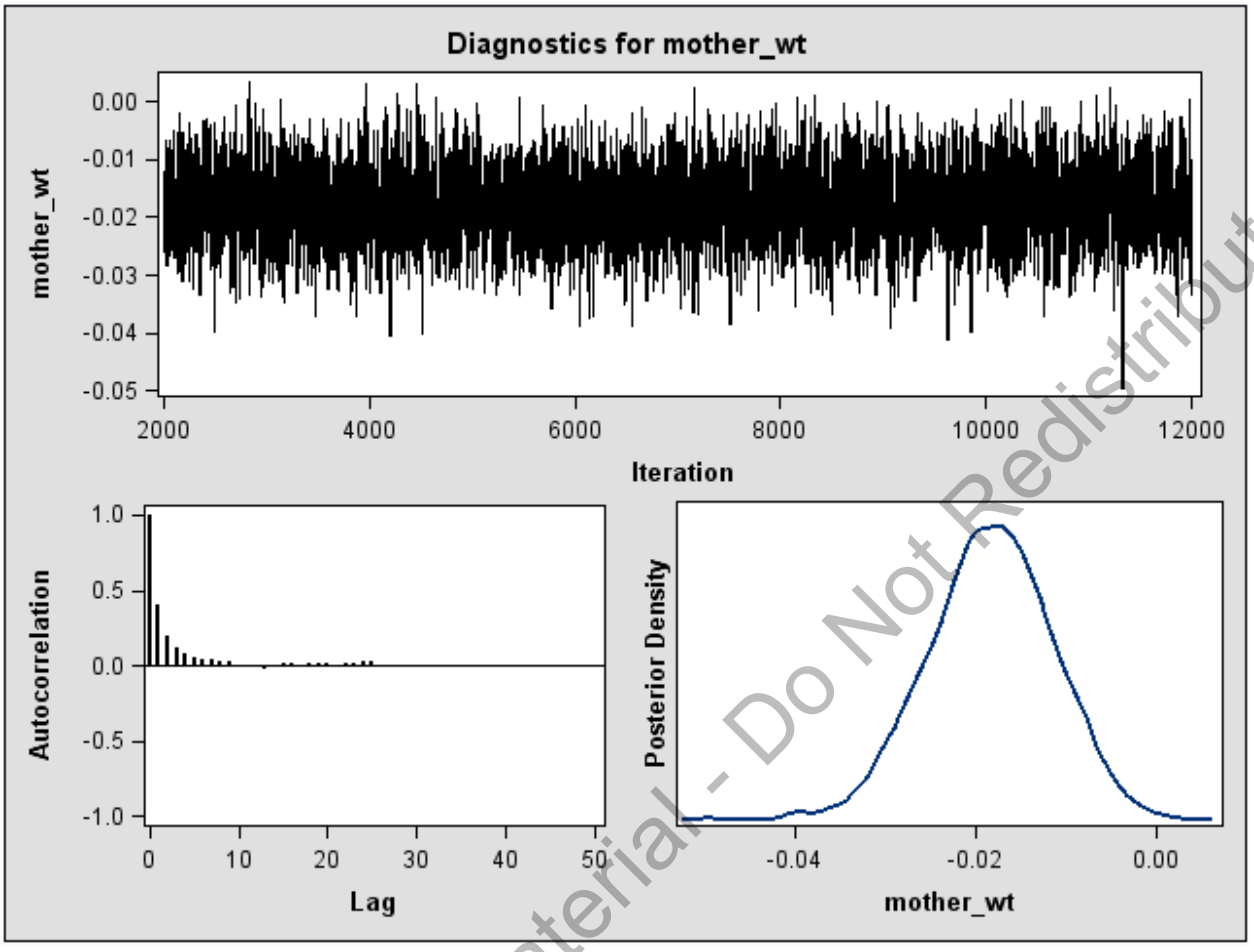
The diagnostic plots for **Intercept** show a converged Markov chain. The trace of the samples centers on 0.92 with a relatively constant mean and variance, and the autocorrelations are quite small after lag 1.



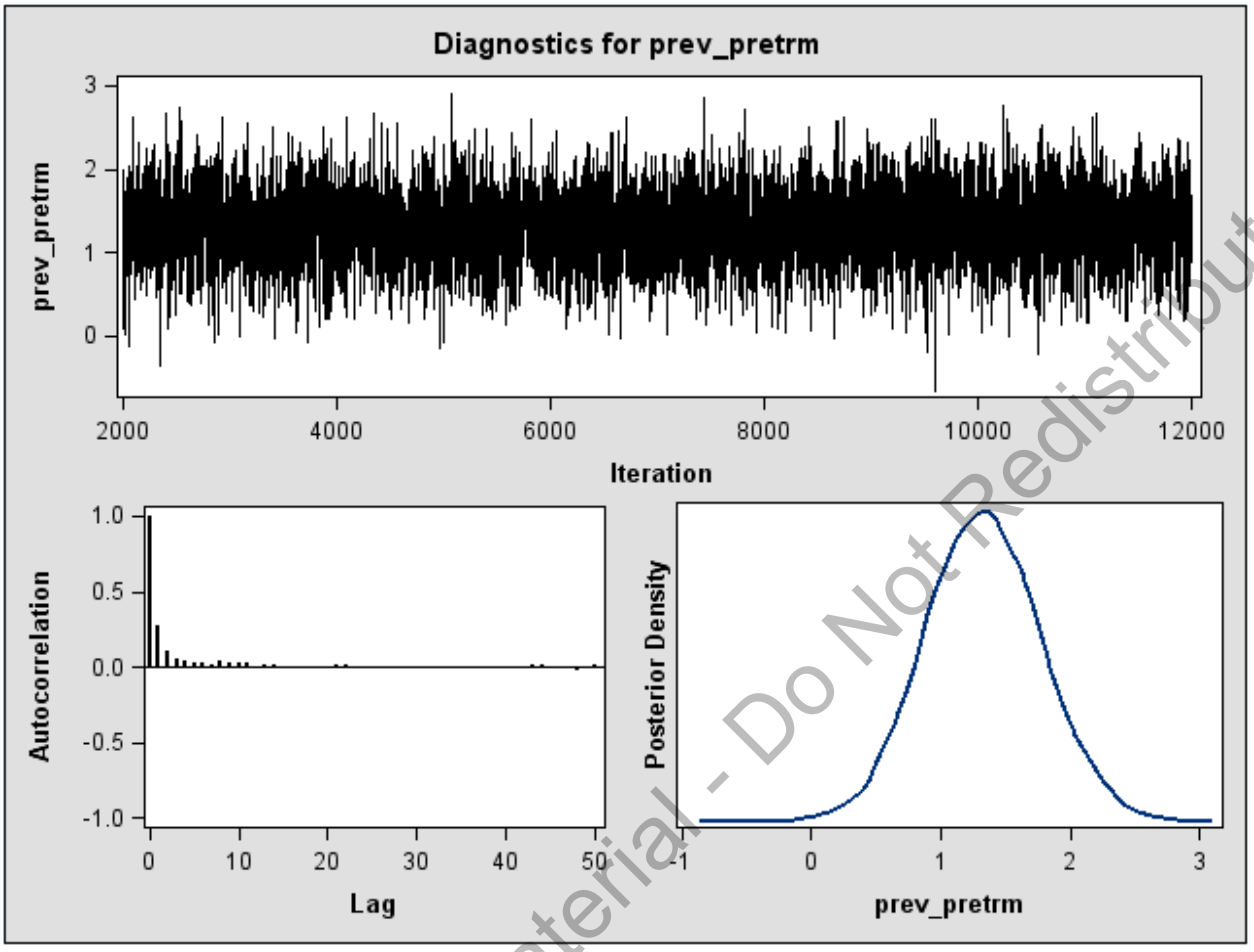
The diagnostic plots for **alcohol** show a converged Markov chain. The trace of the samples centers on 0.52 with a relatively constant mean and variance, and the autocorrelations are quite small. The distribution of the effect of **alcohol** shows that 0 is relatively far from the tail of the distribution.



The diagnostic plots for **hist_hyp** show a converged Markov chain. The trace of the samples centers on 1.94 with a relatively constant mean and variance, and the autocorrelations are quite small. The distribution of the effect of **hist_hyp** shows that 0 is in the far left tail of the distribution and therefore **hist_hyp** seems to be an important effect.



The diagnostic plots for **mother_wt** show a converged Markov chain. The trace of the samples centers on -0.02 with a relatively constant mean and variance, and the autocorrelations are quite small. The distribution of the effect of **mother_wt** shows that 0 is in the far right tail of the distribution and therefore **mother_wt** seems to be an important effect.



The diagnostic plots for **prev_pretrm** show a converged Markov chain. The trace of the samples centers on 1.33 with a relatively constant mean and variance, and the autocorrelations are quite small. The distribution of the effect of **prev_pretrm** shows that 0 is in the far left tail and therefore **prev_pretrm** seems to be an important effect.

Example: Create a data set with the generated posterior samples from the Bayesian model of the low birth weight data set. Then generate the probability that the parameter estimate for each variable is greater than zero.

```
ods select none;
proc genmod data=sasuser.birth desc;
   model low=alcohol hist_hyp mother_wt prev_pretrm
                / dist=binomial link=logit;
   bayes seed=27513 plots=none outpost=bayes_prob;
   title 'Bayesian Analysis of Low Birth Weight Model';
run;
```

Selected BAYES statement option:

OUTPOST= names the SAS data set that contains the posterior samples.

```
ods select all;
proc print data=bayes_prob(obs=10);
   title "Line Listing of Generated Posterior Samples";
run;
```

Line Listing of Generated Posterior Samples

Obs	Iteration	Intercept	alcohol	hist_hyp	mother_wt	prev_pretrm	LogLike	LogPost
1	2001	1.778634	1.054257	2.301187	-0.02579	1.335683	-106.637	-106.637
2	2002	1.778634	1.054257	2.301187	-0.02579	1.335683	-106.637	-106.637
3	2003	-0.07851	1.146923	1.293549	-0.01349	0.844106	-107.393	-107.393
4	2004	1.222932	0.174	2.297145	-0.01969	1.240425	-104.641	-104.641
5	2005	1.763731	1.235926	2.87369	-0.02414	0.09239	-111.5	-111.5
6	2006	1.644417	0.486832	1.831663	-0.02108	1.985427	-109.006	-109.006
7	2007	1.555075	0.09132	1.11072	-0.02231	1.42067	-106.072	-106.072
8	2008	1.048408	0.227555	1.74466	-0.0177	1.510527	-104.485	-104.485
9	2009	0.722584	0.840227	1.269096	-0.01778	1.777872	-105.649	-105.649
10	2010	0.623701	0.236253	2.195146	-0.01199	0.876272	-107.126	-107.126

The data set shows the parameter estimates for the predictor variables at each iteration of the sampling algorithm. The iterations start at 2001 because the default number of burn-in iterations before the chain is saved is 2000.

```
data bayes_prob;
   set bayes_prob;
   alc=(alcohol gt 0);
   hist=(hist_hyp gt 0);
   wt=(mother_wt gt 0);
   pretrm=(prev_pretrm gt 0);
run;

proc means data=bayes_prob mean maxdec=8;
   var alc hist wt pretrm;
   title "Proportion of Parameter Estimates Greater than Zero";
run;
```

Proportion of Parameter Estimates Greater than Zero

The MEANS Procedure

Variable	Mean
alc	0.92690000
hist	0.99660000
wt	0.00190000
pretrm	0.99860000

The results computed from the iterations from the sampling algorithm show that 92.7% of the parameters estimates for alcohol, 99.7% of the parameter estimates for history of hypertension, 0.2% of the parameter estimates for mother's weight and 99.9% of the parameter estimates for previous preterm delivery are greater than 0. This shows the advantage of Bayesian analysis as you can compute the probabilities directly instead of using *p*-values.

End of Demonstration

1.03 Multiple Choice Poll

Which of the following statements is true regarding posterior intervals?

a. The 95% equal-tail interval corresponds to the 5th and 95th percentiles of the posterior distribution.

b. The posterior probability of the region that corresponds to the 95% highest posterior density interval is 95%.

c. You can make the claim that the parameter is inside the interval with a measurable probability for only the highest posterior density interval.

d. The 95% highest posterior density interval is always larger than the 95% equal-tail interval.

63

sas

Bayesian Statistics with Informative Prior

Subject-Matter Knowledge: Mothers who drink during pregnancy will have three times the odds of having a low birth weight baby with confidence bounds of 2.8 to 3.2.

Bounds for Odds Ratio:

$$2.80 < e^{\hat{\beta}_{alcohol}} < 3.20$$

Bounds for Coefficient for Alcohol:

$$1.0296 < \hat{\beta}_{alcohol} < 1.1632$$

65

sas

Information gathered from expert opinion showed that women who drink during pregnancy will have three times the odds of having a low birth weight baby with a confidence interval of 2.80 and 3.20. Therefore, the plausible range that you believe the coefficient for alcohol can take ranges from 1.0296 to 1.1632.

Bayesian Statistics with Informative Prior

For an informative normal prior, we need to estimate μ and σ^2. If you assume that $\mu = 1.0986$ and $\mu \pm 2\sigma \approx (1.0296, 1.1632)$, then

$$\sigma^2 = \left(\frac{1.1632 - 1.0296}{1.96 * 2}\right)^2 = 0.00116$$

66

sas

If the expected odds ratio is 3.00, then the mean of prior distribution of the parameter is 1.0986. If you assume a normal distribution where the majority of the prior distribution mass falls within the plausible range, the variance can be computed using the formula in the slide above. For the other parameters, you can use a normal prior distribution with a mean of 0 and a variance of 10^6.

Bayesian Analysis with an Informative Prior

Example: Create a data set called **Prior** that specifies the means and variances of the prior distributions of the parameters. Then fit the previous logistic regression model but specify **alcohol** as a class variable and reverse the sort order of the classification variable. Use the LSMEANS statement and specify **alcohol** as the model effect and request differences of the least-square means, report difference of least square means in terms of odds ratios, request confidence bounds, and request all of the default plots that correspond to the LSMEANS statement. Finally, use the BAYES statement to specify a SAS data set containing the mean and covariance information of the prior distribution, create an output data set with the posterior samples, display a fitted penalized B-spline curve for each trace plot, increase the number of iterations after burn-in to 25,000, specify the adaptive rejection Metropolis sampling, and request all the diagnostic convergence statistics.

```
/* stbay01d02.sas */
data Prior;
   input _TYPE_ $ alcohol1 hist_hyp mother_wt prev_pretrm;
datalines;
Mean 1.0986 0 0 0
Var 0.00116 1e6 1e6 1e6
;
run;
```

The DATA step creates a data set called **Prior** with the estimated means and variances of the prior distributions of the parameters in the model. The variable **_TYPE_** is needed to identify the statistic.

```
proc genmod data=sasuser.birth desc;
   class alcohol(desc);
   model low=alcohol hist_hyp mother_wt prev_pretrm
             / dist=binomial link=logit;
   lsmeans alcohol / diff oddsratio plots=all cl;
   bayes seed=27513 coeffprior=normal(input=Prior) sampling=arms
         outpost=bayes_prob1 plots(smooth)=all diag=all nmc=25000;
   title 'Bayesian Analysis of Low Birth Weight Model';
run;
```

Selected CLASS statement option:

DESC reverses the sorting order of the classification variable (sorted highest to lowest).

Selected LSMEANS statement options:

DIFF requests that differences of the LS-means be displayed.

ODDSRATIO requests that LS-mean differences are also reported in terms of odds ratios.

PLOTS= requests that graphics related to least squares means be produced via ODS Graphics, provided the plot-request does not conflict with other options in the LSMEANS statement. PLOTS=ALL requests that the default plots that correspond to this LSMEANS statement be produced.

CL requests confidence limits of the least squares means.

Selected BAYES statement options:

COEFFPRIOR specifies the prior distribution for the regression coefficients. The default is COEFFPRIOR=UNIFORM, and the normal prior is specified by COEFFPRIOR=NORMAL. Jeffrey's prior can be specified with COEFFPRIOR=JEFFREYS, which is a noninformative prior that is invariant under reparameterization of the parameter vector. This makes it of special interest for use with scale parameters.

INPUT= 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. For an independent normal prior, the variances can be specified with *_TYPE_*='VAR'.

DIAG= controls the number of diagnostics displayed.

SAMPLING= specifies an algorithm used to sample the posterior distribution. The default is the Gamerman algorithm (GAM). Other methods are the adaptive rejection Metropolis sampling (ARMS) and the independent Metropolis algorithm (IM).

NMC= specifies the number of iterations after the burn-in.

Selected global plot option:

SMOOTH displays a fitted penalized B-spline curve for each trace plot.

```
                Bayesian Analysis of Low Birth Weight Model

                              Bayesian Analysis

                             Model Information

Data Set                    SASUSER.BIRTH
Burn-In Size                         2000
MC Sample Size                      25000
Thinning                                1
Sampling Algorithm                   ARMS
Distribution                      Binomial
Link Function                        Logit
Dependent Variable                     low    Indicator for Birth Weight

                Number of Observations Read         189
                Number of Observations Used         189
                Number of Events                     59
                Number of Trials                    189

                       Class Level Information

                Class        Levels    Values

                alcohol           2    1 0

                     Response Profile

                 Ordered              Total
                   Value    low    Frequency
```

```
                                    1    1            59
                                    2    0           130

PROC GENMOD is modeling the probability that low='1'.

                              Parameter Information

                         Parameter       Effect         alcohol

                         Prm1            Intercept
                         Prm2            alcohol        1
                         Prm3            alcohol        0
                         Prm4            hist_hyp
                         Prm5            mother_wt
                         Prm6            prev_pretrm

         Algorithm converged.

                     Analysis Of Maximum Likelihood Parameter Estimates

                                                 Standard     Wald 95% Confidence
           Parameter           DF    Estimate       Error          Limits

           Intercept            1      0.7843      0.8670     -0.9150      2.4835
           alcohol        1     1      0.5129      0.3450     -0.1633      1.1892
           alcohol        0     0      0.0000      0.0000      0.0000      0.0000
           hist_hyp             1      1.8458      0.7044      0.4651      3.2265
           mother_wt            1     -0.0170      0.0068     -0.0302     -0.0037
           prev_pretrm          1      1.2876      0.4366      0.4319      2.1433
           Scale                0      1.0000      0.0000      1.0000      1.0000

NOTE: The scale parameter was held fixed.
```

The predictor variable **alcohol** was specified as a classification variable because the LSMEANS statement allows only classification variables. The LSMEANS statement was used to compute the odds ratios because the ESTIMATE statement is ignored when you specify a Bayesian analysis. The maximum likelihood parameter estimate for **alcohol** yields an odds ratio of 1.67 (exp (0.5129)) with 95% confidence bounds of 0.849 and 3.284.

```
                    Independent Normal Prior for Regression Coefficients

                          Parameter          Mean    Precision

                          Intercept             0         1E-6
                          alcohol1         1.0986      862.069
                          hist_hyp              0         1E-6
                          mother_wt             0         1E-6
                          prev_pretrm           0         1E-6

         Algorithm converged.

                                Initial Values of the Chain

                                                                                  prev_
      Chain         Seed    Intercept    alcohol1    alcohol0    hist_hyp    mother_wt     pretrm

          1        27513     0.525437    1.092982           0    1.847699     -0.01694    1.195192
```

2	2.979976	0.991289	0	-0.24249	-0.03734	-0.12993
3	-1.9291	1.194675	0	3.937888	0.003448	2.520316

Fit Statistics

DIC (smaller is better)	218.807
pD (effective number of parameters)	4.059

Posterior Summaries

Parameter	N	Mean	Standard Deviation	Percentiles 25%	50%	75%
Intercept	25000	0.6297	0.8725	0.0362	0.6127	1.2069
alcohol1	25000	1.0932	0.0339	1.0700	1.0932	1.1160
hist_hyp	25000	1.9244	0.7327	1.4323	1.9069	2.3987
mother_wt	25000	-0.0179	0.00693	-0.0225	-0.0177	-0.0132
prev_pretrm	25000	1.2219	0.4481	0.9173	1.2186	1.5204

Posterior Intervals

Parameter	Alpha	Equal-Tail Interval		HPD Interval	
Intercept	0.050	-1.0529	2.4211	-1.0992	2.3569
alcohol1	0.050	1.0275	1.1599	1.0290	1.1612
hist_hyp	0.050	0.5276	3.4031	0.5067	3.3719
mother_wt	0.050	-0.0321	-0.00483	-0.0315	-0.00435
prev_pretrm	0.050	0.3541	2.1134	0.3475	2.1011

The 95% equal-tail interval corresponds to the 2.5th and 97.5th percentiles of the posterior distribution. As was stated before, some statisticians prefer this interval because it is invariant under transformations. The highest posterior density interval is an interval in which most of the distribution lies. Some statisticians prefer this interval because it is the smallest interval.

Posterior Correlation Matrix

Parameter	Intercept	alcohol1	hist_hyp	mother_wt	prev_ pretrm
Intercept	1.000	-0.008	0.271	-0.972	-0.142
alcohol1	-0.008	1.000	0.011	-0.011	-0.006
hist_hyp	0.271	0.011	1.000	-0.336	0.020
mother_wt	-0.972	-0.011	-0.336	1.000	0.041
prev_pretrm	-0.142	-0.006	0.020	0.041	1.000

Posterior Autocorrelations

Parameter	Lag 1	Lag 5	Lag 10	Lag 50
Intercept	0.0834	0.0014	-0.0053	0.0117
alcohol1	-0.0014	0.0047	0.0094	0.0121
hist_hyp	0.1150	-0.0032	-0.0047	0.0028
mother_wt	0.1409	-0.0001	-0.0107	0.0101
prev_pretrm	0.0040	-0.0007	0.0020	0.0053

Gelman-Rubin Diagnostics

	97.5%

Parameter	Estimate	Bound
Intercept	1.0001	1.0002
alcohol1	1.0000	1.0002
hist_hyp	1.0000	1.0002
mother_wt	1.0001	1.0002
prev_pretrm	1.0000	1.0002

The Gelman-Rubin diagnostics use parallel chains with dispersed initial values to test whether they all converge to the same target distribution. Failure to converge could indicate the presence of a multimode posterior distribution or the need to run a longer chain. The diagnostic values are one-sided tests based on a variance ratio test statistic with large values indicating rejection. The results shown above are very close to 1.0, which indicates no evidence of a failure to converge.

Geweke Diagnostics

Parameter	z	Pr > |z|
Intercept	-0.4638	0.6428
alcohol1	-0.6459	0.5184
hist_hyp	-0.2208	0.8252
mother_wt	0.3499	0.7264
prev_pretrm	-0.3567	0.7213

The Geweke diagnostics test whether the mean estimates have converged by comparing means from the early and latter part of the Markov chain. The diagnostic values are two-sided tests based on a *z*-score statistic. None of the parameters show converge problems.

Raftery-Lewis Diagnostics

Quantile=0.025 Accuracy=+/-0.005 Probability=0.95 Epsilon=0.001

	Number of Samples			Dependence
Parameter	Burn-In	Total	Minimum	Factor
Intercept	2	3853	3746	1.0286
alcohol1	1	3752	3746	1.0016
hist_hyp	2	3865	3746	1.0318
mother_wt	3	4029	3746	1.0755
prev_pretrm	2	3778	3746	1.0085

The Raftery-Lewis diagnostics evaluate the accuracy of the estimated percentiles by reporting the number of samples needed to reach the desired accuracy of the percentiles. If the total samples needed are greater than the Markov chain sample, then the test indicates rejection. The results above indicates no problem with the desired accuracy of the estimated percentiles. It should be noted that when the number of iterations after burn-in was set to the default value of 10,000, two of the parameters did not meet the required sample size.

Heidelberger-Welch Diagnostics

		Stationarity Test		
	Cramer-von		Test	Iterations
Parameter	Mises Stat	p-Value	Outcome	Discarded
Intercept	0.3445	0.1018	Passed	0
alcohol1	0.0501	0.8754	Passed	0
hist_hyp	0.2947	0.1399	Passed	0
mother_wt	0.3309	0.1109	Passed	0
prev_pretrm	0.2157	0.2391	Passed	0

Heidelberger-Welch Diagnostics

Parameter	Half-Width	Mean	Half-Width Test Relative Half-Width	Test Outcome
Intercept	0.0128	0.6297	0.0203	Passed
alcohol1	0.000406	1.0932	0.000372	Passed
hist_hyp	0.0106	1.9244	0.00551	Passed
mother_wt	0.000112	-0.0179	-0.00627	Passed
prev_pretrm	0.00547	1.2219	0.00448	Passed

The Heidelberger-Welch stationarity test shows whether the Markov chain is a covariance stationary process. The values are one-sided tests based on a Cramér-von Mises statistic. None of the *p*-values indicate rejection.

The Heidelberger-Welch half-width test reports whether the sample size is adequate to meet the required accuracy for the mean estimate. If the relative half-width statistic is greater than a predetermined accuracy measure (the default is 0.1), then the test indicates rejection. All of the parameters showed no indication of rejection.

Effective Sample Sizes

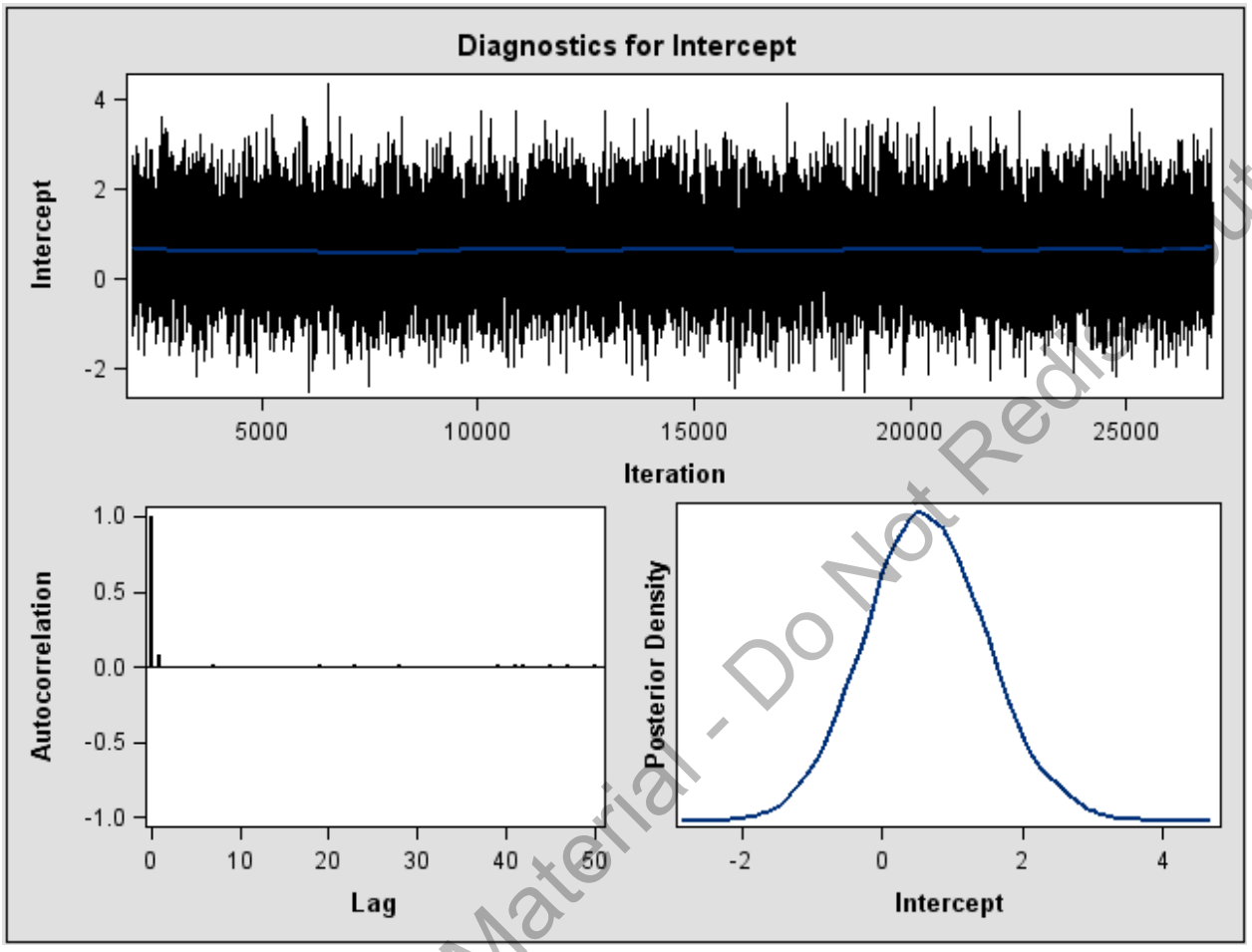
Parameter	ESS	Autocorrelation Time	Efficiency
Intercept	21425.7	1.1668	0.8570
alcohol1	25000.0	1.0000	1.0000
hist_hyp	19809.6	1.2620	0.7924
mother_wt	18997.2	1.3160	0.7599
prev_pretrm	25000.0	1.0000	1.0000

Monte Carlo Standard Errors

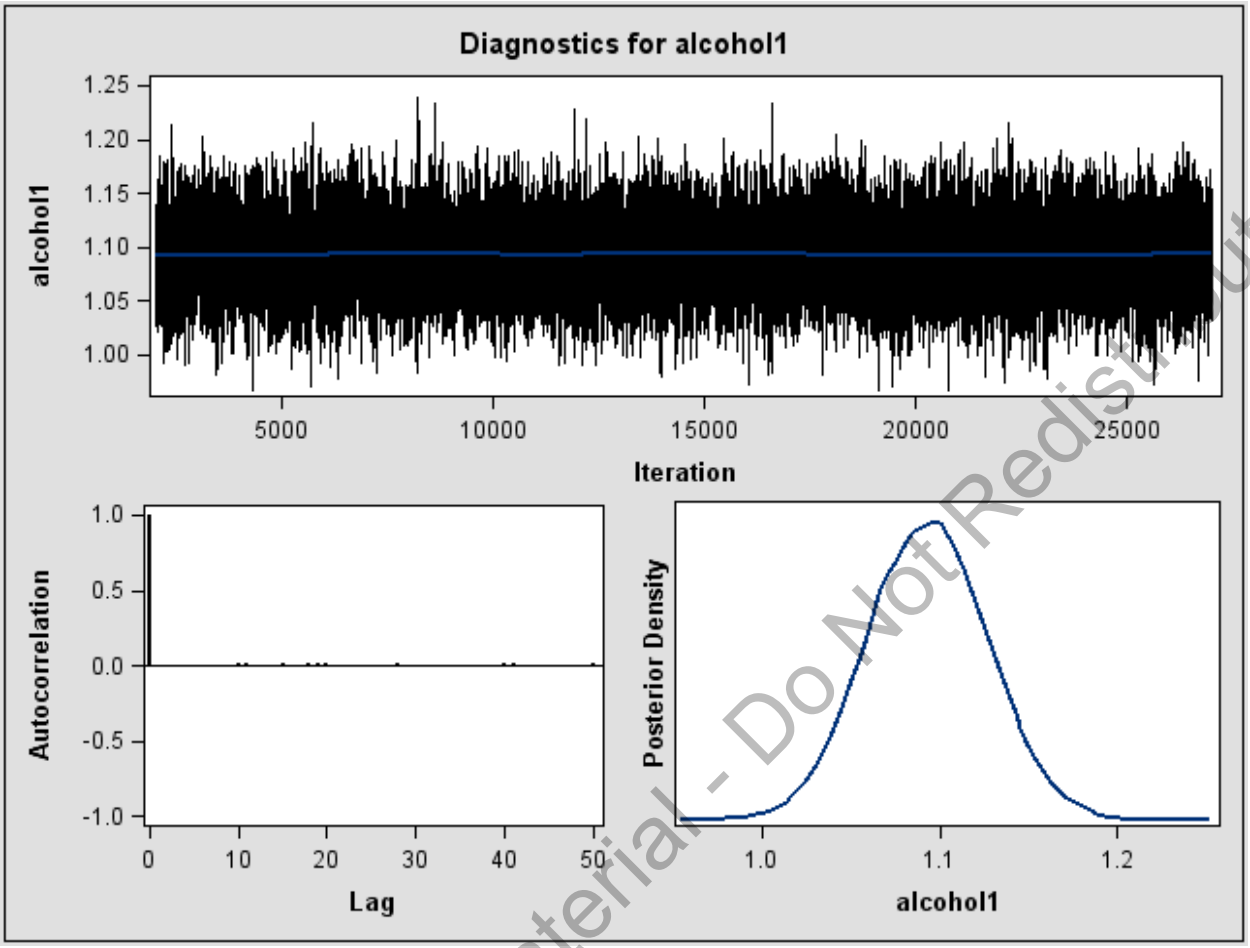
Parameter	MCSE	Standard Deviation	MCSE/SD
Intercept	0.00596	0.8725	0.00683
alcohol1	0.000214	0.0339	0.00632
hist_hyp	0.00521	0.7327	0.00710
mother_wt	0.000050	0.00693	0.00726
prev_pretrm	0.00283	0.4481	0.00632

The Monte Carlo Standard Errors table shows the Monte Carlo standard errors, the posterior sample standard deviation, and the ratio of the two. The table indicates that the standard errors of the mean estimates for each of the parameters are relatively small, with respect to the posterior standard deviations. The ratios are small, which means that only a fraction of the posterior variability is due to the simulation.

Partial Graphics Output:



The diagnostic plots for **Intercept** show no apparent problems with the Markov chain. The trace of the samples centers on the mean for **Intercept** with a relatively constant mean and variance, and the autocorrelations are quite small.



The diagnostic plots for **alcohol1** show no apparent problems with the Markov chain. The trace of the samples centers on the mean for **alcohol1** with a relatively constant mean and variance, and the autocorrelations are quite small.

Note: The remaining diagnostic plots (not shown) show patterns of convergence.

With the diagnostic plots showing convergence of the Markov chain, the Bayesian analysis inferences should be fairly accurate.

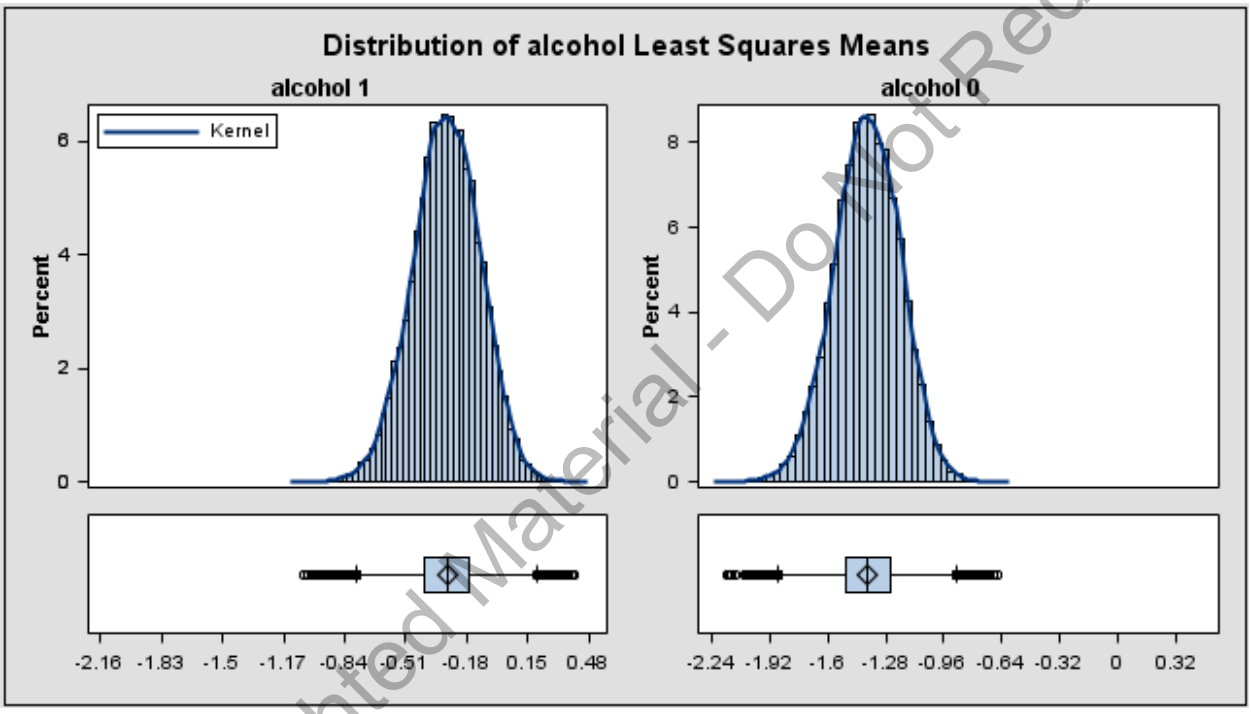
Sample alcohol Least Squares Means

Did the mother drink during pregnancy?	N	Estimate	Standard Deviation	Percentiles 25th	Percentiles 50th	Percentiles 75th	Alpha
1	25000	-0.2887	0.1811	-0.4088	-0.2873	-0.1641	0.05
0	25000	-1.3820	0.1810	-1.5027	-1.3800	-1.2572	0.05

Sample alcohol Least Squares Means

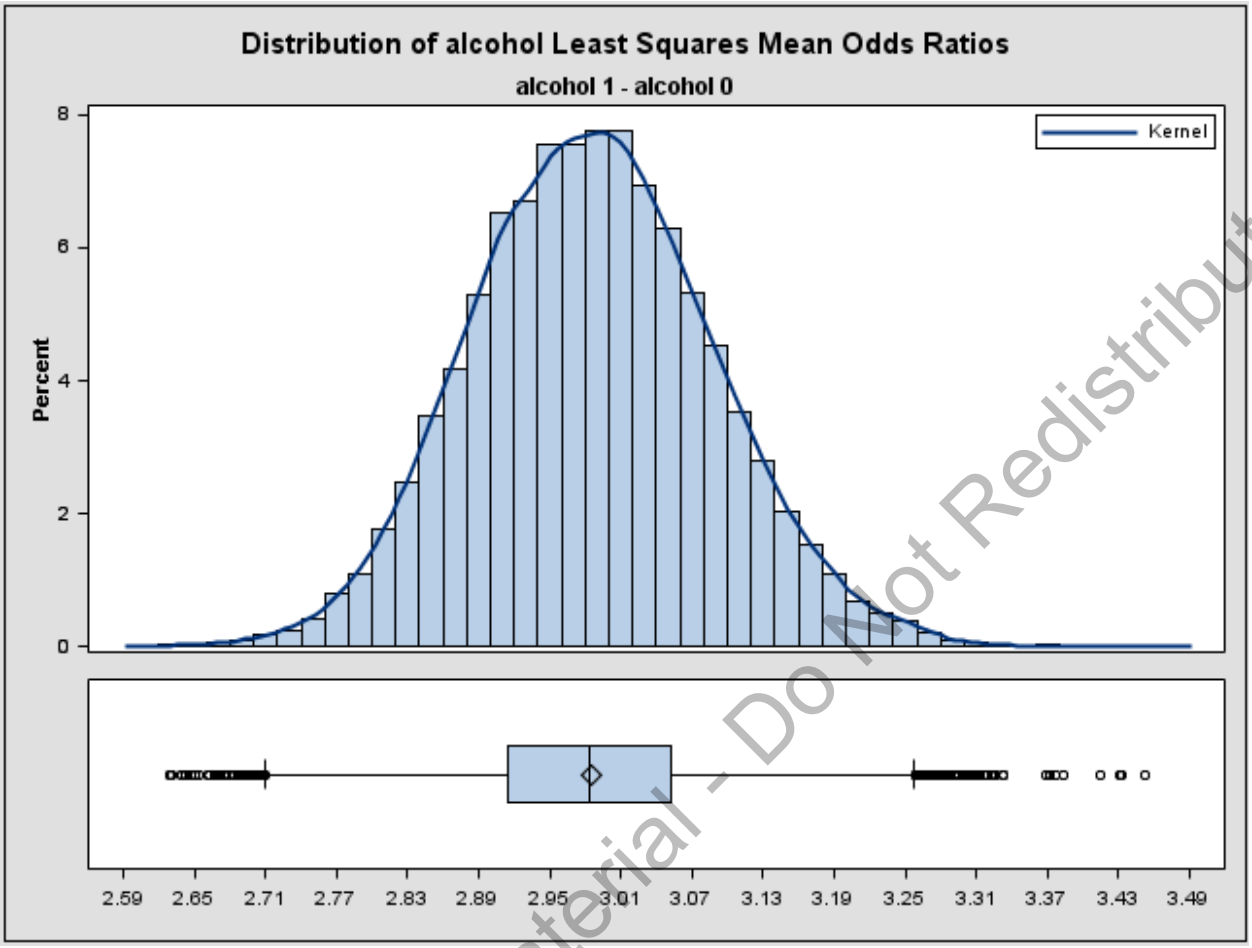
Did the mother drink during pregnancy?	Lower HPD	Upper HPD
1	-0.6505	0.05366
0	-1.7374	-1.0332

The LSMEANS statement produced output showing the least squares means of the posterior distribution for alcohol equaling 1 and alcohol equaling 0. The 25th, 50th, and 75th percentiles and the highest posterior density intervals are also shown.



ODS Graphics generated a histogram of the least squares means from the posterior distribution with a kernel density overlaid.

A box-plot is also created for the least squares means of the posterior distribution.



A histogram of the odds ratios from the posterior distribution is produced along with a box plot.

Sample Differences of alcohol Least Squares Means

Did the mother drink during pregnancy?	Did the mother drink during pregnancy?	N	Estimate	Standard Deviation	Percentiles 25th	Percentiles 50th	Percentiles 75th	Alpha
1	0	25000	1.0932	0.03390	1.0700	1.0932	1.1160	0.05

Sample Differences of alcohol Least Squares Means

Did the mother drink during pregnancy?	Did the mother drink during pregnancy?	Lower HPD	Upper HPD	Odds Ratio	Standard Error Odds Ratio	-Percentiles for Odds Ratios- 25th	50th	75th
1	0	1.0290	1.1612	2.986	0.1013	2.9153	2.9837	3.0525

Sample Differences of alcohol
Least Squares Means

Did the mother drink during pregnancy?	Did the mother drink during pregnancy?	Lower HPD of Odds Ratio	Upper HPD of Odds Ratio
1	0	2.789	3.184

The difference in the least squares means of the posterior distribution is shown with the related percentiles. With the ODDSRATIO option, the odds ratio is shown along with the related percentiles and the highest posterior density interval. The odds ratio of 2.986 and a highest posterior density interval of 2.789 and 3.184 correspond with the informative prior rather than the maximum likelihood estimates. This example shows the potential dominance of the informative prior.

Example: Create side-by-side histograms of the posterior density distributions with one based on the noninformative prior distribution and the other based on the informative prior distribution.

```
data plot;
   length plottype $ 14;
   set bayes_prob bayes_prob1(in=inform rename=(alcohol1=alcohol));
   if inform=1 then plottype="Informative";
   else plottype="Noninformative";
run;
```

Selected DATA step statement:

LENGTH defines the length of a character variable.

Selected DATA step option:

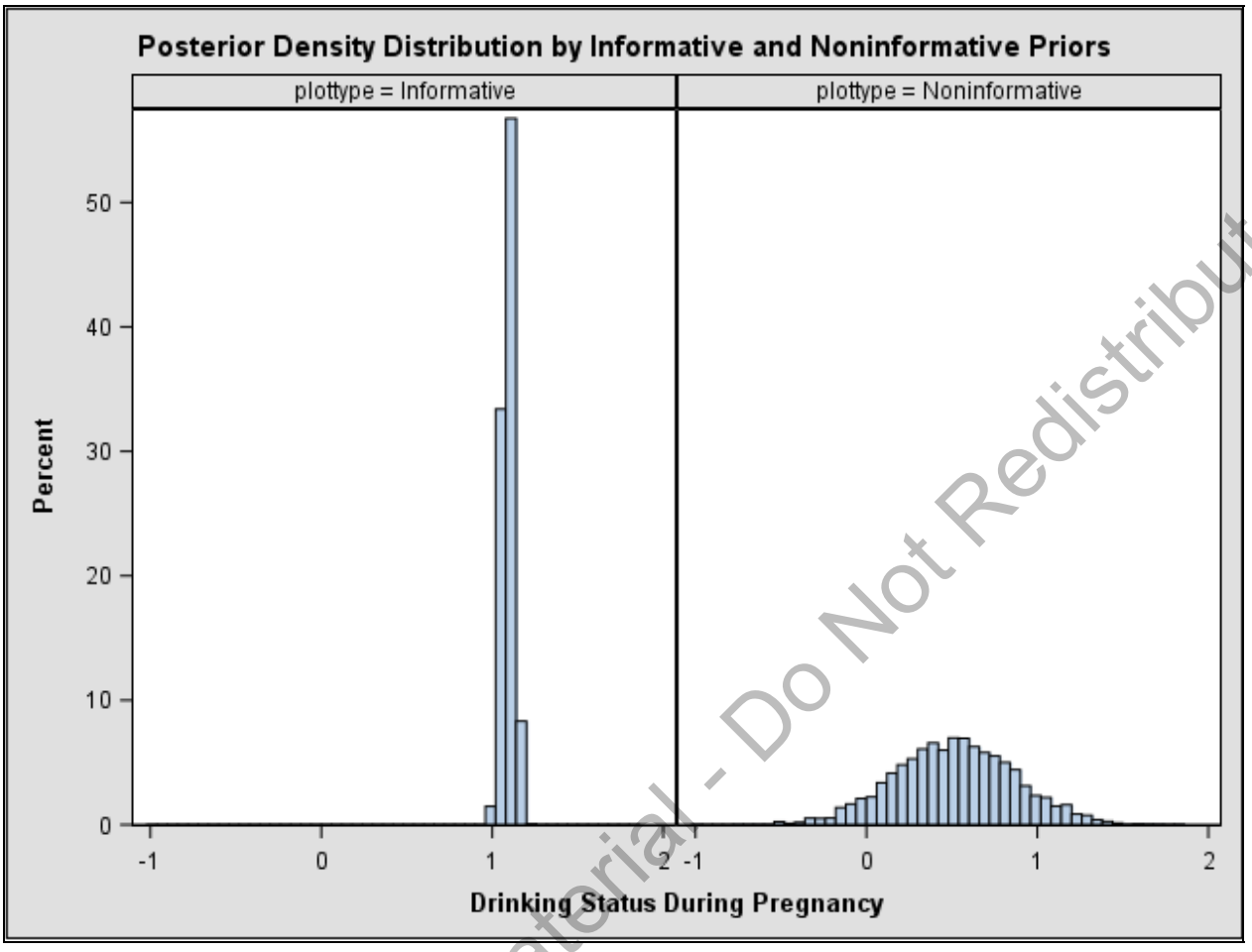
IN=variable detects whether the data set contributed to an observation when you read multiple SAS data sets in one DATA step. The specified variable is a temporary numeric variable with values of 0 (indicates that the data set did not contribute to the current observation) or 1 (indicates that the data set did contribute to the current observation).

```
proc sgpanel data=plot;
   panelby plottype;
   histogram alcohol;
   rowaxis label="Percent";
   colaxis label="Drinking Status During Pregnancy";
   title "Posterior Density Distribution by Informative and "
      "Noninformative Priors";
run;
```

Selected PROC SGPANEL statements:

PANELBY specifies one or more classification variables for the panel, the layout type, and other options for the panel.

HISTOGRAM creates a histogram that displays the frequency distribution of a numeric value.



The graphs show how an informative prior distribution can potentially dominate the data in the posterior distribution.

End of Demonstration

1.04 Multiple Choice Poll

Which of the following statements is true regarding Bayesian models?

a. If the ratio of the Monte Carlo standard errors to the posterior sample standard deviation is large, then only a fraction of the posterior variability is due to simulation.

b. When the effective sample size is much larger than the actual sample size, slower mixing of the Markov chain could be evident.

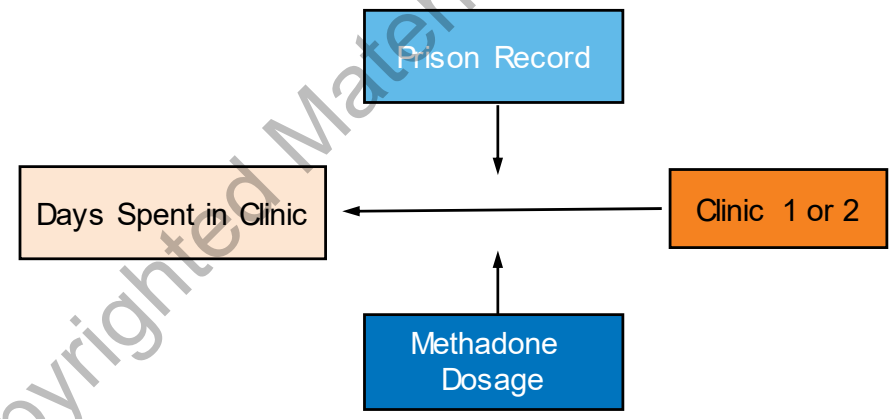
c. The Deviance Information Criterion can be used only to compare nested models.

d. If the Raftery-Lewis diagnostics indicate rejection, then one solution is to increase the number of Markov chain iterations.

68

sas

Methadone Treatment Data



70

sas

Example: A study was conducted to investigate the differences in survival experience between two clinics. The outcome variable is the number of days heroin addicts spent in a methadone clinic. The predictor variables that are believed to affect survival time are prison record and methadone dose. The data are stored in a SAS data set called **sasuser.methadone**.

These are the variables in the data set:

Clinic clinic (1 or 2)

Status survival status (0=censored, 1=departed from clinic)

Time survival time in days spent in clinic

Prison prison record (0=no, 1=yes)

Dose methadone dosage (mg/day).

Note: The data were obtained with permission from the OZDATA website. This website is a collection of data sets and is maintained in Australia. The study in which the data was collected is described in Caplehorn et al. (1991).

Bayesian Statistics with Informative Prior

Subject-Matter Knowledge: 10 mg/day increase in methadone dosage will change the hazard ratio from 0.50 (strong negative effect) to 1.01 (minor positive effect)

Bounds for Hazard Ratio:

$$0.50 < e^{10\hat{\beta}_{dose}} < 1.01$$

Bounds for Coefficient for Dose:

$$-0.069315 < \hat{\beta}_{dose} < 0.000995$$

71

sas

Information gathered from previous studies showed that a 10-unit increase in methadone dosage changed the hazard ratio from 0.50 (the rate of leaving the clinic decreased) to 1.01 (minor increase in the rate of leaving the clinic). Therefore, the plausible range that you believe the coefficient for dose can take ranges from -0.069315 to 0.000995 (take the log of the hazard ratio and divide by 10).

Bayesian Statistics with Informative Prior

For an informative normal prior, we need to estimate μ and σ^2. If you assume that $\mu \pm 2\sigma \approx (-0.069315, 0.000995)$, then

$$\mu = \frac{-0.069315 + 0.000995}{2} = -0.03416$$

$$\sigma^2 = \left(\frac{0.000995 - (-0.069315)}{1.96 * 2}\right)^2 = 0.0003217$$

72

If you assume a normal distribution where the majority of the prior distribution mass falls within the plausible range, the mean and variance can be computed using the formulas in the slide above. For the other parameters, you can use a normal prior distribution with a mean of 0 and a variance of 10^6.

Bayesian Analysis in PROC PHREG

Example: Generate a Bayesian analysis of the methadone data with a Cox proportional hazards model fit in PROC PHREG. Create a data set called **Prior** that specifies the means and variances of the prior distributions of the parameters. Specify **clinic** as a class variable and use reference cell coding with a reference cell of 2. Use the exact method for handling ties. Use the BAYES statement to specify a SAS data set containing the mean and covariance information of the prior distribution and to specify the random walk Metropolis algorithm, a thinning rate of 10, 200000 Markov chain iterations, all the posterior statistics, and to display a fitted penalized B-spline curve for each trace plot. Furthermore, compute the hazard ratios for each predictor variable (specify a ten-unit increase for **dose**).

```
/* stbay01d03.sas  */
data Prior;
   input _TYPE_ $ dose clinic1 prison;
datalines;
Mean -0.034160 0 0
Var .0003217 1e6 1e6
;
run;
```

The DATA step creates a data set called **Prior** with the estimated means and variances of the prior distributions of the parameters in the model. The variable **_TYPE_** is needed to identify the statistic.

```
proc phreg data=sasuser.methadone;
   class clinic (param=ref ref='2');
   model time*status(0)=clinic dose prison / ties=exact;
   bayes seed=27513 coeffprior=normal(input=Prior) diag=all
   plots(smooth)=all sampling=rwm thin=10 nmc=200000 statistics=all;
   hazardratio "HR1" clinic;
   hazardratio "HR2" dose / units=10;
   hazardratio "HR3" prison;
   title "Bayesian Analysis with Informative Prior for Methadone "
         "Data";
run;
```

Selected PHREG procedure statement:

HAZARDRATIO enables you to request hazard ratios for any variable in the model at customized settings. With the BAYES statement, the HAZARDRATIO statements will compute hazard ratios for the predictor variables based on the Bayesian analysis along with summaries of the posterior distribution of the corresponding hazard ratios.

Selected CLASS statement option:

PARAM= specifies the parameterization method for the categorical variable.

REF= specifies the reference level for the categorical variable.

Selected BAYES statement option:

COEFFPRIOR	specifies the prior distribution for the regression coefficients. The default is COEFFPRIOR=UNIFORM, and the normal prior is specified by COEFFPRIOR=NORMAL. The Zellner's *g* prior, a multivariate normal prior distribution, is specified using COEFFPRIOR=ZELLNER.
INPUT=	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. For an independent normal prior, the variances can be specified with *_TYPE_*='VAR'.
DIAG=	controls the number of diagnostics displayed.
NMC=	specifies the number of iterations after burn-in.
THIN=	controls the thinning of the Markov chain.
SAMPLING=	specifies the sampling algorithm used in the Markov chain Monte Carlo (MCMC) simulations. The two sampling algorithms that are available are the adaptive rejection Metropolis sampling (ARMS) and the random walk Metropolis (RWM). The default is ARMS.
STATISTICS=	controls the number of posterior statistics produced.

Selected global plot option:

SMOOTH	displays a fitted penalized B-spline curve for each trace plot.

```
Bayesian Analysis with Informative Prior for Methadone Data

                    The PHREG Procedure

                     Bayesian Analysis

                    Model Information

Data Set                  SASUSER.METHADONE
Dependent Variable        Time
Censoring Variable        Status
Censoring Value(s)        0
Model                     Cox
Ties Handling             EXACT
Sampling Algorithm        Metropolis
Burn-In Size              2000
MC Sample Size            200000
Thinning                  10


  Number of Observations Read         238
  Number of Observations Used         238
```

Class Level Information

Class	Value	Design Variables
Clinic	1	1
	2	0

Summary of the Number of Event and Censored Values

Total	Event	Censored	Percent Censored
238	150	88	36.97

Regression Parameter Information

Parameter	Effect	Clinic
Clinic1	Clinic	1
Dose	Dose	
Prison	Prison	

Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits	
Clinic1	1	1.0098	0.2149	0.5886	1.4310
Dose	1	-0.0354	0.00638	-0.0479	-0.0229
Prison	1	0.3266	0.1672	-0.00120	0.6543

Independent Normal Prior for Regression Coefficients

Parameter	Mean	Precision
Clinic1	0	1E-6
Dose	-0.03416	3108.486
Prison	0	1E-6

Initial Values of the Chains

Chain	Seed	Clinic1	Dose	Prison
1	27513	1.0098	-0.0354	0.3266
2	.	0.3651	-0.0545	-0.1751
3	.	1.6544	-0.0162	0.8283

Starting values (or initial values) can be specified in the INITIAL= data set in the BAYES statement. If INITIAL= option is not specified, PROC PHREG selects its own initial values for the chains. If the prior distribution of the parameter is proper, the starting values for the Markov chain are based on the estimated mean and standard deviation of the posterior distribution given the MLE.

Tuning History

Phase	Scale	Acceptance Rate
1	2.3800	0.3100

Burn-In History

Scale	Acceptance Rate
2.3800	0.3105

Sampling History

Scale	Acceptance Rate
2.3800	0.3120

The random walk Metropolis algorithm was used to sample an entire parameter vector from the posterior distribution. PROC PHREG uses a multivariate normal proposal distribution. One key factor in achieving high efficiency of a Metropolis-based Markov chain is finding a good proposal distribution. This process is referred to as *tuning*. PROC PHREG generates trial samples and automatically modifies the variance of the proposal distribution as a result of the acceptance rate. The acceptance probability is the percentage of candidate iterations in the proposal tuning phase that have been accepted.

The acceptance rate is closely related to the sampling efficiency of a Metropolis chain. For a random walk Metropolis, high acceptance rate means that most new samples occur right around the current data point. Their frequent acceptance means that the Markov chain is moving rather slowly and not exploring the parameter space fully. On the other hand, a low acceptance rate means that the proposed samples are often rejected. Hence, the chain is not moving much. An efficient Metropolis sampler has an acceptance rate that is neither too high nor too low. Roberts, Gelman, and Gilks (1997) showed that if both the target and proposal densities are normal, the optimal acceptance probability for the Markov chain should be around 0.45 in a single dimensional problem, and asymptotically approach 0.234 in higher dimensions.

Fit Statistics

DIC (smaller is better)	1338.390
pD (Effective Number of Parameters)	2.863

Posterior Summaries

Parameter	N	Mean	Standard Deviation	Percentiles 25%	50%	75%
Clinic1	20000	1.0288	0.2156	0.8818	1.0230	1.1731
Dose	20000	-0.0353	0.00598	-0.0394	-0.0352	-0.0312
Prison	20000	0.3266	0.1671	0.2130	0.3283	0.4394

Posterior Intervals

Parameter	Alpha	Equal-Tail Interval		HPD Interval	
Clinic1	0.050	0.6163	1.4614	0.6073	1.4503
Dose	0.050	-0.0469	-0.0237	-0.0468	-0.0236
Prison	0.050	-0.00169	0.6527	0.000528	0.6540

Posterior Covariance Matrix

Parameter	Clinic1	Dose	Prison
Clinic1	0.0465	0.0000	0.0052
Dose	0.0000	0.0000	-.0001
Prison	0.0052	-.0001	0.0279

Posterior Correlation Matrix

Parameter	Clinic1	Dose	Prison
Clinic1	1.0000	0.0194	0.1432
Dose	0.0194	1.0000	-.0919
Prison	0.1432	-.0919	1.0000

Posterior Autocorrelations

Parameter	Lag 1	Lag 5	Lag 10	Lag 50
Clinic1	0.1663	0.0089	-0.0068	-0.0060
Dose	0.1309	0.0040	-0.0034	-0.0108
Prison	0.1565	0.0038	-0.0010	-0.0023

Gelman-Rubin Diagnostics

Parameter	Estimate	97.5% Bound
Clinic1	1.0000	1.0000
Dose	1.0001	1.0000
Prison	1.0000	1.0000

Geweke Diagnostics

Parameter	z	Pr > |z|
Clinic1	0.5128	0.6081
Dose	-0.6413	0.5213
Prison	0.4591	0.6462

Raftery-Lewis Diagnostics

Quantile=0.025 Accuracy=+/-0.005 Probability=0.95 Epsilon=0.001

Parameter	Number of Samples Burn-In	Total	Minimum	Dependence Factor
Clinic1	3	4326	3746	1.1548
Dose	3	4112	3746	1.0977
Prison	3	4410	3746	1.1773

Heidelberger-Welch Diagnostics

	Stationarity Test			Half-Width Test			
Parameter	Cramer-von Mises Stat	p-Value	Test Outcome	Iterations Discarded	Half-Width	Mean	Relative Half-Width	Test Outcome
Clinic1	0.0411	0.9274	Passed	0	0.00355	1.0288	0.00346	Passed
Dose	0.2351	0.2087	Passed	0	0.000085	-0.0353	-0.00241	Passed
Prison	0.1666	0.3428	Passed	0	0.00265	0.3266	0.00812	Passed

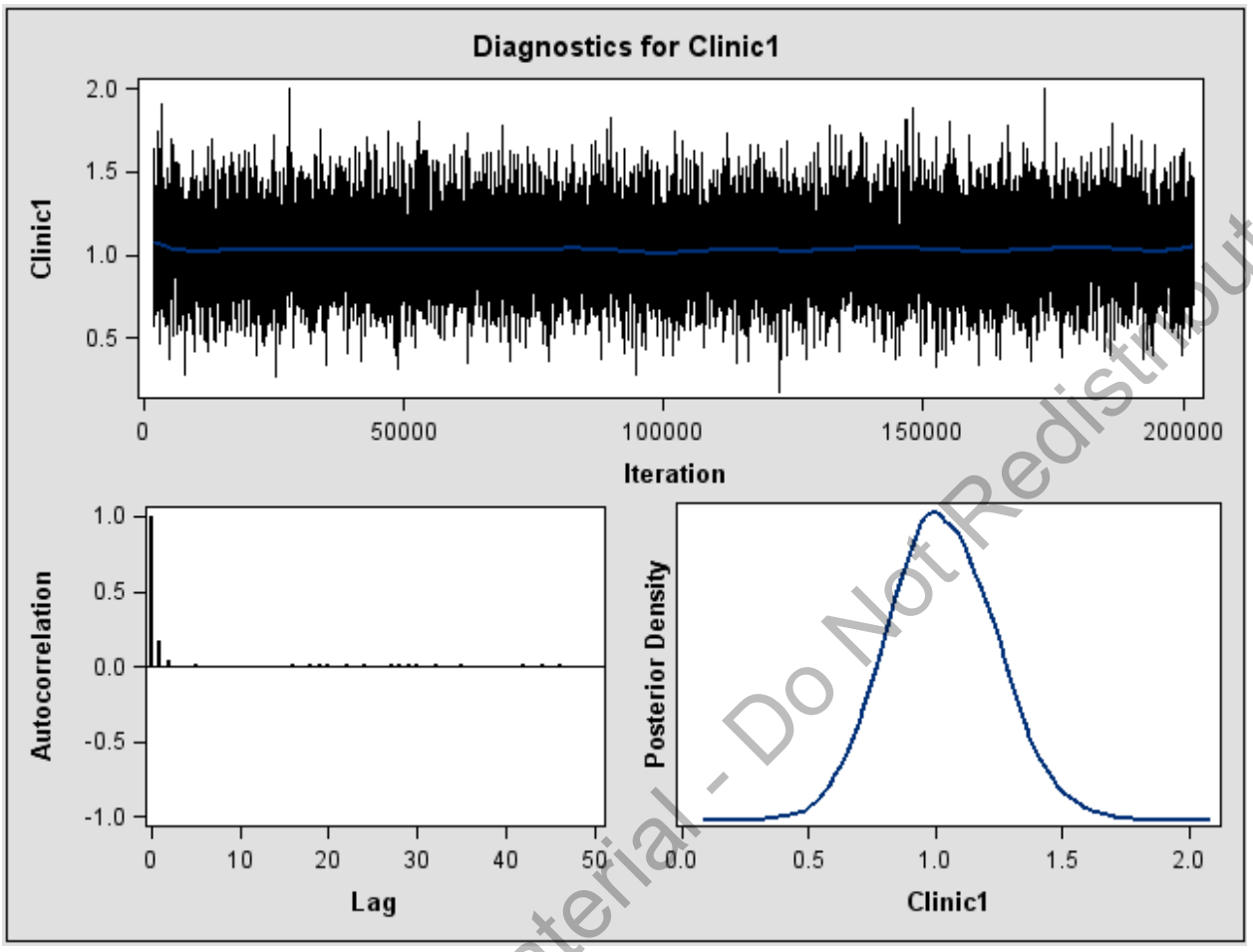
Effective Sample Sizes

Parameter	ESS	Autocorrelation Time	Efficiency
Clinic1	14274.4	1.4011	0.7137
Dose	15377.4	1.3006	0.7689
Prison	14355.8	1.3932	0.7178

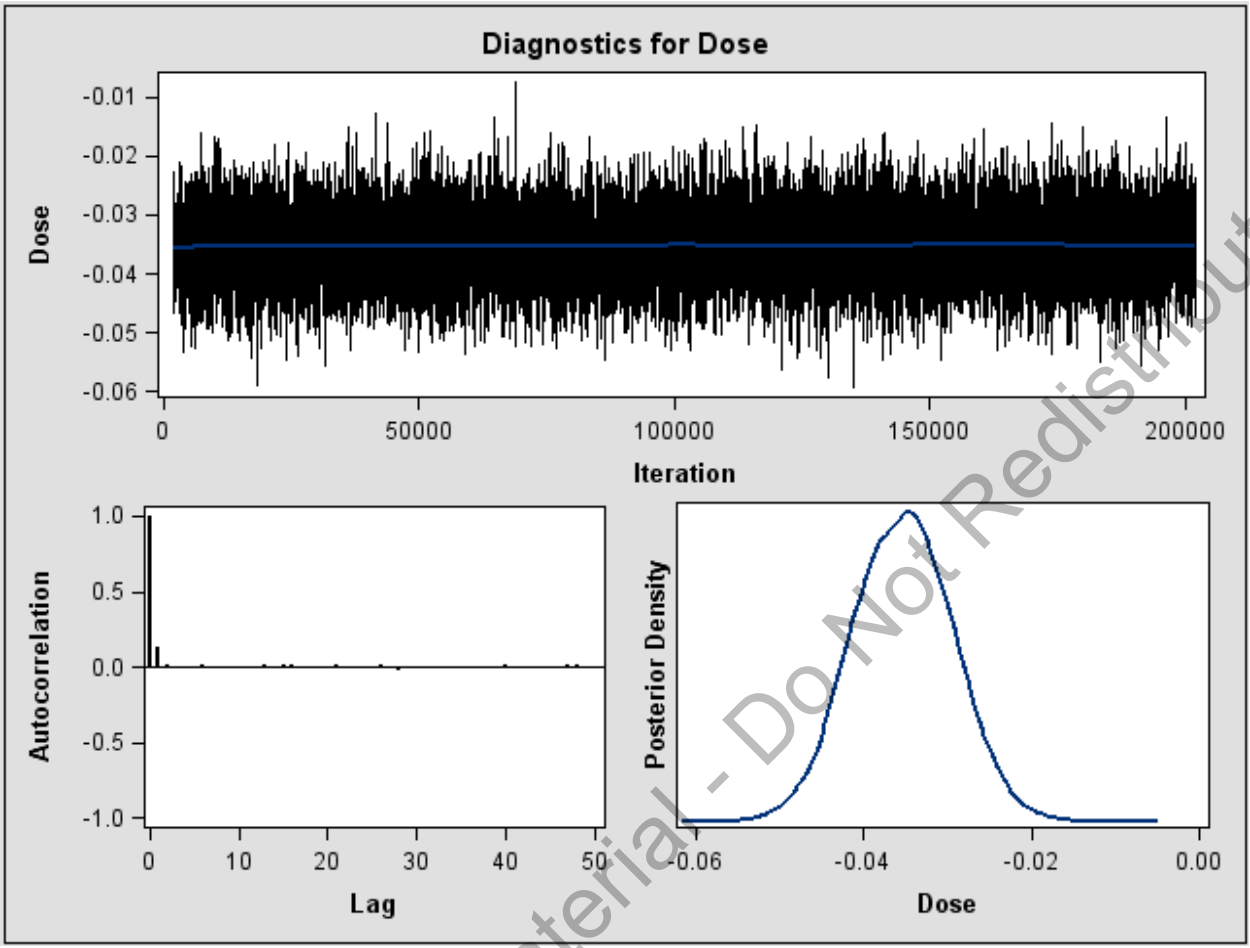
Monte Carlo Standard Errors

Parameter	MCSE	Standard Deviation	MCSE/SD
Clinic1	0.00180	0.2156	0.00837
Dose	0.000048	0.00598	0.00806
Prison	0.00139	0.1671	0.00835

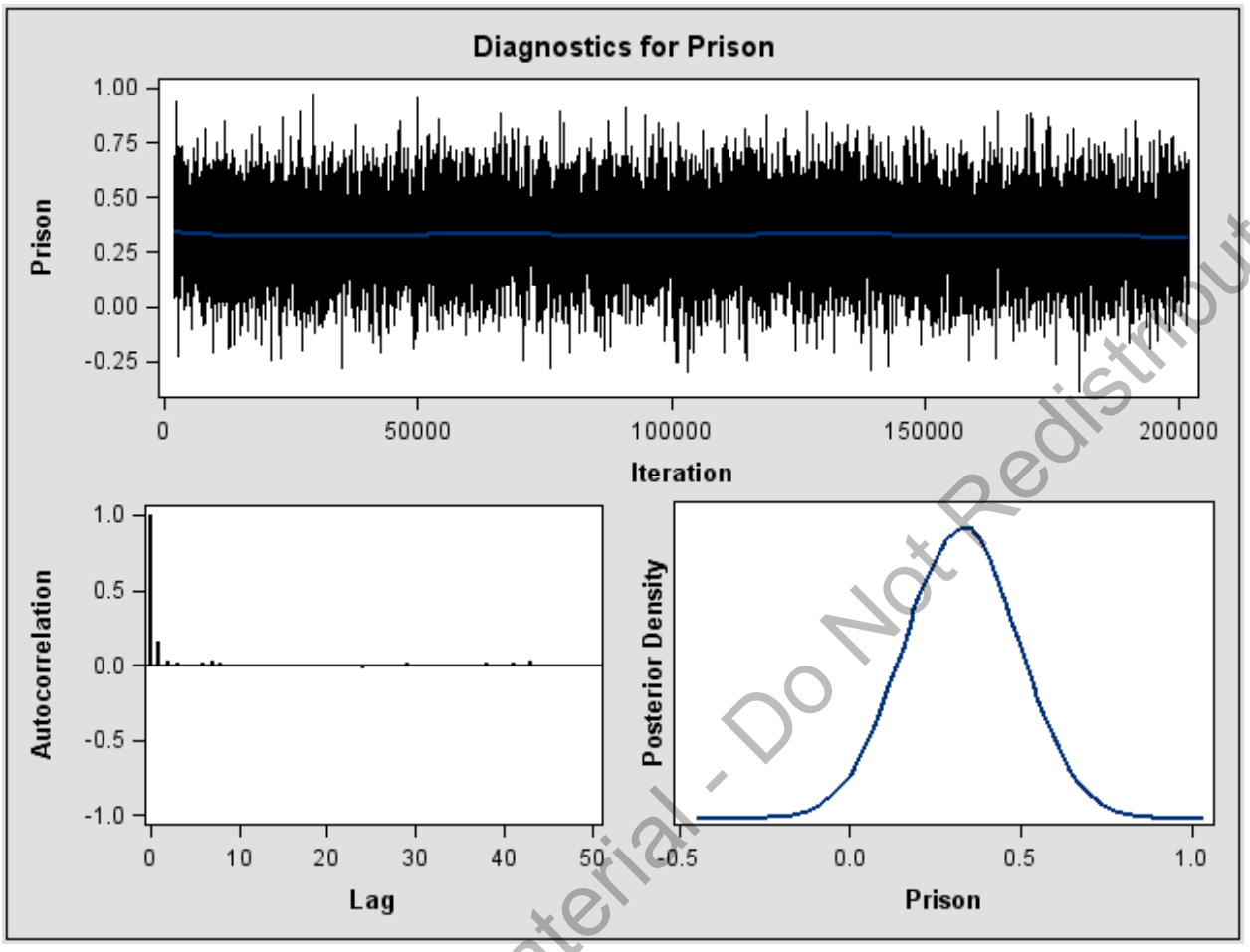
With the autocorrelations near 0 after lag 10, and the convergence diagnostic statistics not significant, the diagnostics indicate a reasonably good mixing of the Markov chain.



The diagnostic plots for **clinic1** show a converged Markov chain. The trace of the samples centers on 1.0 with a relatively constant mean and variance, and the autocorrelations are quite small. The distribution of the effect for **clinic1** shows that 0 is in the far left tail, so the effect for **clinic1** seems to be important.



The diagnostic plots for **dose** show a converged Markov chain. The trace of the samples centers on -0.035 with a relatively constant mean and variance, and the autocorrelations are quite small. The distribution of the effect for **dose** shows that 0 is in the far right tail, so the effect for **dose** seems to be important.



The diagnostic plots for **prison** show a converged Markov chain. The trace of the samples centers on 0.33 with a relatively constant mean and variance, and the autocorrelations are quite small. The distribution of the effect for **prison** shows that 0 is closer to the center of the distribution compared to **clinic1** and **dose**.

HR1: Hazard Ratios for Clinic

Description	N	Mean	Standard Deviation	Quantiles 25%	50%	75%	95% Equal-Tail Interval	
Clinic 1 vs 2	20000	2.8641	0.6329	2.4152	2.7816	3.2319	1.8521	4.3120

HR1: Hazard Ratios for Clinic

95% HPD Interval

1.7558 4.1341

The hazard ratio comparing clinic 1 to clinic 2 is 2.86 with a 95% credible interval of 1.85 to 4.31. In other words, there is a 95% chance that the hazard ratio is between 1.85 and 4.31.

HR2: Hazard Ratios for Dose

Description	N	Mean	Standard Deviation	Quantiles 25%	50%	75%	95% Equal-Tail Interval	
Dose Unit=10	20000	0.7040	0.0421	0.6745	0.7030	0.7319	0.6255	0.7891

HR2: Hazard Ratios for Dose

95% HPD Interval	
0.6223	0.7857

The hazard ratio for a 10-unit increase in methadone dosage is 0.70 and there is a 95% chance that the hazard ratio is between 0.63 and 0.79.

HR3: Hazard Ratios for Prison

Description	N	Mean	Standard Deviation	Quantiles 25%	50%	75%	95% Equal-Tail Interval	
Prison Unit=1	20000	1.4057	0.2362	1.2374	1.3886	1.5518	0.9983	1.9208

HR3: Hazard Ratios for Prison

95% HPD Interval	
0.9599	1.8652

The hazard ratio comparing patients with prison records to patients with no prison records is 1.41 and there is a 95% chance that the hazard ratio is between 0.998 and 1.921.

End of Demonstration

1.05 Multiple Choice Poll

Did the informative prior increase or decrease the posterior mean estimate for glucose?

a. The informative prior increased the posterior mean estimate for glucose.

b. The informative prior decreased the posterior mean estimate for glucose.

c. There was no change in the posterior mean estimate for glucose.

75

1.3 Chapter Summary

Bayesian analysis uses conditional probability to quantify the uncertainty of parameters of interest. In general, Bayesian statistical methods start with a prior distribution for all unknown parameters, updates this prior distribution in the light of the data (for example, using likelihood) to construct the posterior distribution, and then uses the posterior distribution for inferential decisions.

Bayesian methods offer alternatives to classical statistical inference. Instead of treating parameters as fixed constants, Bayesian methods treat parameters as random variables. These parameters cannot be determined exactly, and uncertainty about the parameter is expressed through probability statements
and distributions. Bayesian inference about the parameters is based on the probability distribution for the parameter.

The steps involved in Bayesian analysis include the following:

1. The probability distribution of the parameter, known as the prior distribution, is formulated.
2. Given the observed data, you choose a statistical model that describes the distribution of the data given the parameters.
3. You update your beliefs about the parameter by combining information from the prior distribution and the data through the calculation of the posterior distribution. This is carried out by using Bayes' theorem; hence the term Bayesian analysis.

Markov Chain Monte Carlo methods (MCMC) enable researchers to directly sample sequences of values from the posterior distribution of interest, foregoing the need for closed-form analytic solutions. With MCMC, you use these samples to estimate the posterior distribution's quantities of interest.

An important aspect of any Bayesian analysis is assessing the convergence of the Markov chains. Inferences based on non-converged Markov chains can be both inaccurate and misleading. 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 and plots help resolve these issues.

The GENMOD, LIFEREG, and PHREG procedures provide Bayesian analysis in addition to the standard frequentist analyses that they have always performed. For all three procedures, the new BAYES statement requests a Bayesian analysis.