Practical Bayesian Computation using SAS®

Fang Chen
SAS Institute Inc.
fangk.chen@sas.com

ASA Conference on Statistical Practices
February 20, 2014

Learning Objectives

Attendees will

- understand basic concepts and computational methods of Bayesian statistics
- be able to deal with some practical issues that arise from Bayesian analysis
- be able to program using SAS/STAT procedures with Bayesian capabilities to implement various Bayesian models.
1 Introduction to Bayesian statistics
- Background and concepts in Bayesian methods
- Prior distributions
- Computational Methods
  - Gibbs Sampler
  - Metropolis Algorithm
- Practical Issues in MCMC
  - Convergence Diagnostics

The GENMOD, PHREG, LIFEREG, and FMM Procedures
- Overview of Bayesian capabilities in the GENMOD, PHREG, LIFEREG, and FMM procedures
- Prior distributions
- The BAYES statement
- GENMOD: linear regression
- GENMOD: binomial model
- PHREG: Cox model
- PHREG: piecewise exponential model (optional)
Statistics and Bayesian Statistics

What is Statistics:
> the science of learning from data, which includes the aspects of collecting, analyzing, interpreting, and communicating uncertainty.

What is Bayesian Statistics:
> a subset of statistics in which all uncertainties are summarized through probability distributions.
The Bayesian Method

Given data \( x \), Bayesian inference is carried out in the following way:

1. You select a model (likelihood function) \( f(x|\theta) \) to describe the distribution of \( x \) given \( \theta \).

2. You choose a prior distribution \( \pi(\theta) \) for \( \theta \).

3. You update your beliefs about \( \theta \) by combining information from \( \pi(\theta) \) and \( f(x|\theta) \) and obtain the posterior distribution \( \pi(\theta|x) \).

The paradigm can be thought as a transformation from the before to the after:

\[
\pi(\theta) \rightarrow \pi(\theta|x)
\]

Bayes’ Theorem

The updating of beliefs is carried out by using Bayes’ theorem:

\[
\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\pi(x)} = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta) d\theta}
\]

The marginal distribution \( \pi(x) \) is an integral that is often ignored (as long as it is finite). Hence \( \pi(\theta|x) \) is often written as:

\[
\pi(\theta|x) \propto f(x|\theta)\pi(\theta) = \mathcal{L}(\theta)\pi(\theta)
\]

All inferences are based on the posterior distribution.
Two Different Paradigms

**Bayesian**
- Probability describes degree of belief, not limiting frequency. It is subjective.
- Parameters cannot be determined exactly. They are random variables, and you can make probability statements about them.
- Inferences about $\theta$ are based on the probability distribution for the parameter.

**Frequentist/Classical**
- Probabilities are objective properties of the real world. Probability refers to limiting relative frequencies.
- Parameters $\theta$ are fixed, unknown constants.
- Statistical procedures should be designed to have well-defined long-run frequency properties, such as the confidence interval.

---

1Wasserman 2004

---

Bayesian Thinking in Real Life

You suspect you might have a fever and decide to take your temperature.

1. A possible prior density on your temperature $\theta$: likely normal (centered at 98.6) but possibly sick (centered at 101).
2. Suppose the thermometer says 101 degrees: $f(x|\theta) \sim N(\theta, \sigma^2)$ where $\sigma$ could be a very small number.
3. You get the posterior distribution. Yes, you are sick.
Estimations

All inference about $\theta$ is based on $\pi(\theta|x)$.

- **Point:** mean, mode, median, any point from $\pi(\theta|x)$. For example, the posterior mean of $\theta$ is $E(\theta|x) = \int_\Theta \theta \cdot \pi(\theta|x) d\theta$
  The posterior mode of $\theta$ is the value of $\theta$ that maximizes $\pi(\theta|x)$.

- **Interval:** credible sets are any set $A$ such that $P(\theta \in A|x) = \int_A \pi(\theta|x) d\theta$
  - Equal tail: $100(\alpha/2)$th and $100(1-\alpha/2)$th percentiles.
  - Highest posterior density (HPD):
    - Posterior probability is $100(1-\alpha)$%
    - For $\theta_1 \in A$ and $\theta_2 \notin A$, $\pi(\theta_1|x) \geq \pi(\theta_2|x)$. The smallest region can be disjoint.

Interpretation: “There is a 95% chance that the parameter is in this interval.” The parameter is random, not fixed.

Prior Distributions

The prior distribution represents your belief *before* seeing the data.

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

- Not everyone agrees with the preceding. Some people would like to obtain results that are objectively valid, such as, “Let the data speak for itself.”. This approach advocates noninformative (flat/improper/Jeffreys) priors.

- Subjective approach advocates informative priors, which can be extraordinarily useful, if used correctly.

- Generally speaking, as the amount of data grows (in a model with fixed number of parameters), the likelihood overwhelms the impact of the prior.
Noninformative Priors

- A prior is *noninformative* if it is *flat* relative to the likelihood function. Thus, a prior $\pi(\theta)$ is noninformative if it has minimal impact on the posterior of $\theta$.

- Many people like noninformative priors because they appear to be more objective. However, it is unrealistic to think that noninformative priors represent total ignorance about the parameter of interest. See Kass and Wasserman (1996): JASA: 91:1343-1370.

- A frequent noninformative prior is $\pi(\theta) \propto 1$, which assigns equal likelihood to all possible values of the parameter.
  - However, flat prior is not invariant: flat on odds ratio is not the same as flat on log of odds ratio.

A Binomial Example

- Suppose that you observe 14 heads in 17 tosses. The likelihood is:
  $$L(p) \propto p^x(1-p)^{n-x}$$
  with $x = 14$ and $n = 17$.
- A flat prior on $p$ is:
  $$\pi(p) = 1$$
- The posterior distribution is:
  $$\pi(p|x) \propto p^{14}(1-p)^3$$
  which is a beta(15, 4).
Flat Prior (Observation I)
If $\pi(\theta|x) \propto L(\theta)$ with $\pi(\theta) \propto 1$, then why not use the flat prior all the time?

- Using a flat prior does not always guarantee a proper (integrable) posterior distribution; that is, $\int \pi(\theta|x) d\theta < \infty$.

The reason is that the likelihood function is only proper w.r.t. the random variable $X$. But a posterior has to be integrable w.r.t. $\theta$, a condition not required by the likelihood function.
Flat Prior (Observation II)

In cases where the likelihood function and the posterior distribution are identical, do we get the same answer?

Classical inference typically uses asymptotic results; Bayesian inference is based on exploring the entire distribution.

You Always Have to Defend Something!

In a sense, everyone (Bayesian and non-Bayesian) is a slave to the likelihood function, which serves as a foundation to both paradigms. Given that,

- in Bayesian paradigm, you need to justify the selection of your prior
- in classical paradigm, you need to justify asymptotics: there exists an infinitely amount of unobserved data that are just like the ones that you have seen.
Flat Prior (Observation III)

Is flat prior noninformative? Suppose that, in the binomial example, you choose to model on $\gamma = \text{logit}(p)$ instead of $p$:

$$\pi(p) = \text{uniform}(0, 1) \iff \pi(\gamma) = \text{logistic}(0, 1)$$

You start with

$$p = \frac{\exp(\gamma)}{1 + \exp(\gamma)} = \frac{1}{1 + \exp(-\gamma)}$$

$$\frac{\partial p}{\partial \gamma} = -\frac{\exp(-\gamma)}{(1 + \exp(-\gamma))^2}$$

Do the transformation of variables, with the Jacobian:

$$\pi(p) = 1 \cdot I\{0 \leq p \leq 1\}$$

$$\Rightarrow \pi(\gamma) = \left| \frac{\partial p}{\partial \gamma} \right| \cdot I\{0 \leq \frac{1}{1 + \exp(-\gamma)} \leq 1\} = \frac{\exp(-\gamma)}{(1 + \exp(-\gamma))^2} \cdot I\{-\infty \leq \gamma \leq \infty\}$$

The pdf for the logistic distribution with location $a$ and scale $b$ is

$$\exp\left(-\frac{\gamma - a}{b}\right) \left/ b \left(1 + \exp\left(-\frac{\gamma - a}{b}\right)\right)^2\right.$$

and $\pi(\gamma) = \text{logistic}(0, 1)$. 
Flat Prior (Observation III)

If you choose to be noninformative on the $\gamma$ dimension, you end up with a very different prior on the original $p$ scale:

$$\pi(\gamma) \propto 1 \iff \pi(p) \propto p^{-1}(1-p)^{-1}$$

**Flat Prior**

- A flat prior implies a unit, a measurement scale, on which you assign equal likelihood
  - $\pi(\theta) \propto 1$: $\theta$ is as likely to be between $(0, 1)$ as between $(1000, 1001)$
  - $\pi(\log(\theta)) \propto 1$ (equivalently, $\pi(\theta) \propto 1/\theta$): $\theta$ is as likely to be between $(1, 10)$ as between $(10, 100)$

- One obvious difficulty in justifying a flat (uniform) prior is to explain the choice of unit which the prior is being noninformative on.

- Can we have a prior that is somewhat noninformative but at the same time is invariant to transformations?
  - Jeffreys' Prior
Jeffreys’ Prior

Jeffreys’ prior is defined as

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

where $| \cdot |$ denotes the determinant and $I(\theta)$ is the expected Fisher information matrix based on the likelihood function $p(x|\theta)$:

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

In the Binomial Example:

$$\pi(p) \propto p^{-1/2}(1 - p)^{-1/2}$$

$$L(p)\pi(p) \propto p^{x-\frac{1}{2}}(1 - p)^{n-x-\frac{1}{2}}$$

$$\sim \text{Beta}(15.5, 4.5)$$

Some Thoughts

Jeffreys’ prior is

- *locally uniform*—a prior that does not change much over the region in which the likelihood is significant and does not assume large values outside that range. Hence it is somewhat noninformative.
- invariant with respect to one-to-one transformations.

The prior also

- can be improper for many models
- can be difficult to construct
- violates the likelihood principle
The Likelihood Principle

The likelihood principle states that, if two likelihood functions are proportional to each other,

\[ L_1(\theta|x) \propto L_2(\theta|x) \]

and one observes the same data \( x \), all inferences (about \( \theta \)) should be the same.

Jeffreys’ prior is in violation of this principle.

Negative Binomial Model

Instead of using a Binomial distribution, you can model the number of heads \( x = 14 \) using a negative binomial distribution:

\[ L(q) = \binom{r + x - 1}{x} q^r (1 - q)^x \]

- \( x \) is the number of failures until \( r = 3 \) successes are observed
- \( q \) is the probability of success (getting a tail), and \( 1 - q \) is the probability of failure (getting a head)
- let \( p = 1 - q \) and the likelihood function is rewritten as

\[ L(p) \propto (1 - p)^r p^x \]

This is the same kernel as the binomial likelihood function.
Jeffreys’ Prior

Same math leads to:

\[
\frac{\partial^2 \ell}{\partial p^2} = -\frac{x}{p^2} - \frac{r}{(1-p)^2}
\]

Under a negative binomial model, \( E(X) = \frac{rp}{1-p} \), and we have the following expected Fisher information:

\[
I(p) = \frac{-r}{p(1-p)^2}
\]

The Jeffreys’ prior becomes

\[
\pi(p) \propto p^{-1/2}(1-p)^{-1}
\]

\[
\sim \text{Beta}\left(\frac{1}{2}, 0\right)
\]

A different prior, a different posterior, different inference on \( p \).

The Cause

The cause to the problem is the expectation \( E(X) \), which depends on how the experiment is designed. In other words, taking the expectation means that we are making an assumption on how all future unobserved \( x \) behave.

Why do Bayesians consider this to be a problem?

- inference is based on yet-to-be-observed data and one might ended up being overly confident with the estimates.
Conjugate Prior

Conjugate prior is a family of prior distributions in which the prior and the posterior distributions are of the same family of distributions.

The Beta distribution is a conjugate prior to the binomial model:

$$\mathcal{L}(p) \propto p^x(1 - p)^{n-x}$$

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

The posterior distribution is also a Beta:

$$\pi(p|\alpha, \beta, x, n) \propto p^{x+\alpha-1}(1 - p)^{n-x+\beta-1}$$

$$= \text{Beta}(x + \alpha, n - x + \beta)$$

One nice feature of the conjugate prior is that you can easily understand the amount information that is contained in the prior:

- the data contains $x$ successes out of $n$ trials
- the prior assumes $\alpha$ successes out of $\alpha + \beta$ trials: Beta(2, 2) clearly means different from Beta(3, 17)

A related concept is the unit information (UI) prior (Kass and Wasserman (1995) JASA: 90:928-934), which is designed to contain roughly the same amount of information as one datum (variance equal to the inverse Fisher information based on one observation).
Bayesian Computation

- The key to Bayesian inferences is the posterior distribution
- Accurate estimation of the posterior distribution can be difficult and require a considerate amount of computation
- One of the most prevalent methods used nowadays is simulation-based:
  - repeatedly draw samples from a target distribution and use the collection of samples to empirically approximate the posterior

Simulation-based Estimation

How to do this for complex models that have many parameters?
Markov Chain Monte Carlo

- **Markov Chain**: a stochastic process that generates conditional independent samples according to some target distribution.

- **Monte Carlo**: a numerical integration technique that finds an expectation:

\[
E(f(\theta)) = \int f(\theta)p(\theta)d\theta \approx \frac{1}{n} \sum_{i=1}^{n} f(\theta_i)
\]

with \(\theta_1, \theta_2, \cdots, \theta_n\) being samples from \(p(\theta)\).

- **MCMC** is a method that generates a sequence of dependent samples from the target distribution and computes quantities by using Monte Carlo based on these samples.

Gibbs Sampler

**Gibbs sampler** is an algorithm that sequentially generates samples from a joint distribution of two or more random variables. The sampler is often used when:

- The joint distribution, \(\pi(\theta|x)\), is not known explicitly
- The full conditional distribution of each parameter—for example, \(\pi(\theta_i|\theta_j, i \neq j, x)\)—is known
Gibbs Sampler

\[
\pi(\theta=(\alpha,\beta) \mid x)
\]

\begin{align*}
\pi(\beta \mid \alpha^{(0)}, x) & \\
\pi(\alpha \mid \beta^{(0)}, x) & 
\end{align*}

\[
\alpha^{(0)}
\]

\[
\beta
\]
Gibbs Sampler

\[ \pi(\theta=(\alpha, \beta)|x) \]

\[ \pi(\beta|\alpha^{(0)}, x) \]

\[ \alpha^{(0)} \]

\[ \beta^{(0)} \]
Gibbs Sampler
Gibbs Sampler
Joint and Marginal Distributions

\[ \pi(\alpha, \beta | \mathbf{x}) \]
-1.1634  -1.0416
-1.0666  0.1060
4.0456   4.3485
-0.2531  0.9495
-0.4675  -0.8278
1.5945   2.5618
-3.1744  -4.5091
-0.3086  -2.6782
:        :

Gibbs enables you draw samples from a joint distribution.

The by-products are the marginal distributions.
Joint and Marginal Distributions

The by-products are the marginal distributions.

Gibbs Sampler

The difficulty in implementing a Gibbs sampler is how to efficiently generate from the conditional distribution, $\pi(\theta_i|\theta_j, i \neq j, x)$?

If each conditional distribution is a well known distribution, then it is easy. Otherwise, you must use general algorithms to generate samples from a distribution:

- Metropolis Algorithm
- Adaptive Rejection Algorithm
- Slice Sampler
- ...

General algorithms typically have minimum requirements that are not distribution-specific, such as the ability to evaluate the objective functions.
The Metropolis Algorithm

1. Let $t = 0$. Choose a starting point $\theta^{(t)}$. This can be an arbitrary point as long as $\pi(\theta^{(t)}|y) > 0$.
2. Generate a new sample, $\theta'$, from a proposal distribution $q(\theta'|\theta^{(t)})$.
3. Calculate the following quantity:
   $$ r = \min \left\{ \frac{\pi(\theta'|y)}{\pi(\theta^{(t)}|y)}, 1 \right\} $$
4. Sample $u$ from the uniform distribution $U(0, 1)$.
5. Set $\theta^{(t+1)} = \theta'$ if $u < r$; $\theta^{(t+1)} = \theta^{(t)}$ otherwise.
6. Set $t = t + 1$. If $t < T$, the number of desired samples, go back to Step 2; otherwise, stop.

The Random-Walk Metropolis Algorithm
The Random-Walk Metropolis Algorithm

\[ \pi(\theta | x) \]

\[ \theta' \sim N(\theta^{(0)}, \sigma) \]

\[ \theta', \theta^{(0)} \]

\[ \pi(\theta' | x) \]

\[ \pi(\theta^{(0)} | x) \]
The Random-Walk Metropolis Algorithm

If $\pi(\theta'|x) > \pi(\theta^{(0)}|x)$, then $\theta^{(1)} = \theta'$

If $\pi(\theta'|x) < \pi(\theta^{(0)}|x)$, accept $\theta'$ with prob $\pi(\theta'|x)/\pi(\theta^{(0)}|x)$
The Random-Walk Metropolis Algorithm

\[ \pi(\theta | x) \]

\[ \theta' \sim N(\theta^{(1)}, \sigma) \]

\[ \theta^{(1)} \quad \theta' \quad \theta^{(0)} \]

The Markov chain always move to areas that have higher density.
The Random-Walk Metropolis Algorithm

can still explore tail areas with lower density

Scale and Mixing in the Metropolis
Markov Chain Convergence

An unconverged Markov chain does not explore the parameter space efficiently and the samples cannot approximate the target distribution well. Inference should not be based upon unconverged Markov chain, or very misleading results could be obtained.

It is important to remember:

- Convergence should be checked for ALL parameters, and not just those of interest.
- There are no definitive tests of convergence. Diagnostics are often not sufficient for convergence.

Convergence Terminology

- **Convergence**: initial drift in the samples towards a stationary (target) distribution
- **Burn-in**: samples at start of the chain that are discarded to minimize their impact on the posterior inference
- **Slow mixing**: tendency for high autocorrelation in the samples. A slow-mixing chain does not traverse the parameter space efficiently.
- **Thinning**: the practice of collecting every $k$th iteration to reduce autocorrelation. Thinning a Markov chain can be wasteful because you are throwing away a $\frac{k-1}{k}$ fraction of all the posterior samples generated.
- **Trace plot**: plot of sampled values of a parameter versus iteration number.
### Various Trace Plots

<table>
<thead>
<tr>
<th>Good Mixing</th>
<th>Burn-In</th>
<th>Nonconvergence</th>
<th>Thinning?</th>
</tr>
</thead>
</table>

### To Thin Or Not To Thin?

The argument for thinning is based on reducing autocorrelations, getting from

![Trace Plot](image1.png)

![Autocorrelation Plot](image2.png)

to

![Trace Plot](image3.png)

![Autocorrelation Plot](image4.png)
To Thin Or Not To Thin?

But at the same time, you are getting from

![Graph showing autocorrelation before thinning](image1)

![Graph showing autocorrelation after thinning](image2)

To

![Graph showing autocorrelation before thinning](image3)

![Graph showing autocorrelation after thinning](image4)

...to

To Thin Or Not To Thin?

Thinning reduces autocorrelations and allows one to obtain seemingly independent samples. But at the same time, you throw away an appalling number of samples that can otherwise be used.

Autocorrelations do not lead to biased Monte Carlo estimates. It is simply an indicator of poor sampling efficiency.

On the other hand, sub-sampling loses information and actually increases the variance of sample mean estimators ($\text{Var}(\bar{\theta})$, not posterior variance). See MacEachern and Berliner (1994, *American Statistician*, 48:188).

Advice: unless storage becomes a problem, you are better off keeping all the samples for estimation.
Some Popular Convergence Diagnostics Tests

- Gelman-Rubin: tests whether multiple chains would convergent to the same target distribution.
- Geweke: tests whether the mean estimates have converged by comparing means from the early and latter part of the Markov chain.
- Heidelberger-Welch stationarity test: tests whether the Markov chain is a covariance (weakly) stationary process.
- Heidelberger-Welch halfwidth test: reports whether the sample size is adequate to meet the required accuracy for the mean estimate.
- Raftery-Lewis: evaluates the accuracy of the estimated (desired) percentiles by reporting the number of samples needed to reach the desired accuracy of the percentiles.

More on Convergence Diagnosis

There are no definitive tests of convergence.

- With experience, visual inspection of trace plots is often the most useful approach.
- Geweke and Heidelberger-Welch sometimes reject 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.
- ESS is frequently a good numerical indicator on the status of mixing.
Effective Sample Size (ESS)

ESS (Kass et al. 1998, *American Statistician*, 52:93) provides a measure on how well a Markov chain is mixing.

\[
\text{ESS} = \frac{n}{1 + 2 \sum_{k=1}^{n-1} \rho_k(\theta)}
\]

where \( n \) is the total sample size and \( \rho_k(\theta) \) is the autocorrelation of lag \( k \) for \( \theta \).

- The closer ESS is to \( n \), the better mixing is in the Markov chain.
- ESS of size around 1,000 is mostly sufficient in estimating the posterior density. You want increase the number for tail percentiles.

I personally prefer to use ESS as a way to judge convergence:

- small numbers of ESSs often indicate “something isn’t quite right.”
- large numbers of ESSs are typically good news
- moves away from the conundrum of dealing with and interpreting hypothesis testing results
- You can summarizes the convergence of multiple parameters by looking at the distribution of all the ESSs, or even the minimum ESS (worst case).
Various Trace Plots and ESSs

<table>
<thead>
<tr>
<th>Trace Plot</th>
<th>ESS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>19431.1</td>
</tr>
<tr>
<td></td>
<td>1919.8</td>
</tr>
<tr>
<td></td>
<td>447.6</td>
</tr>
<tr>
<td></td>
<td>231.0</td>
</tr>
<tr>
<td></td>
<td>159.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Trace Plot</th>
<th>ESS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>87.6</td>
</tr>
<tr>
<td></td>
<td>71.5</td>
</tr>
<tr>
<td></td>
<td>36.4</td>
</tr>
<tr>
<td></td>
<td>31.3</td>
</tr>
<tr>
<td></td>
<td>24.2</td>
</tr>
<tr>
<td></td>
<td>22.3</td>
</tr>
</tbody>
</table>
More on ESS

- ESS is not significance test-based, and you can think of it as more of a numerical criterion, similar to convergence criteria used in optimizations.

- You can still get good ESSs in “unconverged” chains, such as a chain that is stuck in a local mode in a multi-mode problem.
  - These are fairly rare (and often there are plenty of other signs to indicate such complex problems).

- Bad ESSs serves as a good indicator when things go bad
  - problems can sometimes be easily corrected (burn-in, longer chain, etc).
  - false rejections (bad ESSs from converged chains) are less common, but do exist (in binary and discrete parameters).

Bernoulli Markov Chains, all with Marginal Prob of 0.2

<table>
<thead>
<tr>
<th>ESS</th>
<th>2133.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESS</td>
<td>131.7</td>
</tr>
<tr>
<td>ESS</td>
<td>5.3</td>
</tr>
<tr>
<td>ESS</td>
<td>6.5</td>
</tr>
</tbody>
</table>
Outline of Part II

- Overview of Bayesian capabilities in the GENMOD, PHREG, LIFEREG, and FMM procedures
- Overview of the BAYES statement and syntax for requesting Bayesian analysis
- Examples
  - GENMOD: linear regression
  - GENMOD: Poisson regression
  - PHREG: Cox model
  - PHREG: piecewise exponential model (optional)

The GENMOD, PHREG, LIFEREG, and FMM Procedures

These four procedures provide:

- The BAYES statement
- A set of frequently used prior distributions (noninformative, Jeffreys'), posterior summary statistics, and convergence diagnostics
- Various sampling algorithms: conjugate, direct, adaptive rejection (Gilks and Wild 1992; Gilks, Best, and Tan 1995), Metropolis, Gamerman algorithm, etc.

Bayesian capabilities include:

- GENMOD: Generalized Linear Models
- LIFEREG: Parametric Lifetime Models
- PHREG: Cox Regression (Frailty) and Piecewise Exponential Models
- FMM: Finite Mixture Models
Prior Distributions in SAS Procedures

- **Uniform** (or flat) prior is defined as:
  \[ \pi(\theta) \propto 1 \]
  This prior is not integrable, but it does not lead to improper posterior in any of the procedures.

- **Improper** prior is defined as:
  \[ \pi(\theta) \propto \frac{1}{\theta} \]
  This prior is often used as a noninformative prior on the scale parameter, and it is uniform on the log-scale.

- **Proper** prior distributions include gamma, inverse-gamma, AR(1)-gamma, normal, multivariate normal densities.

- **Jeffreys’** prior is provided in PROC GENMOD.

### Syntax for the BAYES Statement

The BAYES statement is used to request all Bayesian analysis in these procedures.

```
BAYES < options > ;
```

The following options appear in all BAYES statements:

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>INITIAL=</td>
<td>initial values of the chain</td>
</tr>
<tr>
<td>NBI=</td>
<td>number of burn-in iterations</td>
</tr>
<tr>
<td>NMC=</td>
<td>number of iterations after burn-in</td>
</tr>
<tr>
<td>OUTPOST=</td>
<td>output data set for posterior samples</td>
</tr>
<tr>
<td>SEED=</td>
<td>random number generator seed</td>
</tr>
<tr>
<td>THINNING=</td>
<td>thinning of the Markov chain</td>
</tr>
<tr>
<td>DIAGNOSTICS=</td>
<td>convergence diagnostics</td>
</tr>
<tr>
<td>PLOTS=</td>
<td>diagnostic plots</td>
</tr>
<tr>
<td>SUMMARY=</td>
<td>summary statistics</td>
</tr>
<tr>
<td>COEFFPRIOR=</td>
<td>prior for the regression coefficients</td>
</tr>
</tbody>
</table>
Consider the model

\[ Y = \beta_0 + \beta_1 \log X_1 + \epsilon \]

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

The default priors that PROC GENMOD uses are:

\[
\begin{align*}
\pi(\beta_0) & \propto 1 \\
\pi(\beta_1) & \propto 1 \\
\pi(\sigma^2) & \sim \text{gamma}(\text{shape} = 2.001, \text{iscale} = 0.0001)
\end{align*}
\]

Regression Example

A subset of the data and statements fit Bayesina regression:

```plaintext
data surg;
  input logy logx1 @@;
datalines;
  199.986 1.90211 100.995 1.62924 203.986 2.00148
  100.995 1.87180 508.979 2.05412 80.002 1.75786
  ...
;
proc genmod data=surg;
  model y = logx1 / dist=normal link=identity;
  bayes seed=4 outpost=post diagnostics=all summary=all;
run;
```

- **SEED** specifies a random seed
- **OUTPOST** saves posterior samples
- **DIAGNOSTICS** requests all convergence diagnostics
- **SUMMARY** requests calculation for all posterior summary statistics
Convergence Diagnostics for $\beta_1$

Mixing
The following are the autocorrelation and effective sample sizes. The mixing appears to be very good, which agrees with the trace plots.

**Bayesian Analysis**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Lag 1</th>
<th>Lag 5</th>
<th>Lag 10</th>
<th>Lag 50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.0062</td>
<td>0.0105</td>
<td>0.0244</td>
<td>-0.0003</td>
</tr>
<tr>
<td>logx1</td>
<td>0.0045</td>
<td>0.0106</td>
<td>0.0269</td>
<td>0.0009</td>
</tr>
<tr>
<td>Dispersion</td>
<td>-0.0077</td>
<td>0.0116</td>
<td>0.0082</td>
<td>-0.0003</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>ESS</th>
<th>Autocorrelation Time</th>
<th>Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>10000.0</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>logx1</td>
<td>10000.0</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>Dispersion</td>
<td>10000.0</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
</tbody>
</table>
Additional Convergence Diagnostics

**Bayesian Analysis**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>97.5% Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.0000</td>
<td>1.0002</td>
</tr>
<tr>
<td>logx1</td>
<td>1.0000</td>
<td>1.0002</td>
</tr>
<tr>
<td>Dispersion</td>
<td>0.9999</td>
<td>0.9999</td>
</tr>
</tbody>
</table>

**Raftery-Lewis Diagnostics**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Number of Samples</th>
<th>Dependence Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Burn-in</td>
<td>Total</td>
<td>Minimum</td>
</tr>
<tr>
<td>Intercept</td>
<td>3789</td>
<td>3746</td>
</tr>
<tr>
<td>logx1</td>
<td>3834</td>
<td>3746</td>
</tr>
<tr>
<td>Dispersion</td>
<td></td>
<td>3746</td>
</tr>
</tbody>
</table>

**Bayesian Analysis**

| Parameter | z     | Pr > |z| |
|-----------|------|------|---|
| Intercept | 1.0623 | 0.2881 |
| logx1     | -1.0554 | 0.2912 |
| Dispersion| 0.6388 | 0.5229 |

**Heidelberger-Welch Diagnostics**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Stationarity Test</th>
<th>Half-width Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cramer-von-</td>
<td>Half-width</td>
</tr>
<tr>
<td></td>
<td>Mises Stat</td>
<td>Mean</td>
</tr>
<tr>
<td></td>
<td>p Test Outcome</td>
<td>Relative</td>
</tr>
<tr>
<td></td>
<td>Iterations</td>
<td>Half-width</td>
</tr>
<tr>
<td></td>
<td>Discarded</td>
<td>Test Outcome</td>
</tr>
<tr>
<td>Intercept</td>
<td>0.0587 0.8223</td>
<td>2.3604 -94.5279</td>
</tr>
<tr>
<td>logx1</td>
<td>0.0611 0.8069</td>
<td>1.4139 169.9</td>
</tr>
<tr>
<td>Dispersion</td>
<td>0.1055 0.5585</td>
<td>67.9392 18478.4</td>
</tr>
</tbody>
</table>
Summarize Convergence Diagnostics

- **Autocorrelation**: shows low dependency among Markov chain samples
- **ESS**: values close to the sample size indicate good mixing
- **Gelman-Rubin**: values close to 1 suggest convergence from different starting values
- **Geweke**: indicates mean estimates are stabilized
- **Raftery-Lewis**: shows sufficient samples to estimate 0.025 percentile within $+/- 0.005$ accuracy
- **Heidelberger-Welch**: suggests the chain has reached stationarity and there are enough samples to estimate the mean accurately

---

Posterior Summary and Interval Estimates

*Bayesian Analysis*

<table>
<thead>
<tr>
<th>Parameter</th>
<th>N</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>10000</td>
<td>-94.5279</td>
<td>119.1</td>
<td>-172.9</td>
<td>-95.0444</td>
<td>-16.1862</td>
</tr>
<tr>
<td>logx1</td>
<td>10000</td>
<td>169.9</td>
<td>68.5847</td>
<td>124.6</td>
<td>170.2</td>
<td>214.4</td>
</tr>
<tr>
<td>Dispersion</td>
<td>10000</td>
<td>18478.4</td>
<td>3670.4</td>
<td>15825.8</td>
<td>17987.2</td>
<td>20596.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Alpha</th>
<th>Equal-Tail Interval</th>
<th>HPD Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.050</td>
<td>-331.1</td>
<td>-320.5</td>
</tr>
<tr>
<td>logx1</td>
<td>0.050</td>
<td>35.3949</td>
<td>35.3925</td>
</tr>
<tr>
<td>Dispersion</td>
<td>0.050</td>
<td>12646.1</td>
<td>11957.9</td>
</tr>
</tbody>
</table>
Posterior Inference

Posterior correlation:

### Bayesian Analysis

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Intercept</th>
<th>logx1</th>
<th>Dispersion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.000</td>
<td>-0.987</td>
<td>-0.007</td>
</tr>
<tr>
<td>logx1</td>
<td>-0.987</td>
<td>1.000</td>
<td>0.006</td>
</tr>
<tr>
<td>Dispersion</td>
<td>-0.007</td>
<td>0.006</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Fit Statistics

PROC GENMOD also calculates the Deviance Information Criterion (DIC)

### Bayesian Analysis

<table>
<thead>
<tr>
<th>Fit Statistics</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>DIC (smaller is better)</td>
<td>690.182</td>
</tr>
<tr>
<td>pD (effective number of parameters)</td>
<td>3.266</td>
</tr>
</tbody>
</table>
Posterior Probabilities

Suppose that you are interested in knowing whether $\log(X_1)$ has a positive effect on survival time. Quantifying that measurement, you can calculate the probability $\beta_1 > 0$, which can be estimated directly from the posterior samples:

$$Pr(\beta_1 > 0 | Y, \log(X_1)) = \frac{1}{N} \sum_{t=1}^{N} I(\beta_1^t > 0)$$

where $I(\beta_1^t > 0) = 1$ if $\beta_1^t > 0$ and 0 otherwise. $N = 10,000$ is the sample size in this example.

The following SAS statements calculate the posterior probability:

```sas
data Prob;
    set Post;
    Indicator = (logX1 > 0);
    label Indicator='log(Blood Clotting Score) > 0';
run;
ods select summary;
proc means data = Prob(keep=Indicator) n mean; run;
```

The probability is roughly 0.9926, which strongly suggests that the slope coefficient is greater than 0.
Binomial model

Consider a study of the analgesic effects of treatments on elderly patients with neuralgia.

- Two test treatments and a placebo are compared.
- The response variable is whether the patient reported pain or not.
- Covariates include the age and gender of 60 patients and the duration of complaint before the treatment began.
The Data

A subset of the data:

```
Data Neuralgia;
   input Treatment $ Sex $ Age Duration Pain $ @@;
   datalines;
   P F 68 1 No B M 74 16 No P F 67 30 No
   P M 66 26 Yes B F 67 28 No B F 77 16 No
   A F 71 12 No B F 72 50 No B F 76 9 Yes
   ...
   P M 67 17 Yes B M 70 22 No A M 65 15 No
   P F 67 1 Yes A M 67 10 No P F 72 11 Yes
   A F 74 1 No B M 80 21 Yes A F 69 3 No
```

Treatment: A, B, P

Sex: F, M

Pain: Yes, No

The Model

A logistic regression is considered for this data set:

\[
pain_i \sim \text{binary}(p_i)
\]

\[
p_i = \logit(\beta_0 + \beta_1 \cdot \text{Sex}_{F,i} + \beta_2 \cdot \text{Treatment}_{A,i} +
               \beta_3 \cdot \text{Treatment}_{B,i} + \beta_4 \cdot \text{Sex}_{F,i} \cdot \text{Treatment}_{A,i} +
               \beta_5 \cdot \text{Sex}_{F,i} \cdot \text{Treatment}_{B,i} + \beta_6 \cdot \text{Age} + \beta_7 \cdot \text{Duration})
\]

where \text{Sex}_{F}, \text{Treatment}_{A}, \text{and} \text{Treatment}_{B} are dummy variables for the categorical predictors.

You might want to consider a normal prior with large variance as a noninformative prior distribution on all the regression coefficients:

\[
\pi(\beta_0, \ldots, \beta_7) \sim \text{normal}(0, \text{var} = 1e6)
\]
Logistic Regression

The following statements fit a Bayesian logistic regression model in PROC GENMOD:

```plaintext
proc genmod data=neuralgia;
   class Treatment(ref="P") Sex(ref="M");
   model Pain= sex|treatment Age Duration / dist=bin link=logit;
   bayes seed=1 cprior=normal(var=1e6) outpost=neuout
      plots=trace;
run;
```

- PROC GENMOD models the probability of no pain (Pain = No)
- The default sampling algorithm is the Gamerman algorithm (Gamerman, D. 1997, *Statistics and Computing*, 7:57). PROC GENMOD offers a couple of alternative sampling algorithms, such as adaptive rejection and independence Metropolis.

Logistic Regression

Trace plots of some of the parameters.
Logistic Regression
Posterior summary statistics:

**Bayesian Analysis**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>N</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
</tr>
</thead>
<tbody>
<tr>
<td>SexF</td>
<td>10000</td>
<td>2.9148</td>
<td>1.7137</td>
<td>1.7348</td>
<td>2.8056</td>
<td>3.9222</td>
</tr>
<tr>
<td>TreatmentA</td>
<td>10000</td>
<td>4.6190</td>
<td>1.7924</td>
<td>3.3880</td>
<td>4.4333</td>
<td>5.6978</td>
</tr>
<tr>
<td>TreatmentB</td>
<td>10000</td>
<td>5.1406</td>
<td>1.8808</td>
<td>3.7928</td>
<td>5.0154</td>
<td>6.2784</td>
</tr>
<tr>
<td>TreatmentASexF</td>
<td>10000</td>
<td>-1.0367</td>
<td>2.3097</td>
<td>-2.4499</td>
<td>-0.9233</td>
<td>0.4706</td>
</tr>
<tr>
<td>TreatmentBsexF</td>
<td>10000</td>
<td>-0.3478</td>
<td>2.2499</td>
<td>-1.7787</td>
<td>-0.3578</td>
<td>1.1129</td>
</tr>
<tr>
<td>Age</td>
<td>10000</td>
<td>-0.3372</td>
<td>0.1155</td>
<td>-0.4141</td>
<td>-0.3276</td>
<td>-0.2531</td>
</tr>
<tr>
<td>Duration</td>
<td>10000</td>
<td>0.00894</td>
<td>0.0366</td>
<td>-0.0160</td>
<td>0.00926</td>
<td>0.0328</td>
</tr>
</tbody>
</table>

Logistic Regression
Posterior interval statistics:

**Bayesian Analysis**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Alpha</th>
<th>Equal-Tail Interval</th>
<th>HPD Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.050</td>
<td>5.9732</td>
<td>35.0404</td>
</tr>
<tr>
<td>SexF</td>
<td>0.050</td>
<td>0.0155</td>
<td>6.9155</td>
</tr>
<tr>
<td>TreatmentA</td>
<td>0.050</td>
<td>1.5743</td>
<td>8.4046</td>
</tr>
<tr>
<td>TreatmentB</td>
<td>0.050</td>
<td>1.7895</td>
<td>9.0056</td>
</tr>
<tr>
<td>TreatmentASexF</td>
<td>0.050</td>
<td>-5.6692</td>
<td>3.4066</td>
</tr>
<tr>
<td>TreatmentBsexF</td>
<td>0.050</td>
<td>-4.7148</td>
<td>4.2324</td>
</tr>
<tr>
<td>Age</td>
<td>0.050</td>
<td>-0.5724</td>
<td>-0.1325</td>
</tr>
<tr>
<td>Duration</td>
<td>0.050</td>
<td>-0.0626</td>
<td>0.0836</td>
</tr>
</tbody>
</table>
Odds Ratio

In the logistic model, the log odds function, \( \text{logit}(X) \), is given by:

\[
\text{logit}(X) \equiv \log\left( \frac{\Pr(Y = 1 \mid X)}{\Pr(Y = 0 \mid X)} \right) = \beta_0 + X\beta_1
\]

Suppose that you are interested in calculating the ratio of the odds for the female patients (\( \text{Sex}_F = 1 \)) to the male patients (\( \text{Sex}_F = 0 \)). The log of the odds ratio is the following:

\[
\log(\psi) \equiv \log(\psi(\text{Sex}_F = 1, \text{Sex}_F = 0)) = \text{logit}(\text{Sex}_F = 1) - \text{logit}(\text{Sex}_F = 0) = (\beta_0 + 1 \times \beta_1) - (\beta_0 + 0 \times \beta_1) = \beta_1
\]

It follows that the odds ratio is:

\[
\psi = \exp(\beta_1)
\]

Note that, by default, PROC GENMOD uses PARAM=GLM parametrization, which codes 1 and -1 to the values of \( \text{Sex}_F \).

In general, suppose the values of \( \text{Sex}_F \) are coded as constants \( a \) and \( b \) instead of 0 and 1.

- The odds when \( \text{Sex}_F = a \) become \( \exp(\beta_0 + a \cdot \beta_1) \)
- The odds when \( \text{Sex}_F = b \) become \( \exp(\alpha + b \cdot \beta_1) \)

The odds ratio is

\[
\psi = \exp[(b - a)\beta_1] = [\exp(\beta_1)]^{b-a}
\]

In other words, for any types of the effect parametrization schemes, as long as \( b - a = 1 \), \( \psi = \exp(\beta_1) \)
Odds Ratio

Odds ratios are functions of the model parameters, which can be obtained by manipulating posterior samples generated by PROC GENMOD. To estimate posterior odds ratios,

- save PROC GENMOD analysis to a SAS item store
- postfit odds ratios using the ESTIMATE statement in PROC PLM

An item store is a special SAS-defined binary file format used to store and restore information with a hierarchical structure.

The PLM procedure performs postprocessing tasks by taking the posterior samples (from GENMOD) and estimate functions of interest.

The ESTIMATE statement provides a mechanism for obtaining custom hypothesis testing (or linear combination of the regression coefficients).

The following statements fit the model in PROC GENMOD and saves the content to a SAS item store (logit_bayes):

```sas
proc genmod data=neuralgia;
   class Treatment(ref="P") Sex(ref="M");
   model Pain= sex|treatment Age Duration / dist=bin link=logit;
   bayes seed=2 cprior=normal(var=1e6) outpost=neuout
      plots=trace;
   store logit_bayes;
run;
```
Odds Ratio

The following statements evoke PROC PLM and estimate the odds ratio between the female group and male group conditional on treatment A:

```sql
proc plm restore=logit_bayes;
   estimate "F vs M, at Trt=A"
      sex 1 -1 treatment*sex [1, 1 1] [-1, 1 2]
         / e exp cl plots=dist;
run;
```

sex 1 -1 : estimates the difference between $\beta_1$ and $\beta_2$, which under the GLM parametrization, is equal to $\beta_1$

treatment * sex ... : assigns 1 to the interaction where “treatment=1” and “sex=1”, and -1 to the interaction where “treatment=1” and “sex=2”

e : requests that the L matrix coefficients be displayed
exp : exponentials and displays estimates (exp $\beta_1$)
cl : constructs 95% credit intervals
plots : generates histograms with kernel density overlaid

L Matrix Coefficients (GLM Parametrization)

<table>
<thead>
<tr>
<th>Estimate Coefficients</th>
<th>Treatment</th>
<th>Sex</th>
<th>Row1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sex F</td>
<td></td>
<td>F</td>
<td>1</td>
</tr>
<tr>
<td>Sex M</td>
<td></td>
<td>M</td>
<td>-1</td>
</tr>
<tr>
<td>Treatment A</td>
<td></td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>Treatment B</td>
<td></td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>Treatment P</td>
<td></td>
<td>P</td>
<td></td>
</tr>
<tr>
<td>Treatment A * Sex F</td>
<td></td>
<td>A</td>
<td>F</td>
</tr>
<tr>
<td>Treatment A * Sex M</td>
<td></td>
<td>A</td>
<td>M</td>
</tr>
<tr>
<td>Treatment B * Sex F</td>
<td></td>
<td>B</td>
<td>F</td>
</tr>
<tr>
<td>Treatment B * Sex M</td>
<td></td>
<td>B</td>
<td>M</td>
</tr>
<tr>
<td>Treatment P * Sex F</td>
<td></td>
<td>P</td>
<td>F</td>
</tr>
<tr>
<td>Treatment P * Sex M</td>
<td></td>
<td>P</td>
<td>M</td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Duration</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Odds Ratio
Female vs. Male, at Treatment = A.

<table>
<thead>
<tr>
<th>Label</th>
<th>N</th>
<th>Estimate</th>
<th>Standard Deviation</th>
<th>Percentiles</th>
<th>Alpha</th>
<th>Lower HPD</th>
<th>Upper HPD</th>
</tr>
</thead>
<tbody>
<tr>
<td>F vs M, at Trt=A</td>
<td>10000</td>
<td>1.8781</td>
<td>1.5260</td>
<td>0.7768, 1.7862</td>
<td>0.05</td>
<td>-0.7442</td>
<td>4.9791</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Label</th>
<th>Exponentiated</th>
<th>Standard Deviation of Exponentiated</th>
<th>Percentiles for Exponentiated</th>
<th>Lower HPD of Exponentiated</th>
<th>Upper HPD of Exponentiated</th>
</tr>
</thead>
<tbody>
<tr>
<td>F vs M, at Trt=A</td>
<td>28.3873</td>
<td>188.824003</td>
<td>2.1744, 5.9664, 18.4925</td>
<td>0.1876</td>
<td>93.1034</td>
</tr>
</tbody>
</table>

Histogram of the Posterior Odds Ratio
Similarly, you can estimate odds ratios conditional on different treatments:

```plaintext
proc plm restore=logit_bayes;
  estimate "F vs M, at Trt=B"
      sex 1 -1 treatment*sex [1, 2 1] [-1, 2 2] /exp;
  estimate "F vs M, at Trt=P"
      sex 1 -1 treatment*sex [1, 3 1] [-1, 3 2] /exp;
run;
```

### Odds Ratio

Female vs. Male, at Treatment = B.

<table>
<thead>
<tr>
<th>Label</th>
<th>N</th>
<th>Estimate</th>
<th>Standard Deviation</th>
<th>Percentiles</th>
<th>Alpha</th>
<th>Lower HPD</th>
<th>Upper HPD</th>
</tr>
</thead>
<tbody>
<tr>
<td>F vs M, at Trt=B</td>
<td>10000</td>
<td>2.5670</td>
<td>1.5778</td>
<td>1.4946 2.4569 3.5345</td>
<td>0.05</td>
<td>-0.1317 5.9040</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Label</th>
<th>Exponentiated</th>
<th>Standard Deviation of Exponentiated</th>
<th>Percentiles for Exponentiated</th>
<th>Lower HPD of Exponentiated</th>
<th>Upper HPD of Exponentiated</th>
</tr>
</thead>
<tbody>
<tr>
<td>F vs M, at Trt=B</td>
<td>60.4417</td>
<td>384.724355</td>
<td>4.4575 11.6684 34.2779</td>
<td>0.1399</td>
<td>195.64</td>
</tr>
</tbody>
</table>
**Odds Ratio**

Female vs. Male, at Treatment = P.

<table>
<thead>
<tr>
<th>Label</th>
<th>N</th>
<th>Estimate</th>
<th>Standard Deviation</th>
<th>Percentiles</th>
<th>Lower HPD</th>
<th>Upper HPD</th>
</tr>
</thead>
<tbody>
<tr>
<td>F vs M, at Trt=P</td>
<td>10000</td>
<td>2.9148</td>
<td>1.7137</td>
<td>1.7348 2.8056 3.9222</td>
<td>0.05 -0.1694 6.5110</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Label</th>
<th>Exponentiated</th>
<th>Standard Deviation of Exponentiated</th>
<th>Percentiles for Exponentiated</th>
<th>Lower HPD of Exponentiated</th>
<th>Upper HPD of Exponentiated</th>
</tr>
</thead>
<tbody>
<tr>
<td>F vs M, at Trt=P</td>
<td>175.97</td>
<td>1642.867153</td>
<td>5.6676 16.5362 50.5135</td>
<td>0.3686 408.55</td>
<td></td>
</tr>
</tbody>
</table>

**Outline**

1. The GENMOD, PHREG, LIFEREG, and FMM Procedures
   - Overview of Bayesian capabilities in the GENMOD, PHREG, LIFEREG, and FMM procedures
   - Prior distributions
   - The BAYES statement
   - GENMOD: linear regression
   - GENMOD: binomial model
   - PHREG: Cox model
   - PHREG: piecewise exponential model (optional)
Cox Model

Consider the data for the Veterans Administration lung cancer trial presented in Appendix 1 of Kalbfleisch and Prentice (1980).

<table>
<thead>
<tr>
<th>Time</th>
<th>Death in days</th>
</tr>
</thead>
<tbody>
<tr>
<td>Therapy</td>
<td>Type of therapy: standard or test</td>
</tr>
<tr>
<td>Cell</td>
<td>Type of tumor cell: adeno, large, small, or squamous</td>
</tr>
<tr>
<td>PTherapy</td>
<td>Prior therapy: yes or no</td>
</tr>
<tr>
<td>Age</td>
<td>Age in years</td>
</tr>
<tr>
<td>Duration</td>
<td>Months from diagnosis to randomization</td>
</tr>
<tr>
<td>KPS</td>
<td>Karnofsky performance scale</td>
</tr>
<tr>
<td>Status</td>
<td>Censoring indicator (1=censored time, 0=event time)</td>
</tr>
</tbody>
</table>

A subset of the data:

<table>
<thead>
<tr>
<th>OBS</th>
<th>Therapy</th>
<th>Cell</th>
<th>Time</th>
<th>Kps</th>
<th>Duration</th>
<th>Age</th>
<th>Ptherapy</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>standard</td>
<td>squamous</td>
<td>72</td>
<td>60</td>
<td>7</td>
<td>69</td>
<td>no</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>standard</td>
<td>squamous</td>
<td>411</td>
<td>70</td>
<td>5</td>
<td>64</td>
<td>yes</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>standard</td>
<td>squamous</td>
<td>228</td>
<td>60</td>
<td>3</td>
<td>38</td>
<td>no</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>standard</td>
<td>squamous</td>
<td>126</td>
<td>60</td>
<td>9</td>
<td>63</td>
<td>yes</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>standard</td>
<td>squamous</td>
<td>118</td>
<td>70</td>
<td>11</td>
<td>65</td>
<td>yes</td>
<td>1</td>
</tr>
</tbody>
</table>

- Some parameters are the coefficients of the continuous variables (KPS, Duration, and Age).
- Other parameters are the coefficients of the design variables for the categorical explanatory variables (PTherapy, Cell, and Therapy).
Cox Model

The model considered here is the Breslow partial likelihood:

\[
L(\beta) = \prod_{i=1}^{k} \frac{e^{\beta' \sum_{j \in \mathcal{D}_i} Z_j(t_i)}}{\left[ \sum_{l \in \mathcal{R}_i} e^{\beta' Z_l(t_i)} \right]^{d_i}}
\]

where

- \( t_1 < \cdots < t_k \) are distinct event times
- \( Z_j(t_i) \) is the vector explanatory variables for the \( j \)th individual at time \( t_i \)
- \( \mathcal{R}_i \) is the risk set at \( t_i \), which includes all observations that have survival time greater than or equal to \( t_i \)
- \( d_i \) is the multiplicity of failures at \( t_i \). It is the size of the set \( \mathcal{D}_i \) of individuals that fail at \( t_i \)

The following statements fit a Cox regression model with a uniform prior on the regression coefficients:

```
proc phreg data=VALung;
   class PTherapy(ref='no') Cell(ref='large') Therapy(ref='standard');
   model Time*Status(0) = KPS Duration Age PTherapy Cell Therapy;
   bayes seed=1 outpost=cout coeffprior=uniform;
run;
```
### Cox Model: Posterior Mean Estimates

#### Bayesian Analysis

<table>
<thead>
<tr>
<th>Parameter</th>
<th>N</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kps</td>
<td>10000</td>
<td>-0.0327</td>
<td>0.00545</td>
<td>-0.0364</td>
<td>-0.0328</td>
<td>-0.0291</td>
</tr>
<tr>
<td>Duration</td>
<td>10000</td>
<td>-0.00170</td>
<td>0.00945</td>
<td>-0.00791</td>
<td>-0.00123</td>
<td>0.00489</td>
</tr>
<tr>
<td>Age</td>
<td>10000</td>
<td>-0.00852</td>
<td>0.00935</td>
<td>-0.0147</td>
<td>-0.00850</td>
<td>-0.00223</td>
</tr>
<tr>
<td>Ptherapyyes</td>
<td>10000</td>
<td>0.0754</td>
<td>0.2345</td>
<td>-0.0776</td>
<td>0.0766</td>
<td>0.2340</td>
</tr>
<tr>
<td>Celladeno</td>
<td>10000</td>
<td>0.7867</td>
<td>0.3080</td>
<td>0.5764</td>
<td>0.7815</td>
<td>0.9940</td>
</tr>
<tr>
<td>Cellsmall</td>
<td>10000</td>
<td>0.4632</td>
<td>0.2731</td>
<td>0.2775</td>
<td>0.4602</td>
<td>0.6435</td>
</tr>
<tr>
<td>Cellsquamous</td>
<td>10000</td>
<td>-0.4022</td>
<td>0.2843</td>
<td>-0.5935</td>
<td>-0.4024</td>
<td>-0.2124</td>
</tr>
<tr>
<td>Therapytest</td>
<td>10000</td>
<td>0.2897</td>
<td>0.2091</td>
<td>0.1500</td>
<td>0.2900</td>
<td>0.4294</td>
</tr>
</tbody>
</table>

### Cox Model: Interval Estimates

#### Bayesian Analysis

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Alpha</th>
<th>Equal-Tail Interval</th>
<th>HPD Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kps</td>
<td>0.050</td>
<td>-0.0433</td>
<td>-0.0219</td>
</tr>
<tr>
<td>Duration</td>
<td>0.050</td>
<td>-0.0216</td>
<td>0.0153</td>
</tr>
<tr>
<td>Age</td>
<td>0.050</td>
<td>-0.0271</td>
<td>0.00980</td>
</tr>
<tr>
<td>Ptherapyyes</td>
<td>0.050</td>
<td>-0.3943</td>
<td>0.5335</td>
</tr>
<tr>
<td>Celladeno</td>
<td>0.050</td>
<td>0.1905</td>
<td>1.3969</td>
</tr>
<tr>
<td>Cellsmall</td>
<td>0.050</td>
<td>-0.0617</td>
<td>1.0039</td>
</tr>
<tr>
<td>Cellsquamous</td>
<td>0.050</td>
<td>-0.9651</td>
<td>0.1519</td>
</tr>
<tr>
<td>Therapytest</td>
<td>0.050</td>
<td>-0.1191</td>
<td>0.6955</td>
</tr>
</tbody>
</table>
Suppose that you are interested in estimating the survival curves for two individuals who have similar characteristics, with one receiving the standard treatment while the other did not. The following is saved in the SAS data set `pred`:

<table>
<thead>
<tr>
<th>OBS</th>
<th>Ptherapy</th>
<th>kps</th>
<th>duration</th>
<th>age</th>
<th>cell</th>
<th>therapy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>no</td>
<td>58</td>
<td>8.7</td>
<td>60</td>
<td>large</td>
<td>standard</td>
</tr>
<tr>
<td>2</td>
<td>no</td>
<td>58</td>
<td>8.7</td>
<td>60</td>
<td>large</td>
<td>test</td>
</tr>
</tbody>
</table>

You can use the following statements to estimate the survival curves and save the estimates to a SAS data set:

```sas
proc phreg data=VALung plots(cl=hpd overlay)=survival;
    baseline covariates=pred out=pout;
    class PTherapy(ref='no') Cell(ref='large')
        Therapy(ref='standard');
    model Time*Status(0) = KPS Duration Age PTherapy Cell Therapy;
    bayes seed=1 outpost=cout coeffprior=uniform;
run;
```

- **plots**: requests survival curves with overlaying HPD intervals
- **baseline**: specifies input covariates data set and saves the posterior prediction to the OUT= data set
Cox Model: Posterior Survival Curves

Estimated survival curves for the two subjects and their corresponding 95% HPD intervals.

Hazard Ratios

The HAZARDRATIO statement enables you to obtain customized hazard ratios, ratios of two hazard functions.

**HAZARDRATIO 'label' variables / options ;**

- For a continuous variable: the hazard ratio compares the hazards for a given change (by default, a increase of 1 unit) in the variable.
- For a CLASS variable, a hazard ratio compares the hazards of two levels of the variable.
Hazard Ratios

The following SAS statements fit the same Cox regression model and request three kinds of hazard ratios.

```sas
proc phreg data=VALung;
  class PTherapy(ref='no') Cell(ref='large') Therapy(ref='standard');
  model Time*Status(0) = KPS Duration Age PTherapy Cell Therapy;
  bayes seed=1 outpost=vout plots=trace coeffprior=uniform;
  hazardratio 'HR 1' Therapy / at(PTherapy='yes' KPS=80 duration=12 age=65 cell='small');
  hazardratio 'HR 2' Age / unit=10 at(KPS=45);
  hazardratio 'HR 3' Cell;
run;
```

The following results are the summary statistics of the posterior hazards between the standard therapy and the test therapy.

<table>
<thead>
<tr>
<th>Description</th>
<th>N</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Therapy standard vs test At Prior=yes Kps=80 Duration=12 Age=65 Cell=small</td>
<td>10000</td>
<td>0.7651</td>
<td>0.1617</td>
<td>0.6509</td>
<td>0.7483</td>
<td>0.8607</td>
</tr>
</tbody>
</table>

Bayesian Analysis

<table>
<thead>
<tr>
<th>HR 1: Hazard Ratios for Therapy</th>
<th>Quantiles</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Equal-Tail 95% Interval</strong></td>
<td>0.4988</td>
</tr>
<tr>
<td><strong>HPD Interval 95%</strong></td>
<td>1.1265</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>HR 1: Hazard Ratios for Therapy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>95% HPD Interval</strong></td>
</tr>
<tr>
<td><strong>95% HPD Interval</strong></td>
</tr>
</tbody>
</table>
Hazard Ratios

The following table lists the change of hazards for an increase in Age of 10 years.

Bayesian Analysis

<table>
<thead>
<tr>
<th>Description</th>
<th>N</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Quantiles</th>
<th>95% Equal-Tail Interval</th>
<th>95% HPD Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age Unit=10 At Kps=45</td>
<td>10000</td>
<td>0.9224</td>
<td>0.0865</td>
<td>0.8633 0.9185 0.9779</td>
<td>0.7629 1.1030</td>
<td>0.7539 1.0904</td>
</tr>
</tbody>
</table>

Bayesian Analysis

<table>
<thead>
<tr>
<th>Description</th>
<th>N</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Quantiles</th>
<th>95% Equal-Tail Interval</th>
<th>95% HPD Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell adeno vs large</td>
<td>10000</td>
<td>2.3035</td>
<td>0.7355</td>
<td>1.7797 2.1848 2.7020</td>
<td>1.2099 4.0428</td>
<td>1.0661 3.7509</td>
</tr>
<tr>
<td>Cell adeno vs small</td>
<td>10000</td>
<td>1.4374</td>
<td>0.4124</td>
<td>1.1479 1.3811 1.6622</td>
<td>0.7985 2.3857</td>
<td>0.7047 2.2312</td>
</tr>
<tr>
<td>Cell adeno vs squamous</td>
<td>10000</td>
<td>3.4376</td>
<td>1.0682</td>
<td>2.6679 3.2903 4.0199</td>
<td>1.8150 5.9733</td>
<td>1.6274 5.6019</td>
</tr>
<tr>
<td>Cell large vs small</td>
<td>10000</td>
<td>0.6530</td>
<td>0.1798</td>
<td>0.5254 0.6311 0.7577</td>
<td>0.3664 1.0636</td>
<td>0.3357 1.0141</td>
</tr>
<tr>
<td>Cell large vs squamous</td>
<td>10000</td>
<td>1.5567</td>
<td>0.4514</td>
<td>1.2367 1.4954 1.8103</td>
<td>0.8591 2.6251</td>
<td>0.7776 2.4679</td>
</tr>
<tr>
<td>Cell small vs squamous</td>
<td>10000</td>
<td>2.4696</td>
<td>0.7046</td>
<td>1.9717 2.3742 2.8492</td>
<td>1.3872 4.1403</td>
<td>1.2958 3.9351</td>
</tr>
</tbody>
</table>
The GENMOD, PHREG, LIFEREG, and FMM Procedures

Outline

2 The GENMOD, PHREG, LIFEREG, and FMM Procedures
   - Overview of Bayesian capabilities in the GENMOD, PHREG, LIFEREG, and FMM procedures
   - Prior distributions
   - The BAYES statement
   - GENMOD: linear regression
   - GENMOD: binomial model
   - PHREG: Cox model
   - PHREG: piecewise exponential model (optional)

Piecewise Exponential Model (Optional)

Let \{(t_i, x_i, \delta_i), i = 1, 2, \ldots, n\} be the observed data. Let \(a_0 = 0 < a_1 < \ldots < a_{J-1} < a_J = \infty\) be a partition of the time axis. The hazard for subject \(i\) is

\[
h(t|x_i; \theta) = h_0(t) \exp(\beta'x_i)
\]

where

\[
h_0(t) = \lambda_j \quad a_{j-1} \leq t < a_j \quad (j = 1, \ldots, J)
\]

The hazard for subject \(i\) in the \(j\)th time interval is

\[
h(t) = \lambda_j \exp(\beta'x_i) \quad a_{j-1} < t < a_j
\]
Piecewise Exponential Model

From the hazard function, first define the baseline cumulative hazard function:

\[
H_0(t) = \sum_{j=1}^{J} \lambda_j \Delta_j(t)
\]

where

\[
\Delta_j(t) = \begin{cases} 
0 & t < a_{j-1} \\
(t - a_{j-1}) & a_{j-1} \leq t < a_j \\
(a_j - a_{j-1}) & t \geq a_j 
\end{cases}
\]

The log likelihood is:

\[
\ell(\lambda, \beta) = \sum_{i=1}^{n} \delta_i \left[ \sum_{j=1}^{J} I(a_{j-1} \leq t_i < a_j) \log \lambda_j + \beta' x_i \right] \\
- \sum_{i=1}^{n} \left[ \sum_{j=1}^{J} \Delta_j(t_i) \lambda_j \right] \exp(\beta' x_i)
\]

where \(\delta_i\) is the event status:

\[
\delta_i = \begin{cases} 
0 & \text{if } t_i \text{ is a censored time} \\
1 & \text{if } t_i \text{ is an event time}
\end{cases}
\]

This model has two parameter vectors: \(\lambda\) and \(\beta\).
Piecewise Exponential Model

PROC PHREG supports the following priors for the piecewise exponential model:

- Regression coefficients ($\beta$): normal and uniform priors
- Hazards ($\lambda$): improper, uniform, independent gamma, and AR(1) priors
- Log hazards ($\alpha = \log(\lambda)$): uniform and normal priors
- Regression coefficients and log hazards: multivariate normal (do not need to be independent)

Consider a randomized trial of 40 rats exposed to carcinogen:

- Drug X and Placebo are the treatment groups.
- Event of interest is death.
- Response is time until death.
- What are the effects of treatment and gender on survival?
Piecewise Exponential Model

A subset of the data:

```sas
proc format;
   value Rx 1='X' 0='Placebo';
data Exposed;
   input Days Status Trt Gender $ @@;
   format Trt Rx. ;
datalines;
179 1 1 F 378 0 1 M
256 1 1 F 355 1 1 M
262 1 1 M 319 1 1 M
256 1 1 F 256 1 1 M
... 
268 0 0 M 209 1 0 F
;```

An appropriate model is the piecewise exponential. In the model:

- Each time interval has a constant hazard
- There are a total of eight intervals (PROC PHREG default)
- Intervals are determined by placing roughly equal number of uncensored observations in each interval
- The log hazard is used. It is generally more computationally stable. There are 8 $\lambda_i$’s and two regression coefficients.
Piecewise Exponential Model

The following programming statements fit a Bayesian piecewise exponential model with noninformative priors on both $\beta$ and $\log(\lambda)$:

```plaintext
proc phreg data=Exposed;
   class Trt(ref='Placebo') Gender(ref='F');
   model Days*Status(0)=Trt Gender;
   bayes seed=1 outpost=eout piecewise=loghazard(n=8);
run;
```

The PIECEWISE= option requests the estimating of a piecewise exponential model with 8 intervals.

Suppose that you have some prior information w.r.t. both $\beta$ and $\log(\lambda)$ that can be approximated well with a multivariate normal distribution. You can construct the following data set:

```plaintext
data pinfo;
   input _TYPE_ $ alpha1-alpha8 trtX GenderM;
   datalines;
Mean 0 0 0 0 0 0 0 0 0 0
cov 90.2 1 3 -1.9 4.1 3.7 14.3 -10.7 -7.2 -4.2
   ...;
```

127 / 295
Piecewise Exponential Model

The following programming statements fit a Bayesian piecewise exponential model with informative prior on both $\beta$ and $\log(\lambda)$:

```
proc phreg data=exposed;
  class Trt(ref='Placebo') Gender(ref='F');
  model Days*Status(0)=Trt Gender;
  bayes seed=1 outpost=eout
     piecewise=loghazard(n=8 prior=normal(input=pinfo))
     cprior=normal(input=pinfo);
run;
```

Bayesian Analysis

<table>
<thead>
<tr>
<th>Model Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Set</td>
</tr>
<tr>
<td>Dependent Variable</td>
</tr>
<tr>
<td>Censoring Variable</td>
</tr>
<tr>
<td>Censoring Value(s)</td>
</tr>
<tr>
<td>Model</td>
</tr>
<tr>
<td>Burn-In Size</td>
</tr>
<tr>
<td>MC Sample Size</td>
</tr>
<tr>
<td>Thinning</td>
</tr>
</tbody>
</table>

Summary of the Number of Event and Censored Values

<table>
<thead>
<tr>
<th>Total</th>
<th>Event</th>
<th>Censored</th>
<th>Percent Censored</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>36</td>
<td>4</td>
<td>10.00</td>
</tr>
</tbody>
</table>
Piecewise Exponential Model

The partition of the time intervals:

### Bayesian Analysis

<table>
<thead>
<tr>
<th>Interval</th>
<th>N</th>
<th>Event</th>
<th>Log Hazard Parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0, 193]</td>
<td>193</td>
<td>5</td>
<td>Alpha1</td>
</tr>
<tr>
<td>(193, 221]</td>
<td>221</td>
<td>5</td>
<td>Alpha2</td>
</tr>
<tr>
<td>(221, 239.5]</td>
<td>239.5</td>
<td>7</td>
<td>Alpha3</td>
</tr>
<tr>
<td>(239.5, 255.5]</td>
<td>255.5</td>
<td>5</td>
<td>Alpha4</td>
</tr>
<tr>
<td>(255.5, 256.5]</td>
<td>256.5</td>
<td>4</td>
<td>Alpha5</td>
</tr>
<tr>
<td>(256.5, 278.5]</td>
<td>278.5</td>
<td>5</td>
<td>Alpha6</td>
</tr>
<tr>
<td>(278.5, 321]</td>
<td>321</td>
<td>4</td>
<td>Alpha7</td>
</tr>
<tr>
<td>(321, ∞]</td>
<td>Infty</td>
<td>5</td>
<td>Alpha8</td>
</tr>
</tbody>
</table>

### Posterior summary statistics:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>N</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpha1</td>
<td>10000</td>
<td>-6.4137</td>
<td>0.4750</td>
<td>-6.7077</td>
<td>-6.3770</td>
<td>-6.0852</td>
</tr>
<tr>
<td>Alpha2</td>
<td>10000</td>
<td>-4.0505</td>
<td>0.4870</td>
<td>-4.3592</td>
<td>-4.0207</td>
<td>-3.7058</td>
</tr>
<tr>
<td>Alpha3</td>
<td>10000</td>
<td>-2.9297</td>
<td>0.5146</td>
<td>-3.2468</td>
<td>-2.8954</td>
<td>-2.5737</td>
</tr>
<tr>
<td>Alpha4</td>
<td>10000</td>
<td>-1.9146</td>
<td>0.6212</td>
<td>-2.3256</td>
<td>-1.8936</td>
<td>-1.4839</td>
</tr>
<tr>
<td>Alpha5</td>
<td>10000</td>
<td>1.2433</td>
<td>0.6977</td>
<td>0.7948</td>
<td>1.2598</td>
<td>1.7255</td>
</tr>
<tr>
<td>Alpha6</td>
<td>10000</td>
<td>-0.8729</td>
<td>0.8040</td>
<td>-1.4033</td>
<td>-0.8692</td>
<td>-0.3276</td>
</tr>
<tr>
<td>Alpha7</td>
<td>10000</td>
<td>-0.9827</td>
<td>0.8346</td>
<td>-1.5247</td>
<td>-0.9646</td>
<td>-0.4223</td>
</tr>
<tr>
<td>Alpha8</td>
<td>10000</td>
<td>0.4771</td>
<td>0.9095</td>
<td>-0.1262</td>
<td>0.4796</td>
<td>1.0952</td>
</tr>
<tr>
<td>TrtX</td>
<td>10000</td>
<td>-1.2319</td>
<td>0.3929</td>
<td>-1.4898</td>
<td>-1.2286</td>
<td>-0.9707</td>
</tr>
<tr>
<td>GenderM</td>
<td>10000</td>
<td>-2.6607</td>
<td>0.5483</td>
<td>-3.0159</td>
<td>-2.6466</td>
<td>-2.2888</td>
</tr>
</tbody>
</table>
### Piecewise Exponential Model

#### Interval estimates:

**Bayesian Analysis**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Alpha</th>
<th>Equal-Tail Interval</th>
<th>HPD Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpha1</td>
<td>0.050</td>
<td>-7.4529</td>
<td>-5.5710 -7.3576 -5.5100</td>
</tr>
<tr>
<td>Alpha2</td>
<td>0.050</td>
<td>-5.0961</td>
<td>-3.1973 -5.0030 -3.1303</td>
</tr>
<tr>
<td>Alpha3</td>
<td>0.050</td>
<td>-4.0327</td>
<td>-2.0130 -3.9950 -1.9843</td>
</tr>
<tr>
<td>Alpha4</td>
<td>0.050</td>
<td>-3.1799</td>
<td>-0.7614 -3.1671 -0.7536</td>
</tr>
<tr>
<td>Alpha5</td>
<td>0.050</td>
<td>-0.1893</td>
<td>2.5585 -0.0872 2.6410</td>
</tr>
<tr>
<td>Alpha6</td>
<td>0.050</td>
<td>-2.4616</td>
<td>0.6875 -2.4942 0.6462</td>
</tr>
<tr>
<td>Alpha7</td>
<td>0.050</td>
<td>-2.6588</td>
<td>0.6248 -2.6383 0.6400</td>
</tr>
<tr>
<td>Alpha8</td>
<td>0.050</td>
<td>-1.3264</td>
<td>2.2243 -1.2867 2.2359</td>
</tr>
<tr>
<td>TrtX</td>
<td>0.050</td>
<td>-2.0147</td>
<td>-0.4735 -2.0195 -0.4849</td>
</tr>
<tr>
<td>GenderM</td>
<td>0.050</td>
<td>-3.7758</td>
<td>-1.6150 -3.7269 -1.5774</td>
</tr>
</tbody>
</table>

#### Hazard ratios of Treatment and Gender:

```plaintext
hazardratio 'Hazard Ratio Statement 1' Trt;
hazardratio 'Hazard Ratio Statement 2' Gender;
```

**Bayesian Analysis**

### Hazard Ratio Statement 1: Hazard Ratios for Trt

<table>
<thead>
<tr>
<th>Description</th>
<th>N</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>95% Equal-Tail Interval</th>
<th>95% HPD Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trt Placebo vs X</td>
<td>10000</td>
<td>3.7058</td>
<td>1.5430</td>
<td>2.6399</td>
<td>3.4164</td>
<td>4.4362</td>
<td>1.6056 -7.4981</td>
<td>1.3129 -6.7830</td>
</tr>
</tbody>
</table>

### Hazard Ratio Statement 2: Hazard Ratios for Gender

<table>
<thead>
<tr>
<th>Description</th>
<th>N</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>95% Equal-Tail Interval</th>
<th>95% HPD Interval</th>
</tr>
</thead>
</table>
The MCMC procedure (SAS/STAT® 9.2, 9.22, 9.3, 12.1) is a simulation procedure that can be used to fit:

- single-level or multilevel (hierarchical) models
- linear or nonlinear models, such as regression, survival, ordinal multinomial, and so on.
- missing data problems

The procedure selects appropriate sampling algorithms for the models that you specified, and it is capable of executing SAS DATA step language for estimation and inference.
PROC MCMC Statements

**PROC MCMC** options;

PARMS; define parameters.
PRIOR; declare prior distributions

**Programming statements:**

MODEL } define log-likelihood function
PREDDIST; posterior prediction
RANDOM; random effects
Run;

Linear Regression

\[
\text{weight}_i \sim \text{normal}(\mu_i, \text{var }= \sigma^2)
\]

\[
\mu = \beta \cdot \text{height}_i
\]

\[
\beta \sim \text{normal}(0, \text{var }= 100)
\]

\[
\sigma^2 \sim \text{inverse Gamma}(\text{shape }= 2, \text{scale }= 2)
\]

The data:

```
data class;
  input height weight;
datalines;
  69.0  112.5
  56.5  84.0
  ...
  66.5  112.0
;
```

MCMC Program:

```
proc mcmc data=class seed=1 nbi=5000
  nmc=10000 outpost=regOut;
parsm beta s2;
prior beta ~ normal(0, var=100);
prior s2 ~ igamma(shape=2, scale=2);
mu = beta * height;
model weight ~ normal(mu, var=s2);
run;
```
Linear Regression

\[ \text{weight}_i \sim t(\mu_i, \text{sd} = \sigma, \text{df} = 3) \]
\[ \mu = \beta \cdot \text{height}_i \]
\[ \beta \sim \text{normal}(0, \text{var} = 100) \]
\[ \sigma \sim \text{uniform}(0, 25) \]

Change the model, parameterization, and so on as you please:

```sas
proc mcmc data=class seed=1 nbi=5000 nmc=10000 outpost=regOut;
   parms beta sig;
   prior beta ~ normal(0, var=100);
   prior sig ~ uniform(0, 25);
   mu = beta * height;
   model weight ~ t(mu, sd=sig, df=3);
run;
```

The Posterior Distribution

PROC MCMC is sampling-based procedure, which is similar to other SAS Bayesian procedures. BUT, you must be more aware of how the posterior distribution is constructed:

\[ \pi(\theta|\mathbf{y}, \mathbf{x}) \propto \pi(\theta) \cdot f(\mathbf{y}|\theta, \mathbf{x}) \]

- The PRIOR statements define the prior distributions: \( \pi(\theta) \).
- The MODEL statement defines the likelihood function for each observation in the data set: \( f(y_i|\theta, x_i) \), for \( i = 1, \cdots, n \)
- The procedure calculates the posterior distribution (on the log scale):
  \[ \log(\pi(\theta|\mathbf{y}, \mathbf{x})) = \log(\pi(\theta)) + \sum_{i=1}^{n} \log(f(y_i|\theta, x_i)) \]

where \( \mathbf{y} = \{y_i\} \) and \( \mathbf{x} = \{x_i\} \)
Calculate of $\log(\pi(\theta|y))$

At each iteration, the programming and MODEL statements are executed for each observation to obtain $\log(\pi(\theta|y))$

<table>
<thead>
<tr>
<th>Obs</th>
<th>Height</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>69.0</td>
<td>112.5</td>
</tr>
<tr>
<td>2</td>
<td>56.5</td>
<td>84.0</td>
</tr>
<tr>
<td>3</td>
<td>65.3</td>
<td>98.0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>19</td>
<td>66.5</td>
<td>112.0</td>
</tr>
</tbody>
</table>

at the top of the data set

$$\log \pi(\theta|y) = \log(f(y_1|\theta))$$

```plaintext
proc mcmc data=input;
prior;
progm stmt;
model ;
run;
```

stepping through the data set

$$\log \pi(\theta|y) = \log \pi(\theta|y) + \log(f(y_2|\theta))$$
Calculate of $\log(\pi(\theta|y))$

At each iteration, the programming and MODEL statements are executed for each observation to obtain $\log(\pi(\theta|y))$

<table>
<thead>
<tr>
<th>Obs</th>
<th>Height</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>69.0</td>
<td>112.5</td>
</tr>
<tr>
<td>2</td>
<td>56.5</td>
<td>84.0</td>
</tr>
<tr>
<td>3</td>
<td>65.3</td>
<td>98.0</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>66.5</td>
<td>112.0</td>
</tr>
</tbody>
</table>

stepping through the data set

$$\log \pi(\theta|y) = \log \pi(\theta|y) + \log(f(y_3|\theta))$$

Calculate of $\log(\pi(\theta|y))$

At each iteration, the programming and MODEL statements are executed for each observation to obtain $\log(\pi(\theta|y))$

<table>
<thead>
<tr>
<th>Obs</th>
<th>Height</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>69.0</td>
<td>112.5</td>
</tr>
<tr>
<td>2</td>
<td>56.5</td>
<td>84.0</td>
</tr>
<tr>
<td>3</td>
<td>65.3</td>
<td>98.0</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>66.5</td>
<td>112.0</td>
</tr>
</tbody>
</table>

at the last observation, the prior is included

$$\log \pi(\theta|y) = \log(\pi(\theta)) + \sum_{i=1}^{n} \log(f(y_i|\theta))$$
PROC MCMC and WinBUGS Syntax are Similar

Both require going through the data set (repeatedly). In WinBUGS, a for-loop and array indices are used to access records in variables; In PROC MCMC, the looping over the data set is hidden behind the scene.

```
height[] weight[]
  69.0  112.5
  56.5  84.0
  65.3  98.0
...
  66.5  112.0
END
```

```
model {
  for(i in 1:19) {
    mu[i] = beta * height[i]
    weight[i] ~ dnorm(mu[i], tau)
  }
  beta ~ dnorm(0, 0.1)
  tau ~ gamma(0.1, 0.1)
}
```

Sampling in PROC MCMC

PROC MCMC recognizes certain configurations of the statistical models and applies sampling methods (conjugate or direct) when appropriate.

In other cases, the default sampling algorithm a normal-kernel-based random walk Metropolis. The proposal distribution is

\[ q(\theta_{\text{new}}|\theta^{(t)}) = \text{MVN}(\theta_{\text{new}}|\theta^{(t)}, c^2\Sigma). \]

Two components in the Metropolis algorithm:
- construction of the proposal distribution—automatically done by PROC MCMC
- evaluation of \( \log(\pi(\theta^{(t)}|y)) \) at each iteration—specified via the PRIOR and MODEL statements
PARMS Statement

PARMS name | (name-list) <= number;

- lists the names of the parameters
- specifies optional initial values
- specifies updating sequence of the parameters

For example:

PARMS alpha 0 beta 1;

declares $\alpha$ and $\beta$ to be model parameters and assigns 0 to $\alpha$ and 1 to $\beta$.

PARMS alpha 0 beta;

assigns 0 to $\alpha$ and leaves $\beta$ uninitialized.

PARMS (alpha beta) 1;

assigns 1 to both $\alpha$ and $\beta$.

When multiple PARMS statements are used, each statement defines a block of parameters, which are updated sequentially in each iteration:

PARMS beta0 beta1;
PARMS sigma2;

At each iteration $t$, PROC MCMC updates $\beta_0$ and $\beta_1$ together, alternatively with $\sigma^2$, each with a Metropolis sampler:

$$
\beta_0^{(t)}, \beta_1^{(t)} \mid \sigma^2_{(t-1)}, \text{Data}
$$

$$
\sigma^2_{(t)} \mid \beta_0^{(t)}, \beta_1^{(t)}, \text{Data}
$$
**PRIOR Statement**

**PRIOR** parameter-list ~ distribution;

specifies the prior distributions of model parameters. For example:

PRIOR alpha ~ normal(0, var=10);
PRIOR sigma2 ~ igamma(0.001, iscale=0.001);
PRIOR beta gamma ~ normal(alpha, var=sigma2);

specifies the following joint prior distribution:

\[
\pi(\alpha, \beta, \gamma, \sigma^2) = \pi(\beta|\alpha, \sigma^2) \cdot \pi(\gamma|\alpha, \sigma^2) \cdot \pi(\alpha) \cdot \pi(\sigma^2)
\]

**MODEL Statement**

**MODEL** dependent-variable-list ~ distribution;

specifies the likelihood function. The dependent variables can be

- data set variables
  
  ```
  MODEL y ~ normal(alpha, var=1);
  ```

- functions of data set variables

  ```
  w = log(y);
  MODEL w ~ normal(alpha, var=1);
  ```

You can specify multiple MODEL statements.
Standard Distributions

Standard distributions in the PRIOR and MODEL statements:

<table>
<thead>
<tr>
<th>beta</th>
<th>binary</th>
<th>binomial</th>
<th>cauchy</th>
<th>chisq</th>
</tr>
</thead>
<tbody>
<tr>
<td>expon</td>
<td>gamma</td>
<td>geo</td>
<td>ichisq</td>
<td>igamma</td>
</tr>
<tr>
<td>laplace</td>
<td>negbin</td>
<td>normal</td>
<td>pareto</td>
<td>poisson</td>
</tr>
<tr>
<td>sichisq</td>
<td>t</td>
<td>uniform</td>
<td>wald</td>
<td>weibull</td>
</tr>
<tr>
<td>dirich</td>
<td>iwish</td>
<td>mvn</td>
<td>mvnar</td>
<td>multinom</td>
</tr>
</tbody>
</table>

Distinction argument can be constants, expressions, or model parameters. For example:

```plaintext
prior alpha ~ cauchy(0, 2);
prior p ~ beta(abs(alpha), constant('pi'));
model y ~ binomial(n, p);
```

2 Only in the MODEL statement

Standard Distributions

Some distributions can be parameterized in different ways:

| expon(scale|s = λ) | expon(iscale|is = λ) |
|-------------|---------------------|
| gamma(a, scale|sc = λ) | gamma(a, iscale|is = λ) |
| igamma(a, scale|sc = λ) | igamma(a, iscale|is = λ) |
| laplace(l, scale|sc = λ) | laplace(l, iscale|is = λ) |
| normal(μ, var=σ^2) | normal(μ, sd=σ) | normal(μ, prec=τ) |
| lognormal(μ, var=σ^2) | lognormal(μ, sd=σ) | lognormal(μ, prec=τ) |
| t(μ, var=σ^2, df) | t(μ, sd=σ, df) | t(μ, prec=τ, df) |

For these distributions, you must explicitly name the ambiguous parameter. For example:

```plaintext
prior beta ~ normal(0, var=sigma2);
prior sigma2 ~ igamma(0.001, is=0.001);
```
Truncated Distributions

Univariate distributions allow for optional LOWER= and UPPER= arguments.

```plaintext
prior p ~ beta(2,3, lower=0.5);
prior b ~ expon(scale=100, lower=100, upper=2000);
```

The bounds can be random variables (parameters):

```plaintext
prior alpha ~ normal(0, sd=1);
prior beta ~ normal(0, sd=1, lower=alpha);
```

Programming Statements

Most DATA step operators, functions, and statements can be used in PROC MCMC:

- assignment and operators: +, -, *, /, <>, <, ...
- mathematical functions: ABS, LOG, PDF, CDF, SDF, LOGPDF, ...
- statements: CALL, DO, IF, PUT, WHEN, ...

The functions enable you to:

- compute functions of parameters
- construct general prior and/or likelihood functions
The MCMC Procedure

- A Primer on PROC MCMC
- Monte Carlo Simulation
- Single-level Model: Hyperparameters
- Generalized Linear Models
- Random-effects models
- Missing Data Analysis
- Survival Analysis (Optional)

Monte Carlo Simulation

\[ p \sim \beta(0.47, 2.35) \]

data a;
run;
proc mcmc data=a seed=17 nmc=20000;
parm p;
prior p ~ beta(0.47, 2.35);
model general(0);
run;

<table>
<thead>
<tr>
<th>Parameter</th>
<th>N</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
<td>20000</td>
<td>0.1662</td>
<td>0.1912</td>
<td>0.0199</td>
<td>0.0906</td>
<td>0.2512</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Alpha</th>
<th>Equal-Tail Interval</th>
<th>HPD Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
<td>0.050</td>
<td>0.000152</td>
<td>0.6866</td>
</tr>
</tbody>
</table>
Simple Simulation

If $p$ represents some process of binary success probabilities, then you might be interested in the outcome of a binomial trial where the probability is not known precisely:

$$p \sim \text{beta}(0.47, 2.35)$$

$$\text{success} \sim \text{binomial}(20, p)$$

```
proc mcmc data=a seed=17 nmc=20000
   outpost=o1;
   parm p success;
   prior p ~ beta(0.47, 2.35);
   prior success ~ binomial(20, p);
   model general(0);
run;
```

Estimates of $p$ and success

<table>
<thead>
<tr>
<th>Parameter</th>
<th>N</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Percentiles</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>25%</td>
</tr>
<tr>
<td>$p$</td>
<td>20000</td>
<td>0.1676</td>
<td>0.1908</td>
<td>0.0208</td>
</tr>
<tr>
<td>success</td>
<td>20000</td>
<td>3.3545</td>
<td>4.0982</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Alpha</th>
<th>Equal-Tail Interval</th>
<th>HPD Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p$</td>
<td>0.050</td>
<td>0.000157</td>
<td>0.6832</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.38E-10</td>
<td>0.5834</td>
</tr>
<tr>
<td>success</td>
<td>0.050</td>
<td>0</td>
<td>14.0000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>12.0000</td>
</tr>
</tbody>
</table>
Estimate Cumulative Probability

\[ p \sim \text{beta}(0.47, 2.35) \]
\[ \text{success} \sim \text{binomial}(20, p) \]

what is the \( \Pr(9 \leq \text{success} \leq 12) \)?

```
proc mcmc data=a seed=17 nmc=20000 outpost=o1 monitor=(prob);
    parm p success;
    prior p ~ beta(0.47, 2.35);
    prior success ~ binomial(20, p);
    prob = (9 <= success <= 12);
    model general(0);
run;
```

monitor keeps track of variables in a program

The estimated probability is 0.083.

First few samples of the OUTPOST data set:

<table>
<thead>
<tr>
<th>Obs</th>
<th>Iteration</th>
<th>p</th>
<th>success</th>
<th>prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.00691</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0.00369</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>0.4078</td>
<td>8.0000</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>0.0280</td>
<td>1.0000</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>0.0875</td>
<td>2.0000</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>N</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Percentiles</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>25%</td>
</tr>
<tr>
<td>prob</td>
<td>20000</td>
<td>0.0830</td>
<td>0.2759</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Alpha</th>
<th>Equal-Tail Interval</th>
<th>HPD Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>prob</td>
<td>0.050</td>
<td>0</td>
<td>1.000</td>
</tr>
</tbody>
</table>
Binomial model

Researchers are interested in evaluating the performance of a medical procedure in a multicenter study. The following statements create a data set for the treatment arm of the trials:

```plaintext
data trials;
   input event n center;
datalines;
   2  86  1
   2  69  2
   1  71  3
   1 113  4
   1 103  5
;
```

- **event**: number of deaths
- **n**: number of patients assigned to the treatment procedure
- **center**: center index
Binomial Example
Consider a simple binomial model

\[
event_i \sim \text{binomial}(n_i, p) \]
\[
p \sim \text{beta}(a, b)
\]

where \( p \) is the parameter of interest and \( a \) and \( b \) are hyper-parameters. Consider the following choices for \( a \) and \( b \):

- uniform: \( \text{beta}(1, 1) \)
- Haldane prior: \( \text{beta}(0, 0) \)

\[
\pi(p) \propto p^{-1}(1 - p)^{-1}
\]

Binomial Model with Flat Prior

```
proc mcmc data=trials seed=17 nmc=20000 outpost=UnifBin;
   parm p;
   prior p ~ beta(1,1);
   model event ~ binomial(n,p);
run;
```

<table>
<thead>
<tr>
<th>Parameter</th>
<th>N</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
<td>20000</td>
<td>0.0180</td>
<td>0.00624</td>
<td>0.0135</td>
<td>0.0174</td>
<td>0.0217</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Alpha</th>
<th>Equal-Tail Interval</th>
<th>HPD Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
<td>0.050</td>
<td>0.00784</td>
<td>0.0321</td>
</tr>
</tbody>
</table>
Binomial Model with Haldane Prior

The following syntax does not work in PROC MCMC,

```
prior p ~ beta(0, 0);
```

because the shape and scale parameters of a beta distribution must be positive.

Use the GENERAL function to construct nonstandard prior distribution.

### Specifying a Nonstandard Distribution

The GENERAL and DGENERAL functions enable you construct your own prior or likelihood function. The “D” stands for discrete.

```
PRIOR alpha ~ dgeneral(lp);
MODEL y ~ general(llike);
```

The expressions \( lp \) and \( llike \) must take the values of the logarithm of the distribution.

The normalizing constant of the distribution can be ignored, as long as it is independent of other parameters in the model.
The GENERAL Distribution

Suppose that you want to use the following prior:

\[ \pi(\sigma^2) \propto \frac{1}{\sigma^2} \]

which is a nonstandard distribution (nonintegrable prior). The logarithm of this prior is

\[ \log(\pi(\sigma^2)) = -\log(\sigma^2) + C \]

You use the following statements to declare this prior:

```plaintext
lp = -log(sigma2);
prior sigma2 ~ general(lp, lower=0);
```

More on the GENERAL Distribution

The function argument can be an expression or a constant. For example, to specify \( \pi(\alpha) \propto 1 \), you use the following statement:

```plaintext
prior alpha ~ general(0);
```

Use these functions with care because PROC MCMC cannot verify that the priors you specify lead to valid (integrable) posterior.

When in doubt, stay with proper distributions.
Binomial Model with Haldane Prior

\[ \pi(p) \propto p^{-1}(1 - p)^{-1} \]
\[ \Rightarrow \log(\pi(p)) = -(\log(p) + \log(1 - p)) \]

```
proc mcmc data=trials seed=17 nmc=20000 outpost=HalBin;
parm p 0.5;
lprior = -(log(p) + log(1-p));
prior p ~ general(lprior, lower=0, upper=1);
model event ~ binomial(n,p);
run;
```

**Posterior Summaries**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>N</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
<td>20000</td>
<td>0.0158</td>
<td>0.00595</td>
<td>0.0114</td>
<td>0.0150</td>
<td>0.0193</td>
</tr>
</tbody>
</table>

**Posterior Intervals**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Alpha</th>
<th>Equal-Tail Interval</th>
<th>HPD Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
<td>0.050</td>
<td>0.00640</td>
<td>0.0294</td>
</tr>
</tbody>
</table>

Binomial Model

Suppose that you do not want to have fixed hyperparameter values and want to consider hyperprior distributions on these parameters:

\[ \pi(a) \propto \text{exponential}(\text{scale} = 100) \]
\[ \pi(b) \propto \text{exponential}(\text{scale} = 100) \]

This prior has mean of 100 and variance 10,000. The following statements fit a hierarchical binomial model:

```
proc mcmc data=trials seed=17 nmc=10000 outpost=bmc;
parms p;
parms a b;
prior a b ~ expon(scale=100);
prior p ~ beta(a,b);
model event ~ binomial(n,p);
run;
```
Posterior Density Comparison

Having hyperprior distributions is essentially equivalent to using a uniform prior on $p$—there is no information in the data that can help with estimating the hyperparameters.

![Graph showing posterior density comparison between likelihood and hierarchical model]

Outline

3 The MCMC Procedure
   - A Primer on PROC MCMC
   - Monte Carlo Simulation
   - Single-level Model: Hyperparameters
   - Generalized Linear Models
     - Random-effects models
     - Missing Data Analysis
     - Survival Analysis (Optional)
Logistic Model

Crowder (1978) reported an experiment on germinating seeds. The data set is a $2 \times 2$ factorial layout with:

- Two types of seeds
- Two root extracts

The experiment included five or six replicates for each combination of seeds and root extracts.

A subset of the data:

<table>
<thead>
<tr>
<th>r</th>
<th>n</th>
<th>seed</th>
<th>extract</th>
<th>ind</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>39</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>23</td>
<td>62</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>23</td>
<td>81</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>26</td>
<td>51</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>17</td>
<td>39</td>
<td>0</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>0</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>53</td>
<td>74</td>
<td>0</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>55</td>
<td>72</td>
<td>0</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>32</td>
<td>51</td>
<td>0</td>
<td>1</td>
<td>9</td>
</tr>
</tbody>
</table>

Visualizing the Data Set

![Graph showing data visualization](image-url)
Logistic Regression

A natural way to model proportion data is to use the logistic regression with normal prior on the coefficients:

\[
\begin{align*}
  r_i & \sim \text{binomial}(n_i, p_i) \\
  \mu_i & = \beta_0 + \beta_1 \cdot \text{seed}_i + \beta_2 \cdot \text{extract}_i + \beta_3 \cdot \text{seed}_i \cdot \text{extract}_i \\
  \logit(p_i) & = \log \left( \frac{p_i}{1 - p_i} \right) = \mu_i \\
  \pi(\beta_0, \beta_1, \beta_2, \beta_3) & \propto \text{normal}(0, \text{sd} = 1000)
\end{align*}
\]

where \( i = \{1, \cdots, 21\} \).

Fitting Logistic Regression in PROC MCMC

```sas
proc mcmc data=seeds outpost=postout1 seed=332786 nmc=20000
   stats=(summary intervals) diag=none;
   parms beta0-beta3;
   prior beta: ~ normal(0, sd=1000);
   mu = beta0 + beta1*seed + beta2*extract + beta3*seed*extract;
   pi = logistic(mu);
   model r ~ binomial(n = n, p = pi);
run;
```

\[
\text{logistic} : p_i = \frac{\exp(\mu)}{1 + \exp(\mu)}
\]
Posterior Summary and Interval Statistics

### Posterior Summaries

<table>
<thead>
<tr>
<th>Parameter</th>
<th>N</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta0</td>
<td>20000</td>
<td>-0.5596</td>
<td>0.1256</td>
<td>-0.6457</td>
<td>-0.5583</td>
<td>-0.4704</td>
</tr>
<tr>
<td>beta1</td>
<td>20000</td>
<td>0.1444</td>
<td>0.2250</td>
<td>-0.00375</td>
<td>0.1458</td>
<td>0.2933</td>
</tr>
<tr>
<td>beta2</td>
<td>20000</td>
<td>1.3190</td>
<td>0.1792</td>
<td>1.1988</td>
<td>1.3189</td>
<td>1.4354</td>
</tr>
<tr>
<td>beta3</td>
<td>20000</td>
<td>-0.7723</td>
<td>0.3107</td>
<td>-0.9786</td>
<td>-0.7727</td>
<td>-0.5592</td>
</tr>
</tbody>
</table>

### Posterior Intervals

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Alpha</th>
<th>Equal-Tail Interval</th>
<th>HPD Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta0</td>
<td>0.050</td>
<td>-0.8175 -0.3210</td>
<td>-0.7989 -0.3059</td>
</tr>
<tr>
<td>beta1</td>
<td>0.050</td>
<td>-0.3015 0.5822</td>
<td>-0.3043 0.5763</td>
</tr>
<tr>
<td>beta2</td>
<td>0.050</td>
<td>0.9672 1.6852</td>
<td>0.9597 1.6752</td>
</tr>
<tr>
<td>beta3</td>
<td>0.050</td>
<td>-1.3891 -0.1716</td>
<td>-1.3474 -0.1414</td>
</tr>
</tbody>
</table>

Fit of the Logistic Model

![Graph showing the fit of the Logistic Model](image_url)
Probit Model

You can change from a logistic to a probit regression:

\[ y_i | p_i \sim \binom{n_i, p_i} \]
\[ \mu_i = \beta_0 + \beta_1 \cdot \text{seed}_i + \beta_2 \cdot \text{extract}_i + \beta_3 \cdot \text{seed}_i \cdot \text{extract}_i \]
\[ p_i = \Phi(\mu_i) \]
\[ \pi(\beta_0, \beta_1, \beta_2, \beta_3) \propto \text{normal}(0, \text{sd} = 1000) \]

```
proc mcmc data=seeds outpost=postout1 seed=332786 nmc=20000
   stats=(summary intervals) diag=none;
   parms beta0-beta3;
   prior beta: ~ normal(0, sd=1000);
   mu = beta0 + beta1*seed + beta2*extract + beta3*seed*extract;
   pi = cdf("normal", mu, 0, 1);
   model r ~ binomial(n = n, p = pi);
run;
```

Poisson Model

Or a Poisson model, if the response variable is count data:

\[ y_i | \lambda_i \sim \text{poisson}(\lambda_i) \]
\[ \mu_i = \beta_0 + \beta_1 \cdot \text{seed}_i + \beta_2 \cdot \text{extract}_i + \beta_3 \cdot \text{seed}_i \cdot \text{extract}_i \]
\[ \lambda_i = \exp(\mu_i) \]
\[ \pi(\beta_0, \beta_1, \beta_2, \beta_3) \propto \text{normal}(0, \text{sd} = 1000) \]

```
proc mcmc data=seeds outpost=postout1 seed=332786 nmc=20000
   stats=(summary intervals) diag=none;
   parms beta0-beta3;
   prior beta: ~ normal(0, sd=1000);
   mu = beta0 + beta1*seed + beta2*extract + beta3*seed*extract;
   lambda = exp(mu);
   model y ~ poisson(lambda);
run;
```
Random-Effects Model

Recall that in the trials analysis, we considered a simple model:

\[ \text{event}_i \sim \text{binomial}(n_i, p) \]
\[ p \sim \text{beta}(a, b) \]

which assumes that all groups share the same characteristic (success or failure probability).

Random-effects models enable you to model group-specific characteristics, such as different trials share similar but different failure probabilities:

\[ \text{event}_i \sim \text{binomial}(n_i, p_i) \]
\[ p_i \sim \text{prior} \]
A Typical Random-Effects Model

A generic setup of a random-effects model:

\[ Y_{ij} = \alpha \cdot X_{ij} + \beta_j + \epsilon_{ij}, \quad j = 1 \cdots J, \quad i = 1 \cdots n_j \] (1)

where

- \( Y_{ij} \) is the response value of the \( i \)th subject in the \( j \)th cluster
- \( J \) is the total number of clusters, and \( n_j \) is the total number of subjects in the \( j \)th cluster.
- \( \alpha \) is the fixed-effects parameter for \( X_{ij} \)
- \( \epsilon_{ij} \) are the i.i.d. errors from a common distribution
- \( \beta_j \) is the varying intercepts

Often it is assumed that \( \beta_j \) arise from the same distribution,

\[ \beta_j \sim \pi(\theta) \] (2)

where \( \theta \) are the hyperparameters.

Different Types of Random-Effects Models

- If \( \epsilon_{ij} \) in model (1) is assumed to have a normal distribution, then the model becomes a linear random-effects model.
- If you choose to model

\[ E(Y_{ij}) = g(\alpha \cdot X_{ij} + \beta_j) \]

where \( Y_{ij} \) is assumed to arise from the exponential family and \( g(\cdot) \) is a one-to-one monotone transformation, then the model becomes a generalized linear random-effects model.
- If \( Y_{ij} \) relates to the regression via nonlinear transformation, the model becomes a more general nonlinear random-effects model.
**Varying-Intercept Models**

\[ Y_{ij} = \alpha \cdot X_{ij} + \beta_j \]

**Varying-Slope Models**

\[ Y_{ij} = \gamma_j \cdot X_{ij} + C \]
Varying-Intercept and Varying-Slope Models

\[ Y_{ij} = \gamma_j \cdot X_{ij} + \beta_j \]

The RANDOM Statement

The RANDOM statement is designed to construct random-effects models in PROC MCMC. The statement

- specifies the random-effects parameters (\( \beta_j, \gamma_j \), and so on).
- makes the following conditional independence assumption:

\[
\beta_j \sim \pi(\theta) \\
\beta_i \perp \beta_j \quad \text{a priori}
\]

where \( \theta \) are the hyperparameters.
Syntax

`RANDOM random-effect ~ distribution SUBJECT= <options> ;`

- `random-effect`: defines the effect
- `distribution`: specifies its prior distribution
  - (beta/binary/gamma/igamma/laplace/normal/mvn/general)
- `SUBJECT=`: identifies group membership
- `options`: control initial values, monitoring list, and so on

Random Effects

- A program can have multiple RANDOM statements:
  - one for the classroom-level
  - one for the school-level
  - ...
- You can fit nested or nonnested models:
  - nested models: levels of one factor must cluster within the levels of another factor, such as students clustered within classrooms.
    - the classroom effect can be the hyperparameters to the student effects
  - nonnested models: levels of factors can cross, such as a student effect and an age effect;
- The effects can enter the model in any linear or nonlinear form
  - $\beta_j + \gamma_k$
  - $\exp(\beta_j)$
  - ...

...
The SUBJECT= Variable

This is a data set variable that indicates clustering of the random effects.

- The number of random-effects parameters in a RANDOM statement is determined by the number of unique values in the SUBJECT= variable.
- The SUBJECT= variable be numeric or character, and doesn’t have to be sorted

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>27513</td>
<td>07/13/1995</td>
<td>John</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>27513</td>
<td>01/31/2003</td>
<td>Mary</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>01440</td>
<td>10/12/1997</td>
<td>Ken</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>17923</td>
<td>08/03/2010</td>
<td>John</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Understand the SUBJECT= Syntax

<table>
<thead>
<tr>
<th>y</th>
<th>x</th>
<th>gender</th>
<th>group</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>13</td>
<td>female</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>male</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>11</td>
<td>female</td>
<td>3</td>
</tr>
<tr>
<td>0</td>
<td>7</td>
<td>male</td>
<td>4</td>
</tr>
<tr>
<td>0</td>
<td>10</td>
<td>female</td>
<td>5</td>
</tr>
</tbody>
</table>

We want to fit a model with two random effects, gender \((\alpha_j)\) and group \((\beta_k)\), that has this general form:

\[
\mu_i = g(\alpha_j \cdot x_i + \beta_k) \\
y_i \sim \text{dist}(\mu_i)
\]

where \(i = \{1, \cdots, 5\}\), \(j = \{1, 2\}\), and \(k = \{1, \cdots, 5\}\).
Understand the SUBJECT= Syntax

random alpha ~ dist() subject=gender;
random beta ~ dist() subject=group;
mu = g(alpha * x + beta);
model y ~ dist(mu);

PROC MCMC internally creates two $\alpha$ parameters ($\alpha_{\text{female}}$, $\alpha_{\text{male}}$), five $\beta$ parameters ($\beta_1, \cdots, \beta_5$), and interprets the input data set as:

<table>
<thead>
<tr>
<th>$y$</th>
<th>$x$</th>
<th>gender</th>
<th>group</th>
<th>equation processed</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>13</td>
<td>$\alpha_{\text{female}}$</td>
<td>$\beta_1$</td>
<td>$\mu = g(\alpha_{\text{female}} \cdot 13 + \beta_1)$</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>$\alpha_{\text{male}}$</td>
<td>$\beta_2$</td>
<td>$\mu = g(\alpha_{\text{male}} \cdot 10 + \beta_2)$</td>
</tr>
<tr>
<td>1</td>
<td>11</td>
<td>$\alpha_{\text{female}}$</td>
<td>$\beta_3$</td>
<td>$\mu = g(\alpha_{\text{female}} \cdot 11 + \beta_3)$</td>
</tr>
<tr>
<td>0</td>
<td>7</td>
<td>$\alpha_{\text{male}}$</td>
<td>$\beta_4$</td>
<td>$\mu = g(\alpha_{\text{male}} \cdot 7 + \beta_4)$</td>
</tr>
<tr>
<td>0</td>
<td>10</td>
<td>$\alpha_{\text{female}}$</td>
<td>$\beta_5$</td>
<td>$\mu = g(\alpha_{\text{female}} \cdot 10 + \beta_5)$</td>
</tr>
</tbody>
</table>

Outline

3 The MCMC Procedure
- A Primer on PROC MCMC
- Monte Carlo Simulation
- Single-level Model: Hyperparameters
- Generalized Linear Models
- Random-effects models
  - Introduction
  - Logistic Regression - Overdispersion
  - Hyperpriors in Random-Effects Models - Shrinkage
  - Repeated Measurements Models
- Missing Data Analysis
- Survival Analysis (Optional)
Recall the Logistic Example

Excessive Variation

Crowder (1978) noted heterogeneity of the proportions between replicates. To account for excessive variation, Brewslow and Clayton (1993) suggested a random-effects logistic regression:

\[
\begin{align*}
  r_i & \sim \text{binomial}(n_i, p_i) \\
  \mu_i & = \beta_0 + \beta_1 \cdot \text{seed}_i + \beta_2 \cdot \text{extract}_i + \beta_3 \cdot \text{seed}_i \cdot \text{extract}_i \\
  p_i & = \text{logistic}(\mu_i + \delta_i) \\
  \pi(\beta_0, \beta_1, \beta_2, \beta_3) & \propto \text{normal}(0, \text{sd} = 1000) \\
  \delta_i & \sim \text{normal}(0, \text{var} = \sigma^2) \\
  \sigma^2 & \sim \text{igamma}(\text{shape} = 0.01, \text{scale} = 0.01)
\end{align*}
\]

where $\delta_i$ is the random-effects parameter, and $\sigma^2$ is the hyperparameter variance.
Random-Effects Logistic Regression

The following program fits a logistic random-effects model:

```
proc mcmc data=seeds outpost=postout seed=332786 nmc=20000
stats=(summary intervals) diag=none;
parms beta0-beta3 s2 1;
prior beta: ~ normal(0, sd=1000);
prior s2 ~ igamma(0.01, s=0.01);
mu = beta0 + beta1*seed + beta2*extract + beta3*seed*extract;
random delta ~ normal(0, var=s2) subject=_obs_;
pi = logistic(mu + delta);
model r ~ binomial(n = n, p = pi);
run;
```

`subject=_obs_` : fits observational level random effects

Model Parameters Information

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Block</th>
<th>Parameter</th>
<th>Sampling Method</th>
<th>Initial Value</th>
<th>Prior Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>s2</td>
<td>Conjugate</td>
<td>0.00990</td>
<td>igamma(0.01, s=0.01)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>beta0</td>
<td>N-Metropolis</td>
<td>0</td>
<td>normal(0, sd=1000)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>beta1</td>
<td></td>
<td>0</td>
<td>normal(0, sd=1000)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>beta2</td>
<td></td>
<td>0</td>
<td>normal(0, sd=1000)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>beta3</td>
<td></td>
<td>0</td>
<td>normal(0, sd=1000)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random Effect Parameters</th>
<th>Parameter</th>
<th>Sampling Method</th>
<th>Subject</th>
<th>Number of Subjects</th>
<th>Subject Values</th>
<th>Prior Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>delta</td>
<td>N-Metropolis</td>
<td><em>obs</em></td>
<td>21</td>
<td>1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 ...</td>
<td>normal(0, var=s2)</td>
</tr>
</tbody>
</table>
Posterior Inference

By default, PROC MCMC does not display posterior estimates of the random-effects parameters (there could be too many):

```
Posterior Summaries

<table>
<thead>
<tr>
<th>Parameter</th>
<th>N</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta0</td>
<td>20000</td>
<td>-0.5519</td>
<td>0.1943</td>
<td>-0.6799</td>
<td>-0.5523</td>
<td>-0.4259</td>
</tr>
<tr>
<td>beta1</td>
<td>20000</td>
<td>0.0986</td>
<td>0.3167</td>
<td>-0.1026</td>
<td>0.1119</td>
<td>0.3075</td>
</tr>
<tr>
<td>beta2</td>
<td>20000</td>
<td>1.3606</td>
<td>0.2815</td>
<td>1.1844</td>
<td>1.3580</td>
<td>1.5385</td>
</tr>
<tr>
<td>beta3</td>
<td>20000</td>
<td>-0.8785</td>
<td>0.4589</td>
<td>-1.1652</td>
<td>-0.8824</td>
<td>-0.5905</td>
</tr>
<tr>
<td>s2</td>
<td>20000</td>
<td>0.1239</td>
<td>0.1099</td>
<td>0.0528</td>
<td>0.0968</td>
<td>0.1619</td>
</tr>
</tbody>
</table>
```

```
Posterior Intervals

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Alpha</th>
<th>Equal-Tail Interval</th>
<th>HPD Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta0</td>
<td>0.050</td>
<td>-0.9347</td>
<td>-0.1547</td>
</tr>
<tr>
<td>beta1</td>
<td>0.050</td>
<td>-0.5636</td>
<td>0.7045</td>
</tr>
<tr>
<td>beta2</td>
<td>0.050</td>
<td>0.8072</td>
<td>1.9280</td>
</tr>
<tr>
<td>beta3</td>
<td>0.050</td>
<td>-1.7696</td>
<td>0.0271</td>
</tr>
<tr>
<td>s2</td>
<td>0.050</td>
<td>0.00895</td>
<td>0.3964</td>
</tr>
</tbody>
</table>
```

Monitoring Random-Effects Parameters

The MONITOR= option enables you to display

- all of these random-effects estimates:

```
random delta ~ n(0, var=s2) subject=_obs_ monitor=(delta);
```

- a subset of these estimates:

```
random delta ~ n(0, var=s2) subject=_obs_ monitor=(1 2 6);
```

- have the procedure choose a subset:

```
random delta ~ n(0, var=s2) subject=_obs_ monitor=(random(3));
```
Monitored Random-Effects Parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>N</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Percentiles</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta0</td>
<td>20000</td>
<td>-0.5519</td>
<td>0.1943</td>
<td>-0.6799 -0.5523 -0.4259</td>
</tr>
<tr>
<td>beta1</td>
<td>20000</td>
<td>0.0986</td>
<td>0.3167</td>
<td>-0.1026 0.1119 0.3075</td>
</tr>
<tr>
<td>beta2</td>
<td>20000</td>
<td>1.3606</td>
<td>0.2815</td>
<td>1.1844 1.3580 1.5385</td>
</tr>
<tr>
<td>beta3</td>
<td>20000</td>
<td>-0.8785</td>
<td>0.4589</td>
<td>-1.1652 -0.8824 -0.5905</td>
</tr>
<tr>
<td>s2</td>
<td>20000</td>
<td>0.1239</td>
<td>0.1099</td>
<td>0.0528 0.0968 0.1619</td>
</tr>
<tr>
<td>delta_7</td>
<td>20000</td>
<td>0.0644</td>
<td>0.2403</td>
<td>-0.0809 0.0595 0.2175</td>
</tr>
<tr>
<td>delta_8</td>
<td>20000</td>
<td>0.1972</td>
<td>0.2508</td>
<td>0.0301 0.1821 0.3555</td>
</tr>
<tr>
<td>delta_19</td>
<td>20000</td>
<td>-0.00081</td>
<td>0.2869</td>
<td>-0.1650 -0.00755 0.1563</td>
</tr>
</tbody>
</table>

Fit of the Random-Effects Model
Caterpillar Plot

- A side-by-side bar plot of the 95% equal-tail posterior intervals for multiple parameters
- Useful in visualizing and comparing parameters.
- Better than overlay kernel density plots.

The `%CATER` autocall macro creates a caterpillar plot:

```
%cater(data=postout, var=delta_:);
```

![Caterpillar Plot](image-url)
Hyperprior in Random-Effects Models

Back to the trials analysis model:

\[ \text{event}_i \sim \text{binomial}(n_i, p_i) \]

where \(i\) indexes the group.

The group-specific \(p_i\) is a weighted average of the pooled estimate (shared \(p\)) and independent estimates (separate analysis). The amount of shrinkage is determined by the hyperprior distribution.

Here we consider two common choices:

\[ p_i \sim \text{beta}(a, b) \]
\[ \logit(p_i) \sim \text{normal}(\mu, \sigma^2) \]
Hyperparameters

If you choose constant values for $a$, $b$, or $\sigma^2$, you decide a priori the amount of shrinkage you want on the $p_i$. For example:

- Choosing $a = 1$ and $b = 1$, or $\sigma^2 = \infty$, implies no shrinkage on the $p_i$. The random-effects model becomes an independent model.
- Choosing $\sigma^2 = 0$ imposes no variation amongst $p_i$. This reduces the random-effects model to the pooled model.

Empirical Bayes offers one way of choosing these hyperparameters.

- find estimates $a$, $b$, $\mu$, or $\sigma^2$ by maximizing the posterior marginal distributions of $\pi(a, b|\mathbf{x}, \mathbf{y})$ or $\pi(\mu, \sigma^2|\mathbf{x}, \mathbf{y})$
- plug in these estimates as the hyperparameters

This provides reasonable inferences if there are enough units or groups in the data to estimate the variance.

But the plug-in approach ignores uncertainty that your data indicates about the amount of shrinkage that should be used.
Hyperprior Distributions

Ideally, the data should decide the right amount of strength you want to borrow from different groups in the analysis. This amounts to placing hyperprior distributions on the hyperparameters.

For example, see Spiegelhalter, Abrams, and Myles (2004), and Gelman et al. (2003) for discussions on how to select such prior distributions. Strategies include:

- noninformative
- elicitation
- summary of evidence

First, let's consider the beta hyperprior model and use proper but diffuse prior distributions on $a$ and $b$:

$$\text{event}_i \sim \text{binomial}(n_i, p_i)$$
$$p_i \sim \text{beta}(a, b)$$
$$a, b \sim \text{exponential}($$scale = 100$$)$$

```plaintext
proc mcmc data=trials nmc=50000 outpost=outm seed=17;
  parm a b;
  prior a b ~ expon(scale=100);
  random p ~ beta(a, b) subject=center;
  model event ~ binomial(n, p);
run;
```
95% HPD intervals and estimates for $p_i$. The solid line is the random-effects model; the dashed line is the independence model (individual analysis); the bottom line is the overall (pooled) model.

**Modeling $\sigma^2$ in Hierarchical Models**

Secondly, we consider the following model:

\[
\begin{align*}
\text{event}_i & \sim \text{binomial}(n_i, p_i) \\
\gamma_i = \logit(p_i) & \sim \text{normal}(\mu, \sigma^2) \\
\mu & \sim \text{normal}(0, \text{precision} = 10^{-6})
\end{align*}
\]

What type of (noninformative) prior distribution should be used on $\sigma^2$?
Some Frequently used Prior Distributions for $\sigma^2$

- The Jeffreys' prior is a popular choice for the variance parameter in a normal model:
  \[ \pi(\sigma^2) \propto \frac{1}{\sigma^2} \]
  which is equivalent to a uniform prior on $\log(\sigma^2)$
  \[
  \text{prior } s2 \sim \text{general}(-\log(s2), \text{lower}=0);
  \]
  BUT, this leads to an improper posterior distribution in the random-effects model and SHOULD NOT BE USED.

- Uniform on variance:
  \[ \pi(\sigma^2) \propto 1 \]
  \[
  \text{prior } s2 \sim \text{general}(0, \text{lower}=0);
  \]
  When there are few groups, $\pi(\sigma) \propto 1$ if often recommended.

- Conjugate prior:
  \[ \pi(\sigma^2) \propto i\Gamma(\text{shape}=\alpha, \text{scale}=\beta) \]
  \[
  \text{prior } s2 \sim \text{igamma}(\text{shape}=\alpha, \text{scale}=\beta);
  \]
  \[ \pi(\sigma^2) \propto i\Gamma(\epsilon, \epsilon) \Leftrightarrow \pi(\tau = \frac{1}{\sigma^2}) \propto \Gamma(\epsilon, \text{iscale} = \epsilon) \]
  with $\epsilon = 0.001$ being used frequently. This is a prior that “mimics” the Jeffreys’ prior (However, it can be highly influential).

For more detailed discussion on noninformative prior selections on $\sigma^2$, see Gelman (2006, *Bayesian Analysis* 1:515).
Fitting the Model in PROC MCMC

For illustrative purposes, consider

$$\sigma^2 \sim \text{igamma}(0.001, \text{scale} = 0.001)$$

```sas
proc mcmc data=trials nmc=50000 outpost=outmc seed=17;
    parms mu s2;
    prior mu ~ n(0, prec=1e-6);
    prior s2 ~ igamma(0.001, s=0.001);
    random gamma ~ n(mu, var=s2) subject=center;
    logitP = logistic(gamma);
    model event ~ binomial(n, logitP);
run;
```

logistic : $logitP = \frac{\exp(\gamma)}{1+\exp(\gamma)}$

Posterior Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>N</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
</tr>
</thead>
<tbody>
<tr>
<td>mu</td>
<td>50000</td>
<td>-3.9977</td>
<td>0.3544</td>
<td>-4.2103</td>
<td>-3.9641</td>
<td>-3.7630</td>
</tr>
<tr>
<td>s2</td>
<td>50000</td>
<td>0.1460</td>
<td>0.5927</td>
<td>0.00578</td>
<td>0.0239</td>
<td>0.1060</td>
</tr>
<tr>
<td>gamma_1</td>
<td>50000</td>
<td>-3.9709</td>
<td>0.3923</td>
<td>-4.2067</td>
<td>-3.9450</td>
<td>-3.7151</td>
</tr>
<tr>
<td>gamma_2</td>
<td>50000</td>
<td>-3.9452</td>
<td>0.4051</td>
<td>-4.1859</td>
<td>-3.9312</td>
<td>-3.6923</td>
</tr>
<tr>
<td>gamma_3</td>
<td>50000</td>
<td>-4.0280</td>
<td>0.4235</td>
<td>-4.2615</td>
<td>-3.9840</td>
<td>-3.7524</td>
</tr>
<tr>
<td>gamma_4</td>
<td>50000</td>
<td>-4.0751</td>
<td>0.4441</td>
<td>-4.2995</td>
<td>-4.0066</td>
<td>-3.7916</td>
</tr>
<tr>
<td>gamma_5</td>
<td>50000</td>
<td>-4.0667</td>
<td>0.4266</td>
<td>-4.2979</td>
<td>-4.0098</td>
<td>-3.7870</td>
</tr>
<tr>
<td>gamma_6</td>
<td>50000</td>
<td>-3.8935</td>
<td>0.3645</td>
<td>-4.1182</td>
<td>-3.8869</td>
<td>-3.6530</td>
</tr>
</tbody>
</table>
Transform $\gamma_i$ to $p_i$ Parameters

```
proc mcmc data=trials nmc=50000 outpost=outmc seed=17
    monitor=(mu s2 p);
array p[6];
parms mu s2;
prior mu ~ n(0, prec=1e-6);
prior s2 ~ igamma(0.001, s=0.001);
random gamma ~ n(mu, var=s2) subject=center;
logitP = logistic(gamma);
model event ~ binomial(n, logitP);
p[center] = logitP;
run;
```

**ARRAY** : allocates an array to store $p_i$, which are functions of $\gamma_i$

**MONITOR** : outputs model parameters and all elements of the array $p$

$p : p[\text{center}]$ saves the correct transformation for each $\gamma_i$ according to its cluster/center membership.

### Posterior Estimates of the $p_i$ Parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>N</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
</tr>
</thead>
<tbody>
<tr>
<td>mu</td>
<td>50000</td>
<td>-3.9977</td>
<td>0.3544</td>
<td>-4.2103</td>
<td>-3.9641</td>
<td>-3.7630</td>
</tr>
<tr>
<td>s2</td>
<td>50000</td>
<td>0.1460</td>
<td>0.5927</td>
<td>0.00578</td>
<td>0.0239</td>
<td>0.1060</td>
</tr>
<tr>
<td>p1</td>
<td>50000</td>
<td>0.0198</td>
<td>0.00757</td>
<td>0.0147</td>
<td>0.0190</td>
<td>0.0238</td>
</tr>
<tr>
<td>p2</td>
<td>50000</td>
<td>0.0205</td>
<td>0.00842</td>
<td>0.0150</td>
<td>0.0192</td>
<td>0.0243</td>
</tr>
<tr>
<td>p3</td>
<td>50000</td>
<td>0.0189</td>
<td>0.00727</td>
<td>0.0139</td>
<td>0.0183</td>
<td>0.0229</td>
</tr>
<tr>
<td>p4</td>
<td>50000</td>
<td>0.0181</td>
<td>0.00677</td>
<td>0.0134</td>
<td>0.0179</td>
<td>0.0221</td>
</tr>
<tr>
<td>p5</td>
<td>50000</td>
<td>0.0182</td>
<td>0.00683</td>
<td>0.0134</td>
<td>0.0178</td>
<td>0.0222</td>
</tr>
<tr>
<td>p6</td>
<td>50000</td>
<td>0.0212</td>
<td>0.00761</td>
<td>0.0160</td>
<td>0.0201</td>
<td>0.0253</td>
</tr>
</tbody>
</table>
Prior Matters!

Top panels are $\pi(\sigma|y)$ based on three different prior distributions: uniform on $\sigma$, $i\Gamma(1,1)$ and $i\Gamma(0.001, 0.001)$ on $\sigma^2$. The popular choice of inverse-Gamma distribution is hardly noninformative.

Posterior Estimates of Probabilities

95% HPD credible intervals and posterior point estimates for each $p_i$. There is an excessive amount of shrinkage using the $i\Gamma(0.001, 0.001)$ prior.
Shrinkage Effects

Outline

3 The MCMC Procedure
- A Primer on PROC MCMC
- Monte Carlo Simulation
- Single-level Model: Hyperparameters
- Generalized Linear Models
- Random-effects models
  - Introduction
  - Logistic Regression - Overdispersion
  - Hyperpriors in Random-Effects Models - Shrinkage
- Repeated Measurements Models
- Missing Data Analysis
- Survival Analysis (Optional)
Repeated Measurements

- Individual subjects have repeated observations, for example, over time (longitudinal), or within subjects (test scores).
- Different from time series data in the sense that the number of measurements per subject is generally not very large.
- Covariates information is available either at subject or measurement level.
- Subjects can have same number of repeated measures (balanced) or uneven number of measures (unbalanced).

Two-Arm Study

- The data comes from a two arms (control vs treatment) experiment that is carried out at eight sites.
- The response variables are success counts, $y_c$ and $y_t$, out of the same number of trials, $n_c = 132$ and $n_t = 148$, respectively.
- Each site sees different numbers of repeats (from 1 up to 13).
- Additional covariates information is withheld (but you can easily add them).
Control Data \((n_c = 132)\)

<table>
<thead>
<tr>
<th>id</th>
<th>yc1</th>
<th>yc2</th>
<th>yc3</th>
<th>yc4</th>
<th>yc5</th>
<th>yc6</th>
<th>yc7</th>
<th>yc8</th>
<th>yc9</th>
<th>yc10</th>
<th>yc11</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>40</td>
<td>26</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>1</td>
<td>7</td>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>2</td>
<td>43</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>2</td>
<td>8</td>
<td>20</td>
<td>1</td>
<td>8</td>
<td>14</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Treatment Data \((n_t = 148)\)

<table>
<thead>
<tr>
<th>id</th>
<th>yt1</th>
<th>yt2</th>
<th>yt3</th>
<th>yt4</th>
<th>yt5</th>
<th>yt6</th>
<th>yt7</th>
<th>yt8</th>
<th>yt9</th>
<th>yt10</th>
<th>yt11</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>57</td>
<td>34</td>
<td>2</td>
<td>3</td>
<td>27</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>2</td>
<td>2</td>
<td>24</td>
<td>3</td>
<td>2</td>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>2</td>
<td>75</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>13</td>
<td>28</td>
<td>2</td>
<td>13</td>
<td>15</td>
<td>3</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>13</td>
<td>6</td>
<td>8</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Input data is in a vector format.

<table>
<thead>
<tr>
<th>id</th>
<th>rep</th>
<th>yc</th>
<th>yt</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>40</td>
<td>57</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>26</td>
<td>34</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>20</td>
<td>27</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>10</td>
<td>24</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>19</td>
<td>19</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>43</td>
<td>75</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>

Simple Model

For every \(i = \{1, \ldots, 39\}\)

\[
\begin{align*}
y_{ci} & \sim \text{binomial} \left(n_c, p_c\right) \\
\text{logit}(p_c) & = \gamma \\
y_{ti} & \sim \text{binomial} \left(n_t, p_t\right) \\
\text{logit}(p_t) & = \gamma + \theta \\
\gamma, \theta & \sim \text{normal}(0, sd = 10)
\end{align*}
\]

where \(\gamma\) is the baseline (for control group) and \(\theta\) is the treatment effect (log of odds ratio).

Some quantities of interest could be:
- Difference in probabilities: \(p_t - p_c\)
- Odds ratio: \(\exp(\theta)\)
Fitting the Population Model

```sas
proc mcmc data=TwoArms nmc=20000 seed=1
    monitor=(p_diff or);
    parm gamma theta;
    prior gamma theta ~ n(0, sd=10);
    pc = logistic(gamma);
    model yc ~ binomial(132, pc);
    pt = logistic(gamma + theta);
    model yt ~ binomial(148, pt);
    or = exp(theta);
    p_diff = pt - pc;
run;
```

Observational-level Model

For every $i = \{1, \cdots, 39\}$

\[
\begin{align*}
  y_{ci} & \sim \text{binomial}(n_c, p_{ci}) \\
  \logit(p_{ci}) & = \gamma_i \\
  \gamma_i & \sim \text{normal}(\mu_\gamma, \tau_\gamma) \\
  y_{ti} & \sim \text{binomial}(n_t, p_{ti}) \\
  \logit(p_{ti}) & = \gamma_i + \theta_i \\
  \theta_i & \sim \text{normal}(\mu_\theta, \tau_\theta) \\
  \mu_\gamma, \mu_\theta & \sim \text{normal}(0, sd = 10) \\
  \tau_\gamma, \tau_\theta & \sim \Gamma(3, iscale = 1)
\end{align*}
\]

The number of parameters jumps from 2 to 82.
Fitting the Observational-Level Model

```plaintext
proc mcmc data=TwoArms nmc=20000 seed=1 outpost=ObsOut;
   parms mu_g mu_t tau_g tau_t;
   prior mu_g mu_t ~ n(0, prec=0.1);
   prior tau_g tau_t ~ gamma(3, iscale=1);
   random gamma ~ n(mu_g, prec=tau_g) subject=_obs_;
   random theta ~ n(mu_t, prec=tau_t) subject=_obs_;

   pc = logistic(gamma);
   model yc ~ binomial(132, pc);
   pt = logistic(gamma + theta);
   model yt ~ binomial(148, pt);
run;

%ess(data=ObsOut, var=mu: tau: gamma: theta:, out=ess);
```

ESSs of All Parameters

- $\mu_t$, smallest ESS of 823.
Observational-level odds ratios are rather similar to each other, indicating maybe an over-simplifying assumption on the random effects.

It is more realistic to consider the site-effect in the model, where the repeated measures are modelled as similar to each other within each site:

\[ y_{ci} \sim \text{binomial} (n_c, p_{ci}) \]
\[ \text{logit}(p_{ci}) = \gamma_i \]
\[ \gamma\{i,(j)\} \sim \text{normal}(\mu_{\gamma}, \tau_{\gamma}) \]
\[ \mu_{\gamma} \sim \text{normal}(0, \text{sd} = 10) \]
\[ \tau_{\gamma}, \tau_{\gamma,j} \sim \Gamma(3, \text{iscale} = 1) \]

\[ y_{ti} \sim \text{binomial} (n_t, p_{ti}) \]
\[ \text{logit}(p_{ti}) = \gamma_i + \theta_i \]
\[ \theta\{i,(j)\} \sim \text{normal}(\mu_{\theta}, \tau_{\theta}) \]
\[ \mu_{\theta} \sim \text{normal}(0, \text{sd} = 10) \]
\[ \tau_{\theta}, \tau_{\theta,j} \sim \Gamma(3, \text{iscale} = 1) \]

where \( i = \{1, \cdots , 39\} \) indexes observations, \( j = \{1, \cdots , 8\} \) indexes sites, and \( \{i,(j)\} \) indexes repeated measures in the \( j \)th site.
The MCMC Procedure

Random-effects models

Fitting the Observational-Level Model

```sas
proc mcmc data=TwoArms nmc=20000 seed=1 outpost=MultOut;
  parms mu_g0 mu_t0 tau_g0 tau_t0;

  prior mu_g0 mu_t0 ~ n(0, prec=0.1);  
  prior tau_g0 tau_t0 ~ gamma(3, iscale=1);

  random mu_g ~ n(mu_g0, prec=tau_g0) subject=id;  
  random tau_g ~ gamma(shape=3, iscale=1) subject=id;  
  random gamma ~ n(mu_g, prec=tau_g) subject=_obs_;  

  random mu_t ~ n(mu_t0, prec=tau_t0) subject=id;  
  random tau_t ~ gamma(shape=3, iscale=1) subject=id;  
  random theta ~ n(mu_t, prec=tau_t) subject=_obs_;  

  pc = logistic(gamma);
  model yc ~ binomial(132, pc);
  pt = logistic(gamma + theta);
  model yt ~ binomial(148, pt);
run;
```

The SUBJECT= variables must be nested.

The MCMC Procedure

Random-effects models

Odds Ratios Estimates by the Multi-level Model

Caterpillar Plot

```
OR_OBS1  OR_OBS3  OR_OBS5  OR_OBS7  OR_OBS9  OR_OBS11  OR_OBS13  OR_OBS15
OR_OBS17  OR_OBS19  OR_OBS21  OR_OBS23  OR_OBS25  OR_OBS27  OR_OBS29  OR_OBS31
OR_OBS33  OR_OBS35  OR_OBS37  OR_OBS39
```

230 / 295
Posterior Densities of $\mu_{\theta_j}$

Outline

3 The MCMC Procedure
- A Primer on PROC MCMC
- Monte Carlo Simulation
- Single-level Model: Hyperparameters
- Generalized Linear Models
- Random-effects models
- Missing Data Analysis
  - Introduction
  - Bivariate Normal with Partial Missing
  - Nonignorable Missing (Selection Model)
  - Survival Analysis (Optional)
Introduction

- Missing data problems arise frequently in practice and are caused by many circumstances.
  - study subjects might fail to answer questions on a questionnaire,
  - data can be lost
  - covariate measurements might be unavailable
  - and so on...

- The impact of missing data on inference is potentially important, especially if subjects that have missing data differ systematically from those that have complete data.

- Coherent estimation and valid inference require adequate modeling of the missing values.

Bayesian Approach

Bayesian methods for missing data problems are straightforward.

- Bayesian paradigm treats any unknown quantities as random variables.
- Missing values are treated as additional variables that need to be estimated.
- The approach is very general and capable of handling complex missing data scenarios.
Notations

- Let \( Y = \{ Y_{\text{obs}}, Y_{\text{mis}} \} \) be the response variable of length \( n \) (\( \{y_i\} \)), where \( Y_{\text{obs}} \) and \( Y_{\text{mis}} \) denote the observed and missing values, respectively.
- The sampling distribution is assumed to have the generic form:
  \[
  y_i \sim f(y_i|x_i, \theta)
  \]
  where \( f(\cdot) \) is a known distribution (e.g. the likelihood), \( x_i \) are the covariates, and \( \theta \) is the parameter of interest.
- Let \( R_Y = (r_1, \ldots, r_n) \) be the missing value indicator, also called the missingness random variable, where \( r_{yi} = 1 \) if \( y_i \) is missing and \( r_{yi} = 0 \) otherwise. \( R \) is known when the \( Y \) are known.

Covariates Missing

Similar to missing in response variables, you can have missing covariates.
- Let \( X = \{ X_{\text{obs}}, X_{\text{mis}} \} \), and \( X \) can be multidimensional.
- Typically, covariates are considered to be fixed constants. But here, \( X \) is treated as a random variable:
  \[
  x_i \sim \pi(x_i|u_i, \eta)
  \]
  where \( \pi(\cdot) \) is a “prior” distribution. The model can have \( u_i \) as additional covariates and \( \eta \) the parameter of interest.
- Similarly, you have \( R_X \) the missing value indicator for each covariate \( X \).
Objective

In the Bayesian approach, you can estimate the joint posterior distribution:

$$\pi(\theta, \eta, y_{mis}, x_{mis} | y_{obs}, x_{obs}, u)$$

The Monte Carlo in MCMC enables you to obtain the posterior marginal distributions for the parameters of interest:

$$\pi(\theta | y_{obs}, x_{obs}, u)$$
$$\pi(\eta | y_{obs}, x_{obs}, u)$$

Uncertainty about the missing values is fully propagated and incorporated in your inferences.

Classifications of Missing Data

Generally speaking, there are three types of missing data models (Rubin 1976):

- Missing Completely at Random
- Missing at Random
- Missing not at Random
  - selection model
  - pattern-mixture model
Missing Complete at Random

- **Missing Complete at Random (MCAR)** – if the failure to observe a value does not depend on any data, observed or missing.
  - The probability of observing a missing $y_i$ is independent of other $y_j$, for $j \neq i$, and is independent of other covariates $x_i$.
- Under the MCAR assumption, you can use only the observed data in the analysis. This is called a **complete-case (CC)** analysis.
  - If the MCAR assumption fails to hold, a CC analysis is biased.
  - If the MCAR assumption holds, a CC analysis is unbiased but less efficient than an analysis that uses the full data.

Missing at Random

- **Missing at Random (MAR)** – if the failure to observe a value is independent of missing values but may depend on observed value.
- MCAR assumes that the observed quantities are no longer random samples and adjustments should be made accordingly (a more realistic assumption than MCAR).
- In MAR, the missing mechanism, $(R_Y)$, does not need to be modeled and can be ignored.
- MAR is sometimes referred to as **ignorable missing**; it is not the missing values but the missing mechanism that can be ignored.
Missing not at Random

- **Missing not at Random (MNAR)** – if the failure to observe a value depends on unobserved observations (the would-have-been values).
- MNAR is the most general and most complex missing data scenario, and is frequently encountered in longitudinal studies with repeated measures.
  - In a Quality of Life (QOL) study, a patient can drop out depends on how sick they are, which is unobserved.
- The missing mechanism is no longer ignored and a model for \( R_Y \) is required. MNAR is sometimes referred to as *nonignorable missing*.

## Modeling MNAR

In MNAR, you have a joint likelihood function over \((R, Y)\):

\[
f_{R,Y}(r, y|x, \theta)
\]

- The **selection model** factors joint distribution into:

\[
f(r, y|x, \theta) \propto f(y|x, \alpha) \cdot f(r|y, x, \beta)
\]

where \( \theta = (\alpha, \beta) \).
  - \( f(y|x, \alpha) \), known as the *outcome model*, is the typical likelihood.
  - \( f(r|y, x, \beta) \): typically a binary model
- The **pattern-mixture model** factors the opposite way:

\[
f(r, y|x, \theta) \propto f(y|r, x, \delta) \cdot f(r|x, \gamma)
\]

where \( \theta = (\gamma, \delta) \).
Selection vs Pattern-Mixture

Some prefer the selection approach:
- It is a natural way of decomposing the joint distribution.
- In MAR analysis, you don’t need to include $R$ in the analysis.
- When MNAR analysis is required, adding the conditional model is easy.

Others prefer the pattern-mixture approach:
- The marginal model can model different patterns in $R$.
- You can build meaningful models for subsets of the response variable conditional on different missing patterns.
- On the other hand, you must always model $R$.

Handling of Missing Values in PROC MCMC

The MODEL statement handles the estimation of all missing values:

```
MODEL variable-list ~ distribution / <options> ;
```

- The distribution is the usual likelihood function when the MODEL statement is applied to a response variable;
- It becomes a prior distribution for a covariate;

The procedure steps through the input data set, identifies all missing values that are in `variable-list`, creates a separate parameter for each missing value, and draw samples from their posterior distributions.
Handling of Missing Values in PROC MCMC

- PROC MCMC models missing values only for variables that are specified in the MODEL statement. For example, suppose that there are missing values in y:

  ```
  MODEL y ~ normal(mu, var=1);
  ```

  Each missing value in y becomes a parameter and is sampled in the Markov chain.

- Records that contain missing values in other data set variables (that are not in the MODEL statement) are discarded. Suppose that there are missing values in x:

  ```
  mu = beta0 + beta1 * x;
  MODEL y ~ normal(mu, var=1);
  ```

  PROC MCMC does not model any missing values in x.

Options in the MODEL Statement

The options in the MODEL statement are available only when there are missing values in the variables:

```
MODEL variable-list ~ distribution / <options> ;
```

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>INITIAL=</td>
<td>specifies the initial values of the missing data, which are used to start the Markov chain.</td>
</tr>
<tr>
<td>MONITOR=</td>
<td>outputs analysis for selected missing data variables.</td>
</tr>
<tr>
<td>NAMESUFFIX=</td>
<td>specifies how to create the names of the missing data variables.</td>
</tr>
<tr>
<td>NOOUTPOST</td>
<td>suppresses the output of the posterior samples of the missing data variables.</td>
</tr>
</tbody>
</table>

246 / 295

247 / 295
Bivariate Normal with Partial Missing

<table>
<thead>
<tr>
<th>x1</th>
<th>1</th>
<th>1</th>
<th>-1</th>
<th>-1</th>
<th>2</th>
<th>2</th>
<th>-2</th>
<th>-2</th>
<th>.</th>
<th>.</th>
<th>.</th>
</tr>
</thead>
<tbody>
<tr>
<td>x2</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>2</td>
<td>2</td>
<td>-2</td>
</tr>
</tbody>
</table>

Table: Bivariate Normal Data with Partial Missing (Murray 1977)

The data are assumed to have zero means, and the parameter of interest is the covariance matrix $\Sigma$ and correlation $\rho$. The likelihood is

$$\pi(x | \mu, \Sigma) \propto \text{MVN}(\mu = 0, \Sigma)$$

The prior on $\Sigma$ is

$$\pi(\Sigma) \propto \text{iWishart}(\nu = 3, S = I)$$

A CC analysis uses first four observations, which produces a rather simple estimate of $\rho$ of 0. But this throws away all of the information that is in the other partially observed data.

Fitting MAR of the bivariate normal data with partial missing using PROC MCMC:

```plaintext
proc mcmc data=binorm nmc=20000 seed=17 outpost=pout
   monitor=(rho);
   array x[2] x1 x2; ! response variable
   array mu[2] (0 0);
   array sigma[2, 2];
   array S[2,2] (1 0, 0 1);
   parms sigma;
   prior sigma ~ iwish(3, S);
   rho = sigma[1,2]/sqrt(sigma[1,1]*sigma[2,2]);
   model x ~ mvn(mu, sigma) monitor=(x); ! outputs X_{mis}
run;
```

This is virtually identical to the program for fitting a bivariate normal data without missing values.
Procedure Outputs

### Missing Data Information Table

<table>
<thead>
<tr>
<th>Variable</th>
<th>Number of Missing Obs</th>
<th>Observation Indices</th>
<th>Sampling Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>x1</td>
<td>4</td>
<td>9 10 11 12</td>
<td>Direct</td>
</tr>
<tr>
<td>x2</td>
<td>4</td>
<td>5 6 7 8</td>
<td>Direct</td>
</tr>
</tbody>
</table>

---

### Posterior Estimates

#### Posterior Summaries

<table>
<thead>
<tr>
<th>Parameter</th>
<th>N</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
</tr>
</thead>
<tbody>
<tr>
<td>rho</td>
<td>20000</td>
<td>0.0270</td>
<td>0.6272</td>
<td>-0.6197</td>
<td>0.0772</td>
<td>0.6519</td>
</tr>
<tr>
<td>x2_5</td>
<td>20000</td>
<td>0.0473</td>
<td>1.8898</td>
<td>-1.3000</td>
<td>0.0657</td>
<td>1.3932</td>
</tr>
<tr>
<td>x2_6</td>
<td>20000</td>
<td>0.0700</td>
<td>1.8839</td>
<td>-1.2648</td>
<td>0.0891</td>
<td>1.4010</td>
</tr>
<tr>
<td>x2_7</td>
<td>20000</td>
<td>-0.0652</td>
<td>1.9042</td>
<td>-1.4005</td>
<td>-0.0659</td>
<td>1.2936</td>
</tr>
<tr>
<td>x2_8</td>
<td>20000</td>
<td>-0.0746</td>
<td>1.9147</td>
<td>-1.4345</td>
<td>-0.0825</td>
<td>1.2706</td>
</tr>
<tr>
<td>x1_9</td>
<td>20000</td>
<td>0.0575</td>
<td>1.8807</td>
<td>-1.2914</td>
<td>0.0834</td>
<td>1.3871</td>
</tr>
<tr>
<td>x1_10</td>
<td>20000</td>
<td>0.0606</td>
<td>1.8876</td>
<td>-1.2808</td>
<td>0.0785</td>
<td>1.3945</td>
</tr>
<tr>
<td>x1_11</td>
<td>20000</td>
<td>-0.0497</td>
<td>1.8942</td>
<td>-1.4079</td>
<td>-0.0661</td>
<td>1.2815</td>
</tr>
<tr>
<td>x1_12</td>
<td>20000</td>
<td>-0.0456</td>
<td>1.8839</td>
<td>-1.3938</td>
<td>-0.0762</td>
<td>1.2909</td>
</tr>
</tbody>
</table>
### Posterior Interval Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Alpha</th>
<th>Equal-Tail Interval</th>
<th>HPD Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>rho</td>
<td>0.050</td>
<td>-0.8845 0.8870</td>
<td>-0.8796 0.8895</td>
</tr>
<tr>
<td>x2_5</td>
<td>0.050</td>
<td>-3.5461 3.6009</td>
<td>-3.5456 3.6009</td>
</tr>
<tr>
<td>x2_6</td>
<td>0.050</td>
<td>-3.5444 3.6455</td>
<td>-3.5311 3.6503</td>
</tr>
<tr>
<td>x2_7</td>
<td>0.050</td>
<td>-3.7211 3.5754</td>
<td>-3.5998 3.6847</td>
</tr>
<tr>
<td>x2_8</td>
<td>0.050</td>
<td>-3.7148 3.5835</td>
<td>-3.6730 3.6130</td>
</tr>
<tr>
<td>x1_9</td>
<td>0.050</td>
<td>-3.5345 3.6513</td>
<td>-3.5372 3.6432</td>
</tr>
<tr>
<td>x1_10</td>
<td>0.050</td>
<td>-3.5735 3.5786</td>
<td>-3.4702 3.6568</td>
</tr>
<tr>
<td>x1_11</td>
<td>0.050</td>
<td>-3.6167 3.5969</td>
<td>-3.7271 3.4786</td>
</tr>
<tr>
<td>x1_12</td>
<td>0.050</td>
<td>-3.5976 3.5840</td>
<td>-3.5339 3.6202</td>
</tr>
</tbody>
</table>

### Estimate of the Correlation Parameter

![Histogram of rho](image-url)
Example Data Set

- The data are based on a double-blind antidepressant clinical trial originally reported by Goldstein et al (2004).
- The Drug Information Association (DIA) working group on missing data have made this data set available at www.missingdata.org.uk.
- To avoid implications for marketed drugs, all patients who took active medication are grouped into a single DRUG group and only a subset of the original trial patients are included.
- There are 171 subjects in the data set, 88 in the control arm, and 83 in the active arm.
Variables in the Data Set

- **patient**: patient ID
- **baseval**: baseline assessment on the Hamilton 17-item rating scale for depression (HAMD$_{17}$, Hamilton 1960).
- **change1–change4**: change in HAMD$_{17}$ at weeks 1, 2, 4, and 6.
- **r1–r4**: missing data indicator for each of the change variables.
- **therapy**: treatment (DRUG vs PLACEBO)
- **poolinv**: blocking information (Groups formed by pooling investigator).
- **last**: week index to last non-missing change value. Patient’s last visit week.
- **wkMax**: maximum number of weeks to be included in the analysis.

The first few observations of the selection data set:

```plaintext
data selection;
input PATIENT baseval change1-change4 r1-r4 THERAPY $ POOLINV $ last wkMax;
datalines;
1503 32 -11 -12 -13 -15 0 0 0 0 DRUG 006 4 4
1507 14 -3 0 -5 -9 0 0 0 0 PLACEBO 006 4 4
...
```

Average Mean Changes of HAMD$_{17}$ by Withdrawl Pattern

![Graph showing average mean changes of HAMD$_{17}$ by withdrawal pattern]
Data Characteristics

- Dropout probabilities appear to be correlated with the observed level of improvement (change in score).
- Patients failing to see improvement (flat or up-swinging lines), are more likely to withdraw.
- The probability of withdrawal could also depend on how they felt at the first unobserved visit - the MNAR part of the model.
- Fit a selection model:

\[
f(change|x, \theta) \cdot f(r|change, \phi)
\]

Outcome Model

For every subject \(i\), \(\text{change}_i = \{\text{change}_{j_i}\}\) is modeled using a \(\text{MVN}(\mu_i, \Sigma)\), where \(j = \{1, 2, 3, 4\}\) is the week index.

The mean variables, \(\mu_i = (\mu_{1i}, \mu_{2i}, \mu_{3i}, \mu_{4i})\), are modeled via:

\[
\mu_{ji} = m_{kj} + \beta_j \cdot (\text{baseval-18}) + \gamma_l
\]

where \(k = \{1, 2\}\) indexes the treatment, \(l\) indexes pooling investigator.

The following prior distributions are used in the analysis:

\[
\pi(m_{kj}, \beta_j, \gamma_l) \propto 1
\]

\[
\Sigma \sim \text{iWishart}(4, I)
\]
The Selection Model

The selection model (Diggle-Kenward model) includes the previous and current (possibly missing) response variables for each week:

\[ r_{kj} \sim \text{binary}(q_{kj}) \]

\[ q_{kj} = \text{logistic}(\phi_k 1 + \phi_k 2 \cdot \text{change}_{(j-1)} i + \phi_3 k \cdot \text{change}_{j_i}) \]

The parameters \( \phi_k \) account for treatment effect in separate regression models. Flat prior is used:

\[ \pi(\phi_k) \propto 1 \]

The MCMC Procedure
Missing Data Analysis

proc mcmc data=selection nmc=20000 seed=176 outpost=seleout;
array Change[4] Change1-Change4; ! response
array mu[4]; ! \mu_i
array Sigma[4,4]; ! \Sigma
array S[4,4] (1 0 0 0, 0 1 0 0, 0 0 1 0, 0 0 0 1);! S = I
array beta[4]; ! \beta_j
array M[2,4] m1-m8; ! m_{kj}
array phi[2,3] phi1-phi6; ! \phi_k.

parms beta: 0 ;
parms m1-m8 0;
parms phi1-phi6 0;
parms Sigma ;
prior beta: m1-m8 phi: ~ general(0); ! \pi(m_{kj}, \beta_j, \phi_k) \propto 1
prior Sigma ~ iwish(4, S); ! \pi(\Sigma) = \text{iWishart}(4, S)

/* outcome model */
random gamma ~ general(0) subject=poolinv zero=first init=0; ! \pi(\gamma) \propto 1
do j=1 to 4;
  if therapy eq "DRUG" then do;
    mu[j] = m[1,j] + gamma + beta[j]*(baseval-18); ! \mu_{k = DRUG}j
  end;
  else do;
    mu[j] = m[2,j] + gamma + beta[j]*(baseval-18); ! \mu_{k = PLACEBO}j
  end;
end;
model Change ~ mvn(mu, Sigma); ! likelihood
MCMC Code for the Selection Model

```sas
/* selection mechanism */
array r[4] r1-r4; ! missing data indicator
llike = 0;
do j = 3 to wkMax
   if therapy eq "DRUG" then do;
      mn = phi[1,1] + phi[1,2] * change[j-1] + phi[1,3] * change[j]; ! q_{k=DRUG}j
      q = logistic(mn); ! q_{k=DRUG}j
   end;
   else do;
      q = logistic(mn); ! q_{k=PLACEBO}j
   end;
   llike = llike + lpdfbern(r[i], q); ! accumulates binary likelihood over weeks
end;
model r2 r3 r4 ~ general(llike); ! declares joint likelihood
run;
```

Variable change1 doesn’t contain any missing values, making r1 irrelevant to the analysis.

Outcome Model Estimates

Comparison of posterior distributions of \( m_{\text{drug},j} \) and \( m_{\text{placebo},j} \) over the weeks:

- The treatment difference at week 1 is negligible.
- The difference becomes larger as the trial progresses, with the predicted score change for the DRUG group declining at a faster pace. The difference (mean difference is -2.42) is largest at the end of the trial.
Selection Model Estimates, When All are Estimated

Posterior distributions of $\phi_k$, which model the change in the probability of dropouts given the score changes in the last and the current, potentially missing, week:

- $\phi_{\text{drug},2}$ (phi2) and $\phi_{\text{placebo},2}$ (phi5) are positive, suggesting that as the patient felt worse (increase in HAMD17 score) in their previous visit, they were more likely to dropout.

- $\phi_{\text{drug},2}$ (phi3) $\phi_{\text{placebo},2}$ (phi6) are negative, suggesting that patients were less likely to withdraw from the trial had they felt worse in the current week.

Sensitivity Analysis Fixing MNAR Parameter Values

The parameters in this complete model are poorly estimated. An idea is to fix the regression on the potentially unobserved values (phi3 and phi6) and observe sensitivity to changing these. The estimated model (1st boxplot) produces similar point estimates (but larger s.d.) to the MAR model (2nd).

- when phi3 < phi6, boxplots shift to the left (3rd, 7th, and 8th). DRUG patients were more likely to drop out if they felt improvement in the current week. This results in stronger estimated treatment effect as the estimate is corrected for these missed patients.

- when phi3 > phi6, boxplots shift to the right (4th, 5th, and 6th), resulting in weaker treatment effect estimates.
Outline

3 The MCMC Procedure
   - A Primer on PROC MCMC
   - Monte Carlo Simulation
   - Single-level Model: Hyperparameters
   - Generalized Linear Models
   - Random-effects models
   - Missing Data Analysis
   - Survival Analysis (Optional)
     - Piecewise Exponential Model with Frailty

Piecewise Exponential Model

Let \( \{(t_i, x_i, \delta_i), i = 1, 2, \ldots, n\} \) be the observed data. Let \( a_0 = 0 < a_1 < \ldots < a_{J-1} < a_J = \infty \) be a partition of the time axis. The hazard for subject \( i \) is

\[
h(t|x_i; \theta) = h_0(t) \exp(\beta'x_i)
\]

where

\[
h_0(t) = \lambda_j \quad a_{j-1} \leq t < a_j \quad (j = 1, \ldots, J)
\]

The hazard for subject \( i \) in the \( j \)th time interval is

\[
h(t) = \lambda_j \exp(\beta'x_i) \quad a_{j-1} < t < a_j
\]
Piecewise Exponential Model

From the hazard function, first define the baseline cumulative hazard function:

\[ H_0(t) = \sum_{j=1}^{J} \lambda_j \Delta_j(t) \]

where

\[ \Delta_j(t) = \begin{cases} 0 & t < a_{j-1} \\ t - a_{j-1} & a_{j-1} \leq t < a_j \\ a_j - a_{j-1} & t \geq a_j \end{cases} \]

The log likelihood is:

\[
\begin{aligned}
    l(\lambda, \beta) &= \sum_{i=1}^{n} \delta_i \left[ \sum_{j=1}^{J} \left( a_{j-1} \leq t_i < a_j \right) \log \lambda_j + \beta' x_i \right] \\
                     &\quad - \sum_{i=1}^{n} \left[ \sum_{j=1}^{J} \Delta_j(t_i) \lambda_j \right] \exp(\beta' x_i)
\end{aligned}
\]

where \( \delta_i \) is the event status:

\[
\delta_i = \begin{cases} 0 & \text{if } t_i \text{ is a censored time} \\
1 & \text{if } t_i \text{ is an event time} \end{cases}
\]

This model has two parameter vectors: \( \lambda \) and \( \beta \).
Road map (what needs to be done)

- Reformulate the likelihood to a Poisson likelihood, which enables us to treat hazards as random effects
- Manipulate the data
- Fit using PROC MCMC
- Extend to frailty model

Recall the hazard function

\[ h(t|x_i; \theta) = h_0(t) \exp(\beta'x_i) \]

Define \( N_i(t) \) to be the number of observed failures of the \( i \)th subject up to time \( t \), then the hazard function is a special case of a multiplicative intensity model (Clayton, 1991, Biometrics, 467-485). And the intensity process for \( N_i(t) \) becomes

\[ l_i(t) = Y_i(t)h_0(t) \exp(\beta'x_i) \]

where

\[ Y_i(t) = \begin{cases} 
1 & \text{if the subject is observed at time } t \\
0 & \text{o.w.}
\end{cases} \]
Under noninformative censoring, the corresponding likelihood is proportional to

\[
\prod_{i=1}^{n} \left[ \prod_{t \geq 0} l_i(t) \right] dN_i(t) \exp \left[ -\int_{t \geq 0} l_i(t) \, dt \right]
\]

where \( dN_i(t) \) is the increment of \( N_i(t) \) over the small time interval \( [t, t + dt) \):

\[
dN_i(t) = \begin{cases} 
1 & \text{if the subject } i \text{ fails in the time interval} \\
0 & \text{o.w.}
\end{cases}
\]

### Poisson Process as the Likelihood Function

This is a Poisson kernel with the random variable being the increments of \( dN_i \) and the means \( l_i(t) \, dt \)

\[
dN_i(t) \sim \text{Poisson}(l_i(t) \, dt)
\]

where

\[
l_i(t) \, dt = Y_i(t) \exp(\beta'x) h_0(t)
\]

and

\[
h_0(t) = \int_{0}^{t} h_0(u) \, du.
\]

The integral is the increment in the integrated baseline hazard function that occurs during the time interval \( [t, t + dt) \).
Alternative Approach

The alternative formulation of the piecewise exponential model

\[ dN_i(t) \sim \text{Poisson}(Y_i(t) \exp(\beta' x) h_0(t)) \]

makes it a random-effects model, with each hazard rate, \( h_0(t) \) being a random effect.

You need to manipulate the data and create \( Y_i(t) \) and \( dN_i(t) \) for each interval.

Piecewise Exponential Model

Consider a randomized trial of 40 rats exposed to carcinogen:

- Drug X and Placebo are the treatment groups.
- Event of interest is death.
- Response is time until death.
- What are the effects of treatment and gender on survival?
Piecewise Exponential Model

A subset of the data:

```plaintext
proc format;
   value Rx 1='X' 0='Placebo';
data Exposed;
   input Days Status Trt Gender $ @@;
   format Trt Rx.;
datalines;
  179 1 1 F 378 0 1 M
  256 1 1 F 355 1 1 M
  262 1 1 M 319 1 1 M
  256 1 1 F 256 1 1 M
  ...
  268 0 0 M 209 1 0 F
;```

The following regression model and prior distributions are used in the analysis:

\[
\beta'x_i = \beta_1 \text{Trt} + \beta_2 \text{Gender} \\
\beta_1, \beta_2 \sim \text{normal}(0, \text{var} = 1e6) \\
h_0(t) \sim \text{gamma(}\text{shape} = 0.5, \text{iscale} = 0.5) \\
\]

where \( t = 1 \cdots 8 \) (time intervals).
A Little Problem

Both Trt and Gender are character variables and PROC MCMC does not support a CLASS statement.

PROC TRANSREG to the rescue:

```plaintext
proc transreg data=exposed design;
  model class(trt gender / zero=first);
  id days status id;
  output out=exposed_d(drop=_: Int:);
run;
```

design : specifies design matrix coding  
class : expands the variables to “dummy” variables  
zero : controls reference level (first sets 0 to the first sorted category)

zero="X" "F"

ID : includes additional variables to the OUT= data set  
OUTPUT : creates a new data set

The New Data Set

<table>
<thead>
<tr>
<th>Obs</th>
<th>Days</th>
<th>Status</th>
<th>Trt</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>179</td>
<td>1</td>
<td>X</td>
<td>F</td>
</tr>
<tr>
<td>2</td>
<td>378</td>
<td>0</td>
<td>X</td>
<td>M</td>
</tr>
<tr>
<td>3</td>
<td>256</td>
<td>1</td>
<td>X</td>
<td>F</td>
</tr>
<tr>
<td>4</td>
<td>355</td>
<td>1</td>
<td>X</td>
<td>M</td>
</tr>
<tr>
<td>5</td>
<td>262</td>
<td>1</td>
<td>X</td>
<td>M</td>
</tr>
<tr>
<td>6</td>
<td>319</td>
<td>1</td>
<td>X</td>
<td>M</td>
</tr>
<tr>
<td>7</td>
<td>256</td>
<td>1</td>
<td>X</td>
<td>F</td>
</tr>
<tr>
<td>8</td>
<td>256</td>
<td>1</td>
<td>X</td>
<td>M</td>
</tr>
<tr>
<td>9</td>
<td>255</td>
<td>1</td>
<td>X</td>
<td>M</td>
</tr>
<tr>
<td>10</td>
<td>171</td>
<td>1</td>
<td>X</td>
<td>F</td>
</tr>
</tbody>
</table>

⇒

<table>
<thead>
<tr>
<th>Obs</th>
<th>TrtX</th>
<th>M</th>
<th>Days</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>179</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>378</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>256</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>355</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>262</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>319</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>0</td>
<td>256</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>1</td>
<td>256</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>1</td>
<td>255</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0</td>
<td>171</td>
<td>1</td>
</tr>
</tbody>
</table>

Automatically created macro variable &_trgind contains the list of independent variables created:

```
%put &_trgind;
TrtX GenderM
```
Partition of the Time Axis

You can find a partition along the time axis using PROC UNIVARIATE, placing roughly the same number of event times in each interval:

```sas
%let npar_e = 7;
%let inc = %sysevalf((1/&npar_e) * 100);
proc univariate data=exposed_d(where=(status=1)) pctldef=4;
    var days;
    output out=interval pctlpre=P_ pctlpts= 0 to 100 by &inc;
run;
```

Alternatively, we use the same partition as PROC PHREG:

```sas
data partition;
    input int_1-int_9;
datalines;
0 193 221 239.5 255.5 256.5 278.5 321 1000
280 / 295
```

![Graph showing time axis partition]
Manipulate the Data

The next step is to create $Y_i(t)$ and $dN_i(t)$ using Days and Status:

$$Y_i(t) = \begin{cases} 
1 & \text{if the subject is observed at time } t \\
0 & \text{o.w.}
\end{cases}$$

$$dN_i(t) = \begin{cases} 
1 & \text{if the subject } i \text{ fails in the time interval} \\
0 & \text{o.w.}
\end{cases}$$

Create $Y_i(t)$ and $dN_i(t)$

<table>
<thead>
<tr>
<th>Days</th>
<th>Status</th>
<th>$Y_i(t)$</th>
<th>$dN_i(t)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>355</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>256</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>378</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>179</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Modify the Data

The following statements calculate $Y_i(t)$ for each observation $i$, at every time point $t$ in the Partition data set. The statements also find the observed failure time interval, $dN_i(t)$, for each observation:

```sas
%let n = 8;
data _a;
  set exposed_d;
  if _n_ eq 1 then set partition;
  array int[*] int_:;
  array Y[&n];
  array dN[&n];
  do k = 1 to &n;
    Y[k] = (days - int[k] + 0.001 >= 0);
    dN[k] = Y[k] * ( int[k+1] - days - 0.001 >= 0) * status;
  end;
  output;
  drop int_: k;
run;
```

First few observations of the new data set:

```
       S   t  D  T  a  T
      0   i  t  y  T  d  d  d  d  d  d  d  d  b  I  m  u  p  r  i  Y  Y  Y  Y  Y  Y  Y  N  N  N  N  N  N  N  N  N
     1  5  46.23 0 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0
     2  5  46.23 0 1 0 2 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0
     3 14  42.50 0 0 1 3 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0
     4 14  31.30 1 0 0 4 1 1 1 1 1 1 0 0 0 0 0 0 1 0
     5 16  42.27 0 0 1 5 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0
     6 16  42.27 0 0 0 6 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0
     ...
```

This being the Partition data set:

```
    0  193  221  239.5  255.5  256.5  278.5  321  1000
```
Input Data Set

- Each observation in the _a data set has 8 Y and 8 dN, meaning that you would need eight MODEL statements in a PROC MCMC call, each for a Poisson likelihood.

  ```
  model dN1 ~ poisson(Y1 * exp(beta * x) * h1);
  model dN2 ~ poisson(Y2 * exp(beta * x) * h2);
  ...
  model dN8 ~ poisson(Y8 * exp(beta * x) * h8);
  ```

- Alternatively, you can expand _a, put one Y and one dN in every observation, and fit the data using a single MODEL statement in PROC MCMC. This enables you to treat the hazards \(h_0(t)\) as random-effects and use the RANDOM statement.

The following statements expand the data set _a and save the results in the data set _b:

```
data _b;
  set _a;
  array y[*] y:;
  array dn[*] dn:;
  do i = 1 to (dim(y));
    y_val = y[i];
    dn_val = dn[i];
    int_index = i;
    output;
  end;
  keep y_: dn_: &_trgind int_index id;
run;

data _b;
  set _b;
  rename y_val=Y dn_val=dN;
run;
```
The data set _b now contains 320 observations. The int_index variable is an index variable that indicates interval membership of each observation.

<table>
<thead>
<tr>
<th>Obs</th>
<th>TrtX</th>
<th>M</th>
<th>id</th>
<th>Y</th>
<th>dN</th>
<th>index</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Further Clean Up the Data

Recall the likelihood is:

\[ dN_i(t) \sim \text{Poisson}(Y_i(t) \exp(\beta'x)h_0(t)) \]

where \( Y_i(t) \) does not contribute to the likelihood calculation when it takes a value of 0, you can remove these observations.

```
data inputdata;
  set _b;
  if Y > 0;
run;
```

This steps reduces the size of the input data set (to 174 observations) and shortens the run time of PROC MCMC.
Fitting Piecewise Exponential Model Using PROC MCMC

The following statements fit a piecewise exponential model in PROC MCMC:

```
proc mcmc data=inputdata nmc=10000 outpost=postout seed=12351
   stats=summary diag=ess;
   parms beta1-beta2 0;
   prior beta: ~ normal(0, var = 1e6);
   random h0 ~ gamma(0.5, iscale = 0.5) subject=int_index;
   bZ = beta1*trtx + beta2*genderM;
   idt = exp(bZ) * h0;
   model dN ~ poisson(idt);
run;
```

Note that the $Y_i(t)$ term is omitted in the assignment statement for the symbol idt because $Y = 1$ in all observations.

### Posterior Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>N</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta1</td>
<td>10000</td>
<td>-0.5659</td>
<td>0.3288</td>
<td>-0.7869</td>
<td>-0.5666</td>
<td>-0.3514</td>
</tr>
<tr>
<td>beta2</td>
<td>10000</td>
<td>-1.5919</td>
<td>0.3520</td>
<td>-1.8310</td>
<td>-1.5916</td>
<td>-1.3586</td>
</tr>
</tbody>
</table>
Hazards Estimates

Frailty Model

Now suppose you want to include patient-level information and fit a frailty model to the exposed data set.

\[
\beta' x_i = \beta_1 \text{Trt} + \beta_2 \text{Gender} + u_{id}
\]

\[
u_{id} \sim \text{normal}(0, \text{var} = \sigma^2)
\]

\[
\sigma^2 \sim \text{igamma}(\text{shape} = 0.01, \text{scale} = 0.01)
\]

\[
\beta_1, \beta_2 \sim \text{normal}(0, \text{var} = 1e6)
\]

\[
h_0(t) \sim \text{gamma}(\text{shape} = 0.5, \text{iscale} = 0.5)
\]

\[
dN_i(t) \sim \text{Poisson}(Y_i(t) \exp(\beta' x) h_0(t))
\]

where \( t = 1 \cdots 8 \) and \( \text{id} \) indexes patient.
The actual coding in PROC MCMC of a piecewise exponential frailty model is rather straightforward:

```sas
proc mcmc data=inputdata nmc=10000 outpost=postout seed=12351
   stats=summary diag=none;
   parms beta1-beta2 0 s2;
   prior beta: ~ normal(0, var = 1e6);
   prior s2 ~ igamma(0.01, scale=0.01);
   random h0 ~ gamma(0.01, iscale = 0.01) subject=int_index;
   random u ~ normal(0, var=s2) subject=id;
   bZ = beta1*trtx + beta2*genderM + u;
   idt = exp(bZ) * lambda;
   model dN ~ poisson(idt);
run;
```

And you are done!

---

**Learning Objectives**

Attendees will

- understand basic concepts and computational methods of Bayesian statistics
- be able to deal with some practical issues that arise from Bayesian analysis
- be able to program using SAS/STAT procedures with Bayesian capabilities to implement various Bayesian models.