Introduction to Bayesian MCMC Models

Glenn Meyers

Presentation to Casualty Loss Reserve Seminar
Austin Texas

September 18, 2019

Outline of Workshop

1. Theory behind Bayesian Markov Chain Monte Carlo (MCMC) models
2. An Example with the Metropolis Hastings Algorithm
3. The CRoss Classified (CRC) Stochastic Loss Reserve Model
   - The "rstan" R package
   - Model convergence statistics
   - Graphical model diagnostics
   - Changing the prior distribution
   - Model comparison with the "loo" R package
4. The Changing Settlement Rate (CSR) Model

Attendee Assumptions

- Completely new to Bayesian MCMC
- Familiarity with R
- Familiarity with RStudio - or equivalent [1]
- Prior to the session, attendees should install the packages, "rstan", "loo", "data.table" and "ChainLadder."
    - Installing "rstan" is a little more involved than installing other packages on CRAN. I suggest going to https://mc-stan.org and follow the instructions there to install "rstan."
    - The scripts that call Stan will read the file "Intro_comauto_pos.csv." You should code the "setwd" command accordingly.
- One can test the installation by running the "Intro_CRC.R" script included in the Course Materials directory.

[1] The in-session examples will use RStudio, but one with their own favorite R editor should be able to run the examples.

Bayesians vs. Frequentists

- Given the model $X \sim f(X|\theta)$
- Given the set of observations x.
  - Frequentists test the hypothesis $\theta = \theta_0$.
  - Bayesians calculate the posterior distribution $f(\theta|x)$.

$$f(\theta|x) = \frac{f(x|\theta) \cdot \pi(\theta)}{\int\limits_{\vartheta} f(x|\vartheta) \cdot \pi(\vartheta) \cdot d\vartheta}$$

- The issue — What is the prior distribution, $\pi(\theta)$?

The Philosophical Issue

- Bayesians select π "subjectively" according to prior opinion.
- Frequentists respond by saying that conclusions should be dictated solely by looking at "the data."
- Some Bayesians respond with "noninformative" priors.
  - Is there such a thing?

The Practical Issue — Can we do the calculations?

- For most of the 20th century, the frequentists were winning.
  - Calculations were easy with quadradic forms needed for the normal distributions.
  - The General Linear Model (PROC GLM in SAS).
  - As computers and numerical analysis progressed we got the **Generalized** Linear Model (PROC GENMOD in SAS).
- Now the Bayesians are winning - with MCMC.

The Problem with Bayesian Analysis

- Let θ be an n-parameter vector — e.g. development factors.
- Let X be a set of observations — e.g. a loss triangle.

$$f(\theta|x) = \frac{f(x|\theta)\cdot\pi(\theta)}{\int\limits_{\vartheta_1}\cdots\int\limits_{\vartheta_n} f(x|\vartheta)\cdot\pi(\vartheta)\cdot d\vartheta}$$

- $f(X|\theta)$ is the likelihood of X given θ.
- $\pi(\theta)$ is the prior distribution of θ.
- $f(\theta|X)$ is the posterior distribution of θ.

The Problem with Bayesian Analysis

- Let θ be an *n*-parameter vector — e.g. development factors.
- Let X be a set of observations — e.g. a loss triangle.

$$f(\theta|x) = \frac{f(x|\theta)\cdot\pi(\theta)}{\int\limits_{\vartheta_1}\cdots\int\limits_{\vartheta_n} f(x|\vartheta)\cdot\pi(\vartheta)\cdot d\vartheta}$$

- $f(X|\theta)$ is the likelihood of X given θ.
- $\pi(\theta)$ is the prior distribution of θ.
- $f(\theta|X)$ is the posterior distribution of θ.
- Calculating the *n*-dimensional integral is intractable.

A New World Order

- This impasse came to an end in 1990 when a simulation-based approach to estimating posterior probabilities was introduced.
- *Sampling Based Approach to Calculating Marginal Densities*
  - Alan E. Gelfand and Adrian F.M. Smith
  - Journal of the American Statistical Association, June 1990

Markov Chains

- Let Ω be a finite state with random events

$$X_1, X_2, \ldots, X_t, \ldots$$

- A Markov chain P satisfies

$$Pr(X_t = y | X_{t-1} = x_{t-1}, \ldots, X_1 = x_1) = Pr(X_t = y | X_{t-1} = x_{t-1})$$

- The probability of an event in the chain depends only on the immediate previous event.

The Markov Convergence Theorem

- There is a branch of probability theory, called Ergodic Theory, that gives conditions for which there exists a unique stationary distribution, π, such that

$$Pr(y|X_{t-1}) \longrightarrow \pi(y)$$

$$\text{as } t \longrightarrow \infty$$

The Metropolis Hastings Algorithm
A Very Important Markov Chain.

1. Time $t = 1$: select a random initial position θ_1 in parameter space.
2. Select a proposal distribution $p(\theta|\theta_{t-1})$ that we will use to select proposed random steps away from our current position in parameter space.
3. Starting at time $t = 2$: repeat the following until you get convergence:
   - At step t, generate a proposal $\theta^* \sim p(\theta|\theta_{t-1})$.
   - Generate $U \sim$ uniform(0,1)
   - Calculate

$$R = \frac{f(\theta^*|x)}{f(\theta_{t-1}|x)} \cdot \frac{p(\theta_{t-1}|\theta^*)}{p(\theta^*|\theta_{t-1})}$$

   - If $U < R$ then $\theta_t = \theta^*$. Else, $\theta_t = \theta_{t-1}$.

Dodging the Intractable Integral

$$R = \frac{f(\theta^*|x)}{f(\theta_{t-1}|x)} \cdot \frac{p(\theta_{t-1}|\theta^*)}{p(\theta^*|\theta_{t-1})}$$

$$R = \frac{\dfrac{f(x|\theta^*)\cdot\pi(\theta^*)}{\int\limits_{\vartheta_1}\cdots\int\limits_{\vartheta_n} f(x|\vartheta)\cdot\pi(\vartheta)\cdot d\vartheta}}{\dfrac{f(x|\theta_{t-1})\cdot\pi(\theta_{t-1})}{\int\limits_{\vartheta_1}\cdots\int\limits_{\vartheta_n} f(x|\vartheta)\cdot\pi(\vartheta)\cdot d\vartheta}} \cdot \frac{p(\theta_{t-1}|\theta^*)}{p(\theta^*|\theta_{t-1})}$$

The integral $\int\limits_{\vartheta_1}\cdots\int\limits_{\vartheta_n} f(x|\vartheta)\cdot\pi(\vartheta)\cdot d\vartheta$ cancels out!

The Metropolis Hastings Algorithm Restated

1. Time $t = 1$: select a random initial position θ_1 in parameter space.
2. Select a proposal distribution $p(\theta|\theta_{t-1})$ that we will use to select proposed random steps away from our current position in parameter space.
3. Starting at time $t = 2$: repeat the following until you get convergence:
   - At step t, generate a proposal $\theta^* \sim p(\theta|\theta_{t-1})$.
   - Generate $U \sim$ uniform(0,1)
   - Calculate

     $$R = \frac{f(x|\theta^*) \cdot \pi(\theta^*)}{f(x|\theta_{t-1}) \cdot \pi(\theta_{t-1})} \cdot \frac{p(\theta_{t-1}|\theta^*)}{p(\theta^*|\theta_{t-1})}$$

   - If $U < R$ then $\theta_t = \theta^*$. Else, $\theta_t = \theta_{t-1}$.

The Relevance of the Metropolis Hastings Algorithm

- Defined in terms of the conditional distribution

$$f(X|\theta)$$

and the prior distribution

$$\pi(\theta)$$

- The limiting distribution is the **posterior distribution!**

The Relevance of the Metropolis Hastings Algorithm

- Defined in terms of the conditional distribution

$$f(X|\theta)$$

and the prior distribution

$$\pi(\theta)$$

- The limiting distribution is the **posterior distribution!**
- Code $f(X|\theta)$ and $\pi(\theta)$ into a Markov chain and let it run for a while, and you have a large sample from the posterior distribution.

The Relevance of the Metropolis Hastings Algorithm

- The theoretical limiting distribution is the same, no matter what proposal distribution, $p(\theta|\theta_{t-1})$, is used.

The Relevance of the Metropolis Hastings Algorithm

- The theoretical limiting distribution is the same, no matter what proposal distribution, $p(\theta|\theta_{t-1})$, is used.
- But as we shall see, a good choice of the proposal distribution will speed up convergence.

The Relevance of the Metropolis Hastings Algorithm

- The theoretical limiting distribution is the same, no matter what proposal distribution, $p(\theta|\theta_{t-1})$, is used.
- But as we shall see, a good choice of the proposal distribution will speed up convergence.
- There is no fundamental limit on the number of parameters in your model!
- The practical limit is within range of stochastic loss reserve models.

A Short History of MCMC

- Originated with the study of nuclear fission.
  - Enrico Fermi, John von Neumann, Nicolas Metropolis and Stanislaw Ulam.
  - Developed the Metropolis algorithm.
- Keith Hastings (1970) recognized the potential of the Metropolis algorithm to solve statistical problems.
- Simulations were not readily accepted by the statistical community at that time.

A Short History of MCMC

- Gelfand and Smith (1990) pulled together the relevant ideas at a time when simulation was deemed OK.
  - Seized upon by scientists in other fields.
  - Used the Gibbs sampler (A one parameter at a time special case of Metropolis Hastings algorithm).
- Statisticians in the UK started the BUGS project to produce software for MCMC.
  - **B**ayesian inference **U**sing the **G**ibbs **S**ampler

Evolution of MCMC Software

- WinBUGS (Original — now discontinued)
- OpenBUGS (Continuation of WinBUGS)
  - Designed mainly for the Windows operating system.
- JAGS — **J**ust **A**nother **G**ibbs **S**ampler
  - Originated by Martyn Plummer.
  - Runs on multiple operating systems.
  - Callable from R ("runjags" package.)
- Stan (in honor of Stanislaw Ulam)
  - Stan team led by Andrew Gelman at Columbia University.
  - Runs on multiple operation systems.
  - Callable from R ("rstan" package) and other languages, e.g. Python and Matlab.

Lognormal Example using Metropolis Hastings

- Given the data below, estimate the cost of a 15,000 xs of 10,000 layer.
- Find predictive distribution of losses in that layer.
- Fit a lognormal distribution with
  - log(mean) $= \mu$
  - Prior distribution — $\mu \sim \text{normal}(8, 1)$
  - log(standard deviation) $= 1$

484	603	631	1189	1229
1407	1565	1894	2140	2244
2262	2654	2672	4019	4318
5015	5354	5464	5598	6060
6500	6747	9143	12782	18349

Introductory Example Coded in R

- Open the file "MH Intro with Lognormal.R" in RStudio
- Exploratory runs
  - "Tune" the proposal distribution"
  - Test convergence by running two chains and comparing results
  - "Thinning" the chains can guarantee convergence.

Introductory Example Coded in R

- Open the file “MH Intro with Lognormal.R” in RStudio
- Exploratory runs
  - “Tune” the proposal distribution”
  - Test convergence by running two chains and comparing results
  - “Thinning” the chains can guarantee convergence.
- Translate the posterior distribution of parameters into statistics of interest to actuaries.

Takeaways from Introductory Example

- We need to run multiple chains to test convergence.
- Adaptation — we need to scale the proposal distribution to get a representative sample in as few iterations as possible.
- Thinning — When adaption does not work well, take every *n*th iteration.

Takeaways from Introductory Example

- We need to run multiple chains to test convergence.
- Adaptation — we need to scale the proposal distribution to get a representative sample in as few iterations as possible.
- Thinning — When adaption does not work well, take every *n*th iteration.
- Visual inspection works fairly well with one-parameter models. Our next model will have 29 parameters. What do we do then?

Takeaways from Introductory Example

- We need to run multiple chains to test convergence.
- Adaptation — we need to scale the proposal distribution to get a representative sample in as few iterations as possible.
- Thinning — When adaption does not work well, take every *n*th iteration.
- Visual inspection works fairly well with one-parameter models. Our next model will have 29 parameters. What do we do then?
- MCMC software — e.g. Stan.

Loss Triangle for Example — From Schedule P

Illustrative Insurer Net Written Premium

AY	1	2	3	4	5	6	7	8	9	10
Premium	5812	4908	5454	5165	5214	5230	4992	5466	5226	4962

Illustrative Insurer Paid Losses Net of Reinsurance

AY \ Lag	1	2	3	4	5	6	7	8	9	10
1988	952	1529	2813	3647	3724	3832	3899	3907	3911	3912
1989	849	1564	2202	2432	2468	2487	2513	2526	2531	
1990	983	2211	2830	3832	4039	4065	4102	4155		
1991	1657	2685	3169	3600	3900	4320	4332			
1992	932	1940	2626	3332	3368	3491				
1993	1162	2402	2799	2996	3034					
1994	1478	2980	3945	4714						
1995	1240	2080	2607							
1996	1326	2412								
1997	1413									

The CRoss Classified (CRC) Model

For Accident Year w, Development Year d and cumulative loss C_{wd}:

1. $logelr \sim \text{normal}(\text{-}0.4, \sqrt{10})$.
2. $\alpha_w \sim \text{normal}(0, \sqrt{10})$ for $w = 2, \ldots, 10$. Set $\alpha_1 = 0$.
3. $\beta_d \sim \text{normal}(0, \sqrt{10})$ for $d = 1, \ldots, 9$. Set $\beta_{10} = 0$.
4. $a_i \sim \text{uniform}(0, 1)$ for $i = 1, \ldots, 10$.
5. Set $\sigma_d^2 = \sum_{i=d}^{10} a_i$ for $d = 1, \ldots, 10$.

Note that this forces $\sigma_1^2 > \ldots > \sigma_{10}^2$.

6. Set $\mu_{wd} = \log(\text{Premium}_w) + logelr + \alpha_w + \beta_d$.
7. Then $C_{wd} \sim \text{lognormal}(\mu_{wd}, \sigma_d)$.

Actuarially Interesting Output

- Open "Intro_CRC.R" and run the script.
- The MCMC output, which is of secondary interest, is a sample of the posterior distribution parameters[2].

$$\{logelr\}, \{\alpha_w\}, \{\beta_d\} \text{ and } \{\sigma_d\}$$

[2] The brackets $\{\cdot\}$ denote a sample of size 10,000 from the posterior

Actuarially Interesting Output

- Open "Intro_CRC.R" and run the script.
- The MCMC output, which is of secondary interest, is a sample of the posterior distribution parameters[2].

$$\{logelr\}, \{\alpha_w\}, \{\beta_d\} \text{ and } \{\sigma_d\}$$

- Of greater interest is a sample from the predictive distribution of ultimate losses, $\{U_w\}$ and $\{U_{Tot}\} = \sum_{w=1}^{10}\{U_w\}$ where:

$$U_w = \exp(\mu_{w,10} + \sigma_{10}^2/2)$$

[2] The brackets $\{\cdot\}$ denote a sample of size 10,000 from the posterior

Actuarially Interesting Output

- Open "Intro_CRC.R" and run the script.
- The MCMC output, which is of secondary interest, is a sample of the posterior distribution parameters[2].

$$\{logelr\}, \{\alpha_w\}, \{\beta_d\} \text{ and } \{\sigma_d\}$$

- Of greater interest is a sample from the predictive distribution of ultimate losses, $\{U_w\}$ and $\{U_{Tot}\} = \sum_{w=1}^{10}\{U_w\}$ where:

$$U_w = \exp(\mu_{w,10} + \sigma_{10}^2/2)$$

- Accident year exhibit
- Predictive Distribution of the Loss Reserve

[2] The brackets $\{\cdot\}$ denote a sample of size 10,000 from the posterior

Actuarially Interesting Output

- Open "Intro_CRC.R" and run the script.
- The MCMC output, which is of secondary interest, is a sample of the posterior distribution parameters[2].

$$\{logelr\}, \{\alpha_w\}, \{\beta_d\} \text{ and } \{\sigma_d\}$$

- Of greater interest is a sample from the predictive distribution of ultimate losses, $\{U_w\}$ and $\{U_{Tot}\} = \sum_{w=1}^{10}\{U_w\}$ where:

$$U_w = \exp(\mu_{w,10} + \sigma_{10}^2/2)$$

- Accident year exhibit
- Predictive Distribution of the Loss Reserve
- Of greatest interest — decisions (e.g. risk margin)

[2] The brackets $\{\cdot\}$ denote a sample of size 10,000 from the posterior

Taking a Sample From the Posterior Distribution

- Stan — A separate software package called into R with the "rstan" package.
- The Stan script in a long character string in R.
  - Required block - "data"

Taking a Sample From the Posterior Distribution

- Stan — A separate software package called into R with the "rstan" package.
- The Stan script in a long character string in R.
  - Required block - "data"
  - Required block - "parameters"

Taking a Sample From the Posterior Distribution

- Stan — A separate software package called into R with the "rstan" package.
- The Stan script in a long character string in R.
  - Required block - "data"
  - Required block - "parameters"
  - Optional block - "transformed parameters"

Taking a Sample From the Posterior Distribution

- Stan — A separate software package called into R with the "rstan" package.
- The Stan script in a long character string in R.
  - Required block - "data"
  - Required block - "parameters"
  - Optional block - "transformed parameters"
  - Required block - "model"

Taking a Sample From the Posterior Distribution

- Stan — A separate software package called into R with the “rstan” package.
- The Stan script in a long character string in R.
  - Required block - “data”
  - Required block - “parameters”
  - Optional block - “transformed parameters”
  - Required block - “model”
  - Optional block - “generated quantities”

Taking a Sample From the Posterior Distribution

- Stan — A separate software package called into R with the "rstan" package.
- The Stan script in a long character string in R.
  - Required block - "data"
  - Required block - "parameters"
  - Optional block - "transformed parameters"
  - Required block - "model"
  - Optional block - "generated quantities"
  - Need to specify output that goes back to R

Taking a Sample From the Posterior Distribution

- Stan — A separate software package called into R with the "rstan" package.
- The Stan script in a long character string in R.
  - Required block - "data"
  - Required block - "parameters"
  - Optional block - "transformed parameters"
  - Required block - "model"
  - Optional block - "generated quantities"
  - Need to specify output that goes back to R
- The "Intro_CRC.R" implementation runs 4 chains in parallel.

Remaining Questions

1. Testing convergence of the model
   - The "traceplot()" command
   - The R-Hat statistic
2. Diagnostic Plots
   - Standardized Residual Boxplots
3. Comparing different models for the same data
   - The $\widehat{elpd}_{loo}$ statistic.

Testing for Convergence of the Markov Chain

- Trace plots of parameters show rapid convergence.
- Gelman-Rubin Convergence Diagnostic
  - Run at least 4 chains
  - Let $\widehat{W}$ = Within Chain Variance
  - Let $\widehat{B}$ = Between Chain Variance
  - Define "Potential Scale Reduction Factor" (PSRF) or R-Hat by:

$$\sqrt{\widehat{R}} = \sqrt{\frac{\widehat{W} + \widehat{B}}{\widehat{W}}} \longrightarrow 1$$

  - Gelman and Rubin suggest that we should accept convergence if R-Hat < 1.1
- All R-Hat statistics are less than 1.01 for "Intro_CRC.R"

Standardized Residual Boxplots

- The models in this paper all assume a lognormal distribution with the parameters μ_{wd} and σ_d. Thus we expect that

$$\frac{log(C_{wd}) - \{\mu_{wd}\}}{\{\sigma_d\}}$$

will have a normal(0,1) distribution.
- To test this graphically we split the residuals, in turn by accident year, development year and calendar year and plot a sample of size 200 in each "year" with the R "boxplot" function.

Expected Results with the R "boxplot" Function

- The gray bars correspond to the interquartile range. Ideally the bars should be centered on 0. The endpoints of those bars should be touching the black lines representing the interquartile range of the standard normal distribution.
- Most of the remaining residuals should be between $\pm$ 2. A few could be in the (-3,-2) or the (2,3) ranges. Very few should be outside the $\pm$ 3 range.
- Now look at the Boxplot in the current MCMC output.

Model Selection With the “loo” Package”

- Given two different models for the same data, how do you select the “better” model?
- “loo” stands for **L**eave **O**ne **O**ut.
- Maintained by members of the stan development team.
- Vehtari, A., Gelman, A., and Gabry, J. (2015). “Efficient implementation of leave-one-out cross validation and WAIC for evaluating fitted Bayesian models.”
- See the documentation of the “loo” package for the latest version of the paper.

Selecting Models Fit By Maximum Likelihood

- If we fit a model, $f(x|\theta)$, by maximum likelihood, define

$$AIC = 2 \cdot p - 2 \cdot L(x|\hat{\theta})$$

- Where:
  - p is the number of parameters in the model.
  - $L(x|\hat{\theta})$ is the maximum log-likelihood of the model specified by f.
- Lower AIC indicates a better fit.
  - Encourages a larger log-likelihood.
  - Penalizes an increase in the number of parameters.

Selecting Bayesian MCMC Models with the LOOIC Statistic

- Given an MCMC model with parameters $\{\theta_i\}_{i=1}^{10,000}$, define

$$LOOIC = 2 \cdot \hat{p}_{LOOIC} - 2 \cdot \overline{\{L(x|\theta_i)\}_{i=1}^{10,000}}$$

- Where
  - $\hat{p}_{LOOIC}$ is the ***effective*** number of parameters.

$$\hat{p}_{LOOIC} = \overline{\{L(x|\theta_i)\}_{i=1}^{10,000}} - \sum_{n=1}^{N} \overline{\{L(x_n|x_{(-n)},\theta_i)\}_{i=1}^{10,000}}$$

  - $x_{(-n)} = x_1, \ldots, x_{n-1}, x_{n+1}, \ldots, x_N$
  - $L(x_n|x_{(-n)},\theta_i)$ is the log-likelihood of x_n from a model fit using all data except x_n.

The $\widehat{elpd}_{loo}$ Statistic

- After some algebra we can see that

$$LOOIC = -2 \cdot \sum_{n=1}^{N} \overline{\{L(x_n|x_{(-n)}, \theta_i)\}}_{i=1}^{10{,}000}$$

- which we like as it favors the model with the largest likelihood, and the smallest LOOIC, on the "holdout" data.

The $\widehat{elpd}_{loo}$ Statistic

- After some algebra we can see that

$$LOOIC = -2 \cdot \sum_{n=1}^{N} \overline{\{L(x_n|x_{(-n)},\theta_i)\}}_{i=1}^{10,000}$$

- which we like as it favors the model with the largest likelihood, and the smallest LOOIC, on the "holdout" data.
- Some "loo" package features.
  - $\sum_{n=1}^{N} \overline{\{L(x_n|x_{(-n)},\theta_i)\}}_{i=1}^{10,000} \equiv \widehat{elpd}_{loo}$ (my preference).
  - "loo" does not calculate each summand in $\widehat{elpd}_{loo}$ by MCMC. Instead it approximates the sum using a 10,000 x N matrix of log-likelihoods.

The $\widehat{elpd}_{loo}$ Statistic

- After some algebra we can see that

$$LOOIC = -2 \cdot \sum_{n=1}^{N} \overline{\{L(x_n|x_{(-n)}, \theta_i)\}}_{i=1}^{10,000}$$

- which we like as it favors the model with the largest likelihood, and the smallest LOOIC, on the "holdout" data.
- Some "loo" package features.
  - $\sum_{n=1}^{N} \overline{\{L(x_n|x_{(-n)}, \theta_i)\}}_{i=1}^{10,000} \equiv \widehat{elpd}_{loo}$ (my preference).
  - "loo" does not calculate each summand in $\widehat{elpd}_{loo}$ by MCMC. Instead it approximates the sum using a 10,000 x N matrix of log-likelihoods.
- Let's examine the script of "Intro_CRC.R"

Adjusting the Prior Distribution

- When fitting Bayesian models I initially use wide, but proper, prior distributions. By "wide" I mean wider than I really believe, but not outlandishly so. I like to leave room for surprises.
- Let's look at "Intro_CRC.R" and adjust the prior for the α parameters.

Adjusting the Prior Distribution

- When fitting Bayesian models I initially use wide, but proper, prior distributions. By "wide" I mean wider than I really believe, but not outlandishly so. I like to leave room for surprises.
- Let's look at "Intro_CRC.R" and adjust the prior for the α parameters.
- Let's discuss the "Cape Cod" model.

The Changing Settlement Rate (CSR) Model

The red text denotes the changes from the CRC model.

1. $logelr \sim$ normal(-0.4,$\sqrt{10}$).
2. $\alpha_w \sim \text{normal}(0, \sqrt{10})$ for $w = 2, \ldots, 10$. Set $\alpha_1 = 0$.
3. $\beta_d \sim \text{normal}(0, \sqrt{10})$ for $d = 1, \ldots, 9$. Set $\beta_{10} = 0$.
4. $\gamma \sim \text{normal}(0, 0.05)$
5. $a_i \sim \text{uniform}(0, 1)$ for $i = 1, \ldots, 10$.
6. Set $\sigma_d^2 = \sum_{i=d}^{10} a_i$ for $d = 1, \ldots, 10$.
7. Set $\mu_{wd} = \log(\text{Premium}_w) + logelr + \alpha_w + \beta_d \cdot (1-\gamma)^{w-1}$.
8. Then $C_{wd} \sim \text{lognormal}(\mu_{wd}, \sigma_d)$.

The Changing Settlement Rate (CSR) Model

The red text denotes the changes from the CRC model.

1. $logelr \sim$ normal(-0.4,$\sqrt{10}$).
2. $\alpha_w \sim$ normal$(0, \sqrt{10})$ for $w = 2, \ldots, 10$. Set $\alpha_1 = 0$.
3. $\beta_d \sim$ normal$(0, \sqrt{10})$ for $d = 1, \ldots, 9$. Set $\beta_{10} = 0$.
4. $\gamma \sim$ normal$(0, 0.05)$
5. $a_i \sim$ uniform$(0, 1)$ for $i = 1, \ldots, 10$.
6. Set $\sigma_d^2 = \sum_{i=d}^{10} a_i$ for $d = 1, \ldots, 10$.
7. Set $\mu_{wd} = \log(\text{Premium}_w) + logelr + \alpha_w + \beta_d \cdot (1-\gamma)^{w-1}$.
8. Then $C_{wd} \sim$ lognormal(μ_{wd}, σ_d).

If $\gamma > 0$ the development "factor" is squeezed toward zero which indicates a speed up in claim settlement. Similarly, $\gamma < 0$ indicates a slowdown in claim settlement.

Running the CSR Model

- Open "Intro_CSR.R" and run it.
- Let's compare the CSR to the CRC model.

Bad CRoss Classified (CRC) Models

For Accident Year w, Development Year d and cumulative loss C_{wd}:

1. $logelr \sim \text{normal}(-0.4, \sqrt{10})$.
2. $\alpha_w \sim \text{normal}(0, \sqrt{10})$ for $w = 2, \ldots, 10$. ~~Set $\alpha_1 = 0$.~~
3. $\beta_d \sim \text{normal}(0, \sqrt{10})$ for $d = 1, \ldots, 9$. ~~Set $\beta_{10} = 0$.~~
4. $a_i \sim \text{uniform}(0, 1)$ for $i = 1, \ldots, 10$.
5. Set $\sigma_d^2 = \sum_{i=d}^{10} a_i$ for $d = 1, \ldots, 10$.
6. Set $\mu_{wd} = \log(\text{Premium}_w) + logelr + \alpha_w + \beta_d$.
7. Then $C_{wd} \sim \text{lognormal}(\mu_{wd}, \sigma_d)$.

The Folk Theorem of Statistical Computing

- When you and your computer have some spare time, run the following models.
  - "Intro_Bad_CRC.R" - Drops the requirement that $\alpha_1 \equiv 0$.
  - "Intro_Really_Bad_CRC.R" - In addition, it drops the requirement that $\beta_{10} \equiv 0$.
  - You will find that these models take considerably longer to run. But

The Folk Theorem of Statistical Computing

- When you and your computer have some spare time, run the following models.
  - "Intro_Bad_CRC.R" - Drops the requirement that $\alpha_1 \equiv 0$.
  - "Intro_Really_Bad_CRC.R" - In addition, it drops the requirement that $\beta_{10} \equiv 0$.
  - You will find that these models take considerably longer to run. But
  - They get good results!

The Folk Theorem of Statistical Computing

- When you and your computer have some spare time, run the following models.
  - "Intro_Bad_CRC.R" - Drops the requirement that $\alpha_1 \equiv 0$.
  - "Intro_Really_Bad_CRC.R" - In addition, it drops the requirement that $\beta_{10} \equiv 0$.
  - You will find that these models take considerably longer to run. But
  - They get good results!
- A warning — The Folk Theorem of Statistical Computing (A. Gelman 2008) says that "When you have computational problems, often there's a problem with your model."

The Folk Theorem of Statistical Computing

- When you and your computer have some spare time, run the following models.
  - "Intro_Bad_CRC.R" - Drops the requirement that $\alpha_1 \equiv 0$.
  - "Intro_Really_Bad_CRC.R" - In addition, it drops the requirement that $\beta_{10} \equiv 0$.
  - You will find that these models take considerably longer to run. But
  - They get good results!
- A warning — The Folk Theorem of Statistical Computing (A. Gelman 2008) says that "When you have computational problems, often there's a problem with your model."
- Be wary of "Divergent Transitions." Always investigate. A few may be OK, but if you get a lot of them — something is wrong.

Three Objectives of This Session

1. Provide a high-level explanation of Bayesian MCMC using the Metropolis Hastings algorithm.
2. Provide some hands-on experience using what I ***currently*** consider to be the best Bayesian MCMC software.
3. Stan makes model building easy. — But as actuaries we need to focus on model formulation and evaluation. My objective was to get you started on those tasks.

More MCMC

- Stochastic Loss Reserving Using Bayesian MCMC Models
  - First Edition — January 2015
  - Second Edition — September 2019
- Case Study - Session AR-3, Today at 11:15 AM in the Verbena Room