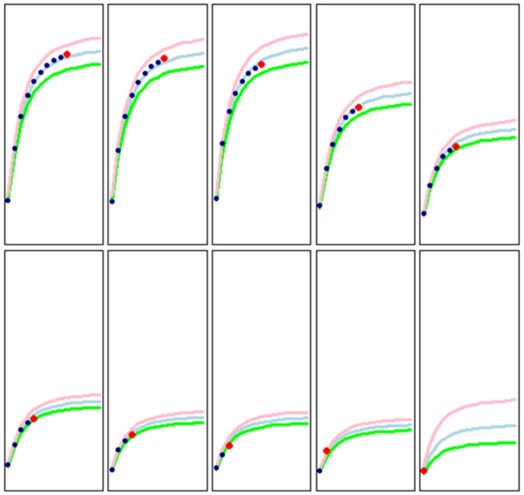
Deloitte.



A Credible Approach to Reserving
Non-Linear Hierarchical Bayesian Models for Loss Reserving

CAS Loss Reserving Seminar
Denver

Jim Guszcza, FCAS, MAAA
Deloitte Consulting LLP

September 6, 2012

Antitrust Notice

The Casualty Actuarial Society is committed to adhering strictly to the letter and spirit of the antitrust laws. Seminars conducted under the auspices of the CAS are designed solely to provide a forum for the expression of various points of view on topics described in the programs or agendas for such meetings.

Under no circumstances shall CAS seminars be used as a means for competing companies or firms to reach any understanding – expressed or implied – that restricts competition or in any way impairs the ability of members to exercise independent business judgment regarding matters affecting competition.

It is the responsibility of all seminar participants to be aware of antitrust regulations, to prevent any written or verbal discussions that appear to violate these laws, and to adhere in every respect to the CAS antitrust compliance policy.

Themes

Motivations

Hierarchical Models

Bayesian Concepts

Our Model

Case Study

Preamble

Why Bayes, Why Now

From John Kruschke, Indiana University:

"An open letter to Editors of journals, Chairs of departments, Directors of funding programs, Directors of graduate training, Reviewers of grants and manuscripts, Researchers, Teachers, and Students":

Statistical methods have been evolving rapidly, and many people think it's time to adopt modern Bayesian data analysis as standard procedure in our scientific practice and in our educational curriculum. Three reasons:

1. Scientific disciplines from astronomy to zoology are moving to Bayesian data analysis. **We should be leaders of the move, not followers.**
2. Modern Bayesian methods provide richer information, with greater flexibility and broader applicability than 20th century methods. Bayesian methods are intellectually coherent and intuitive. **Bayesian analyses are readily computed with modern software and hardware.**
3. Null-hypothesis significance testing (NHST), with its reliance on *p* values, has many problems. **There is little reason to persist with NHST now that Bayesian methods are accessible to everyone.**

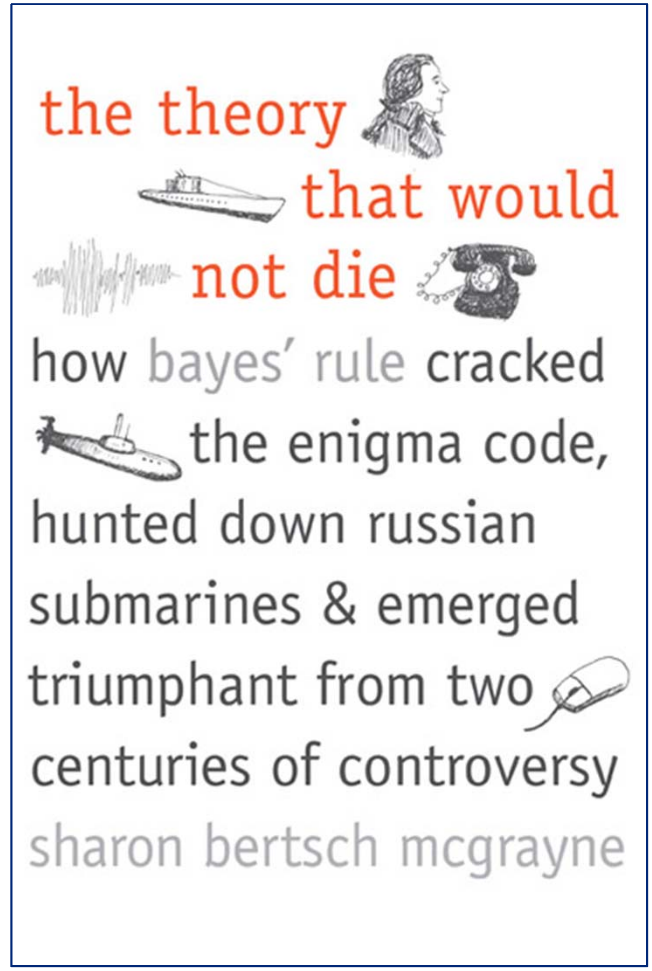
My conclusion from those points is that we should do whatever we can to encourage the move to Bayesian data analysis.

(I couldn't have said it better myself…)

Why Bayes, Why Now

From an Interview with Sharon Bertsch McGrayne in *Chance* Magazine:

"When I started research on [my] book, I could Google the word 'Bayesian' and get 100,000 hits. Recently I got 17 million."



Our Profession's Bayesian Heritage: Early

- Late 18th Century: Thomas Bayes and Pierre-Simon Laplace formulate the principles of "inverse probability"
  - Probabilistic inference from data to model parameters
  - Bayes' intellectual executor, Richard Price, became perhaps the world's first consulting actuary (Equitable Life Assurance company, London)
  - Price's – and perhaps Bayes' – thinking was influenced by the publication of David Hume's Treatise on Human Nature (1740)

- 1918: A. W. Whitney "The Theory of Experience Rating".
  - Advocated combining the claims experience of a single risk with that of a cohort (class, portfolio, …) of similar risks.

$$\bar{\mu} = Z \cdot \hat{\mu}_{risk} + (1 - Z) \cdot \hat{\mu}_{class} \quad , \quad Z = \frac{w}{w + k}$$

  - Estimated pure premium should be a weighted average of the individual risk's claim experience with that of the cohort… *k* is judgmentally determined.

Our Profession’s Bayesian Heritage: Early-Modern

- 1950: Arthur Bailey publishes “Credibility Procedures: Laplace’s Generalization of Bayes’ Rule and the Combination of Collateral Knowledge with Observed Data”.
  - Anticipates Hans Bühlmann's subsequent work.
  - Quoted Richard Price on making inferences from available data.

“At present, practically all methods of statistical estimation appearing in textbooks… are based on an equivalent to the assumption that any and all collateral information or a priori knowledge is worthless. There have been rare instances of rebellion against this philosophy by practical statisticians who have insisted that they actually had a considerable store of knowledge apart from the specific observations being analyzed… However it appears to be only in the actuarial field that there has been an organized revolt against discarding all prior knowledge when an estimate is to be made using newly acquired data.”

Our Profession's Bayesian Heritage: Mid-Century Modern

- 1967: Bühlmann's "greatest accuracy" Bayes credibility model.
  - Let X_{ij} denote dollars of loss associated with risk *i* at time *j*.
  - Assume $X_1, \ldots, X_m$ are iid, conditional on a parameter (vector) θ
  - Let $m(\theta_i)$ denote "risk premium": $m(\theta_i) \equiv E[X_{ij}|\theta_i]$

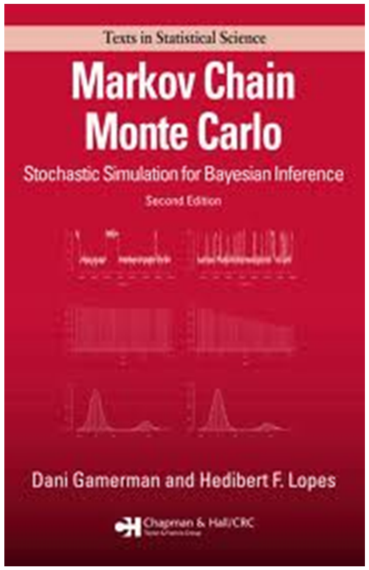
- Bühlmann minimizes mean squared errors: $$E\left[m(\theta_i) - \alpha - \sum_j \beta_j X_{ij}\right]^2$$

- … to arrive at an estimator for $m(\theta_i)$: $$z_i \cdot \bar{X}_i + (1 - z_i) \cdot \mu$$

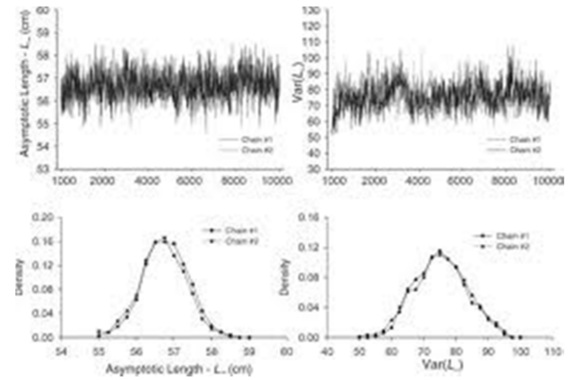
- … where: $$z_i = \frac{n_i}{n_i + k} \quad , \quad k = \frac{E\left[Var\left(X_{ij} \mid \theta_i\right)\right]}{Var\left(m\left(\theta_i\right)\right)}$$

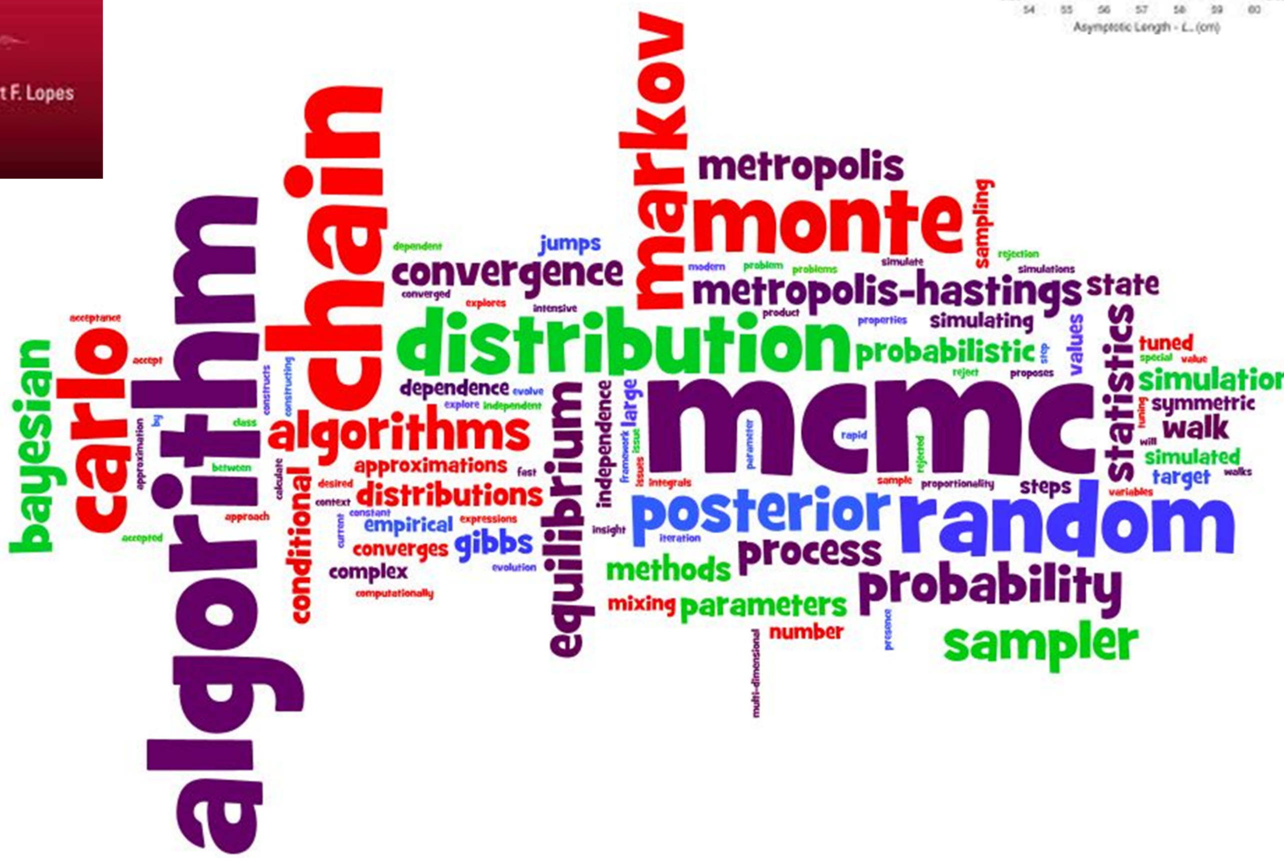
- The within/between variances in *k* are estimated from the data.

Our Profession's Bayesian Heritage: Modern



?





Loss Reserving and its Discontents

Loss Reserving and its Discontents

- Much loss reserving practice is still "pre-theoretical" in nature.
  - Techniques like chain ladder, BF, and Cape Cod aren't performed in a statistical modeling framework.
  - (Do people agree with this statement?)

- Traditional methods aren't necessarily optimal from a statistical POV.
  - Potential of over-fitting small datasets
  - Difficult to assess goodness-of-fit, compare nested models, etc
  - Often no concept of out-of-sample validation or diagnostic plots

- Related point: traditional methods produce point estimates only.
  - Reserve variability estimates are often ad-hoc

Models vs Methods

- Rather than promulgating a collection of loss reserving **methods**, we build statistical **models** of loss development.
  - Attempt to place loss reserving practice on a sound scientific footing.
  - Field is developing rapidly

- Today: Sketch non-linear hierarchical Bayesian models
  - Natural, parsimonious models of the loss development process
  - Initially motivated by Dave Clark's [2003] paper as well as hierarchical Bayesian modeling theory.

- By the way: the debate over "models vs methods" is misleading
  - Rather we want to have a flexible and extensible **modeling methodology**
  - A framework that can always be tailored to the specifics of a given situation
  - Spreadsheets aren't the only way to accomplish this

Four Essential Features of Loss Reserving

- Repeated measures
  - Loss reserving is longitudinal data analysis

Cumulative Losses in 1000's

AY	*premium*	*12*	*24*	*36*	*48*	*60*	*72*	*84*	*96*	*108*	*120*
1988	2,609	404	986	1,342	1,582	1,736	1,833	1,907	1,967	2,006	2,036
1989	2,694	387	964	1,336	1,580	1,726	1,823	1,903	1,949	1,987	
1990	2,594	421	1,037	1,401	1,604	1,729	1,821	1,878	1,919		
1991	2,609	338	753	1,029	1,195	1,326	1,395	1,446			
1992	2,077	257	569	754	892	958	1,007				
1993	1,703	193	423	589	661	713					
1994	1,438	142	361	463	533						
1995	1,093	160	312	408							
1996	1,012	131	352								
1997	976	122									

- A “bundle” of time series
  - A loss triangles is a collection of time series that are “related” to one another
  - … but no guarantee that the same development pattern is appropriate to all

- Non-linear
  - Each year’s development patter is inherently non-linear
  - Ultimate loss (ratio) is an asymptote

- Incomplete information
  - Few loss triangles contain all of the information needed to make forecasts
  - Most reserving exercises must incorporate judgment and/or background information

➔ ***Loss reserving is inherently Bayesian!***

Towards a More Realistic Loss Reserving Framework

- How many stochastic reserving techniques reflect all of these considerations?
  1. Repeated Measures (isn't loss reserving a type of longitudinal data analysis?)
  2. Multiple time series
  3. Non-linear (are GLMs really appropriate?)
  4. Incomplete Information ("Bayes or Bust"!)

- 1-2 ➔ We need **hierarchical** models
- 3 ➔ They should use **growth curves**
- 4 ➔ Non-linear hierarchical models should be **Bayesian**

Cumulative Losses in 1000's

AY	*premium*	*12*	*24*	*36*	*48*	*60*	*72*	*84*	*96*	*108*	*120*
1988	2,609	404	986	1,342	1,582	1,736	1,833	1,907	1,967	2,006	2,036
1989	2,694	387	964	1,336	1,580	1,726	1,823	1,903	1,949	1,987	
1990	2,594	421	1,037	1,401	1,604	1,729	1,821	1,878	1,919		
1991	2,609	338	753	1,029	1,195	1,326	1,395	1,446			
1992	2,077	257	569	754	892	958	1,007				
1993	1,703	193	423	589	661	713					
1994	1,438	142	361	463	533						
1995	1,093	160	312	408							
1996	1,012	131	352								
1997	976	122									

Another Big Motivation: Predictive Distributions

> **"Given any value (estimate of future payments) and our current state of knowledge, what is the probability that the final payments will be no larger than the given value?"**
> **-- Casualty Actuarial Society**
> **Working Party on Quantifying Variability in Reserve Estimates, 2004**

- This can be read as a request for a Bayesian analysis
  - Bayesians (unlike frequentists) are willing to make probability statements about unknown parameters
  - Ultimate losses are "single cases" – difficult to conceive as random draws from a "sampling distribution in the sky".
    - Frequentist probability involved repeated trials of setups involving physical randomization.
    - In contrast it is meaningful to apply Bayesian probabilities to "single case events"
  - The Bayesian analysis yields an entire posterior probability distribution – not merely moment estimates

➔ ***Bayesian statistics is the ideal framework for loss reserving!***

The Bayesian Perspective

"For Bayesians as much as for any other statistician, parameters are (typically) fixed but unknown. It is the knowledge about these unknowns that Bayesians model as random…

… typically it is the Bayesian who makes the claim for inference in a particular instance and the frequentist who restricts claims to infinite populations of replications."

- Andrew Gelman and Christian Robert

Origin of the Approach: Dave's Idea + Random Effects

LDF Curve-Fitting and Stochastic Reserving: A Maximum Likelihood Approach

or

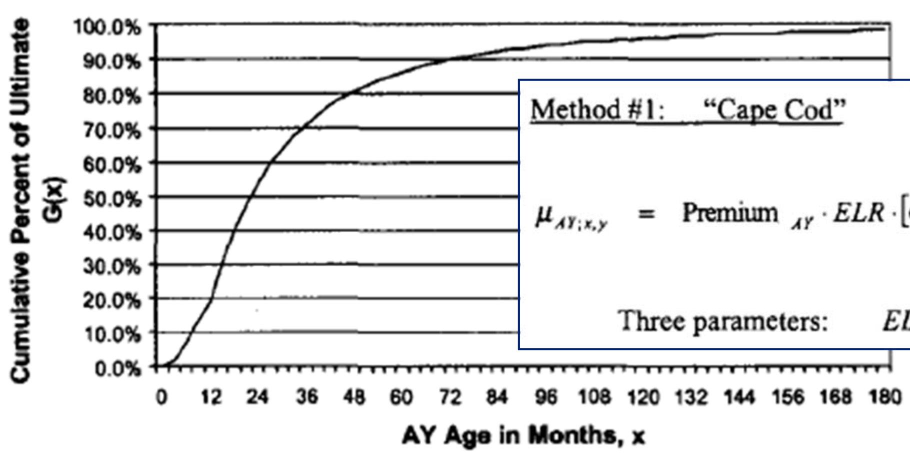
How to Increase Reserve Variability with Less Data

David R. Clark
American Re-Insurance

2003 Reserves Call Paper Program

+

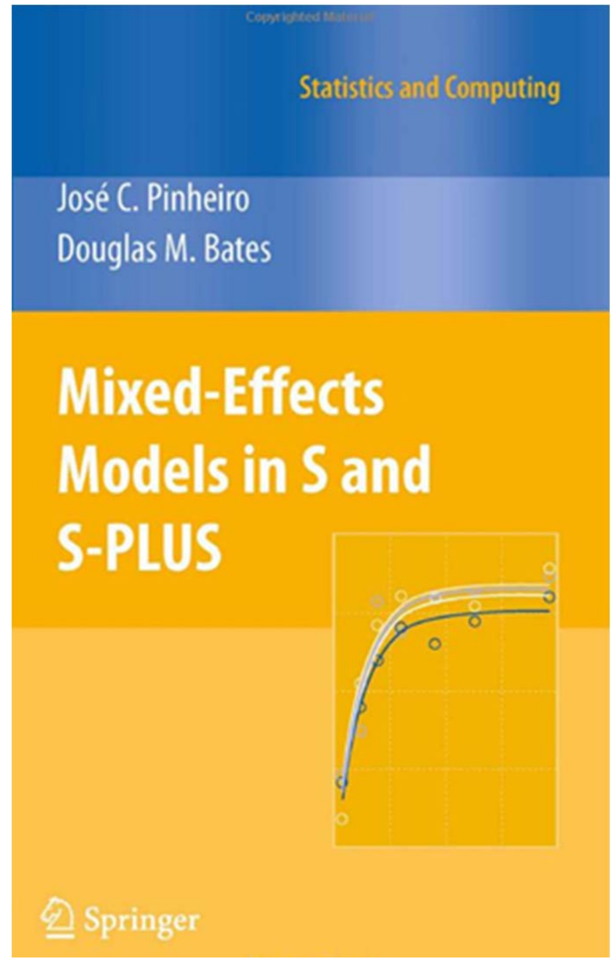
$G(x) \;=\; 1/LDF_x$ = cumulative % reported (or paid) as of time x



Method #1:   "Cape Cod"

$$\mu_{AY;x,y} \;=\; \text{Premium}_{AY} \cdot ELR \cdot [G(y \mid \omega,\theta) - G(x \mid \omega,\theta)]$$

Three parameters: $ELR,\ \omega,\ \theta$



Current State of… uh… Development

Journal of the
Royal Statistical Society
SERIES A
Statistics in Society

J. R. Statist. Soc. A (2012)
175, *Part* 2, *pp.*

A Bayesian non-linear model for forecasting insurance loss payments

Yanwei Zhang,

CNA Insurance Company, Chicago, USA

Vanja Dukic

University of Colorado—Boulder, USA

and James Guszcza

University of Wisconsin—Madison, USA

[Received August 2010. Revised July 2011]

Components of Our Approach

- **Growth curves** to model the loss development process (Clark 2003)
  - Parsimony – obviates the need for tail factors
- Loss reserving treated as **longitudinal data analysis** (Guszcza 2008)
  - A type of hierarchical modeling
  - Parsimony; similar approach to non-linear mixed effects models used in biological/social sciences
- Further using the hierarchical modeling framework to simultaneously model **multiple loss triangles** (Zhang-Dukic-Guszcza 2012)
  - "Borrow strength" from other loss reserving triangles
  - Similar in spirit to credibility theory
  - Insufficient time to cover this aspect today
- Building a **fully Bayesian** model by assigning prior probability distributions to all hyperparameters (Zhang-Dukic-Guszcza 2012)
  - Provides formal mechanism for incorporating background knowledge and expert opinion with data-driven indications.
  - Results in full predictive distribution of all quantities of interest
  - Conceptual advantages: Bayesian paradigm treats data as fixed and parameters are randomly varying

The Notion of Hierarchical Structure is Key

- **NB: Bayesian models ≠ Hierarchical models!**
  - E.g. we could fit a Bayesian chain ladder by putting priors on the parameters of an overdispersed Poisson regression model… but this wouldn't make it hierarchical.
  - Similarly non-Bayesian hierarchical models are a useful way to quickly fit "exploratory" models while gearing up to do a fully Bayesian analysis.

BRAINSTORMS

Predicting Loss Ratios with a Hierarchical Bayesian Model

By Glenn Meyers

At the recent CAS Annual Meeting in Washington, I saw concurrent session presentations by Peng Shi, Jim Guszcza, and Wayne Zhang that applied hierarchical Bayesian models to the problem of stochastic loss reserving.[1] These three presentations reinforced my opinion that that the time has come to include these models in loss reserving. To help the rest of us catch up, this column describes an example of a hierarchical Bayesian model in a simpler setting, that of estimating loss ratios. I will try to describe the example at a high level, with the details given in the R code that accompanies the Web version of this article.

Hierarchical Models

What is Hierarchical Modeling?

- Hierarchical modeling is used when one's data is **grouped** in some important way.
  - Claim experience by state or territory
  - Workers Comp claim experience by class code
  - Claim severity by injury type
  - Churn rate by agency
  - Multiple years of loss experience by policyholder.
  - **Multiple observations of a cohort of claims over time**

- Often grouped data is modeled either by:
  - Building separate models by group
  - Pooling the data and introducing dummy variables to reflect the groups

- Hierarchical modeling offers a "middle way".
  - Parameters reflecting group membership enter one's model through appropriately specified **probability sub-models**.

Common Hierarchical Models

- **Classical Linear Model**

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

  - Equivalently: $Y_i \sim N(\alpha + \beta X_i, \sigma^2)$
  - Same α, β for each data point

- Random Intercept Model

$$Y_i = \alpha_{j[i]} + \beta X_i + \varepsilon_i$$

  - Where: $Y_i \sim N(\alpha_{j[i]} + \beta X_i, \sigma^2)$
  - And: $\alpha_j \sim N(\mu_\alpha, \sigma^2_\alpha)$
  - Same β for each data point; but α varies by group j

- Random Intercept and Slope Model

$$Y_i = \alpha_{j[i]} + \beta_{j[i]} X_i + \varepsilon_i$$

  - Both α and β vary by group

$$Y_i \sim N\left(\alpha_{j[i]} + \beta_{j[i]} \cdot X_i, \sigma^2\right) \quad where \quad \begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} \sim N\left(\begin{bmatrix} \mu_\alpha \\ \mu_\beta \end{bmatrix}, \Sigma\right) \quad , \quad \Sigma = \begin{bmatrix} \sigma_\alpha^2 & \sigma_{\alpha\beta} \\ \sigma_{\alpha\beta} & \sigma_\beta^2 \end{bmatrix}$$

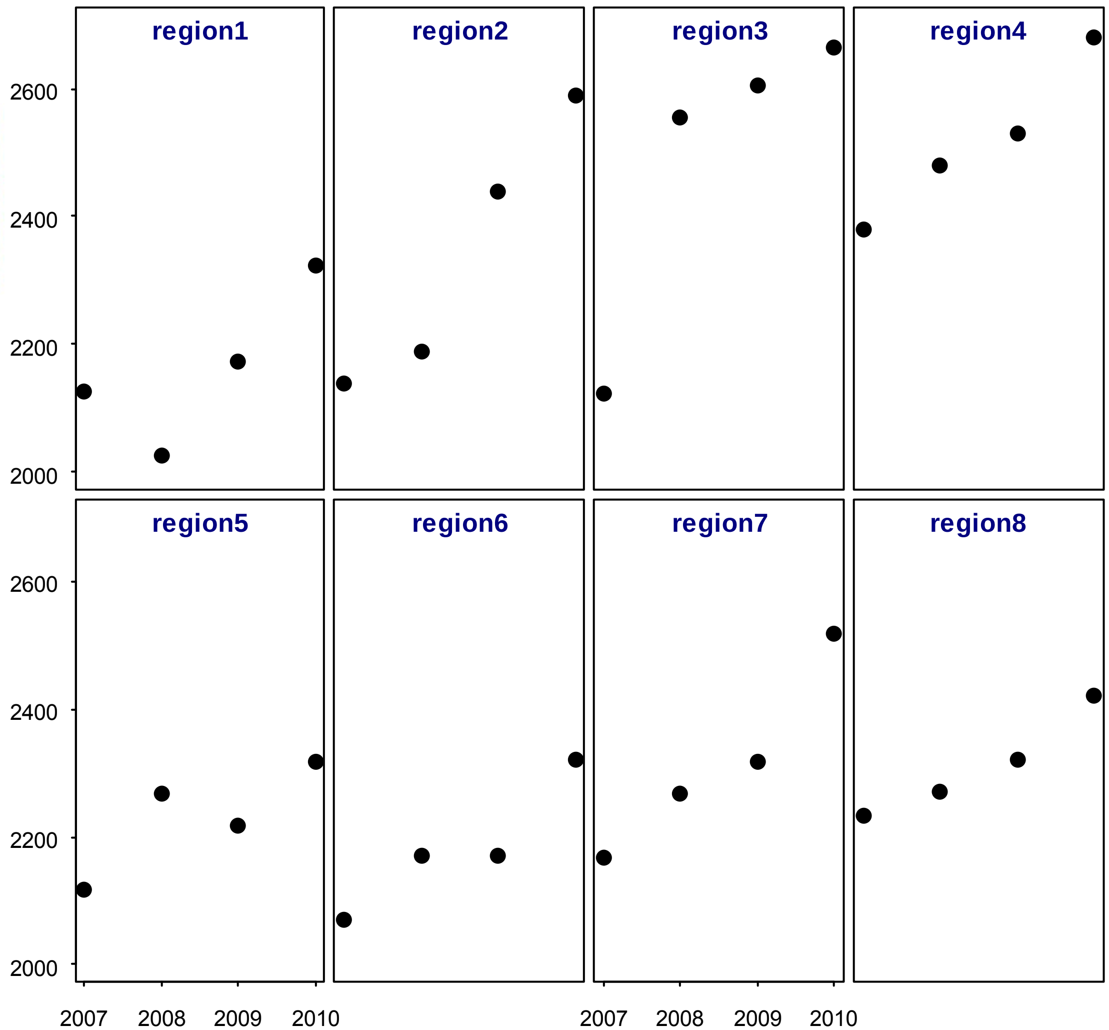
Simple Example: PIF by Region

- Simple example: Change in PIF by region from 2007-10
- 32 data points
  - 4 years
  - 8 regions

region	2005	2006	2007	2008
1	2124	2024	2174	2324
2	2138	2188	2438	2588
3	2121	2554	2604	2666
4	2380	2480	2530	2680
5	2118	2268	2218	2318
6	2070	2170	2170	2320
7	2167	2267	2317	2517
8	2232	2272	2322	2422

- But we could as easily have 80 or 800 regions
- Our model would not change
- We view the dataset as a bundle of very short time series

PIF Growth by Region



Classical Linear Model

- Option 1: the classical linear model

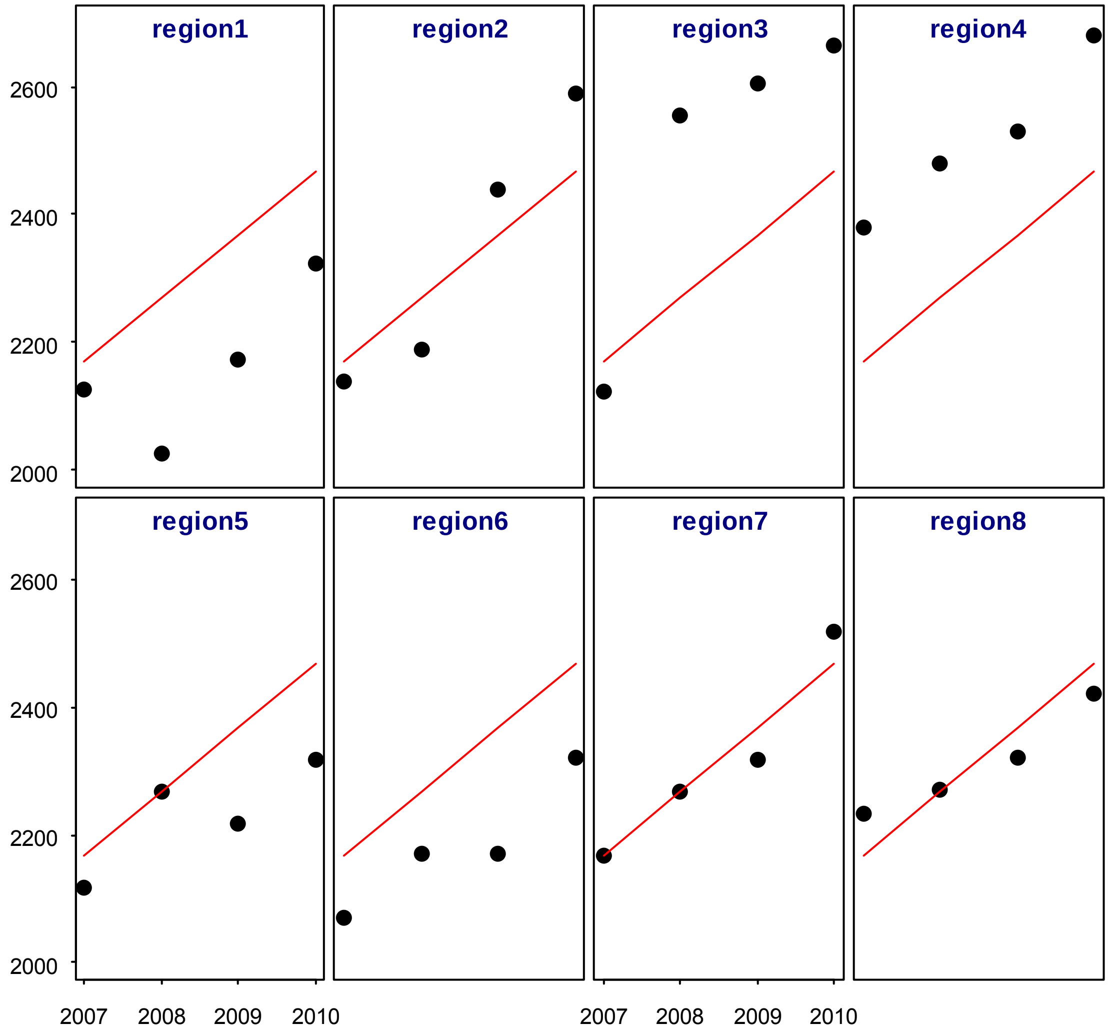
- Complete Pooling
  - Don't reflect region in the model design
  - Just throw all of the data into one pot and regress

- Same α and β for each region.

- This obviously doesn't cut it.
  - But filling 8 separate regression models or throwing in region-specific dummy variables isn't an attractive option either.
  - Danger of over-fitting
  - i.e. "credibility issues"

$$Y_i \sim N(\alpha + \beta X_i, \sigma^2)$$

PIF Growth by Region



Randomly Varying Intercepts

$$Y_i \sim N(\alpha_{j[i]} + \beta X_i, \sigma^2) \qquad \alpha_j \sim N(\mu_\alpha, \sigma_\alpha^2)$$

- Option 2: random intercept model

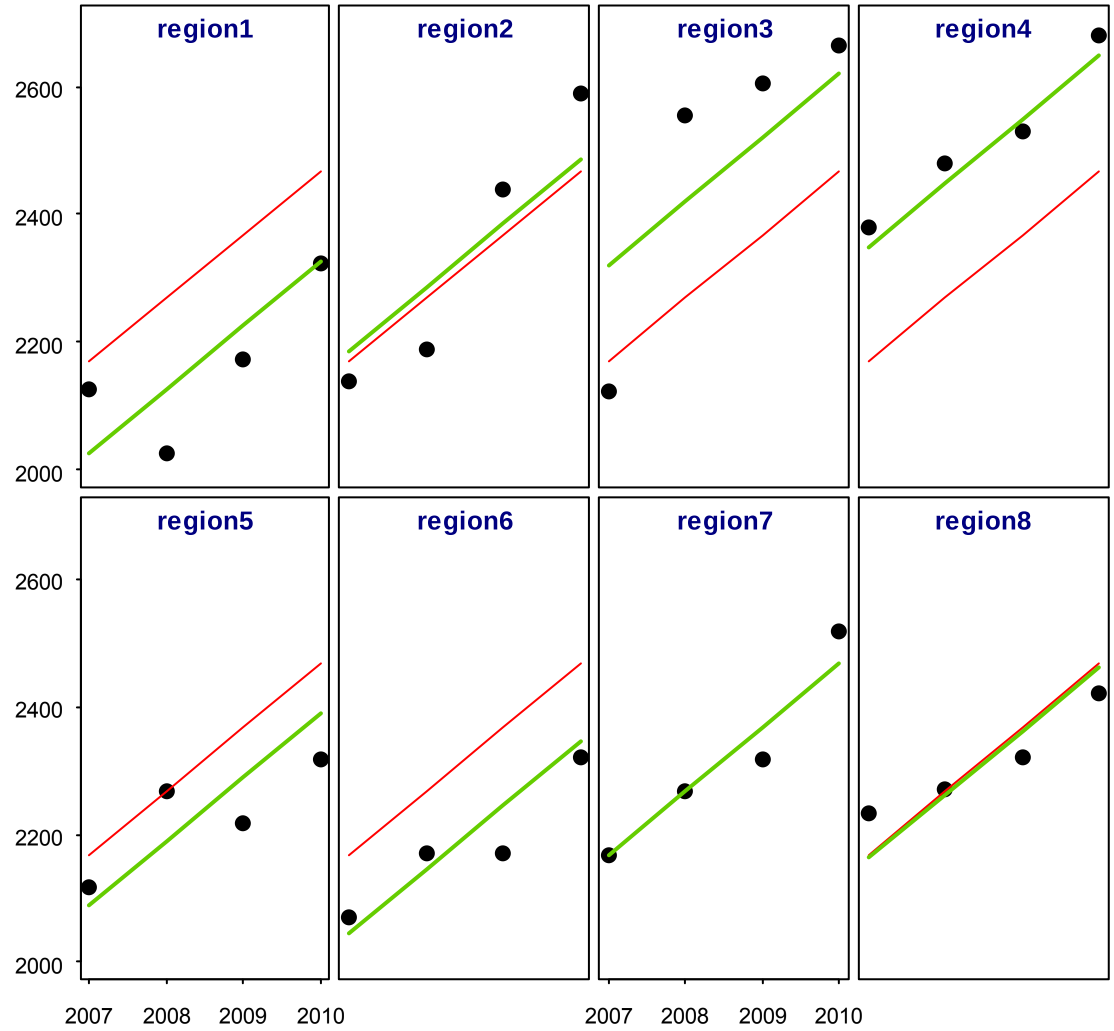
- $Y_i = \alpha_{j[i]} + \beta X_i + \varepsilon_i$

- This model has 9 parameters:
  $\{\alpha_1, \alpha_2, \ldots, \alpha_8, \beta\}$

- And it contains 4 hyperparameters:
  $\{\mu_\alpha, \beta, \sigma_\alpha, \sigma\}$

- A major improvement

PIF Growth by Region

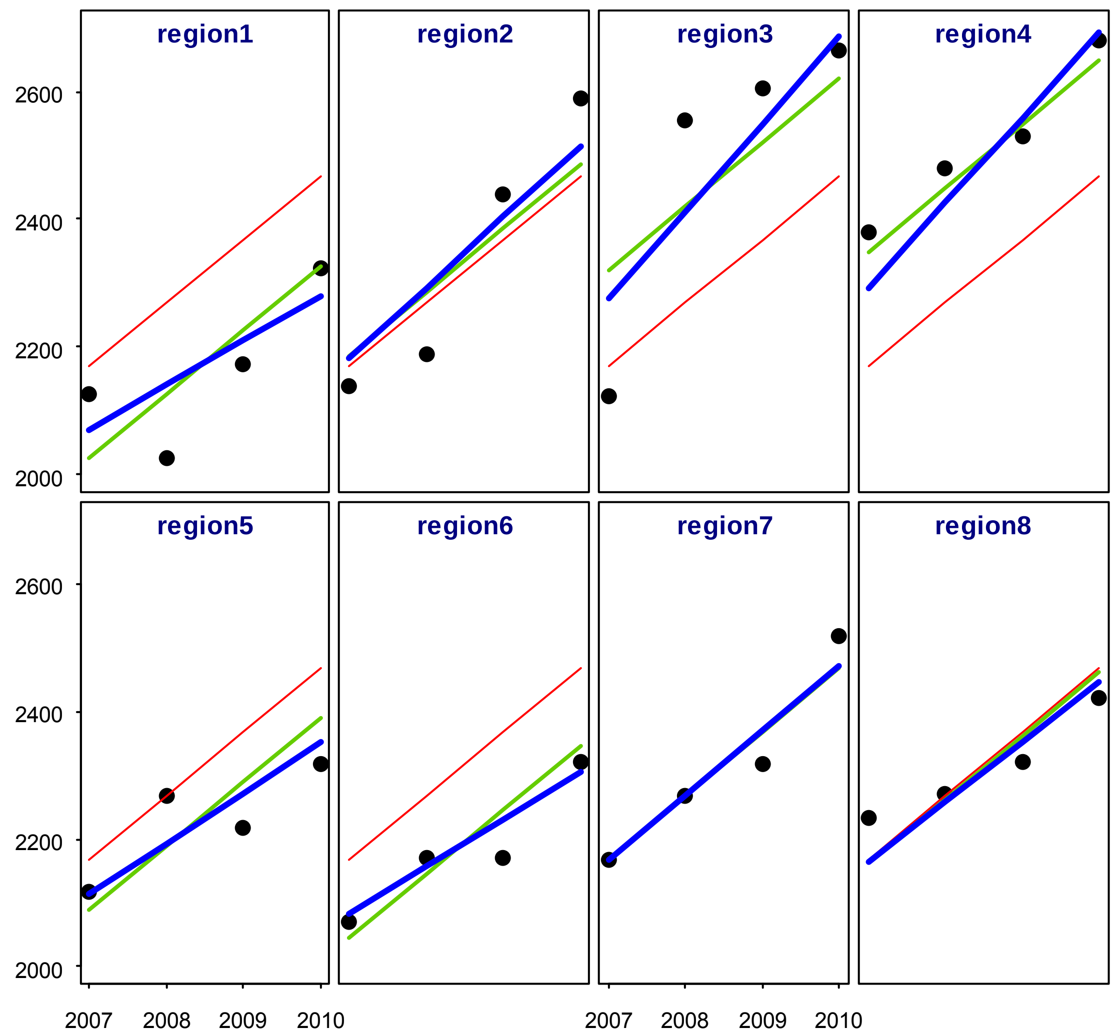


Randomly Varying Intercepts and Slopes

$$Y_i \sim N\left(\alpha_{j[i]} + \beta_{j[i]} \cdot X_i, \sigma^2\right) \quad where \quad \begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} \sim N\left(\begin{bmatrix} \mu_\alpha \\ \mu_\beta \end{bmatrix}, \Sigma\right) \quad , \quad \Sigma = \begin{bmatrix} \sigma_\alpha^2 & \sigma_{\alpha\beta} \\ \sigma_{\alpha\beta} & \sigma_\beta^2 \end{bmatrix}$$

- Option 3: the random slope and intercept model
- $Y_i = \alpha_{j[i]} + \beta_{j[i]} X_i + \varepsilon_i$
- This model has 16 parameters:
  $\{\alpha_1, \alpha_2, \ldots, \alpha_8,$
  $\beta_1, \beta_2, \ldots, \beta_8\}$
  - (note that 8 separate models also contain 16 parameters)
- And it contains 6 hyperparameters:
  $\{\mu_\alpha, \mu_\beta, \sigma, \sigma_\alpha, \sigma_\beta, \sigma_{\alpha\beta}\}$
- To repeat: the same number of hyperparameters if we had 80 or 800 regions

PIF Growth by Region



A Compromise Between Complete Pooling and No Pooling

$$PIF = \alpha + \beta t + \varepsilon$$

Complete Pooling

- Ignore group structure altogether

$$\left\{PIF = \alpha^k + \beta^k t + \varepsilon^k\right\}_{k=1,2,..,8}$$

No Pooling

- Estimate a separate model for each group

Compromise

Hierarchical Model

- Estimates parameters using a compromise between complete pooling and no pooling.

$$Y_i \sim N\left(\alpha_{j[i]} + \beta_{j[i]} \cdot X_i, \sigma^2\right) \quad where \quad \begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} \sim N\left(\begin{bmatrix} \mu_\alpha \\ \mu_\beta \end{bmatrix}, \Sigma\right) \quad , \quad \Sigma = \begin{bmatrix} \sigma_\alpha^2 & \sigma_{\alpha\beta} \\ \sigma_{\alpha\beta} & \sigma_\beta^2 \end{bmatrix}$$

A Credible Approach

- For illustration, recall the random intercept model:

$$Y_i \sim N(\alpha_{j[i]} + \beta X_i, \sigma^2) \qquad \alpha_j \sim N(\mu_\alpha, \sigma_\alpha^2)$$

- This model can contain a large number of parameters $\{\alpha_1, \ldots, \alpha_J, \beta\}$.

- Regardless of *J*, it contains 4 hyperparameters $\{\mu_\alpha, \beta, \sigma, \sigma_\alpha\}$.

- Here is how the hyperparameters relate to the parameters:

$$\hat{\alpha}_j = Z_j \cdot (\bar{y}_j - \beta \bar{x}_j) + (1 - Z_j) \cdot \hat{\mu}_\alpha \quad where \quad Z_j = \frac{n_j}{n_j + \sigma^2 / \sigma_\alpha^2}$$

Bühlmann credibility is a special case of hierarchical models.

A Fully Bayesian Model

With a Case Study

Case Study Data

- A garden-variety Workers Comp Schedule P loss triangle:

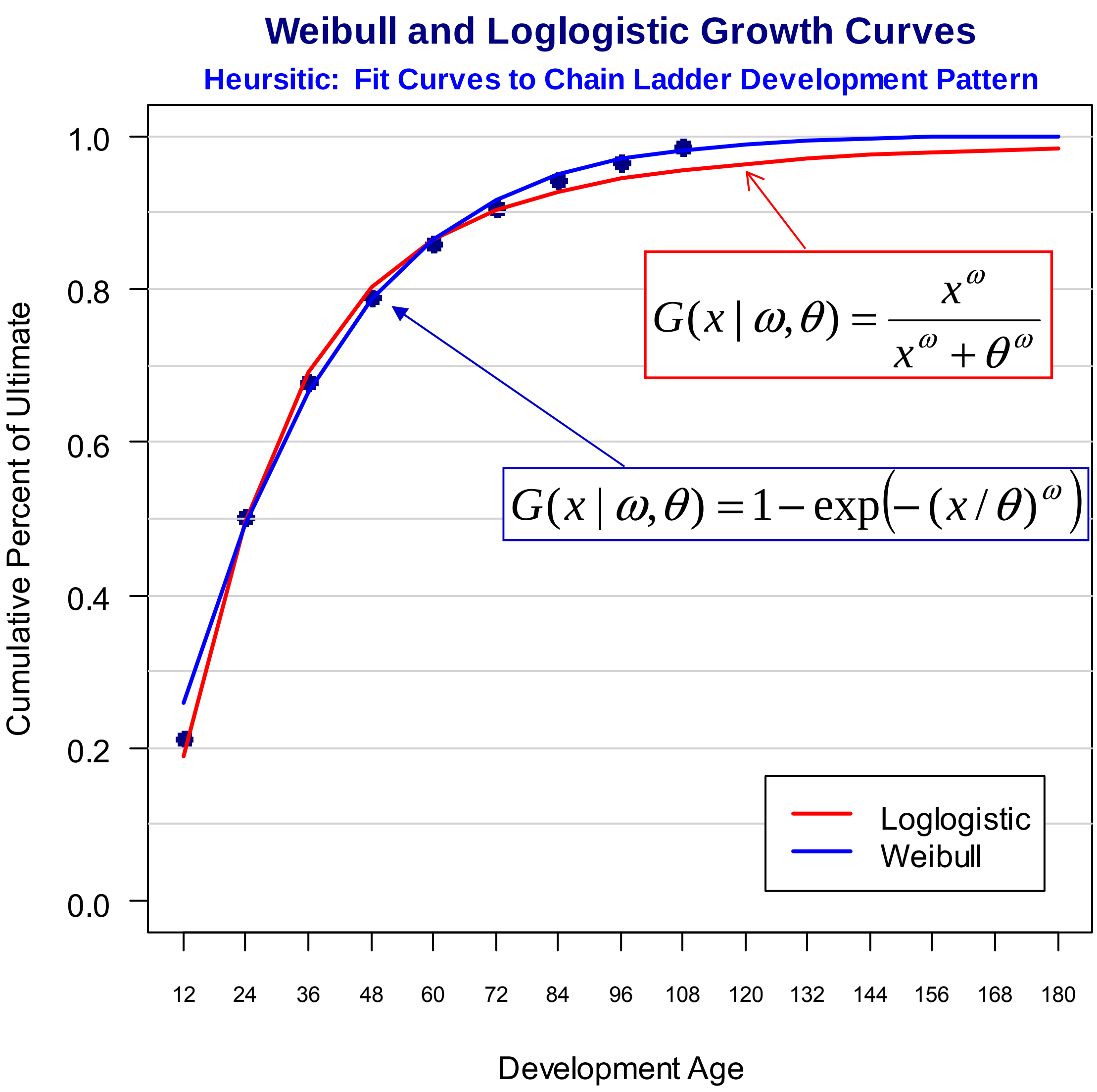
Cumulative Losses in 1000's														
AY	*premium*	*12*	*24*	*36*	*48*	*60*	*72*	*84*	*96*	*108*	*120*	*CL Ult*	*CL LR*	*CL res*
1988	2,609	404	986	1,342	1,582	1,736	1,833	1,907	1,967	2,006	2,036	2,036	0.78	0
1989	2,694	387	964	1,336	1,580	1,726	1,823	1,903	1,949	1,987		2,017	0.75	29
1990	2,594	421	1,037	1,401	1,604	1,729	1,821	1,878	1,919			1,986	0.77	67
1991	2,609	338	753	1,029	1,195	1,326	1,395	1,446				1,535	0.59	89
1992	2,077	257	569	754	892	958	1,007					1,110	0.53	103
1993	1,703	193	423	589	661	713						828	0.49	115
1994	1,438	142	361	463	533							675	0.47	142
1995	1,093	160	312	408								601	0.55	193
1996	1,012	131	352									702	0.69	350
1997	976	122										576	0.59	454
chain link		2.365	1.354	1.164	1.090	1.054	1.038	1.026	1.020	1.015	1.000	12,067		1,543
chain ldf		4.720	1.996	1.473	1.266	1.162	1.102	1.062	1.035	1.015	1.000			
growth curve		21.2%	50.1%	67.9%	79.0%	86.1%	90.7%	94.2%	96.6%	98.5%	100.0%			

- Let's model this as a longitudinal dataset.
- Grouping dimension: Accident Year (AY)

- We can build a parsimonious non-linear model that uses random effects to allow the model parameters to vary by accident year.

Growth Curves – At the Heart of the Model

- We want our model to reflect the **non-linear** nature of loss development.
  - GLM shows up a lot in the stochastic loss reserving literature…
  - … but are GLMs natural models for loss triangles?
- Growth curves (Clark 2003)
  - γ = ultimate loss ratio
  - θ = scale
  - ω = shape ("warp")
- Heuristic idea
  - We judgmentally select a growth curve form
  - Let γ vary by year (hierarchical)
  - Add priors to the hyperparameters (Bayesian)



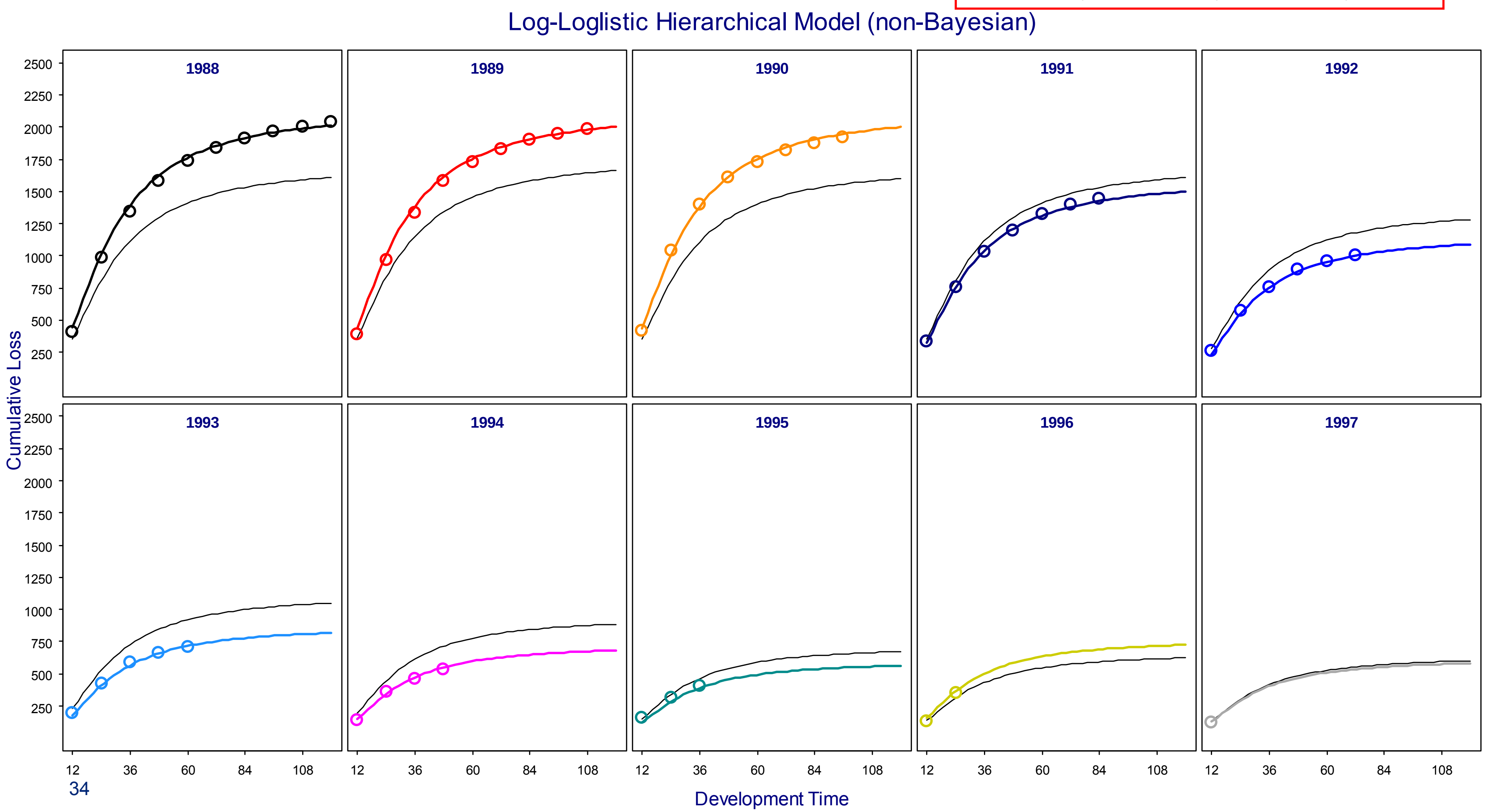
An Exploratory Non-Bayesian Hierarchical Model

- It is easy to fit non-Bayesian hierarchical models as a data exploration step.

$$y_i(t_j) = \gamma_i * p_i * \left(\frac{t^{\omega}}{t^{\omega} + \theta^{\omega}} \right) + \varepsilon_i(t_j)$$

$$\gamma_i \sim N\left(\gamma, \sigma_{\gamma}^2\right)$$

$$\varepsilon_i(t_j) = \rho \varepsilon_i(t_{j-1}) + \delta_i(t_j)$$

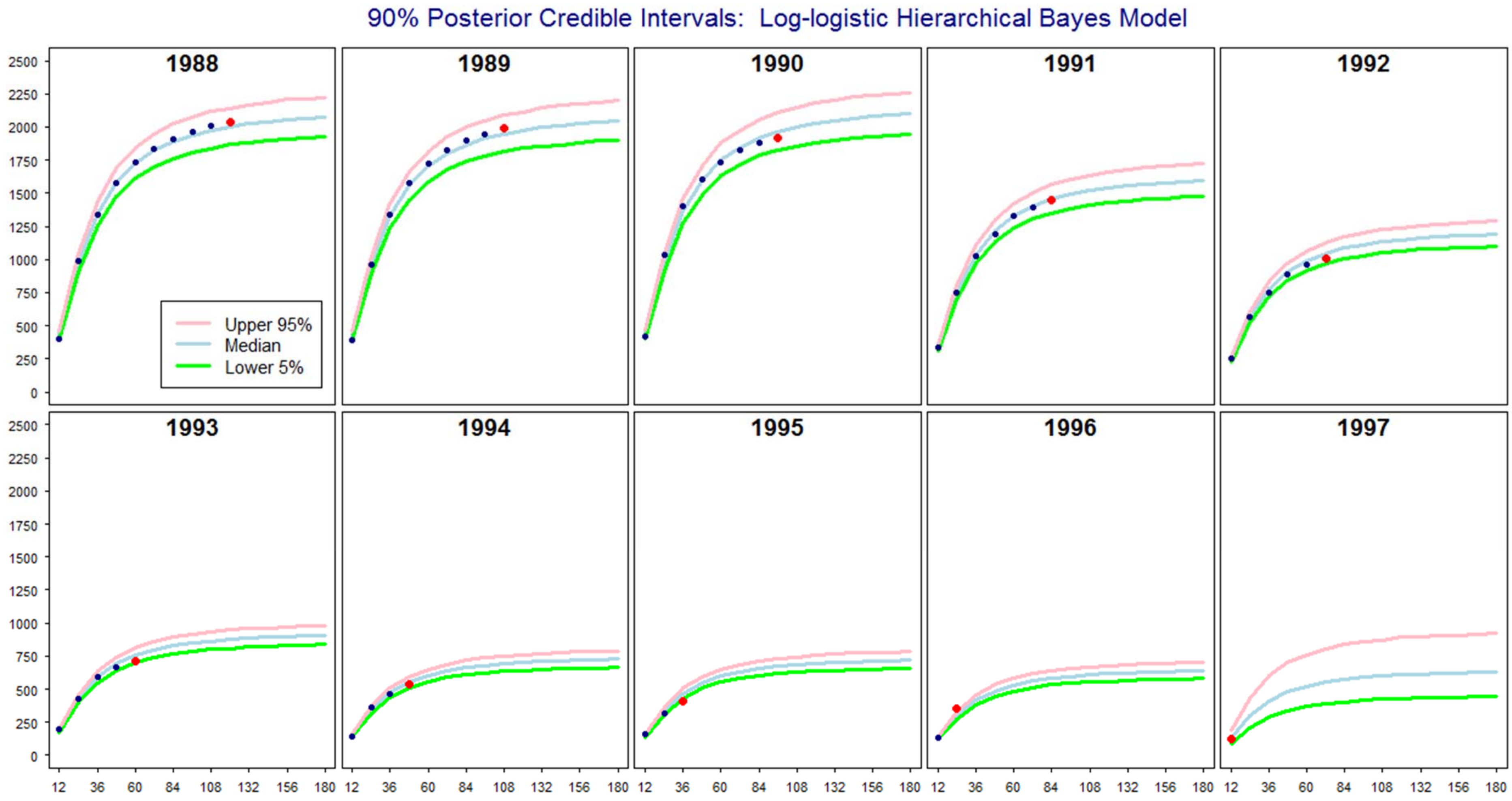


Adding Bayesian Structure

- Our hierarchical model is “half-way Bayesian”
  - On the one hand, we place probability sub-models on certain parameters
  - But on the other hand, various (hyper)parameters are estimated directly from the data.
- To make this fully Bayesian, we need to put probability distributions on **all** quantities that are uncertain.
- We then employ Bayesian updating: the model (“likelihood function”) together with the prior results in a posterior probability distribution over **all** uncertain quantites.
  - Including ultimate loss ratio parameters and hyperparameters!
  - ➔ We are directly modeling the ultimate quantity of interest.
- This is not as hard as it sounds:
  - We do **not** explicitly calculate the high-dimensional posterior probability distribution.
  - We **do** use Markov Chain Monte Carlo [MCMC] simulation to sample from the posterior.
  - Technology: JAGS (“Just Another Gibbs Sampler”), called from within R.

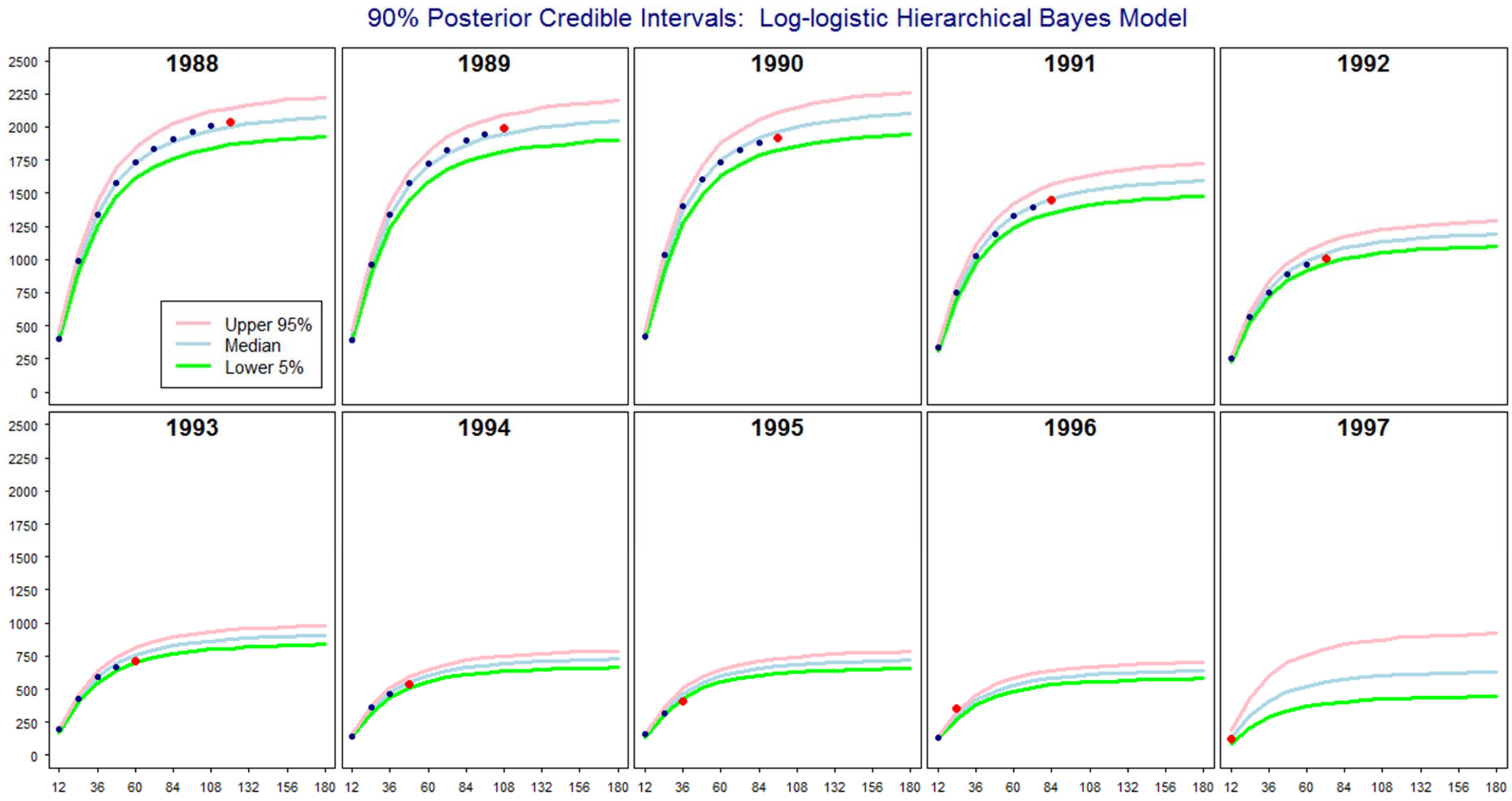
Results of a Fully Bayesian Hierarchical Model

- Now we fit a fully Bayesian version of the model by providing **prior distributions** for all of the model hyperparameters, and simulating the posterior distribution.



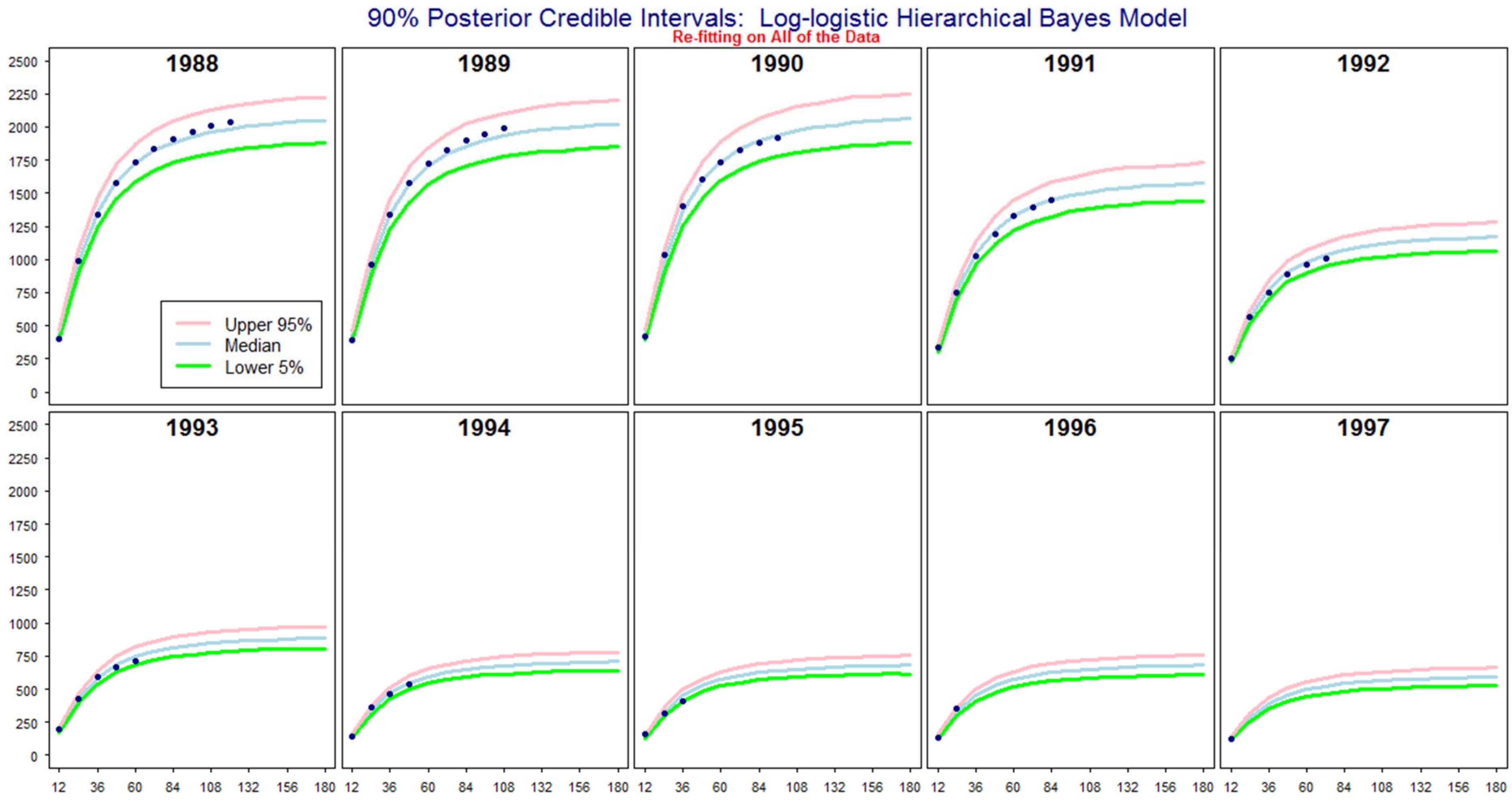
Results of a Fully Bayesian Hierarchical Model

- Here we are using the most recent Calendar Year (red) as a holdout sample.
- The model fits the holdout well.



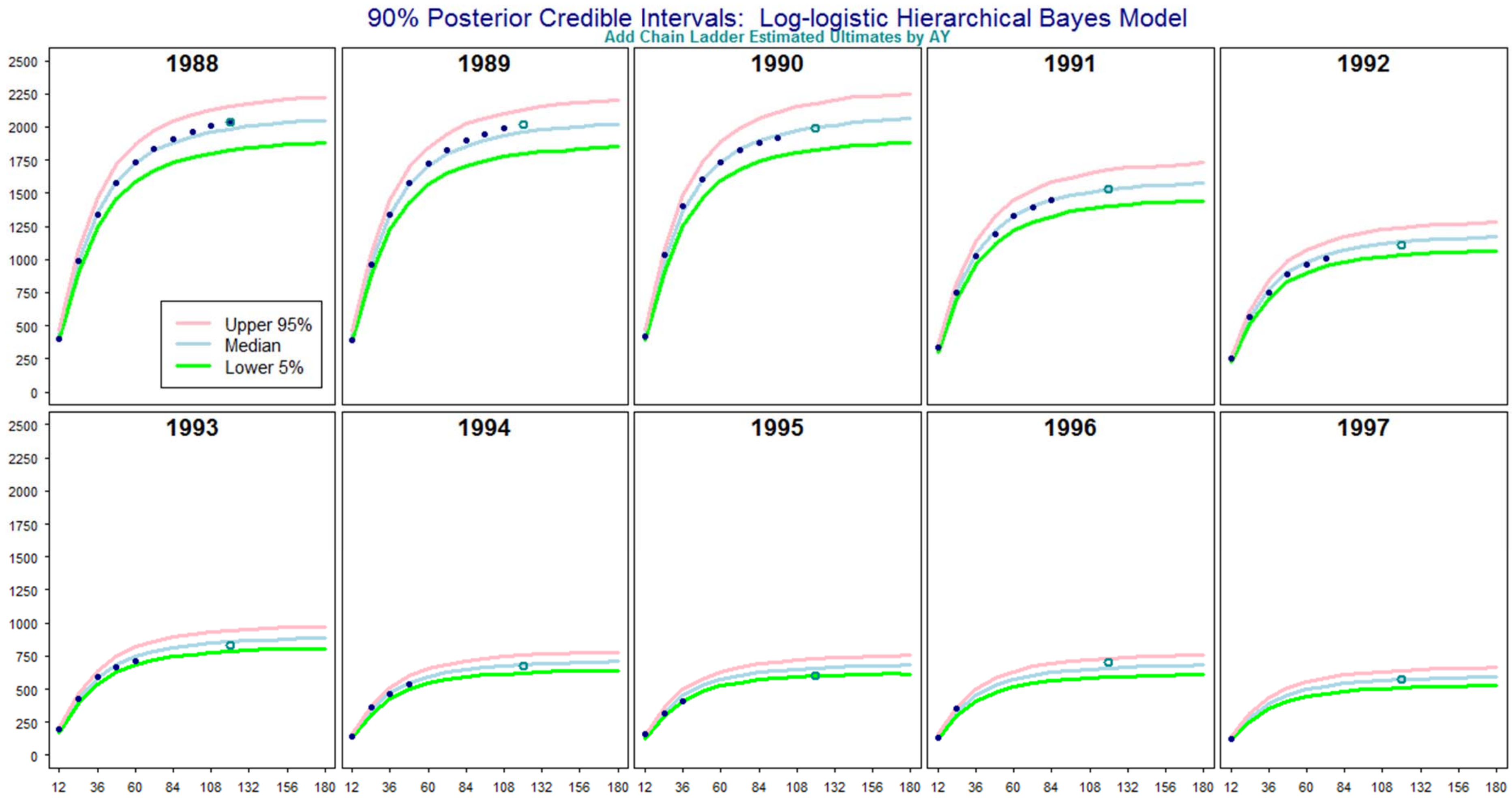
Bayesian Credible Intervals

- Now refit the model on all of the data and re-calculate the posterior credible intervals.



Comparison with the Chain Ladder

- For comparison, superimpose the "at 120 months" chain ladder estimates on the posterior credible intervals.



Posterior Distribution of Aggregate Outstanding Losses

- In the top two images, we sum up the projected losses for all estimated AY's evaluated at 120 (180) months; then subtract losses to date (LTD).
  - For the 120 month estimate, the posterior median (1519) comes very close to the chain ladder estimate (1543)
- In the bottom image, we multiply the estimated ultimate loss ratio parameters by premium and subtract LTD.
- Deciding which of these options is most appropriate is akin to selecting a tail factor.