Deloitte.

Introduction to Hierarchical Modeling

CAS Predictive Modeling Seminar
San Diego
October, 2008

Jim Guszcza
Bill Stergiou
Deloitte Consulting

Topics

- Hierarchical Data Structures
- Hierarchical Modeling Theory
- Sample Hierarchical Model
- Hierarchical Models and Credibility Theory
- Final Example: Loss Reserving

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
  - Income by profession
  - Claim severity by injury type
  - Churn rate by agency
  - Multiple years of loss experience by policyholder.
  - …
- Often grouped data is modeled either by:
  - Pooling the data and introducing dummy variables to reflect the groups
  - Building separate models by group
- Hierarchical modeling offers a "third way".
  - Parameters reflecting group membership enter one's model through appropriately specified *probability sub-models*.

What's in a Name?

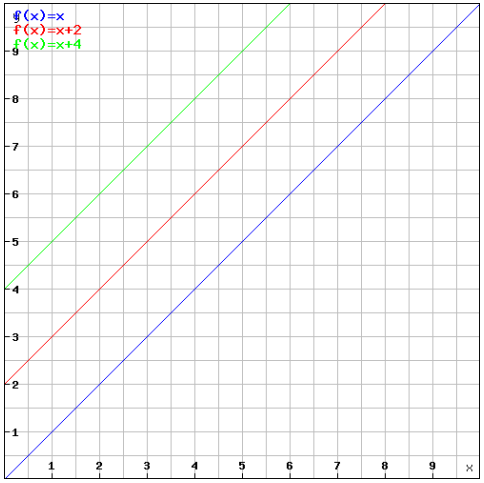
- Hierarchical models go by many different names
  - Mixed effects models
  - Random effects models
  - Multilevel models
  - Longitudinal models
  - Panel data models

- We prefer the "hierarchical model" terminology because it evokes the way models-within-models are used to reflect levels-within-levels of ones data.

- An important special case of hierarchical models involves multiple observations through time of each unit.
  - Here group membership is the repeated observations belonging to each individual.
  - Time is the covariate.

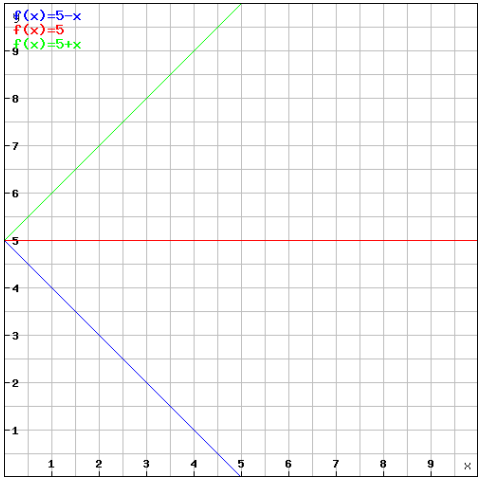
Varying Slopes and Intercepts

Random Intercept Model



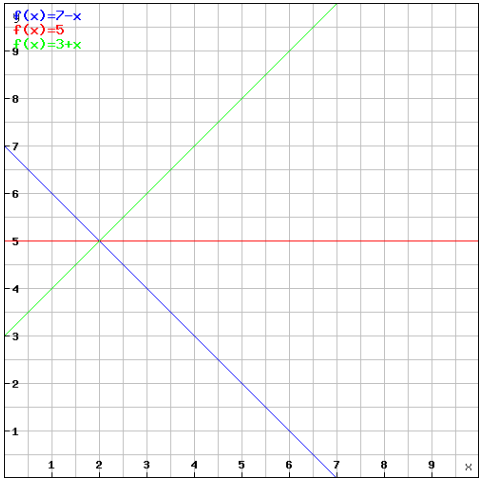
- Intercept varies with group
- Slope stays constant

Random Slope Model



- Intercept stays constant
- Slope varies by group

Random Intercept / Random Slope Model



- Intercept and slope vary by group

- Each line represents a different group

Common Hierarchical Models

- Notation:
  - Data points $(\boldsymbol{X}_i, Y_i)_{i=1\ldots N}$
  - $j[i]$: data point i belongs to group j.

- Classical Linear Model $\quad \boldsymbol{Y_i = \alpha + \beta X_i + \varepsilon_i}$
  - Equivalently: $Y_i \sim N(\alpha + \beta X_i, \sigma^2)$
  - Same α and β for every data point

- Random Intercept Model $\quad \boldsymbol{Y_i = \alpha_{j[i]} + \beta X_i + \varepsilon_i}$
  - Where $\alpha_j \sim N(\mu_\alpha, \sigma^2_\alpha)$ & $\varepsilon_i \sim N(0, \sigma^2)$
  - Same β for every data point; but α varies by group

- Random Intercept and Slope Model $\quad \boldsymbol{Y_i = \alpha_{j[i]} + \beta_{j[i]} X_i + \varepsilon_i}$
  - Where $(\alpha_j, \beta_j) \sim N(\mathrm{M}, \Sigma)$ & $\varepsilon_i \sim N(0, \sigma^2)$
  - Both α and β vary by group

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

Parameters and Hyperparameters

- We can rewrite the random intercept model this way:

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

- This model contains 9 parameters: $\{\alpha_1, \alpha_2, \ldots, \alpha_8, \beta\}$.
- And it contains 4 hyperparameters: $\{\mu_\alpha, \beta_2, \sigma, \sigma_\alpha\}$.

- Here is how the hyperparameters relate to the parameters:

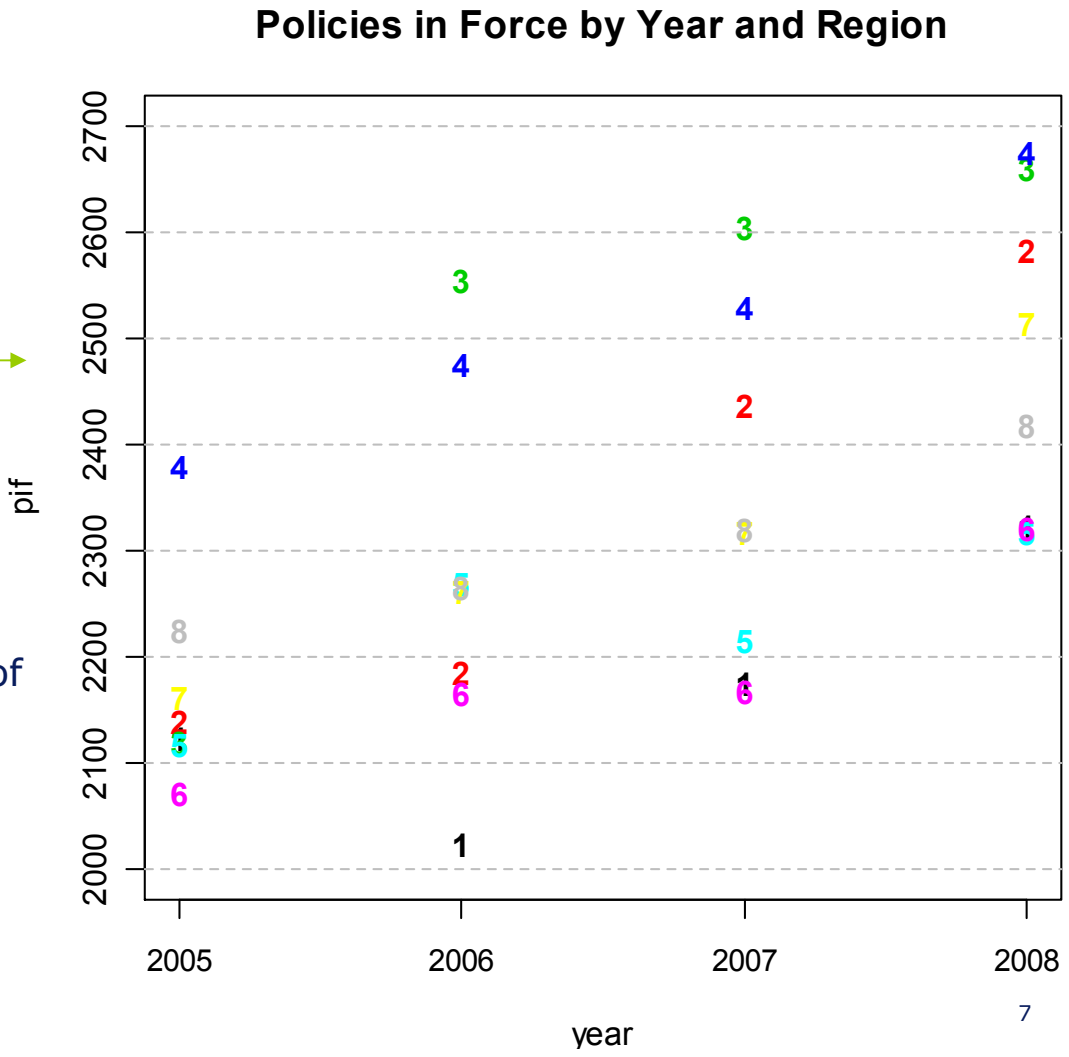
$$\hat{\alpha}_j = Z_j \cdot (\bar{y}_j - \beta \bar{x}_j) + (1 - Z_j) \cdot \hat{\mu}_\alpha \quad \textit{where} \quad Z_j = \frac{n_j}{n_j + \sigma^2 / \sigma_\alpha^2}$$

- Does this formula look familiar?

Example

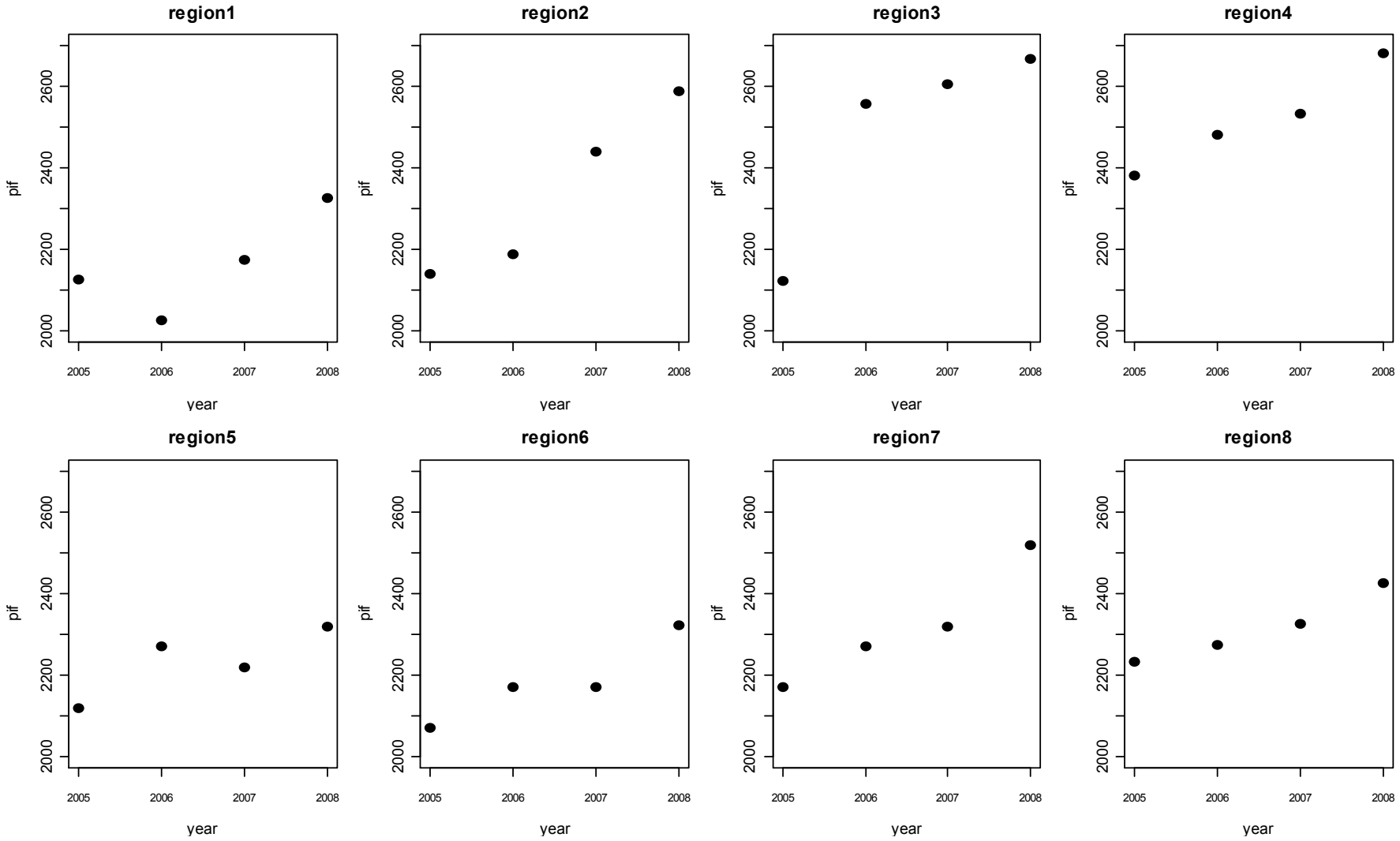
- Suppose we wish to model a company's policies in force, by region, for the years 2005-08.
- 8 * 4 = 32 data points.

- One way to visualize the data:
  - Plot all of the data points on the same graph, use different colors/symbols to represent region.
- Alternate way:
  - Use a trellis-style display, with one plot per region
  - More immediate representation of the data's hierarchical structure.
  - (see next slide)



Trellis-Style Data Display

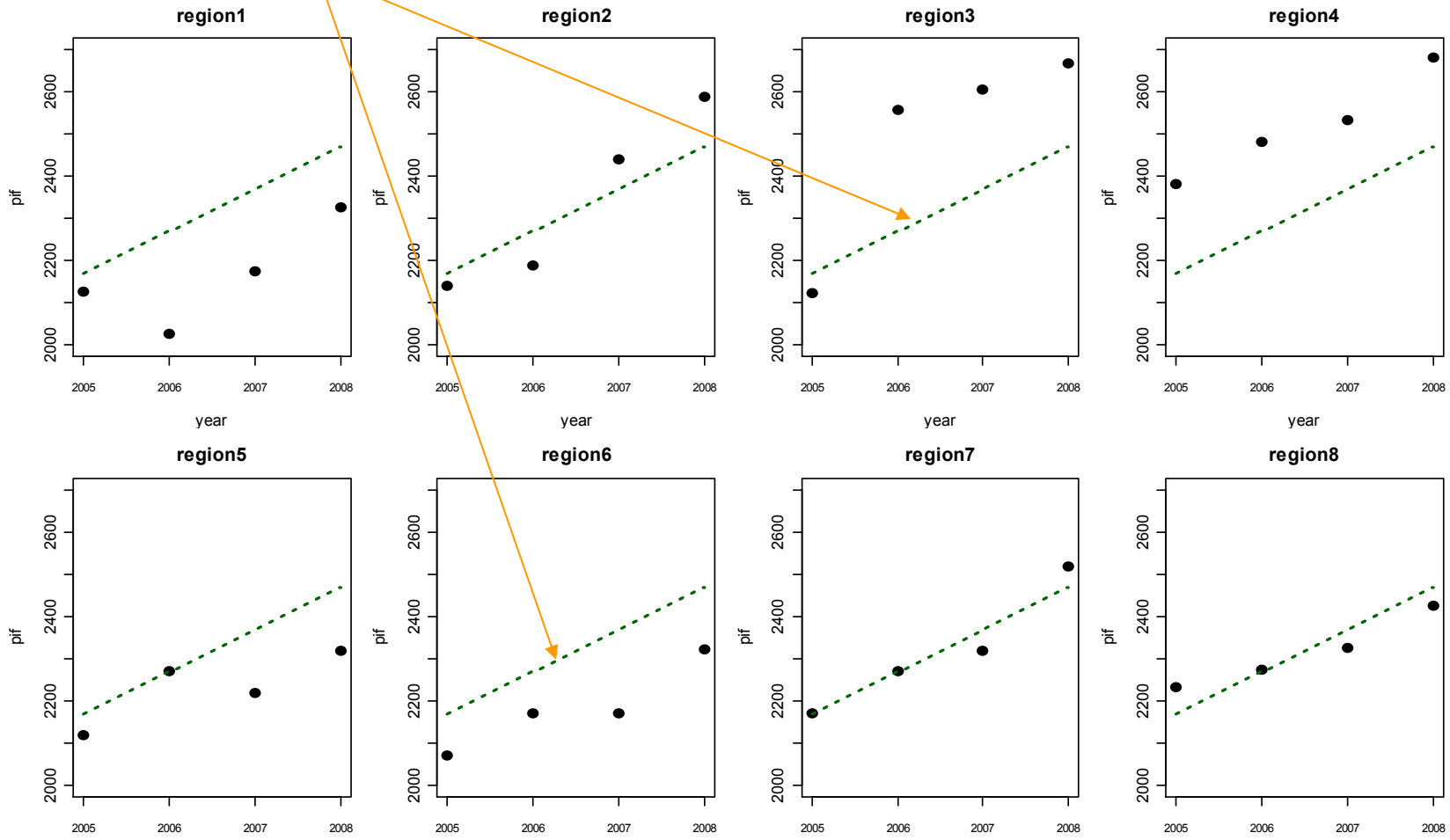
- We wish to build a model that captures the change in PIF over time.
- We must reflect the fact that PIF varies by region.



Option 1: Simple Regression

- The easiest thing to do is to pool the data across groups -- **i.e. simply ignore region**
- Fit a simple linear model
- Alas, this model is <u>not</u> appropriate for all regions

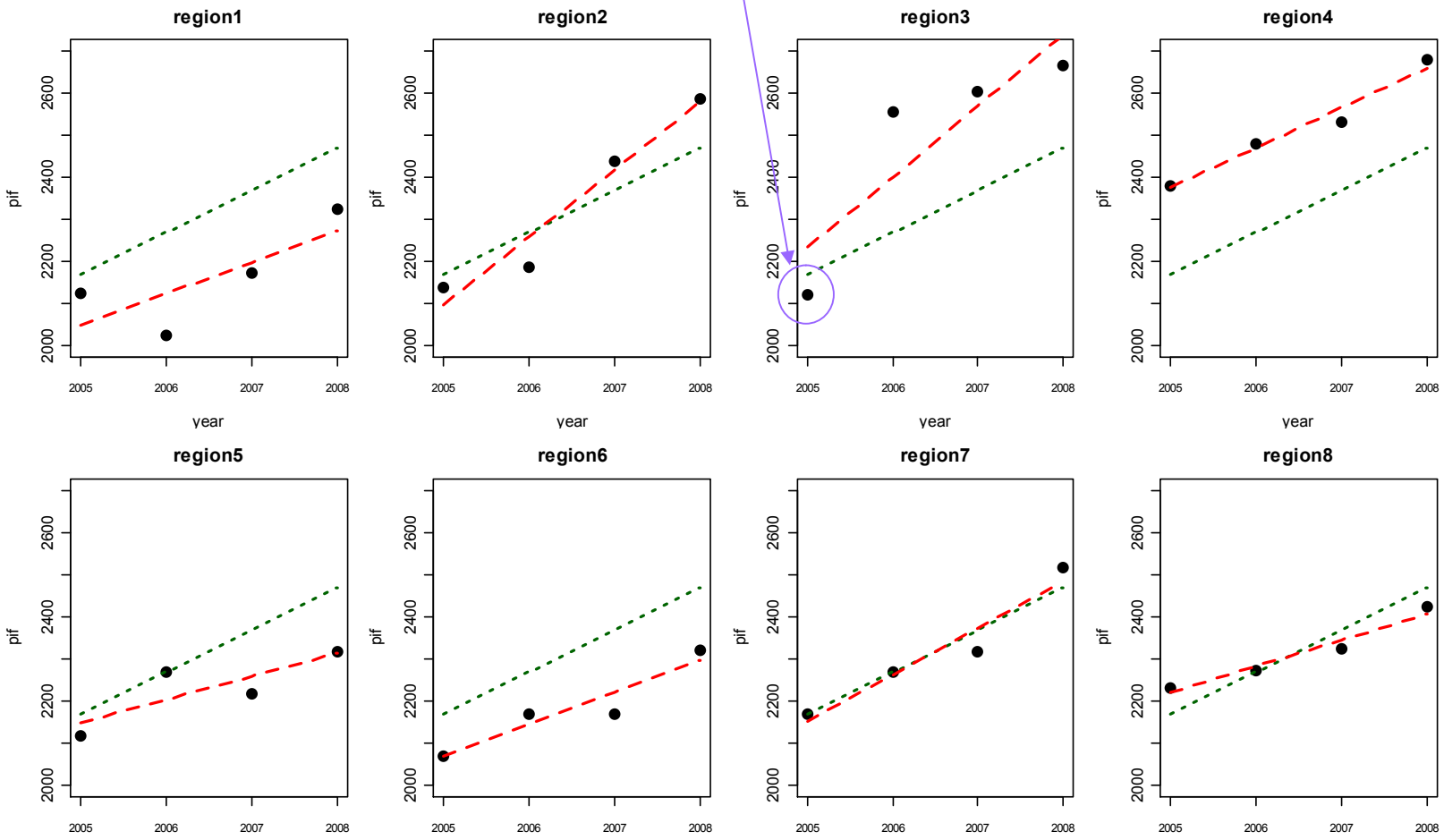
$$PIF = \alpha + \beta t + \varepsilon$$



Option 2: Separate Models by Region

- At the other extreme, we can fit a separate simple linear model for each region.
- Each model is fit with 4 data points.
- Introduces danger of over-fitting the data.

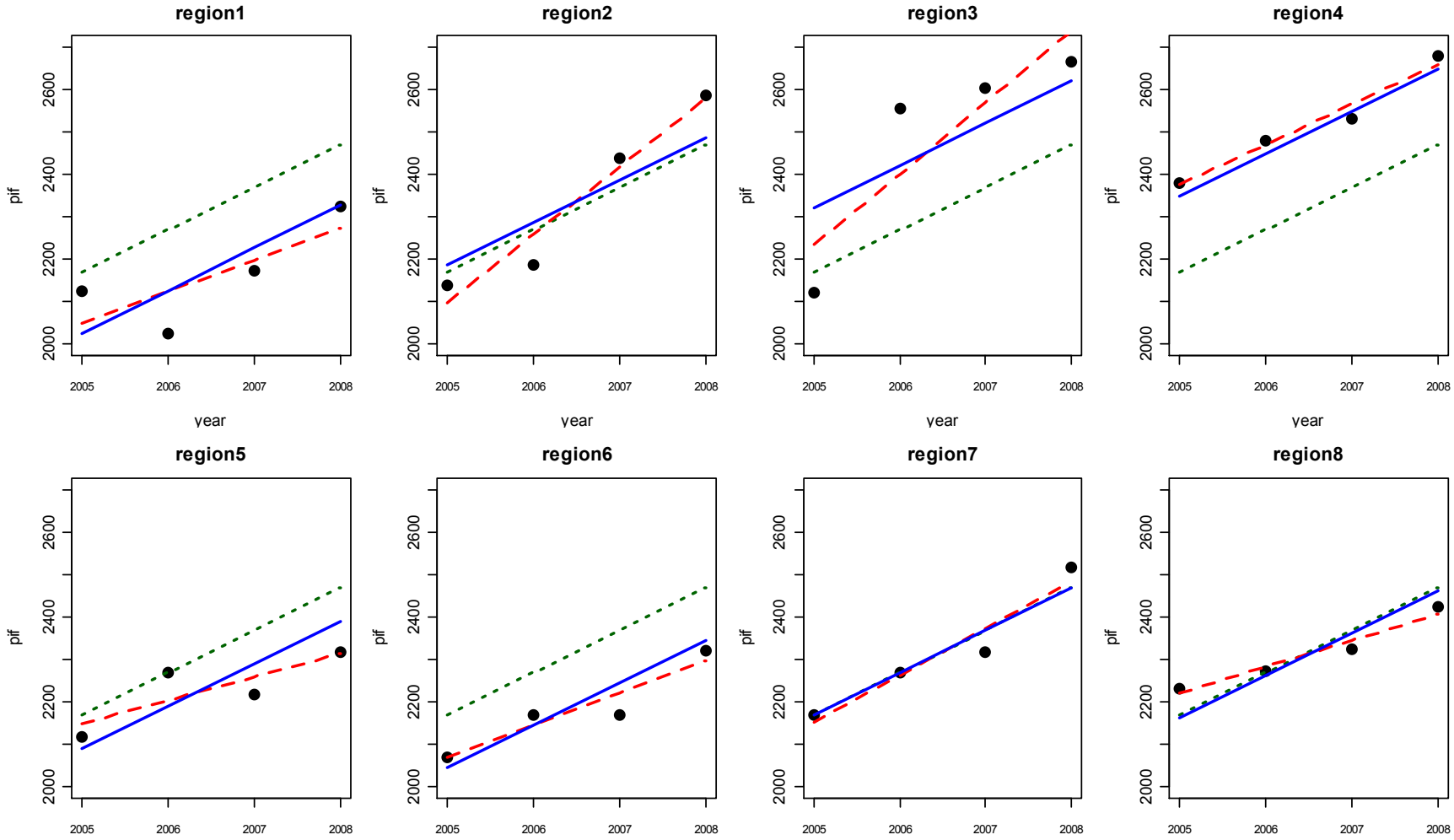
$$\left\{PIF = \alpha^k + \beta^k t + \varepsilon^k\right\}_{k=1,2,..,8}$$



Option 3: Random Intercept Hierarchical Model

- Compromise: Reflect the region group structure using a hierarchical model.

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



Compromise Between Complete Pooling & 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

- Estimating one model for each group

Compromise

Hierarchical Model

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

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

Option 1b: Adding Dummy Variables

- Question: of course it'd be crazy to fit a separate SLR for each region.
- But what about adding 8 region dummy variables into the SLR?

$$PIF = \gamma_1 R_1 + \gamma_2 R_2 + \ldots + \gamma_8 R_8 + \beta t + \varepsilon$$

- If we do this, we need to estimate 9 parameters instead of 2.
- In contrast, the random intercept model contains 4 hyperparameters: μ_α, β, σ, σ_α
- Now suppose our example contained 800 regions. If we use dummy variables, our SLR potentially requires that we estimate 801 parameters.
- But the random intercept model will contain the same 4 hyperparameters.

Varying Slopes

- The random intercept model is a compromise between a "pooled" SLR and a separate SLR by region.

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

- But there is nothing sacred about the intercept term: **we can also allow the slopes to vary by region**.

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

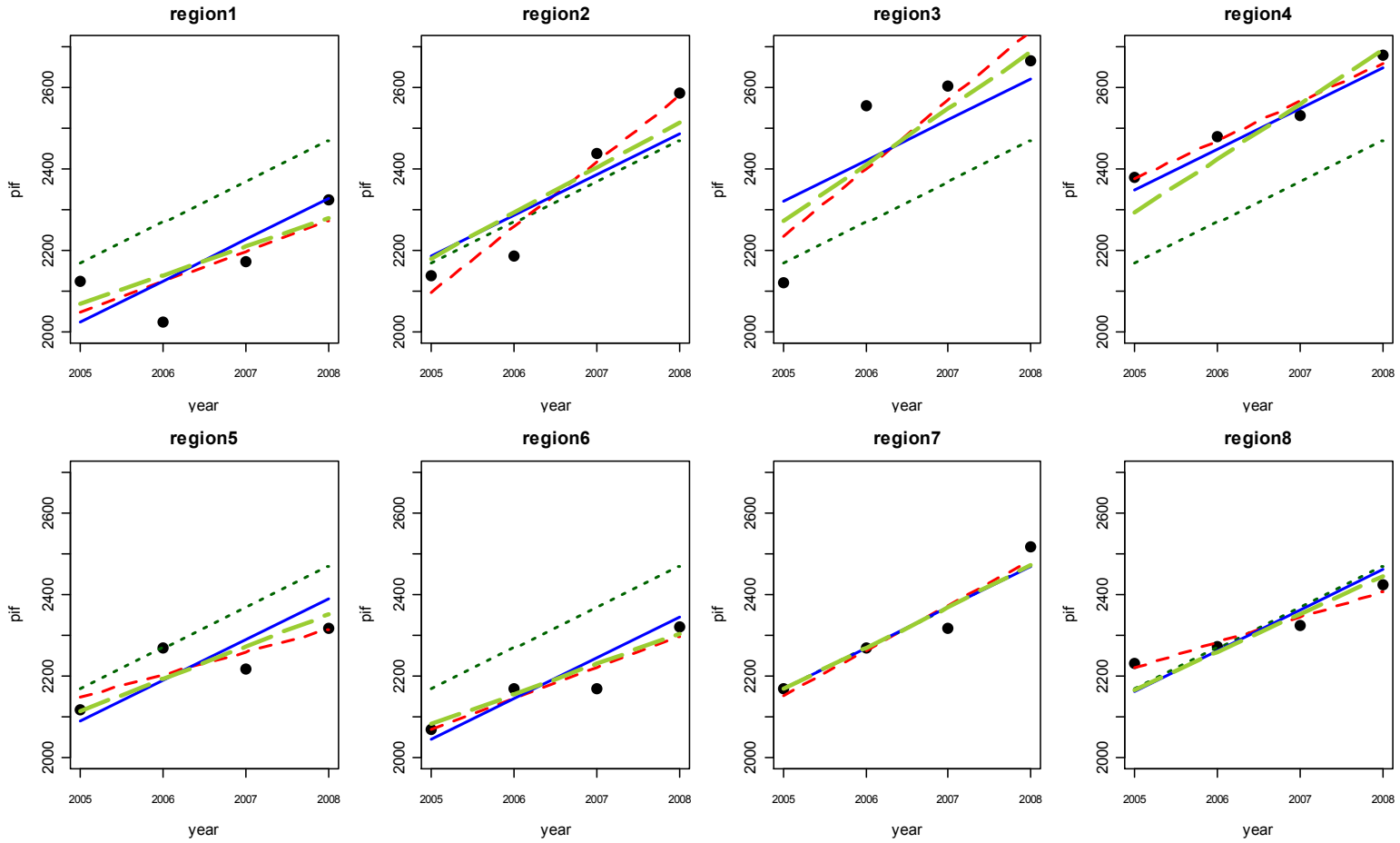
- In the dummy variable option (1b) this would require us to interact region with the time *t* variable… i.e. it would return us to option 2.
  - Great danger of overparameterization.

- Adding random slopes adds considerable flexibility at the cost of only two additional hyperparameters.
  - Random slope only: μ_α, β, σ, σ_α
  - Random slope & intercept: μ_α, μ_β, σ, σ_α, σ_β, $\sigma_{\alpha\beta}$

Option 4: Random Slope & Intercept Hierarchical Model

- We can similarly include a sub-model for the slope β.

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



Does Adding Random Slopes Improve the Model?

- How do we determine whether adding the random slope term improves the model?

1. Graphical analysis and judgment:
    - the random slopes arguably yield an improved fit for Region 5.
    - but it looks like the random slope model might be overfitting Region 3.
    - Other regions a wash

2. Out of sample lift analysis.

3. Akaike information Criterion [AIC]: -2*LL + 2*d.f.
    - Random intercept AIC: 380.40
    - Random intercept & slope AIC: 380.64
    - Slight deterioration ➔ better to select the random intercept model.

- Random slopes don't help in this example, but it is a very powerful form of variable interaction to consider in one's modeling projects.

Parameter Comparison

- It is important to distinguish between each model's *parameters* and *hyperparameters*.

	α, β		$\mu_\alpha, \beta, \sigma, \sigma_\alpha$		$\mu_\alpha, \mu_\beta, \sigma, \sigma_\alpha, \sigma_\alpha, \sigma_{\alpha\beta}$	
	SLR		random intercept		random intercept & slope	
region	intercept	slope	intercept	slope	intercept	slope
1	2068.0	100.1	1911.3	100.1	1999.3	70.3
2	2068.0	100.1	2087.8	100.1	2070.2	111.2
3	2068.0	100.1	2236.1	100.1	2137.0	137.4
4	2068.0	100.1	2267.3	100.1	2159.6	133.2
5	2068.0	100.1	1980.3	100.1	2033.1	79.3
6	2068.0	100.1	1932.3	100.1	2008.9	73.8
7	2068.0	100.1	2066.8	100.1	2066.3	101.2
8	2068.0	100.1	2061.8	100.1	2069.5	94.1

- SLR: 2 parameters and 2 hyperparameters
- Random intercept: 11 parameters and 4 hyperparameters
- Random intercept & slope: 20 parameters and 6 hyperparameters

- **How do the hyperparameters relate to the parameters?**

Hierarchical Models and Credibility Theory

- Let's revisit the random intercept model.

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

- This is how we calculate the random intercepts $\{\alpha_1, \alpha_2, \ldots, \alpha_8\}$:

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

- Therefore: each random intercept is a **credibility-weighted average** between:
  - The intercept for the pooled model (option 1)
  - The intercept for the region-specific model (option 2)

Hierarchical Models and Credibility Theory

- This makes precise the sense in which the random intercept model is a compromise between the pooled-data model (option 1) and the separate models for each region (option 2).

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

- As $\sigma_\alpha \rightarrow 0$, the random intercept model → option 1
- As $\sigma_\alpha \rightarrow \infty$, the random intercept model → option 2

- Aside: what happens to the above formula if we remove the covariate *t* from our random intercept model?

Bühlmann's Credibility and Random Intercepts

- If we remove the time covariate (t) from the random intercepts model, we are left with a very familiar formula:

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

- **Therefore: Bühlmann's credibility model is a specific instance of hierarchical models.**

- The theory of hierarchical models gives one a practical way to integrate credibility theory into one's GLM modeling activities.

Sample Applications

- Territorial ratemaking or including territory in a GLM analysis.
    - The large number of territories typically presents a problem.
- Vehicle symbol analysis
- WC or Bop business class analysis
- Repeated observations by policyholder
- Experience rating
- Loss reserving
    - Short introduction to follow

Summing Up

- Hierarchical models are applicable when one's data comes grouped in one or more important ways.
- A group with a large number of levels might be regarded as a "massively categorical value"…
  - Building separate models by level or including one dummy variable per level is often impractical or unwise from a credibility point of view.
- Hierarchical models offer a compromise between complete pooling and separate models per level.
- This compromise captures the essential idea of credibility theory.
- **Therefore hierarchical model enable a practical unification of two pillars of actuarial modeling:**
  - **Generalized Linear Models**
  - **Credibility theory**

Other thoughts

- The "credibility weighting" reflected in the calculation of the random effects represents a "shrinkage" of group-level parameters (α_j, β_j) to their means (μ_α, μ_β).
- The lower the "between variance" (σ_α^2) the greater amount of "shrinkage" or "pooling" there is.
- There is more shrinkage for groups with fewer observations (n).
- Panel data analysis is a type of hierarchical modeling ➔ this is a natural framework for analyzing longitudinal datasets.
    - Multiple observations of the same policyholder
    - Loss reserving: loss development is multiple observations of the same AY claims

Parting Shot: Hierarchical Modeling for Loss Reserving

- A garden variety loss triangle (Dave Clark CAS *Forum* 2003):

Cumulative Losses in 1000's													
AY	*12*	*24*	*36*	*48*	*60*	*72*	*84*	*96*	*108*	*120*	*reported*	*est ult*	*reserve*
1991	358	1,125	1,735	2,183	2,746	3,320	3,466	3,606	3,834	3,901	3,901	3,901	0
1992	352	1,236	2,170	3,353	3,799	4,120	4,648	4,914	5,339		5,339	5,434	95
1993	291	1,292	2,219	3,235	3,986	4,133	4,629	4,909			4,909	5,379	470
1994	311	1,419	2,195	3,757	4,030	4,382	4,588				4,588	5,298	710
1995	443	1,136	2,128	2,898	3,403	3,873					3,873	4,858	985
1996	396	1,333	2,181	2,986	3,692						3,692	5,111	1,419
1997	441	1,288	2,420	3,483							3,483	5,672	2,189
1998	359	1,421	2,864								2,864	6,787	3,922
1999	377	1,363									1,363	5,644	4,281
2000	344										344	4,971	4,627
chain link	3.491	1.747	1.455	1.176	1.104	1.086	1.054	1.077	1.018	1.000	34,358	53,055	18,697
chain ldf	14.451	4.140	2.369	1.628	1.384	1.254	1.155	1.096	1.018	1.000			
growth curve	6.9%	24.2%	42.2%	61.4%	72.2%	79.7%	86.6%	91.3%	98.3%	100.0%			

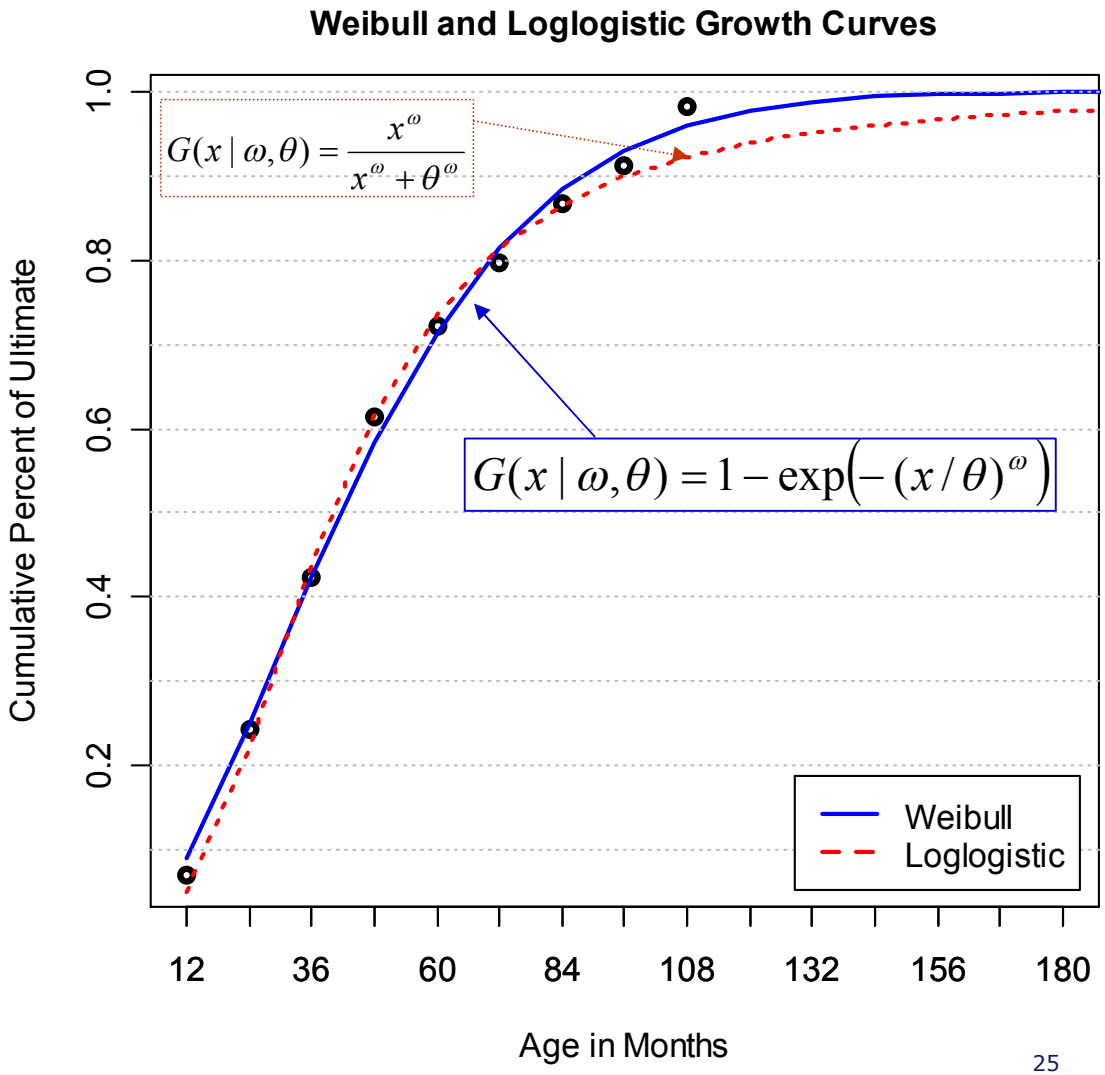
- We can regard this as a longitudinal dataset.

- 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

- Let's build a **non-linear** model of the loss triangle.
  - Are GLMs natural models for loss triangles?
- Uses growth curve to model the loss development process
  - 2-parameter curves
  - θ = scale
  - ω = shape
- Roughly speaking, we fit these curves to the LDFs and add random effects to θ and/or ω to allow the curves to vary by year.



Hierarchical Growth Curve Model

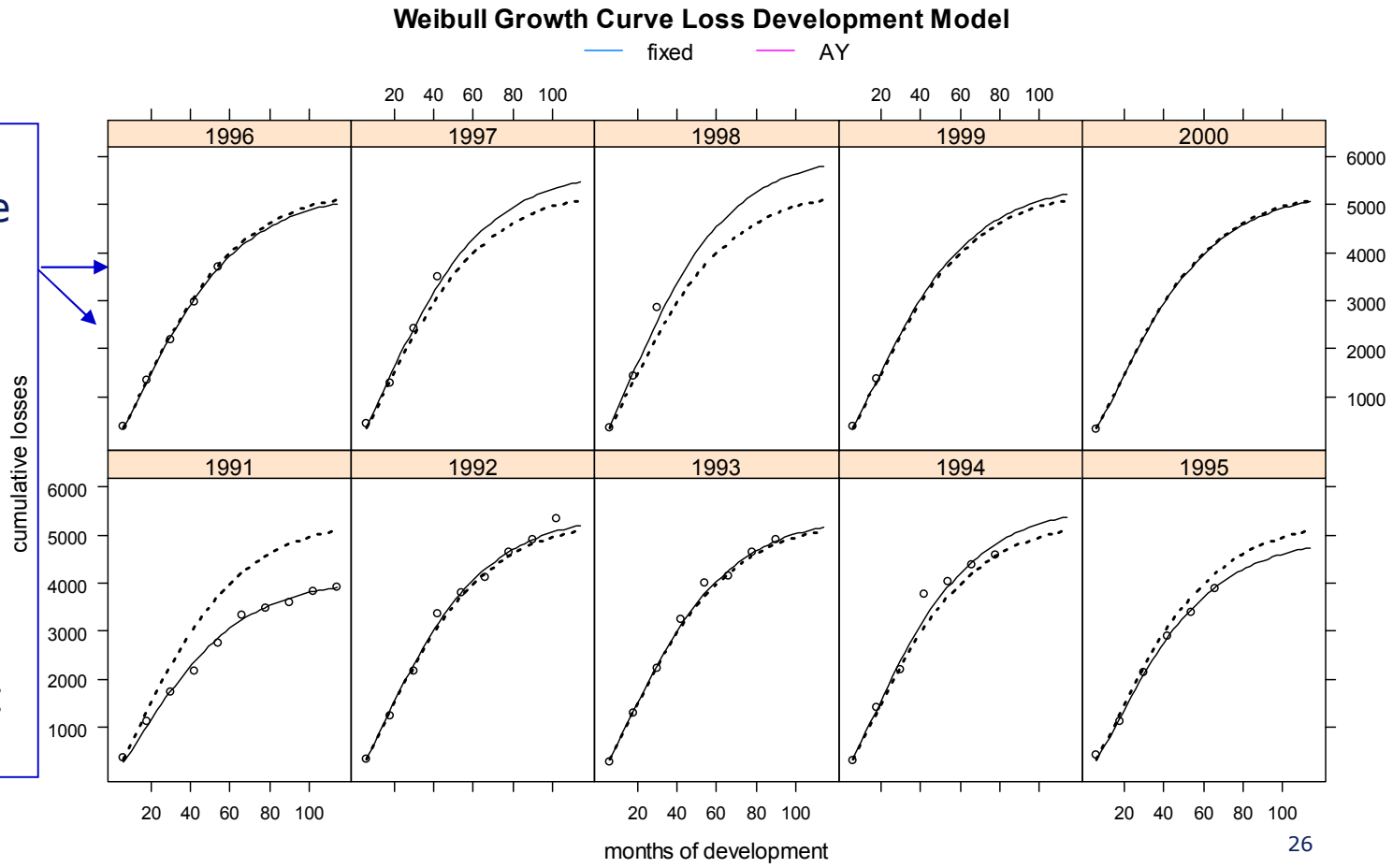
Cumulative losses @ *dev* =
(**Ult losses) * (modeled growth)**

We must estimate the parameters:

$\{\mu_{ULT}; \omega; \theta; \sigma_{ULT}; \sigma\}$

$$CumLoss_{AY,dev} = ULT_{AY}\left[1-\exp\left(-(dev/\theta)^{\omega}\right)\right]+\varepsilon_{AY,dev}$$
$$ULT_{AY} \sim N\left(\mu_{ULT}, \sigma^2_{ULT}\right)$$
$$Var(\varepsilon_{AY,dev}) = \sigma^2 \hat{CL}_{AY,dev}$$

- Random effects added to ultimate loss (*ULT*) parameter.
  - Analogous to random intercepts
- Random shape (ω), scale (θ) effects were tested, found not to be significant.



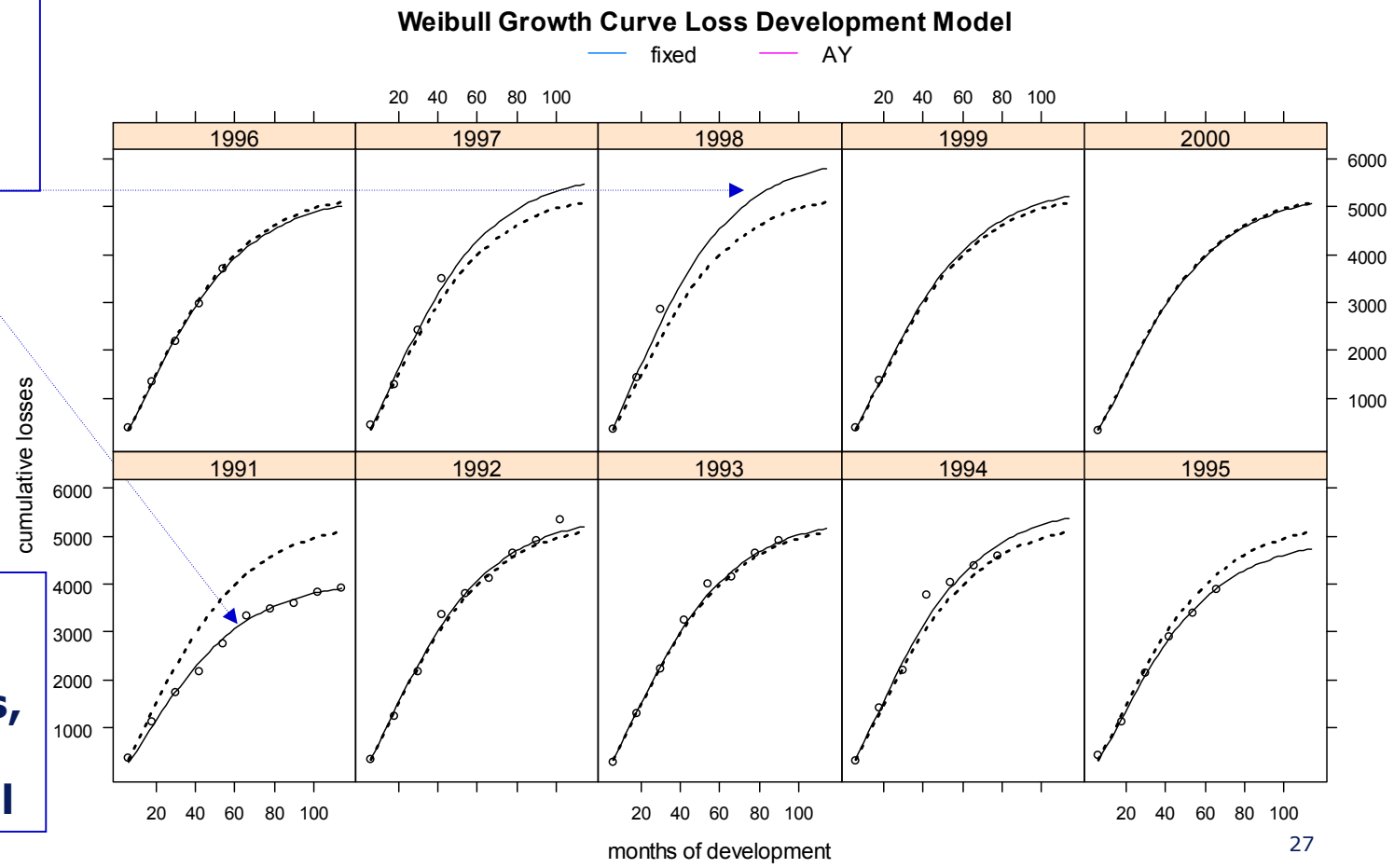
Hierarchical Growth Curve Model

$$CumLoss_{AY,dev} = ULT_{AY}\left[1-\exp\left(-(dev/\theta)^{\omega}\right)\right]+\varepsilon_{AY,dev}$$

$$ULT_{AY} \sim N\left(\mu_{ULT},\sigma^2_{ULT}\right)$$

$$Var(\varepsilon_{AY,dev}) = \sigma^2 \hat{CL}_{AY,dev}$$

The random effects allow a "custom fit" growth curve for each AY while maintaining parsimony.



The model contains only 5 hyperparameters, but fits the loss triangle very well

Some References

Frees, Edward (2006). Longitudinal and Panel Data Analysis and Applications in the Social Sciences. New York: Cambridge University Press.

Gelman, Andrew and Hill, Jennifer (2007). Data Analysis Using Regression and Multilevel / Hierarchical Models. New York: Cambridge University Press.

Guszcza, J. C. (2008). "Hierarchical Growth Curve Models for Loss Reserving". CAS *Forum*.

Pinheiro, J. C. and D. M. Bates. Mixed-Effects Models in S and S-Plus. New York: Springer-Verlag.