Semiparametric regression with hierarchical models

Yanwei (Wayne) Zhang

Statistical Research
CNA Insurance Company
New Orleans

March 17, 2011

Antitrust Notice

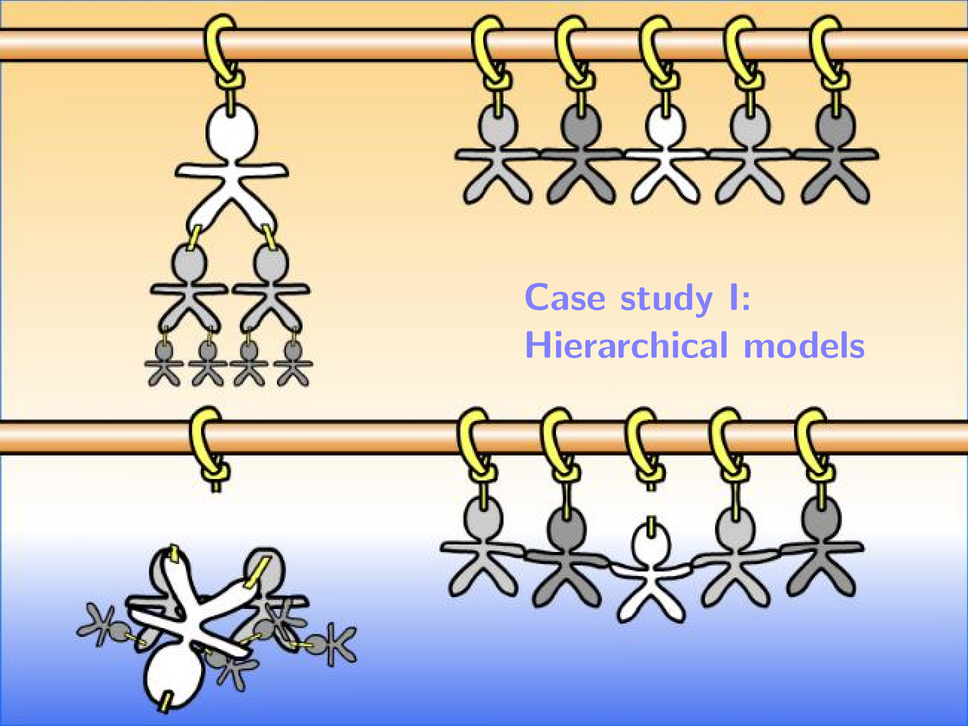
- The Casualty Actuarial Society is committed to adhering strictly to the letter and spirit of the antitrust laws. Seminars conducted under the auspices of the CAS are designed solely to provide a forum for the expression of various points of view on topics described in the programs or agendas for such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

Outline

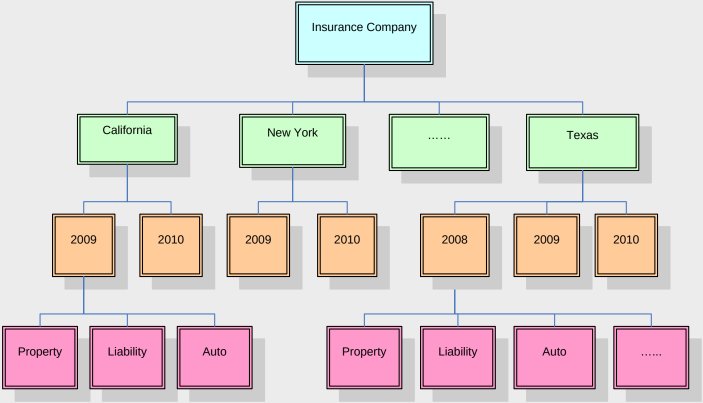
- Case study I: Review basic concepts and theories in hierarchical models
- Case study II: Build connection between penalized splines and hierarchical models
- Case study III: Geo-spatial smoothing with bivariate penalized splines



Case study I: Hierarchical models

Hierarchies in insurance data

- Insurance data often come with an inherent hierarchy (classification)
- Homogeneity VS stability?



Hierarchical models

Three methods to deal with data with inherent hierarchies:

- Complete pooling, assuming all groups are exactly the same
- No pooling, assuming complete heterogeneity
- Partial pooling (hierarchical), a compromise between the two extremes

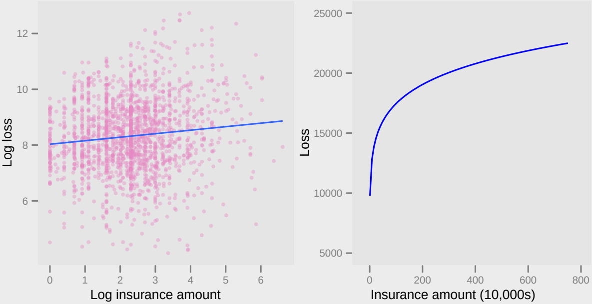
Advantages using hierarchical models:

- Using all data to make robust inference (group with small sample size)
- Inference of group level variation
- Inclusion of group level predictors
- Prediction for new group is available, and accounts for group variation

Example: contents loss due to theft/burglary

Suppose we have reported loss data (severity) for contents coverage due to theft/burglary in California:

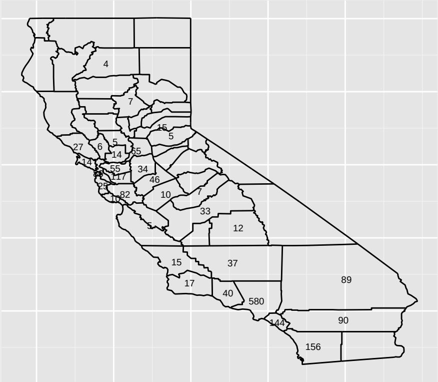
- Y: reported loss for each claim, X: contents insurance amount
- Can build a simple model $\log E(Y_i) = \alpha + \beta \log X_i$ for severity
- Exponentiating it will lead to $E(Y_i) = \exp(\alpha) X_i^{\beta}$
- $\exp(\alpha)$ will be the rate per insurance amount (or per 1,000,...)
- β determines the curvature of the curve



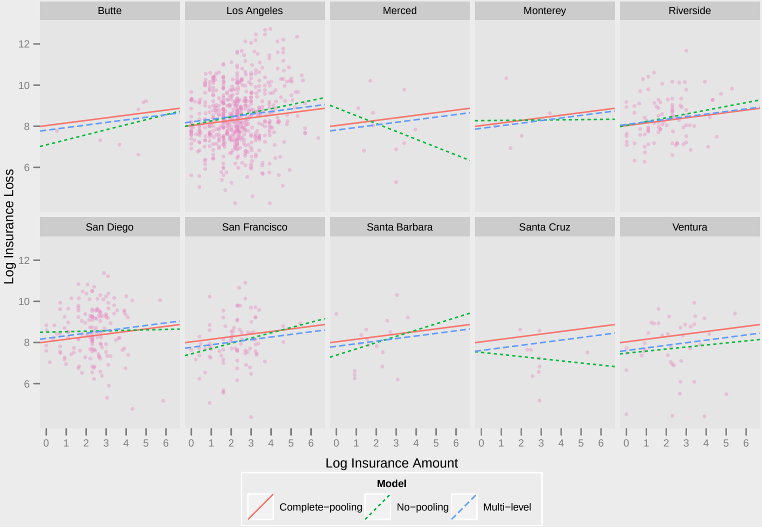
Data and model

How would you determine the rate $\exp(\alpha)$ for each county?

- Run one big regression using all data? Does not fit well, and can not get rate for each county!
- Run separate regression for each county? Estimate is so volatile for small county, and even get slope reversal!
- Hierarchical model[random intercepts]:
  $\log E(Y_i) = \alpha_{j[i]} + \beta \log X_i$



Visualization of the three models



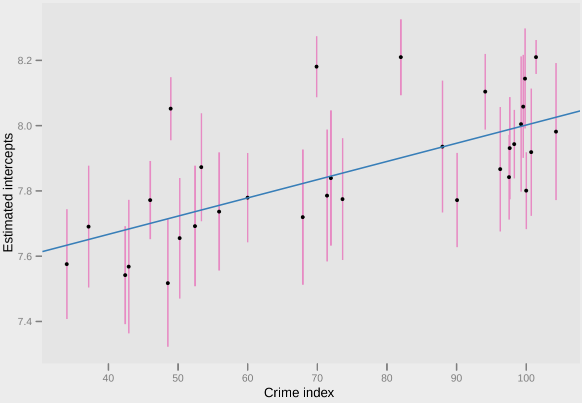
Adding group level predictor

Improve the model by adding county-level predictors (Z)- the crime index:

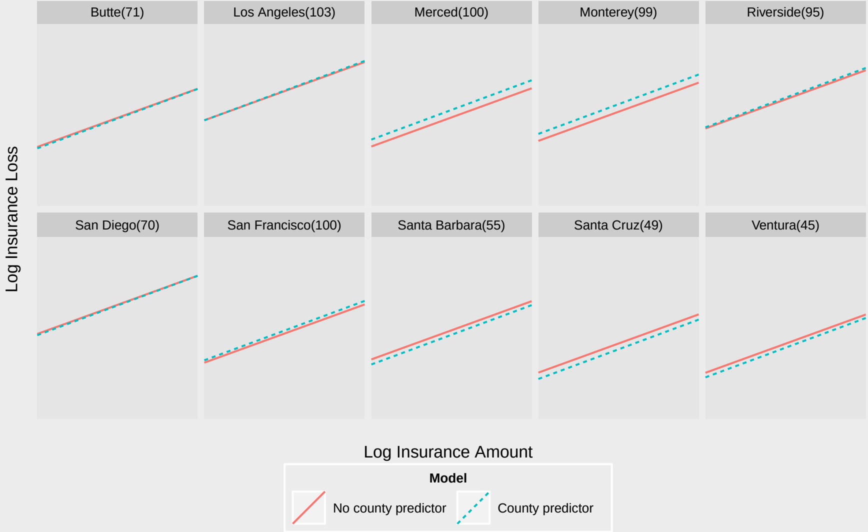
- $\log E(Y_i) = \alpha_{j[i]} + \beta \log X_i$ and $\alpha_j = a + bZ_j$
- Reduce group-level variation
- Make groups conditionally exchangeable

```
Models:
M3: logloss ~ 1 + logamt + (1 | county)
M4: logloss ~ 1 + logamt + crime + (1 | county)
   Df    AIC    BIC  logLik  Chisq Chi Df Pr(>Chisq)
M3  4 6228.9 6251.0 -3110.5
M4  5 6226.1 6253.7 -3108.1 4.8153      1    0.02821 *
```

Visualizing group-level regression

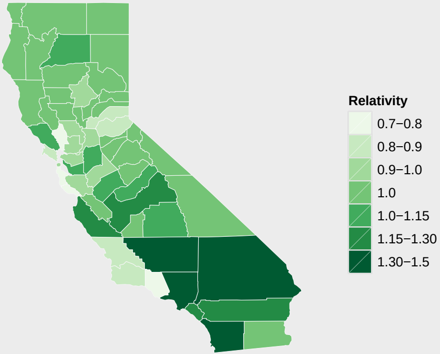


Comparison to hierarchical model with no group-level predictors



Rate map

- Can produce rate relativity (average county loss / average state loss) for a fixed insurance amount, assuming the modeled frequency is flat
- If a county is not available in the data, it is automatically set to be state average
- The right is a rate map at 10,000 insurance amount



Inference on linear models

Suppose

$$\mathbf{y}|\mathbf{u} \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \mathbf{R}) \quad (1)$$

$$\mathbf{u} \sim N(0, \mathbf{G}) \quad (2)$$

Maximum likelihood estimation leads to minimizing the following:

$$\boxed{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})^T\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}) + \mathbf{u}^T\mathbf{G}^{-1}\mathbf{u}} \quad (3)$$

This yields the GLS estimator

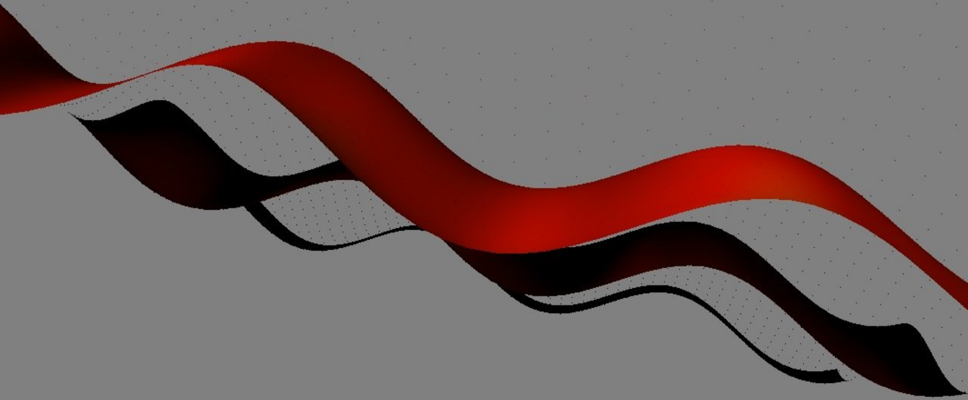
$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y},\ \mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R}, \quad (4)$$

and the best linear unbiased predictor

$$\hat{\mathbf{u}} = \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \quad (5)$$

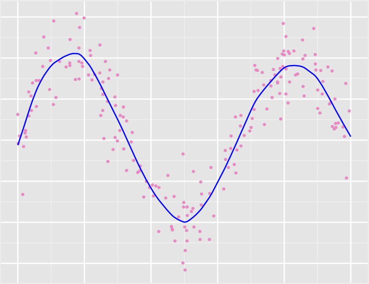
Using these to maximize the profile likelihood to get estimate for **V** and **R** and plug back into (4) and (5).

Case study II: Semiparametric models



Motivation

- Flexible modeling of nonlinear pattern
  - Hard to find a parametric nonlinear model
  - Even found, hard to estimate
- Rely on basis functions (e.g., two knots κ_1, κ_2)
  - Linear: $1, x, (x - \kappa_1)_+, (x - \kappa_2)_+$
  - Quadratic: $1, x, x^2, (x - \kappa_1)_+^2, (x - \kappa_2)_+^2$
  - Cubic: $1, x, x^2, x^3, (x - \kappa_1)_+^3, (x - \kappa_2)_+^3$



Penalized splines

With the basis functions, the model can be written as

$$Ey_i = \beta_0 + \beta_1 x_i + \sum_{k=1}^{K} u_k (x_i - \kappa_k)_+ \tag{6}$$

Or, using matrix notation,

$$E\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} \tag{7}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1)'$, $\mathbf{u} = (u_0, \cdots, u_K)'$, $\mathbf{X}_i = (1, x_i)$ and
$\mathbf{Z}_i = [(x_i - \kappa_1)_+, \cdots, (x_i - \kappa_K)_+]$.
Impose the constraints $\mathbf{u}^T\mathbf{u} < C$ to avoid wiggly fit. Using Lagrange multiplier, this is equivalent to minimize

$$\boxed{(\mathbf{Y} - \mathbf{X}\beta - \mathbf{Z}\mathbf{u})^T \frac{1}{\sigma_y^2} (\mathbf{Y} - \mathbf{X}\beta - \mathbf{Z}\mathbf{u}) + \mathbf{u}^T \frac{\lambda}{\sigma_y^2} \mathbf{u}} \tag{8}$$

This is the same as the hierarchical model in (3)!

Why P-splines

- The above shows that the P-splines can be estimated using hierarchical models, for which many softwares are available
  - R: lme4, nlme
  - SAS: PROC MIXED, %GLIMIX
  - WinBUGS for Bayesian analysis
- Compared to the Generalized Additive Model (GAM), which uses all knots but penalizes the second derivative, P-splines are much easier to fit
- Compared to other spline models such as B-splines, the number and the positioning of the knots in P-splines are not important given that the set of knots is relatively dense with respect to the x.
- Easy generalization to include parametric components to form semi-parametric models
- Easy generalization to other spline forms, such as $(x-\kappa_1)_+^p, |x-\kappa_1|^p$.

Claimant loss data

The following data is from Frees (2010). It includes automobile injury claims data from the Insurance Research Council (IRC), and contains information on age information about the claimant, attorney involvement and the economic loss (LOSS, in thousands), among other variables.

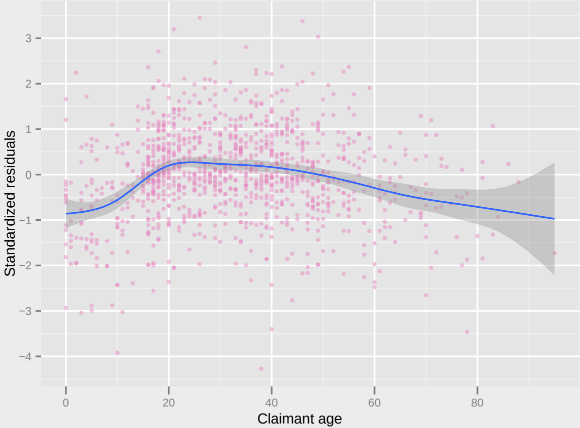
LOSS	ATTORNEY	SEATBELT	CLMAGE
34,940	1	1	50
10,892	0	1	28
330	0	1	5
⋮	⋮	⋮	⋮

Fit a regression model on log(LOSS):

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.2750	0.2884	25.23	0.0000
CLMAGE	0.0154	0.0022	7.12	0.0000
ATTORNEY:1	1.3667	0.0741	18.45	0.0000
SEATBELT:1	-0.9866	0.2787	-3.54	0.0004

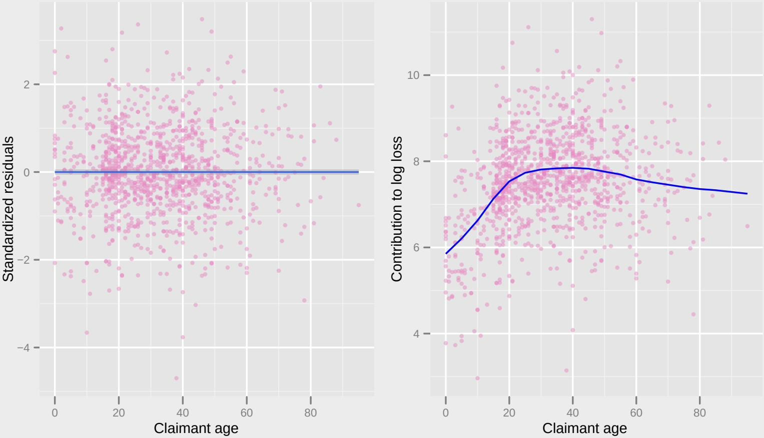
Diagnostics

The model makes sense, but what happens to the residuals?

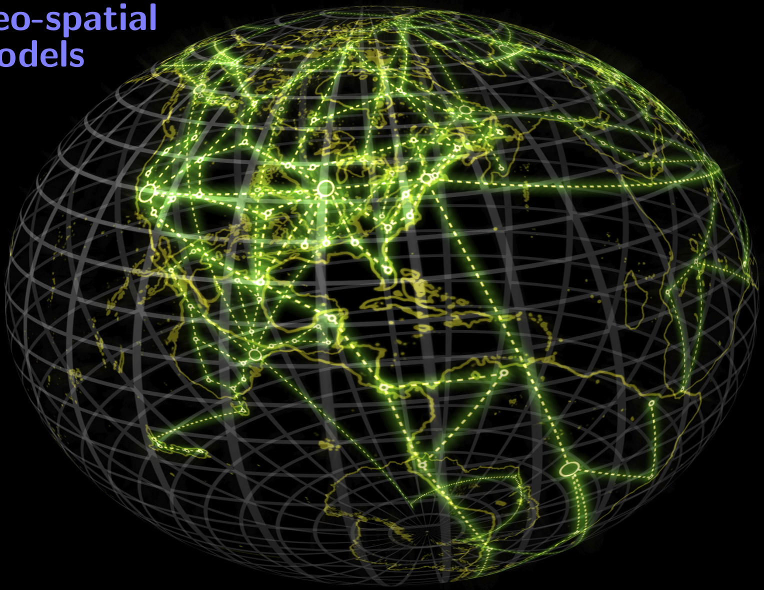


Model improvement

- Can model the curve with linear splines
- Knots at seq(5,85,by=5), and estimated using hierarchical models



Case study III: Geo-spatial models



Introduction

- The univariate P-spline model can be extended to the multivariate setting, f(longitude, latitude)
- This could explain spatial dependency and allow spatial interpolation
- Such an extension is more straightforward when the spline basis is Radial, $|x - \kappa_1| \rightarrow ||\mathbf{x} - \boldsymbol{\kappa}_1||$, since this distance is invariant to rotation of coordinate systems
- Selection of knots is harder - can resort to space filling algorithm
- The efficiency gain of using hierarchical models in computing is enormous

Example: California house value

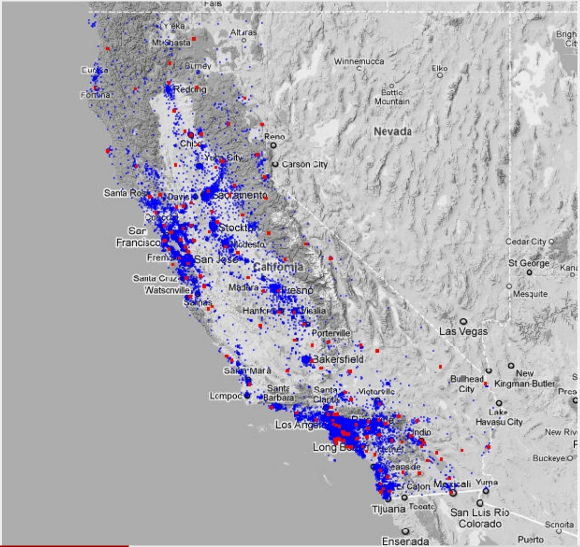
- From Pace and Barry (1997).
- Attempt to predict median house value using predictors such as median income, number of bedrooms, median house age and etc.

```
log(value) ~ income + I(income^2) + I(income^3) +
       log(house.age) + log(rooms) + log(bedrooms) +
       log(population/households) + log(households)
Multiple R-squared: 0.6078
```

- Pace and Barry (1997) used a Spatial Autoregressive (SAR) model where the R^2 is improved to 0.8594.
- Here, we model the spatial dependency through a spline term f(longitude, latitude)

Data and knots

- Run space filling algorithm to select knots (red)

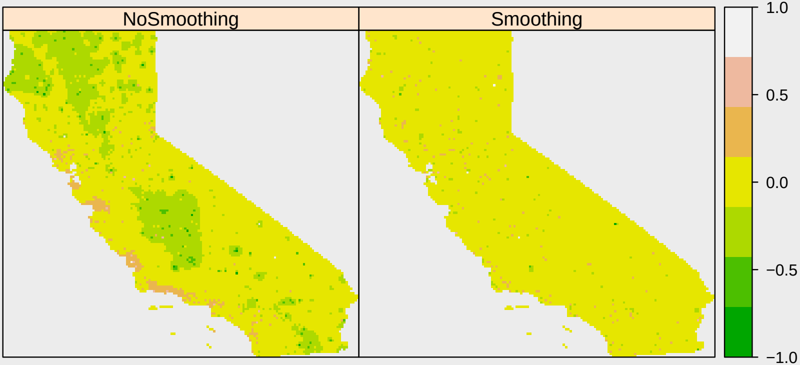


Results

- The bivariate spline models results in better R^2 than OLS

```
lme(log(value)~-1+X,random=pdIdent(~-1+Z))
Multiple R-squared: 0.8099
```

- Also resolve the spatial dependency and allow surface estimation.



Summary

- Hierarchical model incorporates actuarial credibility, a compromise between two extremes- complete pooling and no pooling
- This existing software can be applied to the inference of penalized splines, where nonparametric non-linear pattern in the underlying insurance data can be readily modeled
- Multivariate extension of the penalized splines can be further applied to model spatial dependencies and perform geo-spatial interpolations