*International Journal of Business and Management Invention*
*ISSN (Online): 2319 – 8028, ISSN (Print): 2319 – 801X*
*www.ijbmi.org Volume 2 Issue 6‖ June. 2013‖ PP.33-40*

# Bayesian Hierarchical Regression (Generalized Linear Model) Analysis of US State-Specific Non-Violent Crime[1]

Xiaolu Wang

*Graduate student, Economics Dept. Duke University, Durham, NC, USA*

***ABSTRACT:*** *I developed both an OLS and a Bayesian model to predict US state-specific non-violent crime rates. Based on Economic theories and the characteristics of my highly multi-collinear and endogenous data, I first tried Stochastic Search Variable Selection (SSVS) in variable selection. I found some problems, and then I used Principal Component Analysis (PCA) and Factor Analysis (FA) to help select variables and did a box-cox transformation to develop a generalized linear regression (assuming homogeneity). Then I developed a Bayesian hierarchical regression model (allowing heterogeneity), which fit the data much better. After checking the convergence and shrinkage, I also conducted model diagnostic tests to compare the difference in the two models. Further normal mixture, clustering, and comparison with frequentist hierarchical regression are tried without concrete results. These results about heterogeneity in different states can be used as an inference indicator for commercial crime rate insurance, which usually uses frequentist method in crime rate prediction and related actuarial pricing, instead of more scientific Bayesian analysis.*

***KEYWORDS:*** *Generalized Linear Model, Bayesian Hierarchical Regression, Crime Analysis, Heterogeneity*

## I. RESEARCH FOCUS AND DATA DESCRIPTION

We are interested in the following **questions**: (1) How to predict state-specific crime rates using social-economic and demographic data? (2) How to develop a generalized linear mixed effect model with strong support in economics and using fancy statistical method? (3) Whether there are obvious state-specific crime characteristics, and how to visualize the difference? (Heterogeneity in both regression coefficients and residual terms)In order to explore a good model to **answer** the questions above, I'm going to: (1) Conduct a simple data mining uses both Bayesian and frequentist methods: SSVS, PCA and FA, compare and make sure the variables are supported by economic theory; (2) Construct a generalized linear model; (3) using bayesian hierarchical method to handle state-specific characteristics: estimate/approximate interested parameters, check shrinkage and convergence, and check heteroskedasticity.

Table.1 summary of communities and crime unnormalized data[2]

| **Number counting** | | **Interested Dependent variables** | | | **Data cleaned** |
|---|---|---|---|---|---|
| ***instances*** | 2215 | | Observed | Missing | ***response:*** nonViolent Crimes / 101k population |
| ***attributes*** | 147 | Violent # | 1994 | 221 | ***check***: identiable GLM ***predictors:*** 101 (potential) |
| ***Missing*** | not M(C)AR | **nonViolent** | 2118 | 97 | ***instances:*** 1884 total; numbers vary across states |
| ***normlized*** | NO | Both | 1902 | 5 | ***characters:*** most multicollinear,some endogenous |

My **dataset** is a typical unnormalized demographic and socio-economic data, which is nasty yet amazing. It combines:

[1] Thanks Dr. Dunson for discussion, and thanks Prof. Joe Hotz for advice in my topic selection. Thanks Prof. Charlie Becker, Prof. Edward Tower, Ye Wang, David Klemish, Ying Guo and Fan Tong for their great advice and encouragement. All errors are my own.

[2] Raw data and description is available at the Machine Learning Department at UCI, online link is: http://archive.ics.uci.edu/ml/datasets/communities+and+crime+unnormalized.

demographic and socio-economic data from the '90 Census, law enforcement data from the 1990 Law Enforcement Management and Admin Stats Survey, Crime data from the 1995 FBI UCR. Number of observations varies a lot across different states. Some state has only one observation (Washington DC), some states have more than hundred observations: California (CA) has 278 observations. So there are strong potential *identification problems* for the preliminary development of a reliable Generalized Linear Model (GLM).And the difference in unit and scale need some transformation of the raw data in order to fit a GLM scientifically.

Strong correlations among many potential predictors (e.g., household income and per capita income, ρ=0.9), indicating strong multicollinearity. There are also many potential endogenous variables and unobserved effects, which can harm regression assumptions but cannot be detected by pure statistic method.

**Unnormalized data** is relatively more complicated to handle in doing project. But normalization will make interpretation less intuitive, and in my case, in order to keep rich economic flavor, I didn't normalize it. The ***response variable*** that I finally choose is the per capita nonviolent crime, which was calculated using the sum of crime variables considered non-violent crimes in the United States: burglaries, larcenies, and auto thefts and arsons. (There are many other types of crimes; these only include FBI 'Index Crimes')

## II. MODEL DEVELOPMENT AND APPROACH

### 2.1. Model Selection and Preliminary Linear Regression

I originally used a ***Stochastic Search Variable Selection*****(SSVS)** to do the variable selection, but since there are $2^{101}$ potential models with highly collinear variables, SSVS works badly, cannot reduce the dimension effectively, and include some bad endogenous variables. So I go back to use exploratory data analysis tools-***Principal Component Analysis*** **(PCA)**&***Factor Analysis*** **(FA)**to reduce the dimension by projecting variables to some components (make a rotation)and selected variables that both significant in economics and statistics. I further conducted a ***Box-Cox Transformation*** to modify the unnormalized data and fit the model. Then I checked the rank condition to ensure ***GLM identifiability*** for all the groups (states) in Hierarchical Regression afterwards. The final generalized linear model that I developed (based on economic theory and statistic method) is:[3]

$$\ln nonViolentCrimes = \beta_1 + \beta_2 \ln population + \beta_3 \ln perCapitaIncome + \beta_4 \ln PopulationDensity + \beta_5 DivorceRate + \beta_6 PectKidBorntoNeverMarried + \beta_7 PctLessthan9grade + \varepsilon$$

Now we have ***explanatory variables*** on: population, population density, per capita income, education, divorce rate, family effect. This makes tremendous sense in economics and public policy. So we got strong theoretical foundation now, and the derived predictive model would probably have robust-inference for policy decisions.

This generalized linear model can be simplified as:

$$Y = X\beta + \varepsilon, \quad where$$
$$X = [1, x_2, x_3, x_4, x_5, x_6, x_7] \quad and \quad \varepsilon | X \sim N(0, \sigma^2 I)$$

### 2.2. Bayesian Hierarchical Generalized Linear Regression Model

Data (Likelihood)：

$$y_{ij} \overset{i.i.d}{\sim} N(X_{ij}\beta_j, \sigma_j^2) \quad parameters \ \ of \ \ interest: \beta_j, \sigma_j^2 \ \ where \ \ j = 1, 2, ..., m = 38$$

**(1) Prior choice:** "borrow" information from the data (across different groups (states): here ***I have 38 states!!!***)

[3]The biggest econometric problem that I cannot totally solve now is unobserved effect. But based on all the statistical inferences, the unobserved effect is already controlled to a desirable level.

⇨ Bayesian regression with weak but unbiased prior information (OLS estimate).

**(2) Posterior Distribution Derivation:**

$$\beta_j \overset{i.i.d}{\sim} Mvt - Normal(\theta, \Sigma) \qquad \sigma_j^2 \overset{i.i.d}{\sim} Inverse - Ga(\frac{v_0}{2}, \frac{v_0\sigma_0^2}{2})$$

$$where \begin{cases} \theta \sim \ mvt\text{-}Normal(\hat{\theta}, \hat{\Sigma}) \\ \Sigma^{-1} \sim inv\text{-}Wishart(p+2, \hat{\Sigma}^{-1}) \end{cases} \qquad here\ \ set \begin{cases} v_0 = 1\ \ non - informative \\ \sigma_0^2 = 1\ \ s.t.\ \ \sigma_0^2 < Var(Y) \end{cases}$$

**(3)** Get **full conditional distribution** of $\theta, \Sigma, \beta_j, \sigma_j^2$

$$\theta \big| \Sigma, \beta_1, ..., \beta_m \sim mvt\text{-}Normal(\mu_m, \Lambda_m) \qquad \Sigma \big| \theta, \beta_1, ..., \beta_m \sim Inv - Wishart(4+m, [\hat{\Sigma} + S]^{-1})$$

$$where \begin{array}{l} \mu_m = \Lambda_m(\hat{\Sigma}^{-1}\hat{\theta} + m\Sigma^{-1}\bar{\beta}) \\ \Lambda_m = (\hat{\Sigma}^{-1} + m\Sigma^{-1})^{-1} \end{array} \qquad where\ \ S_\theta = \sum_{j=1}^{m} (\beta_j - \theta)(\beta_j - \theta)^T$$

$$\beta_j \big| \theta, \Sigma, 1/\sigma_j^2, X, Y \sim mvt\text{-}Normal(\mu_0, \Lambda_0) \qquad \sigma_j^2 \big| \theta, \Sigma, \beta_1, ..., \beta_m, v_0, \sigma_0^2, X, Y \sim Ga(a_j, b_j)$$

$$where \begin{array}{l} \mu_0 = \Lambda_0(\Sigma^{-1}\theta + \frac{X_j^T Y_j}{\sigma_j^2}) \\ \Lambda_0 = (\Sigma^{-1} + \frac{X_j^T X_j}{\sigma_j^2})^{-1} \end{array} \qquad where \begin{cases} a_j = \frac{(v_0 + n_j)}{2} \\ b_j = \frac{[v_0\sigma_0^2 + (Y_j - X_j\beta_j)^T (Y_j - X_j\beta_j)]}{2} \end{cases}$$

## III. CONCRETE APPROACH: MARKOV CHAIN MONTE CARLO

MCMC methods are implemented. More concretely, use Gibbs Sampler for all the parameters with conditional distribution derived above, I run 50000 iterations and only use the last 10000 iterations as effective "post burn-in" simulation samples. I did this in order to guarantee a good convergence and make the result more reliable. An interesting "extra take away" in developing this model is that: I originally try to update $v_0$ and $sigma_0^2$ too, using Metropolis-Hastings Algorithm (Metropolis random walk) for $v_0$and MC sample from inverse-Gamma for$sigma_0^2$. However, it's hard to pin down the range for v0 and guarantee $sigma_0^2$ is smaller than the total variance of the data (y). (The ACF and traceplots for $sigma_0^2$ are horrible!) After several trails and fails, I finally successfully make it work well by fixing both the values as non-informative hyper-priors.

### 3.1. Extensions: Mixture, Cluster, and Comparison with Frequentist Hierarchical[4]

There are many fantastic extensions that I tried, even though I didn't get satisfactory results. Since my main model is a hierarchical generalized linear regression model, I extended it as follow:

(1) **To capture outliers** by using a scale-mixture of two normal distributions for the error term. First, draw probability weights for the two components (two normal distributions) from a *Dirichlet distribution*, and then incorporate them to build a scale-mixture of Gaussians. This modification can mimic/fit the data better by capturing the outliers. In this case, the outliers will be "thrown" into the normal component with larger variance.

(2) Another possible thinking is to do some **clustering**. Since now I have 38 groups in the hierarchical generalized regression analysis (group based on different states), it is intuitively reliable to make some clustering for the 38 states (e.g., using hierarchical clustering method, or just based on typical Euclidian distance).

(3) A third possible thought is to increase the credibility of prediction by using **time series data**. Here it's just my general thinking, not necessarily on this problem, since this highly relies on your data.

(4) Compare ***frequentist hierarchical regression model*** vs. ***bayesian hierarchical regression model***: I attempted to

[4]They are now work as "further research", due to strong time conflicts and constrain. I've tried most of them.

build a frequentist hierarchical model and compare it with the bayesian one, using the R packages *lmer4* and *nlme*. However, I cannot make them work well for my data. I also tried some *simple cross-section validation* by running a simple "naïve regression" using only the training data in CA, and predict MLE for the 50 testing values. *The Mean Square Prediction Error (MSPE) for the Bayesian model is smaller than the corresponding one for simple linear regression.* Yet, it's not very scientific to compare a hierarchical result with a pure non-hierarchical linear regression using single group data.

## IV. RESULTS SUMMARY

### 4.1. Model Diagnostics Test:

The residual diagnostic tests show us: the hierarchical regression residual fit the normal assumption better, especially when we use the transformed plot in the third graph.

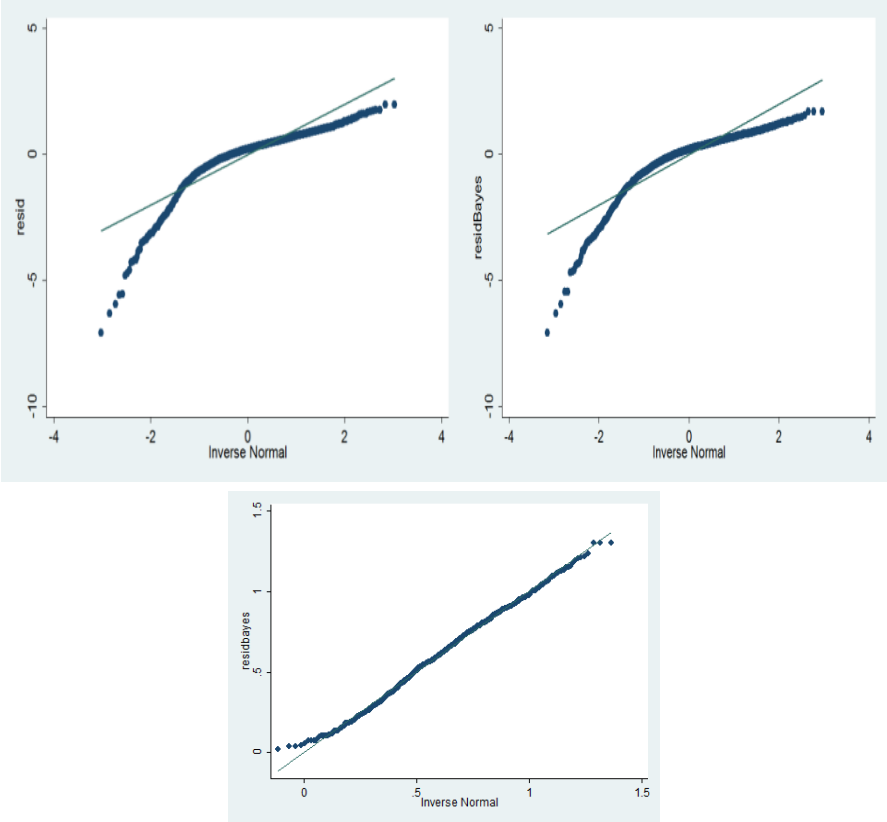


Figure.1 model diagnostic test (QQ plot): for OLS(1) and hierarchical residuals (2&3)

### 4.2. Convergence Check: ACF and Traceplot

As we see, the auto-correlation functions for the seven posterior thetas are all look nice, with ACFs jump to 0 around lag 10. And the traceplots for all the seven posterior thetas are nice as well, with nice static convergence after throwing away the burn-ins. (10000 posterior simulations after 40000 burn-ins).

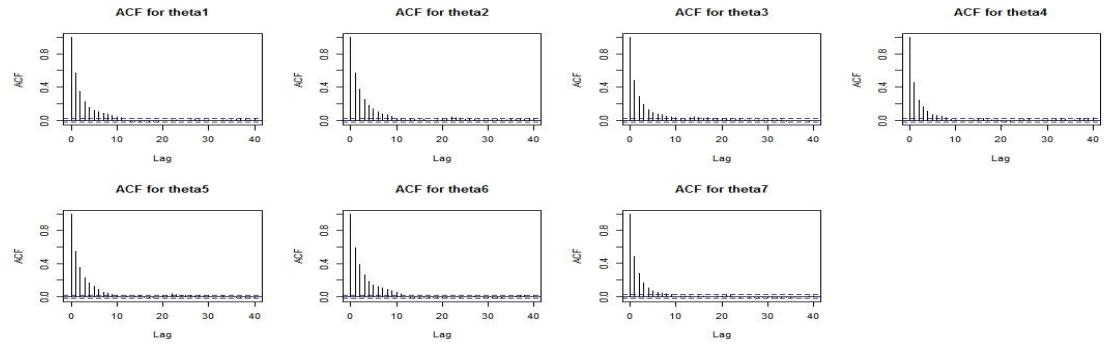


Figure.2convergence check: ACF for posterior thetas

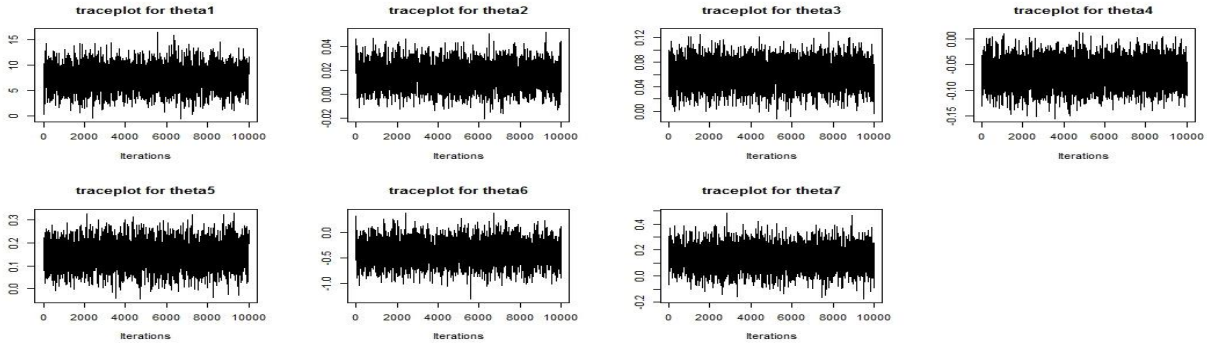


Figure.3 convergence check: traceplots for posterior thetas

As it was shown in Figure.2-3, the ACF and Traceplots for all the thetas perform well. Now I display the posterior inferences below. Output for all seven betas and 6 sigma$^2$ (out of 38 groups) are reported. As we can see from table.2, all the 95% Confidence Interval and 95% HPD interval cover 0; and the ranges are rather large, indicating heterogeneity across different states. We can further check this in Figure.3.

Table.2 summary of posterior inferences for beta

| $\beta_j$ | **mean** | **95% credible interval** | | **95% HPD interval** | |
|---|---|---|---|---|---|
| $\beta_1$ | 7.38061848 | -5.54665647 | 21.03542371 | -5.67429699 | 20.88633312 |
| $\beta_2$ | 0.01527041 | -0.04756808 | 0.07637843 | -0.04676158 | 0.07715483 |
| $\beta_3$ | 0.06000084 | -0.08447064 | 0.18038883 | -0.07721382 | 0.18659419 |
| $\beta_4$ | -0.06725023 | -0.26589530 | 0.08984892 | -0.26096499 | 0.09410068 |
| $\beta_5$ | 0.14923470 | -0.22155682 | 0.47743207 | -0.21350488 | 0.48449148 |
| $\beta_6$ | -0.39684236 | -1.61310163 | 0.82974282 | -1.60200393 | 0.83918587 |
| $\beta_7$ | 0.13917707 | -0.38717285 | 0.72168719 | -0.38255008 | 0.72562964 |

Table.3posterior inferences for $\sigma_j^2$ in 6 groups (out of 38 states)

| $\sigma_j^2$ | **mean** | **95% credible interval** | | **95% HPD interval** | |
|---|---|---|---|---|---|
| $\sigma_1^2$ | 0.8159499 | 0.3778829 | 1.6952233 | 0.3176422 | 1.5097465 |
| $\sigma_2^2$ | 1.7064460 | 0.9204576 | 3.0483144 | 0.8200803 | 2.8180583 |

| | | | | | |
|---|---|---|---|---|---|
| $\sigma_3^2$ | 0.4362865 | 0.2187859 | 0.8547357 | 0.1910964 | 0.7667413 |
| $\sigma_4^2$ | 0.4270706 | 0.3602668 | 0.5037091 | 0.3597551 | 0.5027664 |
| $\sigma_5^2$ | 0.4502799 | 0.2474726 | 0.8080682 | 0.2280811 | 0.7528439 |
| $\sigma_6^2$ | 0.5360850 | 0.3792335 | 0.7613718 | 0.3602619 | 0.7307799 |

From table.3, we can have a deeper look at the state-specific (within group) variance, and found that the mean and CI, HPD of them are so different, confirm our belief in state-specific variance/effects. Since not even a single pair of states has the same or much closed variance, it seems true that we should admit strong heteroskedasticity in our case, among the 38 different states.

### 4.3. Shrinkage Check& Heteroskedasticity Check

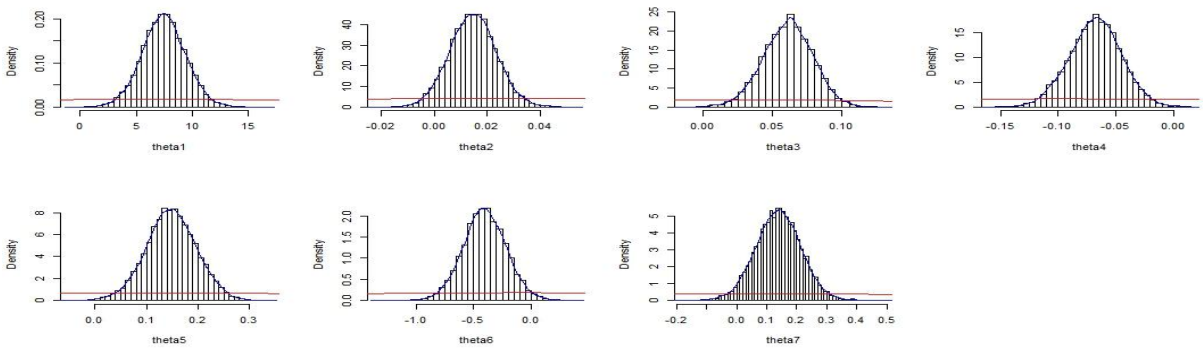


Figure.3 shrinkage check: probability density distributions of posterior thetas

The posterior deviation for theta is much more concentrated than the prior Ordinary Least Square (OLS) estimates, as it was shown in Figure.3. So, we obtained smaller Mean Square Error (MSE) from the posterior estimates. Thus, we gain evidence that Bayesian hierarchical regression model did much better than the OLS priors since the estimates go closer to the true parameter values. Shrinkage estimate is very popularly used in Bayesian analysis, and it can usually give you stronger prediction power.

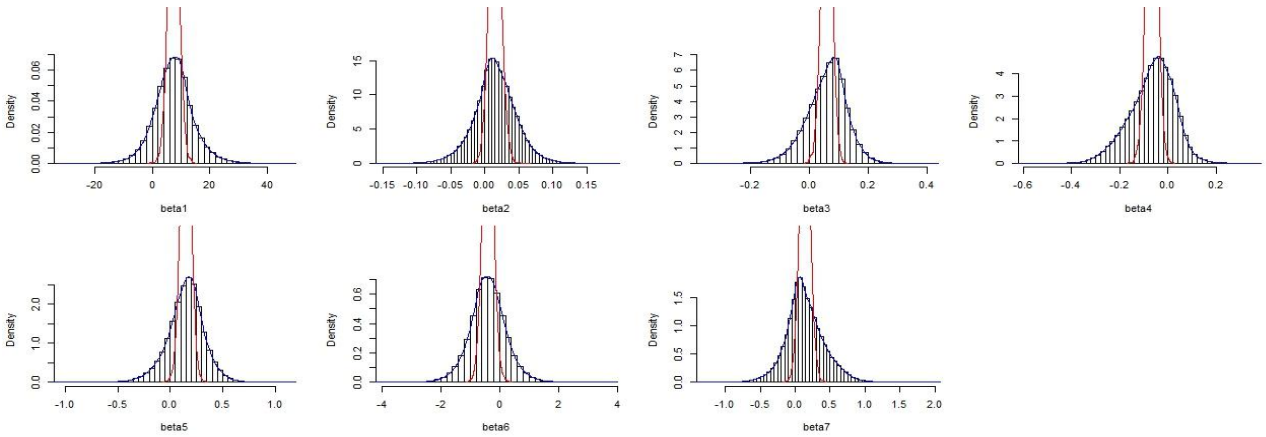


Figure.4 Heterogeneity in regression coefficients

In Figure.4, all posterior means showed much more variance than corresponding prior means (the same red one).This is quite reliable since it matches our heterogeneity analysis before.

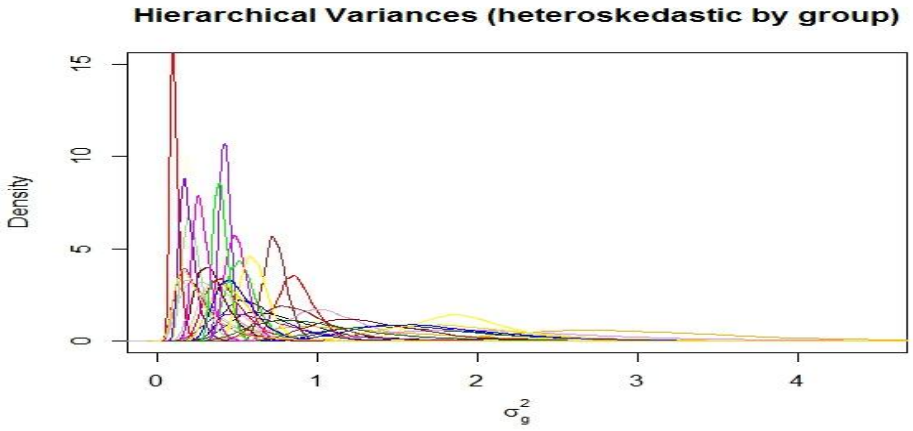


Figure.5 heteroskedasticity in hierarchical group variances

$Sigma_g^2$ for different states show quite different posterior densities in Figure.5. Obvious heteroskedasticity again provide evidence that a Bayesian hierarchical regression performs much better than using a simple OLS method by assuming homogeneity.

### 4.4. Prediction for Decision

Obtain coefficients for state-specific predictive models. We can directly see the heterogeneity. I've approximated coefficients for all the 38 states, here I choose 4 of them to display. (For example, you want to choose relatively safer places to live or to apply for a PhD program there)

Table.3 approximated coefficients for state-specific predictive models

| Estimate of $\beta_j$ | California (CA) | Massachusetts (MA) | North Carolina (NC) | Pittsburgh (PA) |
|---|---|---|---|---|
| $\beta_1$ | 9.734732826 | 10.036811094 | 6.34207986 | 6.08510760 |
| $\beta_2$ | 0.006498597 | -0.001816541 | 0.01412628 | 0.01941027 |
| $\beta_3$ | 0.098004382 | 0.104080608 | 0.02852369 | 0.12665660 |
| $\beta_4$ | -0.071662298 | 0.026416059 | -0.05685982 | 0.01930357 |
| $\beta_5$ | 0.219581726 | 0.269747044 | 0.18038866 | 0.17331998 |
| $\beta_6$ | -0.661732038 | -0.889825003 | -0.27940858 | -0.20035850 |
| $\beta_7$ | 0.028475190 | 0.152349597 | 0.14041911 | -0.11676510 |

## V. RESULTS AND CONCLUSIONS

Answering the several questions I rose at the beginning. I conclude them as follow:

(1) In variable selection and model building. I tried SSVS, PCA and FA together to conduct a conservative data mining. Due to the special dataset I have, I need to reduce the dimensions, try to avoid endogeneity problems and match related economic theories. There are some trade-off between fancy and accuracy. And I finally chose the smallest number of variables in a nested model, that can best explain/predict log(non-ViolentCrimeRate/100kpopulation).

(2) In comparing the preliminary generalized linear model, and the bayesian hierarchical regression model:

The use of a weak but unbiased prior (OLS estimate) is not only nice for policy reasons, but also make it easier to compare the results from frequentist to bayesian by just comparing the prior and posterior.

I confirmed my "educated guess" about the heterogeneity of crime determination in different states, instead of the super strong assumption of homogeneity and homoscedasticity among all the states in GLM.

The accuracy of MLE estimate is constrained by the limited information (observations) in several states. Bayesian hierarchical regression did pretty well by borrowing information from all the data, and so we get shrinkage estimates, which are very good for prediction.

(3) In other extensions: The trial of a more scientific comparison between frequentist hierarchical regression model and bayesian hierarchical regression model, even though not complete, gave me some interesting flavor of the advantage of bayesian approach.

Some of my preliminary try of model mixture and cluster gave me new dimensional thinking about the resourceful application of bayesian method in dealing with real world research topics. And I'm happy and excited to explore more in the further.

(4) All these discussed above showed good ideas and practice for the design of a more scientific commercial crime insurance. The application of bayesian hierarchical model could be more interesting and intriguing than the general application of frequentist hierarchical model and simple regression by state. So I recommend insurance companies to take more detailed consideration to the usage of bayesian analysis in their pricing processes.

## REFERENCES

[1] Peter D.Hoff, A First Course in Bayesian Statistical Methods, Springer 2009

[2] David Dunson, Bayesian and Modern Statistics (PhD Core course) Lecture Notes, 2013 Spring

* Stata code for data clearing and R code for Bayesian hierarchical regression are available upon request via email: Xiaolu.wang@duke.edu. Many delicate steps are involved, and can be derived from the introduction part, by dealing with all the data problems (e.g., missing not at random (not MAR or MCAR), and identifications check). Welcome to discuss.