Bayesian Data Analysis

Sourish Das[1]

[1]Mathematics,
Chennai Mathematical Institute, INDIA

Bayesian Hierarchical Model

CHENNAI MATHEMATICAL INSTITUTE

Introduction

- When there are few parameters, posterior inference in nonconjugate multiparameter models can be obtained by simulation methods.
- sophisticated models can often be represented in a hierarchical for which effective computational strategies are available

CHENNAI MATHEMATICAL INSTITUTE

General Approach to Bayesian Modeling

1. Write the likelihood part of the model, $p(y|\theta)$, ignoring any factors that are free of θ.

2. Write the posterior density, $p(\theta|y) \propto p(\theta)p(y|\theta)$. If prior information is well-formulated, include it in $p(\theta)$. Otherwise use non-informative prior

3. Create a crude estimate of the parameters, θ, for use as a starting point and a comparison to the computation in the next step.

4. Draw simulations from $\theta^1, ..., \theta^S$, from the posterior distribution. Use the sample draws to compute the posterior density of any functions of θ that may be of interest. For non-conjugate models this step could be difficult.

General Approach to Bayesian Modeling

5. If any predictive quantities, $\tilde{y}$, are of interest simulate $\tilde{y}^1, ..., \tilde{y}^S$ by drawing each $\tilde{y}^s$ from the sampling distribution conditional on the drawn value θ^s, $p(\tilde{y}|\theta^s)$.

- Various methods (such as Markov Chain Monte Carlo) have been developed to draw posterior simulations in complicated models.

- If θ has only one or two components, it is possible to draw simulations by computing on a grid.

Hierarchical Models

- Hierarchical data is ubiquitous in the social sciences where measurement occurs at different levels of aggregation.
  - e.g. we collect measurements of individuals who live in a certain locality or belong to a particular race or social group.
- When this occurs, standard techniques either assume that these groups belong to entirely different populations or ignore the aggregate information entirely.
- Hierarchical models provide a way of pooling the information for the disparate groups without assuming that they belong to precisely the same population.

Hierarchical Models

- Suppose we have collected data about some random variable Y from m different populations with n observations for each population.
- Let Y_{ij} represent observation j from population i
- Suppose $Y_{ij} \sim f(\theta_i)$ where θ_i is a vector of parameters of population i.
- Further $\theta_i \sim f(\Theta)$ where Θ may be a vector.
- Note, until this point this is just a standard Bayesian setup where we are assigning some prior distribution for the parameters θ that govern the distribution of y.

Hierarchical Models

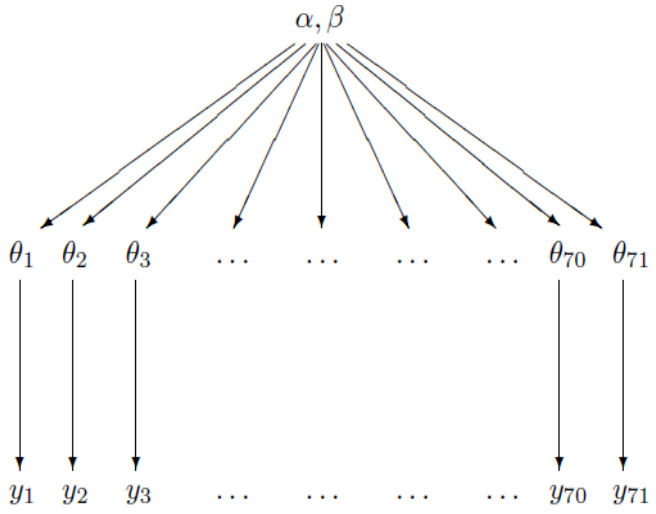
- Now we extend the model, and assume that the parameters Θ_{11}, Θ_{12} that govern the distribution of the Θ's are themselves random variables and assign a prior distribution to these variables as well:

$$\Theta \sim f(\alpha, \beta)$$

- Θ is called hyperprior. The parameters a, b, c, d for the hyperprior may be "known" and represent our prior beliefs about Θ.

- In theory, we can also assign a probability distribution for these quantities as well, and proceed to another layer of hierarchy.

Hierarchical Models



Exchangeability

Exchangeability : Formal

The parameters $\theta_1, \theta_2, ..., \theta_n$ are exchangeable in their joint distribution if $p(\theta_1, \theta_2, ..., \theta_n)$ is invariant to permutations in the index $1, 2, ..., n$.

Exchangeability : Informal

If no information other than the data is available to distinguish any of the θ_j's form any of the others, and no ordering of the parameters can be made, one must assume symmetry among the parameters in the prior distribution.

This concept is closely related to the concept of identically and independent random variables where, conditional on the data, each observation is treated the same.

Application: Poisson-Gamma Model

- Robert etal. (2004) presents data of the number of failures (y_i) for each of 10 pumps in a nuclear plant?
- We also have the times (t_i) at which each pump was observed.
- To model this process, we assume that the number failure follows a Poisson distribution.

$$\text{failure}_i \sim Poisson(\lambda_i t_i)$$

Q1 How would we address this question?

Q2 Why might we model this as a hierarchical process?

cmi | CHENNAI MATHEMATICAL INSTITUTE

Exchangeability

- Exchangeability means that we can treat the parameters for each sub-population as exchangeable units.
- In its simplest form, each parameter θ_j is treated as an independent sample from a distribution governed by unknown parameter vector Θ.

$$p(\theta_1, \theta_2, ..., \theta_n|\Theta) = \prod_i p(\theta_i|\Theta)$$

- In a more general form, we may also condition on data that we have about the different sub-populations.

Exchangeability

- We can write the joint prior distribution as:

$$p(\theta_1, \theta_2, ..., \theta_n, \Theta) = p(\theta_1, \theta_2, ..., \theta_n|\Theta)p(\Theta)$$

- By Baye's Rule

$$p(\theta_1, ..., \theta_n, \Theta|Y) \propto \text{prior} \times \text{likelihood for } Y$$

Application: Poisson-Gamma Model

- We consider the data model to be:

$$y_i \sim Poisson(\lambda_i t_i) \quad \text{for } i = 1, ..., 10.$$

- To model this as a hierarchical process, we assume that each of the absence λ_i are exchangeable draws from a common distribution.

- In this case, the gamma distribution has desirable properties.

$$\lambda_i \sim Gamma(\alpha, \beta) \quad \text{for } i = 1, ..., 10.$$

  Note that $\alpha = 1.8$ and β are unknown parameters.

Application: Poisson-Gamma Model

- To satisfy the requirement of exchangeability, what must we assume about the data generating process?

- Finally, to complete the hierarchical structure, we must assign "hyperpriors" for the parameters on β. Again, the gamma distribution has nice properties, so we assume that:

$$\beta \sim Gamma(\nu, \delta)$$

- assign $Gamma(\nu, \delta)$ on unknow β with $\nu = 0.001$ and $\delta = 1$.

Application: Poisson-Gamma Model

- The joint posterior distribution is:

$$p(\lambda_i, \beta | y, t) \propto \prod_i Pois(\lambda_i t_i | y_i) \times Gamma(\lambda_i | \alpha, \beta) Gamma(\beta | \nu, \delta)$$

- Using our trick for conditional distributions, we know

$$p(\lambda_i | \beta, \alpha, y, t) \sim Gamma(y_i + \alpha, t_i + \beta)$$

- $p(\beta | \lambda, y, t) \sim Gamma(10\alpha + \nu, \delta + \sum_{i=1}^{10} \lambda_i)$

Application: Poisson-Gamma Model

Implemented a Gibbs Sampler

```
Posterior mean of lambda's :

 [1] 0.07054569 0.15350870 0.10329934 0.12324020 0.65301834 0.62245
 [7] 0.85017017 0.85507109 1.34955910 1.91816030

Posterior mean of beta :

[1] 2.405042
```

Application: Poisson-Gamma Model

- As we assume $\lambda_i \sim Gamma(\alpha = 1.8, \beta)$ and estimated $\hat{\beta} = 2.39$
- 95% CI of $Gamma(\alpha = 1.8, \hat{\beta} = 2.39)$

```
[1] 0.07631031 2.17514481
```

- It indicates λ_1 might be an outlier

Hierarchical Regression Models

1. The intuition behind hierarchical regression models
2. Hierarchical models provide a way of examining differences across populations. They pool the information for the disparate groups without assuming that they belong to precisely the same population.
3. In the context of regression analyses, hierarchical models allow us to examine whether the extent to which regression coefficients vary across different sub-populations, while borrowing strength from the full sample.

cmi | CHENNAI MATHEMATICAL INSTITUTE

Example

- The importance of uncertainty about the Democratic Party's ideology for its electoral success during the Jacksonian era (circa 1840).

- **Dependent variable**:
  - Percentage of seats won by the Democratic Party in the House of Representatives in United States in state i in election t.

- **Independent variable**:
  - Level of ideological conflict within state i's Democratic Party delegation to the House in period $t-1$.

  - Possible control variables include dummy variables for the various states measuring their preference for the Democratic Party and for each election.

- Key modeling question: Does the sample pool?

Parameters of Pooled OLS Model of Democratic Electoral Success Due to Intra-Party Unity

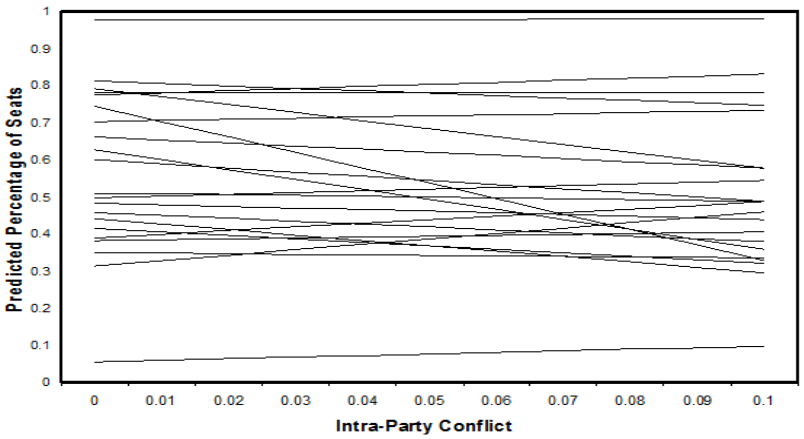
Dep. Var. Democratic Electoral Success	Posteriod Mean	Posterior standard deviation
Intercept	0.578*	0.024
Ideological Conflict	−3.512*	1.192

- Mean Squared Error: 0.08516

  * Denotes statistical significance

cmi | CHENNAI MATHEMATICAL INSTITUTE

Unpooled OLS Model : (Different state-specific intercepts and slopes)



- F-tests reject the unpooled model as statistically unwarranted; however, there were significant state-specific intercepts and coefficients suggesting that there was causal heterogeneity in the model. What to do?

Example

- F-tests reject the unpooled model as statistically unwarranted; however, there were significant state-specific intercepts and coefficients suggesting that there was causal heterogeneity in the model.

- In a context like this, hierarchical structures are perfect.
  - Where differences are not statistically important, the state-specific coefficients are shrunk back toward the national average.

  - Where differences are statistically meaningful, the state-specific effects remain markedly different from the national average.

The Hierarchical Probability Model

- Electoral Success$_{it} \sim N(m_{it}, \tau)$
- where $m_{it} = a_i + b_i * \textit{Intra-Party Conflict}_{i,t-1}$,
- $a_i \sim N(A, \tau_A)$ for all i
- $A \sim N(0, 0.01)$ and $\tau_A \sim \textit{Gamma}(0.1, 0.1)$
- $b_i \sim N(B, 0.01)$ for all i
- $B \sim N(0, 0.01)$ and $\tau_B \sim \textit{Gamma}(0.1, 0.1)$
- and $\tau \sim \textit{Gamma}(0.1, 0.1)$

Comments

- The crucial difference between unpooled OLS and the hierarchical model is that the state-specific intercept terms and the coefficients for intra-party conflict are now treated as exchangeable draws from a common probability model with unknown mean and variance.
- The posterior distributions of these state-specific parameters convey information about local effects.

cmi | CHENNAI MATHEMATICAL INSTITUTE

Comments

- The hyper-parameter A represents the average level of Democratic electoral success while τ_A measures the variation in the party's fortunes across states.
- Similarly, B is the average impact of intra-party conflict, while τ_B indicates the variation in the influence of party unity across states.

Comments

- If the posterior distribution of the hyper-parameters reveal that $\tau_A = \tau_B = \infty$ then pooled OLS is a special case.
- This is because if there is no variance (i.e. infinite precision) in the intercept or coefficient across states, then one should conclude that there are no regime effects.
- Similarly, if $\tau_A = \tau_B = 0$, then unpooled OLS is a special case because there is no underlying structure to the data across states.

Hyper-Parameters for Model of Democratic Electoral Success Due to Intra-Party Unity

Parameter for State specific Intercept	Posteriod Mean	Posterior standard deviation
Mean of Stat specific Intercept	0.54*	0.054
Precision of State Specific Intercept	21.2	8.101

* denotes statistical significance

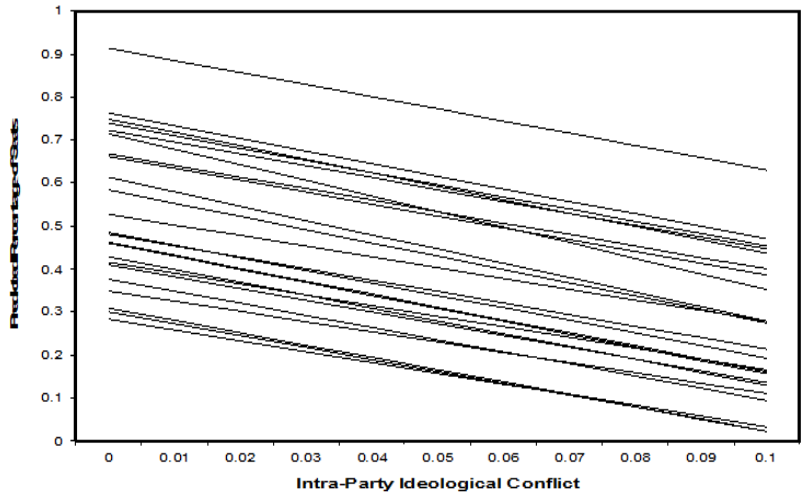
CHENNAI MATHEMATICAL INSTITUTE

Parameter for State specific Intra-Part Coefficients	Posteriod Mean	Posterior standard deviation
Mean of Stat specific Coefficients	-2.85^*	1.216
Precision of State Specific Coefficients	3.071	4.81

* denotes statistical significance

CHENNAI MATHEMATICAL INSTITUTE

State-Specific Predicted Values



CHENNAI MATHEMATICAL INSTITUTE

What sort of voodoo is this?

- The explanation for why the random coefficient model had such a substantial impact on the parameter estimates for intra-party conflict was precisely because our pooling tests rejected the joint significance of state-specific effects.
- The wild variations observed from unpooled OLS were an artifact of over-fitting the data based on a small number of observations.
- To prevent this over-fitting, the random coefficient model "borrowed strength" from the overall effect of the independent variable in order to make inferences about the state-specific effects.

CHENNAI MATHEMATICAL INSTITUTE

What sort of voodoo is this?

- The extent of this borrowing is contingent on the relative precision of the state-specific and overall effects.
- Thus, the regression lines became approximately parallel with the introduction of the random coefficient model because there was relatively little information provided by the state-specific data regarding the effect of intra-party conflict relative to that provided by the entire sample.
- Meanwhile, the intercepts remained variant across regression lines, because there was sufficient state-specific data to establish that each state had a different predisposition in favor or against the Democratic Party.

Dynamic Pricing with Hierarchical Regression

- conside the data set called `cheese` from the `bayesm` package.
- The data set contains marketing data of certain brand name processed cheese, such as the weekly sales volume (`VOLUME`), unit retail price (`PRICE`), and display activity level (`DISP`) in various regional retailer accounts.
- A list of 88 Retailer accounts are there

Dynamic Pricing with Hierarchical Regression

- For each account, we can define the following linear regression model of the log sales volume, where β_1 is the intercept term, β_2 is the display measure coefficient, and β_3 is the log price coefficient.

$$\log(\mathit{Volume}) = \beta_1 + \beta_2 * \mathit{Display} + \beta_3 * \log(\mathit{Price}) + \epsilon$$

  ϵ relies on regional market conditions, and we would not expect it to have the same dispersion among retailers.

- For the same reason, we cannot expect identical regression coefficients for all accounts, or attempt to define a single linear regression model for the entire data set.

Dynamic Pricing with Hierarchical Regression

- We expect regression coefficients of the retailer accounts to be related. A common approach to simulate the relationship is the hierarchical linear model, which treats the regression coefficients as random variables of yet another linear regression at the system level.

- **Problem**:

  Fit the hierarchical linear model, and estimate the average impact on sales volumes of the retailers if the unit retail price is to be raised by 5%

cmi | CHENNAI MATHEMATICAL INSTITUTE

Dynamic Pricing with Hierarchical Regression

- Let i to be an integer between 1 and the number of retailer accounts. We define a filter for the i^{th} account as follows.

```
> library(bayesm)
> data(cheese)
> retailer<-levels(cheese$RETAILER)
> nreg<-length(retailer)
```


Dynamic Pricing with Hierarchical Regression

- We now loop through the accounts, and create a list of data items consisting of the X and y components of the linear regression model in each account. The columns of X below contains the intercept placeholder, the display measure, and log price data.

```
> regdata<-NULL
> for (i in 1:nreg) {
+     filter <- cheese$RETAILER==retailer[i]
+     y <- log(cheese$VOLUME[filter])
+     X <- cbind(1,      # intercept placeholder
+             cheese$DISP[filter],
+             log(cheese$PRICE[filter]))
+     regdata[[i]] <- list(y=y, X=X)
+ }
```

Dynamic Pricing with Hierarchical Regression

- We wrap the `regdata` and the iteration parameters in lists, and invoke the `rhierLinearModel` method of the bayesm package. It takes about half a minute for 2,000 MCMC iterations on an average CPU.

```
> Data <- list(regdata=regdata)
> Mcmc <- list(R=2000)
> set.seed(7831)
> system.time(out <- bayesm::rhierLinearModel(
+             Data=Data,
+             Mcmc=Mcmc))
Z not specified -- putting in iota

Starting Gibbs Sampler for Linear Hierarchical Model
    88  Regressions
    1  Variables in Z (if 1, then only intercept)

Prior Parms:
Deltabar
     [,1] [,2] [,3]
```

CHENNAI MATHEMATICAL INSTITUTE

Dynamic Pricing with Hierarchical Regression

- A 5% increase of the unit price amounts to an expected drop in .
- Can we estimate the same for specific retailer?
- The answer is yes and it can help us in dynamic pricing.

CHENNAI MATHEMATICAL INSTITUTE

400 (0.0)

Dynamic Pricing with Hierarchical Regression

We can estimate retailer specific intercept and slope

	retailer	int	b_dip	b_price
1	ALBANY,NY - PRICE CHOPPER	10.52	1.21	-3.70
2	ATLANTA - KROGER CO	10.17	1.11	-1.66
3	ATLANTA - WINN DIXIE	9.95	-0.60	-1.99
4	BALTI/WASH - GIANT FOOD INC	12.95	0.37	-3.47
5	BALTI/WASH - SAFEWAY	11.24	2.14	-2.20
6	BALTI/WASH - SUPER FRESH	10.96	1.84	-2.91

CHENNAI MATHEMATICAL INSTITUTE

Dynamic Pricing with Hierarchical Regression

- For the first retailer if we drop the price by 5%, then expected increase in volume would be

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
12.35	19.10	20.97	20.93	22.74	30.31

CHENNAI MATHEMATICAL INSTITUTE

Dynamic Pricing with Hierarchical Regression

- How much this change is going to affect in the revenue?

```
current expected revenue
                3759.832
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   4013    4254    4321    4319    4384    4654
```

CHENNAI MATHEMATICAL INSTITUTE

Hierarchical Regression Models

- Suppose we have a standard multiple regression model where observations i cluster across sub-populations j.
  where j indexes, for example, geographic location, social group, period in history.

- But, we do not want to assume that regression coefficients are identical across sub-populations.

- We also want to allow for unequal variances across sub-populations.

- We assume that each observation is distributed normally with an expected value determined by both observation-specific and sub-population characteristics and level of aggregation-specific variance. Thus,

$$y_{ij} \sim N(m_{ij}, t_j)$$

CHENNAI MATHEMATICAL INSTITUTE

The Random Coefficient Model

- Suppose
$$y_{ij} \sim N(m_{ij}, t_j)$$
then
$$m_{ij} = \beta_{0j} + \beta_{1j}X_{1j} + ... + \beta_{kj}X_{kj}$$

- For a **random coefficient** (hierarchical) model, we assume that:
$$b_{kj} = \gamma_k + \delta_{kj},$$
where γ_k represnts overall effect of β_k
δ_j represents the difference in the coefficient between sub-population j and the overall coefficient, where $E[\delta_{kj}] = 0$

Random Coefficient Model and its Prior

- Suppose that $y_{ij} \sim N(m_{ij}, t_j)$
$$m_{ij} = b_{0j} + b_{1j}X_{1j} + ... + b_{mj}X_{mj}$$
then
$$m_{ij} = (\gamma_0 + \delta_{0j}) + (\gamma_1 + \delta_{1j})X_{1j} + ... + (\gamma_1 + \delta_{1j})X_{mj}$$

- We shall assume that $t_j \sim Gamma(0.001, 0.001)$ for all j

Random Coefficient Model and its Prior

- Two basic strategies for defining priors for the coefficients:
  1. Specify priors for both γ_k and δ_{kj} as follows: $\gamma_k \sim N(\text{prior mean}, \text{prior prec})$ and $\delta_{kj} \sim N(0, \tau_k)$ where $\tau_k \sim Gamma(0.001, 0.001)$
  2. Use "Hierarchical-centering" as follows: $\beta_{kj} \sim N(\gamma_k, \tau_k)$ where $\gamma_k \sim N(\text{prior mean}, \text{prior prec})$ and $\tau_k \sim Gamma(0.001, 0.001)$
- Method 2 improves MCMC markedly in some cases (see Gilks and Roberts, "Strategies for improving MCMC" in MCMC in Practice)

Hierarchical Gaussian Linear Regression Model

Aka. **Linear Mixed Effect Models**. The model takes the following form:

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + W_i b_i + \varepsilon_i,$$

where each group i have k_i observations.

The random effects:

$$b_i \sim \mathcal{N}_q(0, V_b).$$

The errors:

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2 I_{k_i})$$

Assume standard, conjugate priors:

$$\beta \sim \mathcal{N}_p(\mu_\beta, V_\beta) \quad \text{and} \quad \sigma^2 \sim \mathcal{IG}amma(\nu, 1/\delta)$$
$$V_b \sim \mathcal{IW}ishart(r, rR)$$

This is the model available in `MCMCpack`.
Chib and Carlin (1999)

Democratic Party's ideology in Jacksonian Era

- **Dependent variable**:

  Percentage of seats won by the Democratic Party in the House of Representatives in state i in election t.

- **Independent variable**:
- Level of ideological conflict within state i's Democratic Party delegation to the House in period $t-1$.

- Control variables include dummy variables for the various states measuring their preference for the Democratic Party and for each election.

Hierarchical Binomial Linear Regression Model using the logit link

- The model takes the following form:

$$y_i \sim Bernoulli(\theta_i)$$

- With latent variables $I(\theta)$, $I(.)$ being the logit link function:

$$I(\theta_i) = X_i * \beta + W_i * b_i + \epsilon_i$$

- where each group i have k_i observations.

- the random effects:

$$b_i \sim N_q(0, V_b)$$

cmi | CHENNAI MATHEMATICAL INSTITUTE

Hierarchical Binomial Linear Regression Model using the logit link

- the over-dispersion terms:

$$\epsilon_i \sim N(0, \sigma^2 \mathbf{I}_{k_i})$$

- We assume standard, conjugate priors:

$$\beta \sim N_p(\mu_\beta, V_\beta)$$

$$\sigma^2 \sim IGamma(\nu, 1/\delta)$$

$$V_b \sim IWishart(r, rR)$$

Hierarchical Binomial Linear Regression Model using the logit link

- It is difficult to have default parameters for the priors on the precision matrix for the random effects.
- When fitting one of these models, it is of utmost importance to choose a prior that reflects your prior beliefs about the random effects.
- Using the `dwish` and `rwish` functions might be useful in choosing these values.

CHENNAI MATHEMATICAL INSTITUTE

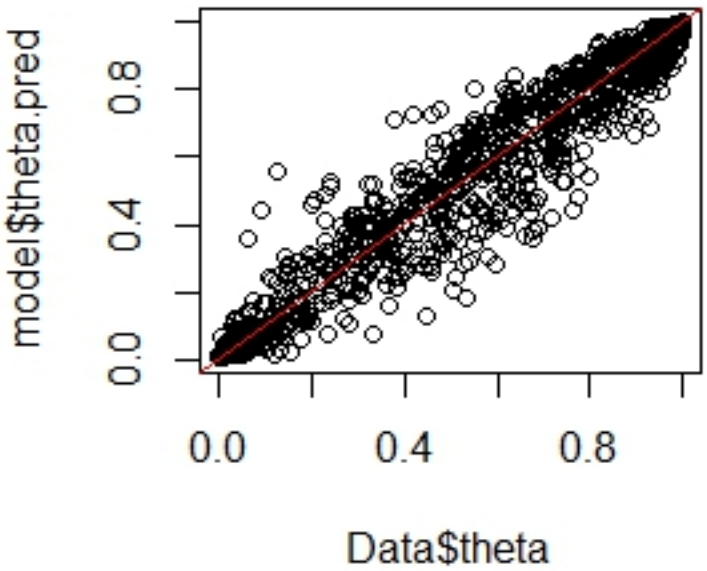
Implementation in R

- In `MCMCpack` R-package the `MCMChlogit` implement the Hierarchical Binomial Linear Regression Model with the logit link function

```
#== Call to MCMChlogit
model <- MCMChlogit(fixed=Y~X1+X2, random=~X1+X2, group="species",
                    data=Data, burnin=1000, mcmc=1000, thin=1,verbose=1,
                    seed=NA, beta.start=0, sigma2.start=1,
                    Vb.start=1, mubeta=0, Vbeta=1.0E6,
                    r=3, R=diag(c(1,0.1,0.1)), nu=0.001, delta=0.001, FixOD=1)
```

CHENNAI MATHEMATICAL INSTITUTE

Implementation in R



CHENNAI MATHEMATICAL INSTITUTE

Hierarchical Poisson Linear Regression Model using the log link function

- The model takes the following form:

$$y_i \sim Poisson(\lambda_i)$$

- With latent variables $I(\theta)$, $I(.)$ being the log link function:

$$I(\theta_i) = X_i * \beta + W_i * b_i + \epsilon_i$$

- where each group i have k_i observations.

- the random effects:

$$b_i \sim N_q(0, V_b)$$

CHENNAI MATHEMATICAL INSTITUTE

Hierarchical Poisson Linear Regression Model using the log link function

- the over-dispersion terms:

$$\epsilon_i \sim N(0, \sigma^2 \mathbf{I}_{k_i})$$

- We assume standard, conjugate priors:

$$\beta \sim N_p(\mu_\beta, V_\beta)$$

$$\sigma^2 \sim IGamma(\nu, 1/\delta)$$

$$V_b \sim IWishart(r, rR)$$

CHENNAI MATHEMATICAL INSTITUTE

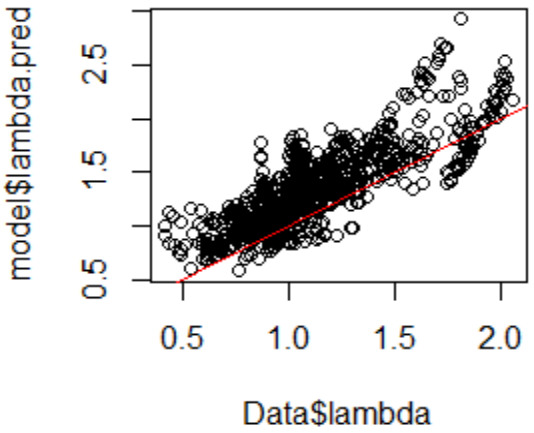
Implementation in R

- In `MCMCpack` R-package the `MCMChpoisson` implement the Hierarchical Poisson Linear Regression Model with the log link function

```
#== Call to MCMChpoisson
model <- MCMChpoisson(fixed=Y~X1+X2, random=~X1+X2, group="species",
                      data=Data, burnin=500, mcmc=1000, thin=1,verbose=1,
                      seed=NA, beta.start=0, sigma2.start=1,
                      Vb.start=1, mubeta=0, Vbeta=1.0E6,
                      r=3, R=diag(c(0.1,0.1,0.1)), nu=0.001, delta=0.001, FixOD=1)
```

CHENNAI MATHEMATICAL INSTITUTE

Implementation in R



λ's are overestimated by the models

CHENNAI MATHEMATICAL INSTITUTE

Thank You

sourish@cmi.ac.in

www.cmi.ac.in/~sourish

CHENNAI MATHEMATICAL INSTITUTE