



# Applied Bayesian Statistics and Estimation for Social Scientists


By Dr Diego Perez Ruiz

School of Social Statistics at the  
University of Manchester

2022

Contact - [diego.perezruiz@Manchester.ac.uk](mailto:diego.perezruiz@Manchester.ac.uk)

# Outline of the Presentation

- Basics of Bayesian Statistics
  - Introduction to Hierarchical Models
    - Logistic Regression
    - Poisson Regression
  - Case Study – Employment in India
  - References
- 

## Frequentist Statistics

Frequentist statistics is a type of statistical inference that draws conclusions from sample data by emphasizing the frequency or proportion of the data. The frequentist approach which considers parameter values as unknown and fixed quantities



## Bayesian Statistics

The Bayesian approach to data analysis differs from the frequentist one in that each parameter of the model is considered as a random variable

The Bayesian approach to data analysis differs from the frequentist one in that each parameter of the model is considered as a random variable and by the explicit use of probability to model the uncertainty (Gelman et al., 2013).

A direct consequence of these two differences is that Bayesian data analysis allows researchers to discuss the probability of a parameter (or a vector of parameters)  $\theta$ , given a set of data  $\mathbf{D}$  :

**Bayes' theorem:**

$$P(\theta | \mathbf{D}) = \frac{P(\mathbf{D} | \theta) \cdot P(\theta)}{P(\mathbf{D})}$$

Using this equation, a probability distribution  $p(\theta | \mathbf{D})$  can be derived (called the posterior distribution).

This distribution is the goal of any Bayesian analysis and contains all the information needed for inference.

# Posterior probability

- The Posterior function depends on the data  $P(\theta | D)$  and reflects the knowledge of the parameter given the data and the prior :

$$P(\theta | D) = \frac{P(D | \theta) \cdot P(\theta)}{P(D)}$$

- $P(D)$  is called the marginal likelihood. It is meant to normalise the posterior distribution, that is, to scale it in the “probability world”. It gives the “probability of the data”
- In turn, the posterior is proportional to the likelihood and the prior (product):

$$P(\theta | D) \propto P(D | \theta) \cdot P(\theta)$$

# Posterior probability

- And we estimate this by multiplying the **Likelihood · Prior.**
- Understanding these words depends on the concept of ‘likelihood’ which is a **function** describing the joint probability of the observed data as a function of the parameters.

# Hierarchical models

One of the important features of a Bayesian approach is the relative ease with which hierarchical models can be constructed and estimated using Gibbs sampling.

One of the key reasons for the recent growth in the use of Bayesian methods in the social sciences is that the use of hierarchical models has also increased dramatically in the last two decades.

Examples: **Kim, Jihye, Olsen, W.K. and Arkadiusz Wiśniowski (2020), Olsen, Wendy, Manasi Bera, Amaresh Dubey, Jihye Kim, Arkadiusz Wiśniowski, Purva Yadav (2020), etc.**

Hierarchical models serve two purposes

**Methodologically**, when units of analysis are drawn from clusters within a population (communities, neighbourhoods, city blocks, etc.), they can no longer be considered independent.

**Substantively** we may believe that there are differences in how predictors in a regression model influence an outcome of interest across clusters, and we may wish to model these differences.

# Multilevel structure

Multilevel modelling allows both fixed and random effects to be incorporated.

In order to use a consistent vocabulary, we follow the recommendations of Gelman and Hill (2007) and avoid these terms. We instead use the more explicit terms **constant** and **varying** to designate effects that are constant, or that vary by groups .

Declaring a simple multilevel model (Gelman and Hill, Chapter 16)

In a simple model,

$$y_i \sim N(\alpha_j + \beta x_i, \sigma_y^2) \text{ and } \alpha_j \sim N(\mu_\alpha, \sigma_\alpha^2)$$

The second normal distribution is acting as a prior for the main model intercepts, which are distributed across Level 2 units whose subscript is j. The Level 1 units have subscript i.

Therefore,  $\mu_\alpha$  and  $\sigma_\alpha$  are the hyperparameters here. The hyperparameters by default are given a uniform distribution each, or a half-t-distribution for variance which is non-negative. *Ibid.*



- Then, we can extend this model to the following multilevel model, adding a varying intercept:

$$y_i \sim N(\alpha_{j[i]} + \beta x_i, \sigma_y^2)$$
$$\alpha_j \sim N(\mu_\alpha, \sigma_\alpha^2)$$

---

- where we use the notation  $\alpha_{j[i]}$  to indicate that each group  $j$  is given a unique intercept, issued from a Gaussian distribution centered on  $\alpha$ , the grand intercept.
- This mean that might be different mean for each group. From this notation we can see that in addition to the residual standard deviation  $\sigma_y^2$ , we are now estimating one more variance component  $\sigma_\alpha^2$ , which is the standard deviation of the distribution of varying intercepts.

# ICC – Intra Class Correlation Coefficient

We can interpret the variation of the parameter  $\alpha$  between groups  $j$  by considering the **intra-class correlation (ICC)**

$$ICC = \frac{\sigma_{\alpha}^2}{\sigma_{\alpha}^2 + \sigma_y^2}$$

which goes to 0, if the grouping conveys no information, and to 1, if all observations in a group are identical (Gelman & Hill, 2007, p. 258).

# Bayesian Random Effects

In the Bayesian framework, every unknown quantity is considered as a random variable that we can describe using probability distributions. Therefore, there is no such thing as a "fixed effect" or a "random effects distribution".

**However, when we write down the model this semantic vanish !!!!**

Suppose we have a dependent continuous variable  $y$  and a dichotomic categorical predictor  $x$  (assumed to be contrast-coded). Let  $y_{ij}$  denote the response of the  $i$ -th participant in the  $j$ -th group.

We can write a "mixed effects" model (as containing both fixed and random effects) as follows:

$$y_{ij} = \alpha_{\text{fix}} + \alpha_i + \beta x_i + e_{ij}, \text{ with}$$

$$e_{ij} \sim \text{Normal}(0, \sigma_e^2), \alpha_i \sim \text{Normal}(0, \sigma_\alpha^2)$$

Where the terms  $\alpha$  and  $\beta$  represent the "fixed effects" and denote the overall mean response and the condition difference in response, respectively.

# from a Bayesian standpoint

We can rewrite this model to make apparent that the so-called "random effects distribution" can be considered a prior distribution, since distributions on unknown quantities are considered as priors:

$$y_{ij} \sim \text{Normal}(\alpha_{ij}, \sigma_e^2)$$

$$\alpha_{ij} = \alpha_i + \beta x_j$$

$$\alpha_i \sim \text{Normal}(\alpha_0, \sigma_\alpha^2)$$

where the parameters of this prior are learned from the data !!!

## Software Programmes

**The brms package** (Bürkner, 2017b), that implements BMLMs in **R**, using Stan under the hood, with a **lme4-like** syntax.

Tsyntax required by brms will not surprise the researcher familiar with lme4, as models of the following form:

$$y_i \sim \text{Normal}(\alpha_i, \sigma_e^2) \text{ with } \alpha_i = \alpha + \alpha * \text{subject}[i] + \beta x_i$$

are specified in brms (as in lme4) with  $y \sim 1 + x + (1 | \text{subject})$ .

In addition to linear regression models, brms allows generalised linear and non-linear multilevel models to be fitted and comes with a great variety of distribution and link functions.

## Modelling Count Data

*Very common in the social sciences*

Number of person in employment

Number of children;

Number of arrests;

Number of flows;

Number of marriages

Number of traffic accidents

Number of deaths

***Modelled by***

Logit regression models the log odds of an underlying propensity of an outcome;

Poisson regression models the log of the underlying *rate* of occurrence of a count.

# Hierarchical Poisson regression

Hierarchical Poisson regression models are expressed as Poisson models with a log link and a normal variance on the mean parameter.

More formally, a hierarchical Poisson regression model is written as :

$$y_j | \lambda_j \sim \text{Poisson}(\lambda_j)$$

$$\log(\lambda_j) = \gamma_j + \beta x_j$$

$$\gamma_j \sim \text{Normal}(0, \sigma^2)$$

*Counts have characteristics*

- Integers and cannot be negative
- Often positively skewed; a 'floor' of zero
- In practice often rare events which peak at 1,2 or 3 and rare at higher values

# Hierarchical Logistic regression

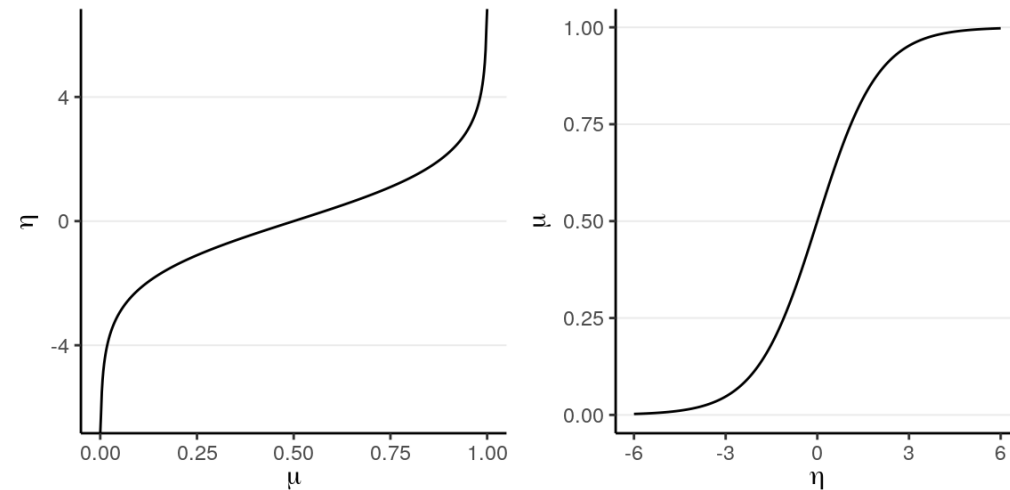
The hierarchical logistic regression model data with group structure and a binary response variable  $y_i = \{0, 1\}$ .

$$y_i \sim \text{Binomial}(n_i, p_i)$$

$$\text{logit}(p_i) = \gamma_i + \beta x_i$$

The logit link convert probabilities to log-odds.

The binomial function is funny in having just one parameter  $p$ , plus the number of trials,  $n$

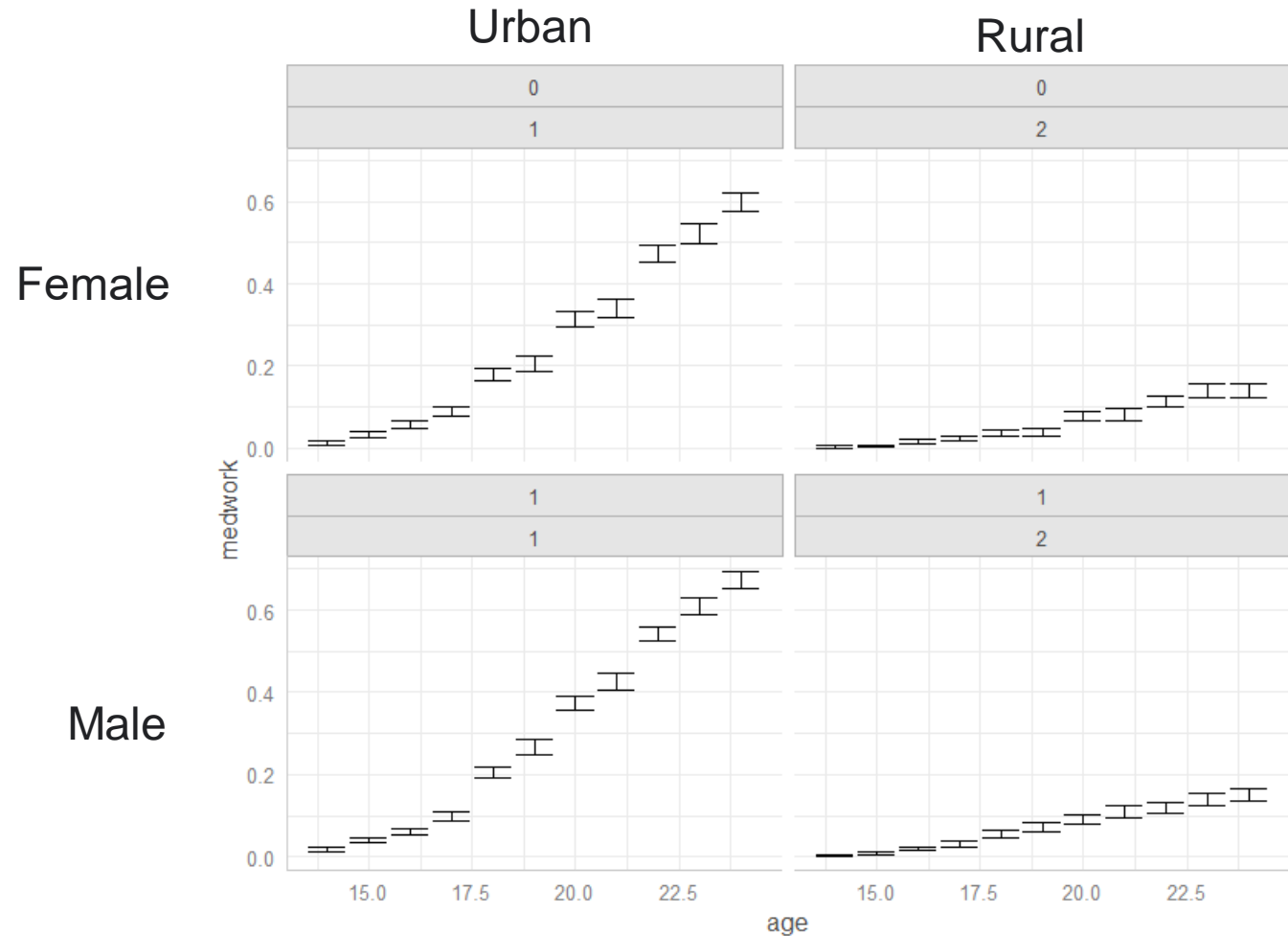




## India

The Periodic Labour Force Survey (**PLFS**) was designed with two major objectives for measurement of employment.

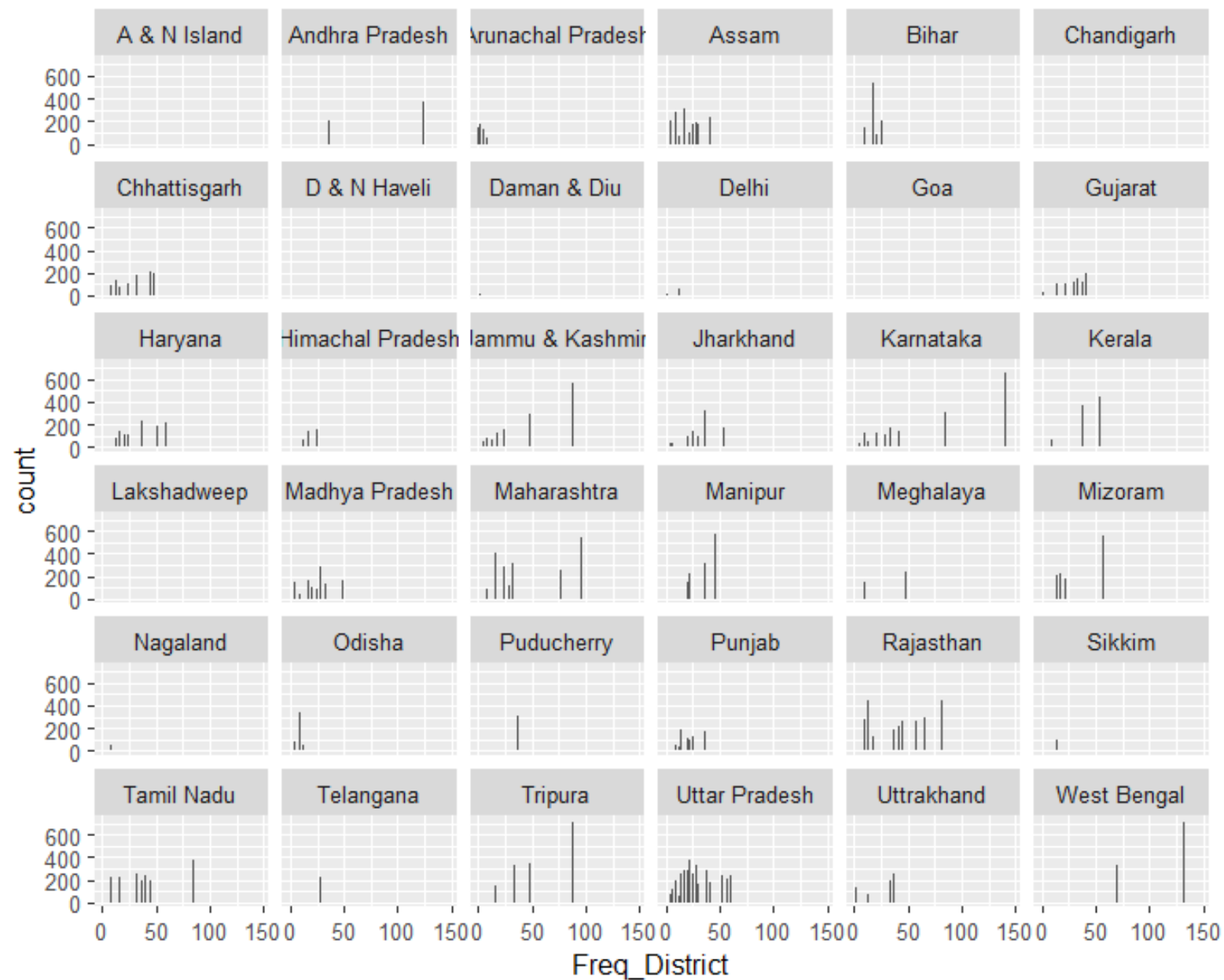
The hierarchical logistic regression model data with group structure and a binary response variable  $y_i = \{ \text{Employ (medwork), Unemploy} \}$ .

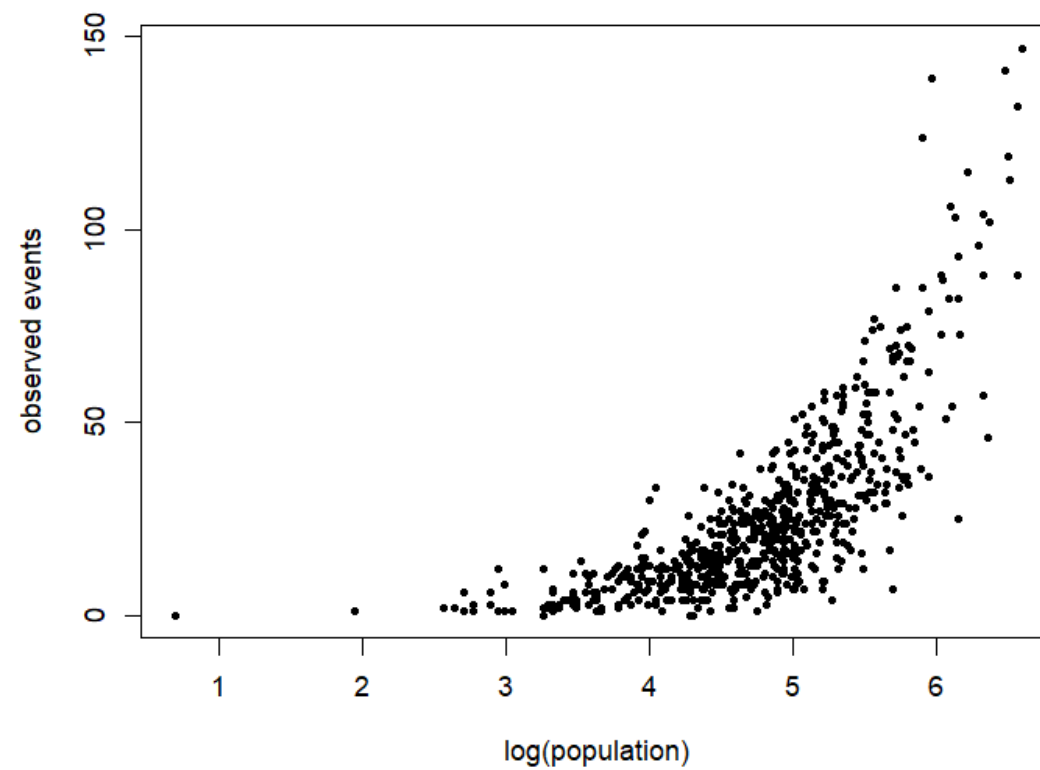
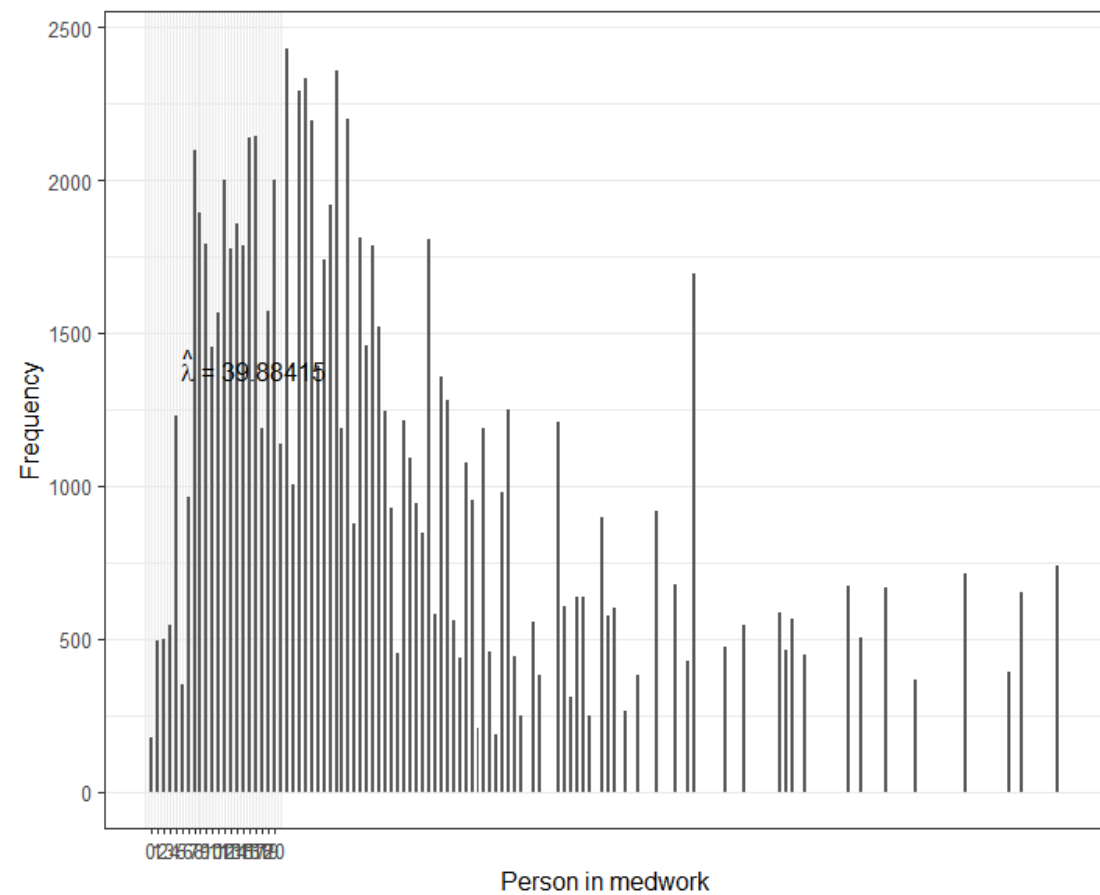


# Hierarchical Poisson regression models

District?

States?



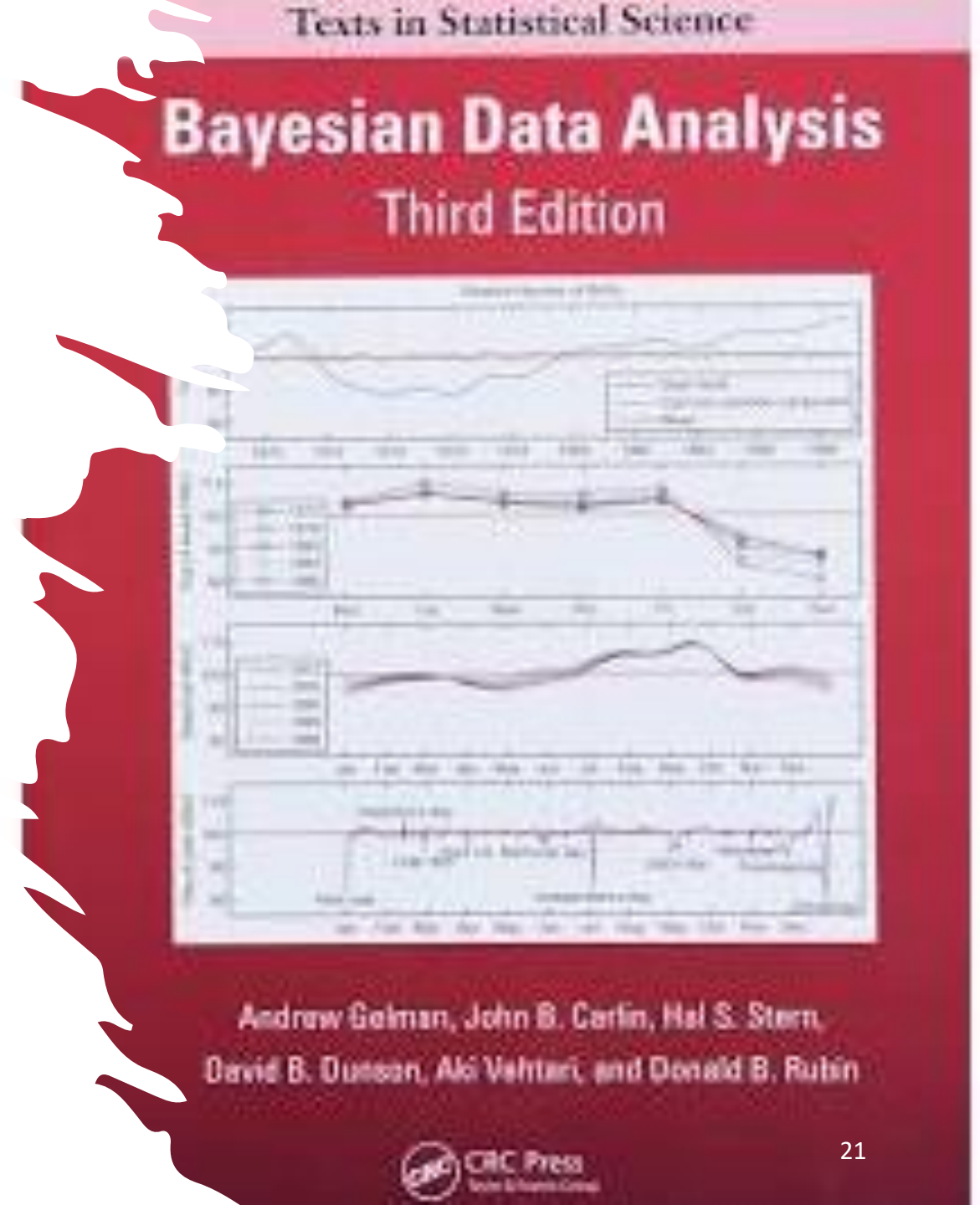


# Case Studies - References

- Kim, Jihye, Olsen, W.K. and Arkadiusz Wiśniowski (2020), A Bayesian Estimation of Child Labour in India, Child Indicators Research, DOI <https://doi.org/10.1007/s12187-020-09740-w>. Online 8 June.
- Olsen, Wendy, Manasi Bera, Amaresh Dubey, Jihye Kim, Arkadiusz Wiśniowski, Purva Yadav (2020). Hierarchical Modelling of COVID-19 Death Risk in India in the Early Phase of the Pandemic, European Journal of Development Research. DOI <https://link.springer.com/article/10.1057/s41287-020-00333-5>, 15 December.
- Raymer, James, & Arkadiusz Wiśniowski (2018) Applying and testing a forecasting model for age and sex patterns of immigration and emigration, *Population Studies*, 72:3, 339-355, DOI: 10.1080/00324728.2018.1469784

# References

- Casella, G., and R.L. Berger, 1990, *Statistical Inference*, Wadsworth, to review maximum likelihood, PDFs, CDFs and various tests.
- Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin, 2013, *Bayesian Data Analysis*, 3<sup>rd</sup> ed., London: CRC Press and Taylor & Francis, Chapman and Hall. Series: Texts in Statistical Science.
- Gelman, A., 2004, "Parameterization and Bayesian Modeling", *Journal of the American Statistical Association*, 99 537-545.
- Crawley, M.J., 2013, *The R Book*, 2<sup>nd</sup> ed., London: Wiley. (Maximum likelihood is covered in Chs 7 and 9, and regression ch 10, Bayesian statistics Ch 22, including BUGS and JAGS in R)



# References

- Gelman and Hill, 2007, *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge University Press.
- Bürkner, P.-C. (2017b). brms: An R package for bayesian multilevel models using Stan. 594 Journal of Statistical Software, 80 (1), 1–28.  
<https://doi.org/10.18637/jss.v080.i01>

