Bayesian statistics

Hierarchical/multilevel regression

Leonardo Egidi

2024/2025

Università di Trieste

Table of contents i

The hierarchical/multilevel framework

Multilevel structures

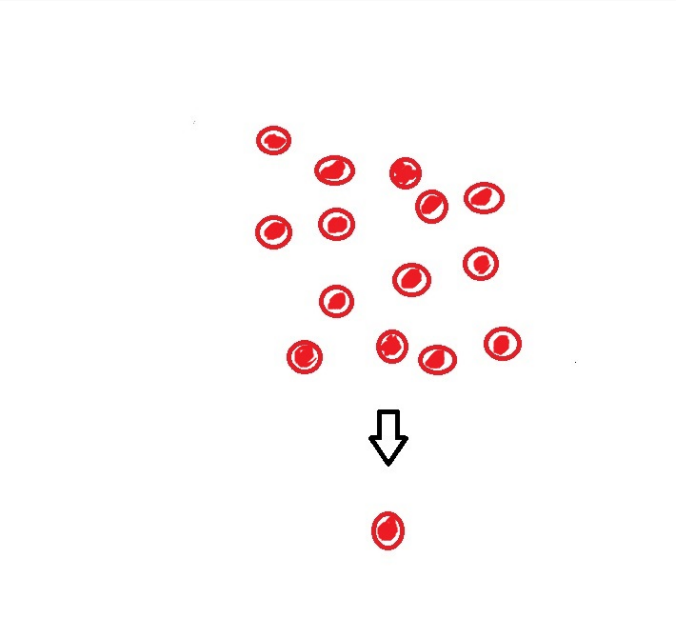
- Hierarchical/Multilevel models are extensions of regression in which data are structured in groups and coefficients can vary by group.
- Example of multilevel structures:
  - Simple grouped data—persons within cities—where some information is available on persons and some information is at the city level.
  - Repeated measurements.
  - Time-series cross sections.
  - Non-nested structures.

Motivations i

- A common problem in applied statistics is modeling individuals/objects of a *population*.
- Within this population, there may be some *subpopulations* sharing some common features. Thus, we should statistically acknowledge for this distinct groups' membership.
- Multilevel/hierarchical models are extensions of regression models in which data are structured in groups and coefficients can vary by group. We start with simple grouped structures—such as people within cities, students within schools, etc—where some information is available on individuals and some information is at the group level.

Motivations ii

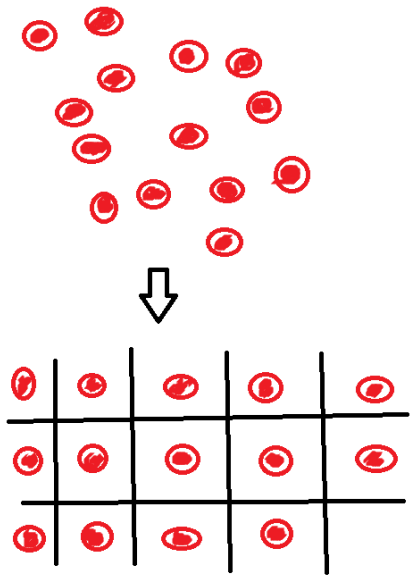
If we assume that every individual is equivalent then we can pool the data, but only at the expense of bias $\Leftrightarrow$ Complete pooling.



$$y_i \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2)$$

Motivations iii

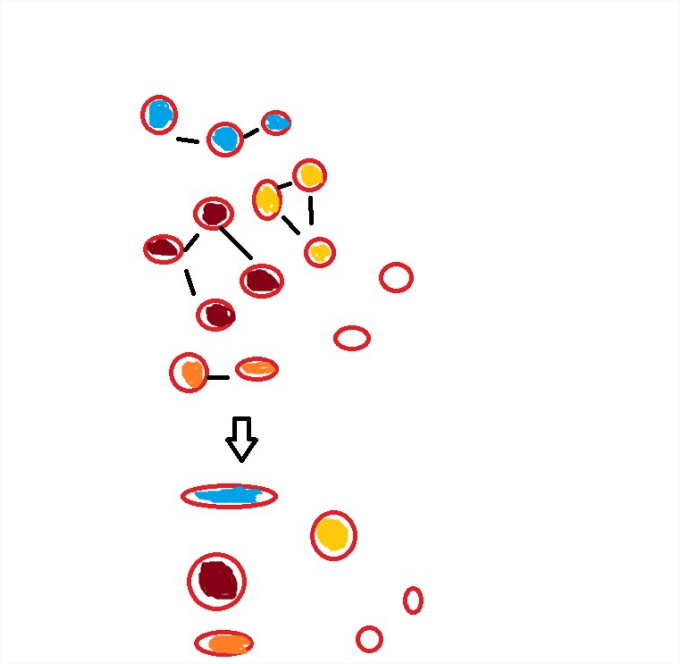
Conversely, modelling every individual separately avoids any bias, but then the data becomes very sparse and inferences weak $\Leftrightarrow$ No pooling.



$$y_i \sim \mathcal{N}(\alpha_i + \beta x_i, \sigma^2)$$

Motivations iv

A compromise between complete pooling and no pooling that could balance bias and variance would be ideal. Thus, hierarchical models allow for this:



$$y_{ij} \sim \mathcal{N}(\alpha_{j(i)} + \beta x_i, \sigma^2)$$

Motivations v

- The common feature of such models is that the observed units y_{ij} are indexed by the statistical unit i in group j (examples: *students within schools*, *players within teams*). In general, these observable outcomes are modeled conditionally on certain *not observable* parameters θ_j, viewed as drawn from a population distribution, which themselves are given a probabilistic (prior) distribution in terms of further parameters, known as *hyperparameters*.
- Simple nonhierarchical models are usually inappropriate for hierarchical data: with few parameters, they generally cannot fit large datasets accurately.
- Conversely, hierarchical models can have enough parameters to fit the data well, while using a population distribution.

The fundamental concept of exchangeability i

- In order to formalize this approach we need to consider the concept of exchangeability, which turns out to be relevant in Bayesian statistics.
- Consider a set of experiments $j = 1, \ldots, J$, in which experiment j has data (vector) y_j and parameter vector θ_j, with likelihood $p(y_j|\theta_j)$. In the linear model, we have $\theta = (\alpha, \beta, \sigma^2)$
- If no information-other than the data y-is available to distinguish any of the θ_j's from any of the others, and no ordering or grouping of the parameters can be made, one must assume symmetry among the parameters in their prior distribution.

The fundamental concept of exchangeability ii

- This symmetry is represented probabilistically by **exchangeability**: the parameters $(\theta_1, \ldots, \theta_J)$ are exchangeable in their joint prior distribution if $\pi(\theta_1, \ldots, \theta_J)$ is invariant to permutations of the indexes $(1, \ldots, J)$.
- In practice, ignorance implies exchangeabilitiy. Consider the analogy to a roll of a dice: we should initially assign equal probabilities to all six outcomes, but if we study the measurements of the dice and weigh the dice carefully, we might eventually notice imperfections, which might make us favour one outcome over the others and thus eliminate the symmetry among the six outcomes.

The fundamental concept of exchangeability iii

- The simplest form of an *exchangeable distribution* has each of the parameters θ_j as an independent sample from a prior (or population) distribution governed by some unknown parameter vector ϕ; thus,

$$\pi(\theta|\phi) = \prod_{j=1}^{J} \pi(\theta_j|\phi). \tag{1}$$

- In general, ϕ is unknown, so our distribution for θ must average over our uncertainty in ϕ:

$$\pi(\theta) = \int \left(\prod_{j=1}^{J} \pi(\theta_j|\phi) \right) \pi(\phi) d\phi. \tag{2}$$

The fundamental concept of exchangeability iv

- In such a way, the joint distribution for y and θ becomes:

$$p(\theta, y) = \prod_{i=1}^{n} p(y_{ij}|\theta_{j(i)})\pi(\theta_{j(i)}|\phi)\pi(\phi), \tag{3}$$

with the nested index $j(i)$ denoting the group membership of the i-th unit, whereas the joint posterior distribution for θ, ϕ is:

$$\pi(\theta, \phi|y) \propto \pi(\phi, \theta)p(y|\theta). \tag{4}$$

- Careful! ϕ is usually not known. Thus, the joint prior distribution $\pi(\phi, \theta)$ may be factorized as

$$\pi(\phi, \theta) = \pi(\phi)\pi(\theta|\phi),$$

where $\pi(\phi)$ is the *hyperprior* distribution.

Example: FdI voters. Role of exchangeability in inference i

FdI voters

Suppose you are an asian guy and let $\theta_1, \ldots, \theta_5$ are the proportions of voters for the party Fratelli d'Italia (FdI) in five Italian regions from the last polls for the next European Elections. The regions, here in a random order, are: Piemonte, Liguria, Umbria, Puglia, Lazio. What can you say about the FdI vote proportion θ_5, in the fifth region?

Since you have no information to distinguish any of the five regions from the others, you must model them exchangeably. You might use a Beta distribution for the five θ_j's, or some other distributions restricted in $[0, 1]$.

I now randomly sample four regions from these five and tell you the polls' proportions (in %): 23.2, 24.3, 18.4, 24.5. Remember, you are asian, you do not know anything about FdI and the Italian politics...what can you say about θ_5?

Example: FdI voters. Role of exchangeability in inference ii

Changing the indexing does not change the joint prior distribution. θ_j are exchangeable, *but they are not independent* as we assume that the voters' proportion θ_5 is probably similar to the observed rates.

However, today you come in Italy for a two-weeks holiday and you start reading *Il Fatto Quotidiano*, *La Repubblica*, *Il Giornale*, *Libero*. Mmh...what a weird nation is Italy! You are getting information.

You reconsider the four voters' proportions. You know that Giorgia Meloni, the FdI leader and the actual Italian Prime Minister, is born in Roma, Lazio, a region headed by Francesco Rocca, supported by the right-parties as well. Maybe the missing proportion θ_5 represents Lazio, where FdI is very strong...You end up with a non-exchangeable prior distribution.

Take-home message: the more you know, the more informative (then, less exchangeable) should be your prior distribution! However. exchangeability is a very good starting point...

Hierarchical models: formalization

Often observations (and/or parameters) are not fully exchangeable, but are *partially* or *conditionally* exchangeable.

- If observations can be grouped, we may make hierarchical modelling, where each group has its own subgroup, but the group properties are unknown.
- If y_i has additional information x_i so that y_i are not exchangeable but (y_i, x_i) still are exchangeable, then we can make a joint model for (y_i, x_i) or a conditional model for $y_i|x_i$.

In general, the usual way to model exchangeability with covariates is through conditional independence:

$$\pi(\theta_1, \ldots, \theta_J | x_1, \ldots, x_J) = \int \left[\prod_{j=1}^{J} \pi(\theta_j | \phi, x_j) \right] \pi(\phi | x) d\phi$$

Hierarchical models: objections to exchangeability

- In virtually any statistical application, it is natural to object to exchangeability on the grounds that the units actually differ.
- That the units differ, implies that the θ_j's differ, but it might be perfectly acceptable to consider them as if drawn from a common distribution.
- As usual in regression, the valid concern is not about exchangeability, but about encoding relevant knowledge as explanatory variables where possible.

Hierarchical models: formalization

We may try to formalize a hierarchical model by acknowledging at least two levels:

- individual level: observed $y_{ij},\ i = 1, \ldots, n,\ j = 1, \ldots J$;

$$y_{ij} \sim p(y|\theta_j) \quad \text{likelihood}$$

- group level: unobserved θ_j, $j = 1, \ldots, J$, depending on an hyperparameter ϕ.

$$\theta_j \sim \pi(\theta|\phi) \quad \text{group-level model}$$

- heterogeneity level: only in the Bayesian framework, we could model the unobserved ϕ

$$\phi \sim \pi(\phi) \quad \text{hyperprior}$$

Hierarchical linear regression

Extending linear models

- Hierarchical regression models are useful as soon as there are predictors at different levels of variation. Some examples may be:
  - In studying scholastic achievement, we may have students within schools, with predictors both at the individual and at the group level.
  - Data obtained by stratified or cluster sampling
- With predictors at multiple levels, the assumption of exchangeability of units or subjects at the lowest level breaks down.
- We can think of a generalization of linear regression, where intercepts, and possibly slopes, are allowed to vary by group.
- A batch of J coefficients is assigned a model, and this group-level model is estimated simultaneously with the data-level regression of y.

The general hierarchical linear model i

- n observations in J groups.
- Within each group, a likelihood $p(y_{ij}|\theta_j)$ fro the individual units is defined.
- At the second stage, a group-level modeling distribution for $\pi(\theta_j|\phi)$ is required. Then, a varying-intercept, varying slope model takes the general form:

$$\begin{aligned} y_{ij} &\sim \mathcal{N}(\alpha_{j(i)} + x_{ij}\beta_{j(i)}, \sigma_y^2), \\ \begin{pmatrix} \alpha \\ \beta \end{pmatrix} &\sim \mathcal{N}\left(\begin{pmatrix} \mu_\alpha \\ \mu_\beta \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta^2 \end{pmatrix} \right), \end{aligned} \tag{5}$$

  where x_{ij} is a given covariate/predictor.
- $\mu_\alpha, \mu_\beta, \sigma_\alpha^2, \sigma_\beta^2$ are hyperparameters for which we require a hyperprior distribution if a Bayesian framework is assumed.

The general hierarchical linear model ii

- When more than two coefficients vary by group, we can write (5) in vector-matrix form as:

$$\begin{aligned} y_{ij} &\sim \mathcal{N}(X_j \beta_{j(i)}, \sigma_y^2), \\ \beta_j &\sim \mathcal{N}(\mu_\beta, \Sigma_\beta), \end{aligned} \tag{6}$$

  where Σ_β is a variance/covariance matrix for β_j.

- In a Bayesian framework, Σ_β needs to be assigned a prior distribution. Canonical and conjugate choice: inverse-Wishart (see BDA, 15.4).

Extending linear models: radon data

Radon data (G&H book, chapter 12)
Suppose to measure radon emissions in more than 80000 houses throughout US. Our goal in analyzing these data is to estimate the distribution of radon levels in each of the approximately 3000 counties, so that homeowners could make decisions about measuring or remediating the radon in their houses.

The data are structured *hierarchically*: houses within counties. As a predictor, we have the floor on which th measurement is taken, either basement or first floor; radon comes from underground and can enter more easily when a house is built into the ground. We fit a model where y_i is the logarithm of the radon measurement in house i, and x is the floor variable (0 if basement, 1 if first floor).

Partial pooling with no predictors i

- Hierarchical (or multilevel) modelling is a compromise between two extremes: complete pooling, in which the group indicators are not included in the model, and no pooling, in which separate models are fit within each group. For such a reason, we may refer to hierarchical modellling as partial pooling.
- We start our journey into hierarchical models with the simplest model ever for the radon data, a hierarchical linear model with no predictors:

$$\begin{aligned} y_{ij} &\sim \mathcal{N}(\alpha_{j(i)}, \sigma^2), \ i = 1, \ldots, n \quad \text{Individual level} \\ \alpha_j &\sim \mathcal{N}(\mu_\alpha, \tau^2), \ j = 1, \ldots, J \ \text{Group level} \end{aligned} \tag{7}$$

where $\alpha_{j(i)} = 1, \ldots, J$ is the intercept for the i-th unit, belonging to the j-th group.

Partial pooling with no predictors ii

- Consider the goal of estimating the distribution of radon levels of the houses within each of 85 counties in Minnesota. One estimate would be the average that completely pools data across all counties. This ignores variation among counties, however, so perhaps a better option would be simply to use the average log radon level in each county. Estimates $\pm$ standard errors are plotted against the number of observations in each county in the next plot, left panel.
- A third option is hierarchical modelling: estimates $\pm$ standard errors are plotted against the number of observations for each county.

Partial pooling with no predictors iii

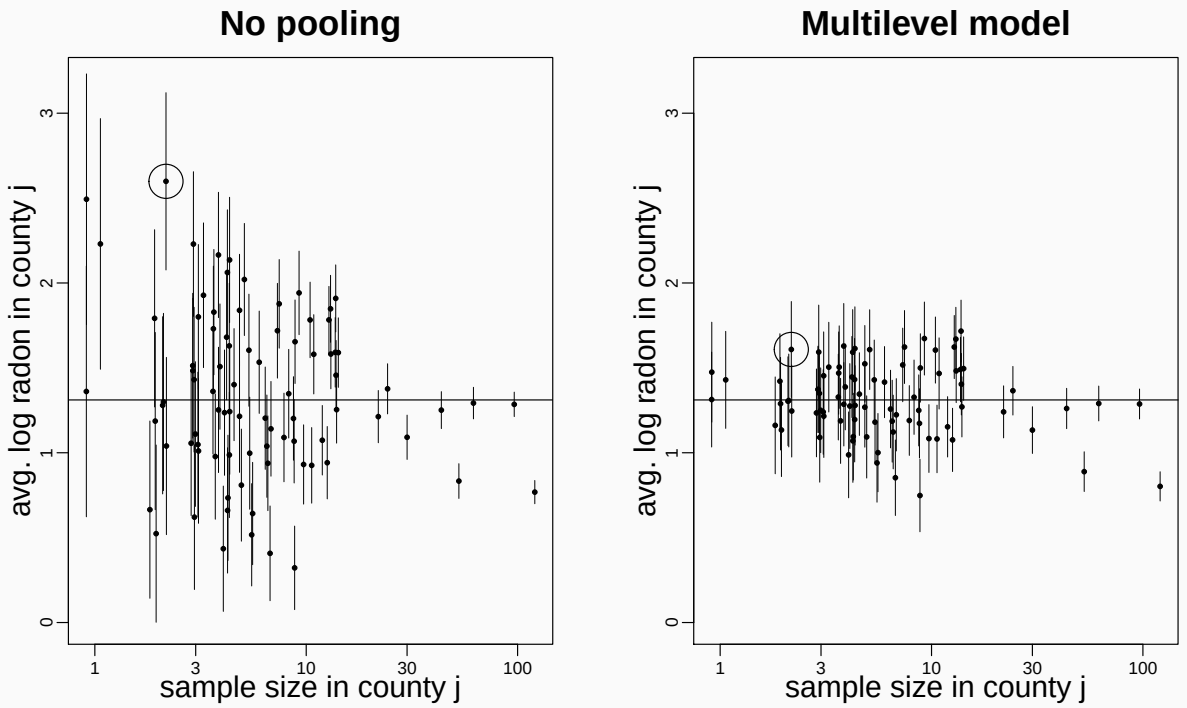


Figure 1: Estimates $\pm$ standard errors for the average log radon levels in Minnesota counties plotted versus the number of observations in the county.

Partial pooling with no predictors iv

- Whereas complete pooling ignores variation between counties, the no-pooling analysis overfits the data within each county.
- In no-pooling analysis, the counties with fewer measurements have more variable estimates and larger higher standard errors. It systematically causes us to think that certain counties are more extreme, just because they have smaller sample sizes!
- The hierarchical estimate for a given county j can be approximated as a weighted average:

$$\hat{\alpha}_j = \frac{\frac{n_j}{\sigma^2}\bar{y}_j + \frac{1}{\tau^2}\bar{y}_{\text{all}}}{\frac{n_j}{\sigma^2} + \frac{1}{\tau^2}} \qquad (8)$$

where n_j is the number of observations in the j-th county, $\bar{y}_j$ is the mean of the observations in the county (unpooled estimate), and $\bar{y}_{\text{all}}$ is the mean over all counties (completely pooled estimate).

Partial pooling with no predictors v

The weighted average (8) reflects the relative amount of information available about the individual county, on one hand, and the average of all counties, on the other:

- Averages from counties with smaller sample sizes carry less information (n_j small), and the weighting pulls the multilevel estimates closer to the overall state average. If $n_j = 0$, $\hat{\alpha}_j = \bar{y}_{\text{all}}$, the overall average.
- Averages from counties with larger sample sizes carry more information. As $n_j \to \infty, \hat{\alpha}_j = \bar{y}_j$, the county average.
- When variation across counties is very small, the weighting pulls the multilevel estimates to the overall mean: as $\tau^2 \to 0$, $\hat{\alpha}_j = \bar{y}_{\text{all}}$.
- When variation across the counties is large, the weighting pulls the multilevel estimates to the county average: as $\tau^2 \to \infty, \hat{\alpha}_j = \bar{y}_j$.

Partial pooling with predictors i

- The same principle of finding a compromise between these two extremes applies for more general models. We consider now the individual-level predictor x, where $x_i = 1$ for the first floor and $x_i = 0$ for the basement.
- Thus, the second model we consider is a *varying-intercept* model:

$$\begin{aligned} y_{ij} &\sim \mathcal{N}(\alpha_{j(i)} + \beta x_i, \sigma^2), \ i = 1, \ldots, n \quad \text{Individual level} \\ \alpha_j &\sim \mathcal{N}(\mu_\alpha, \tau^2), \ j = 1, \ldots, J \qquad \text{Group level} \end{aligned} \tag{9}$$

- To appreciate hierarchical modelling, we start plotting some estimates according to complete and no pooling.

Partial pooling with predictors ii

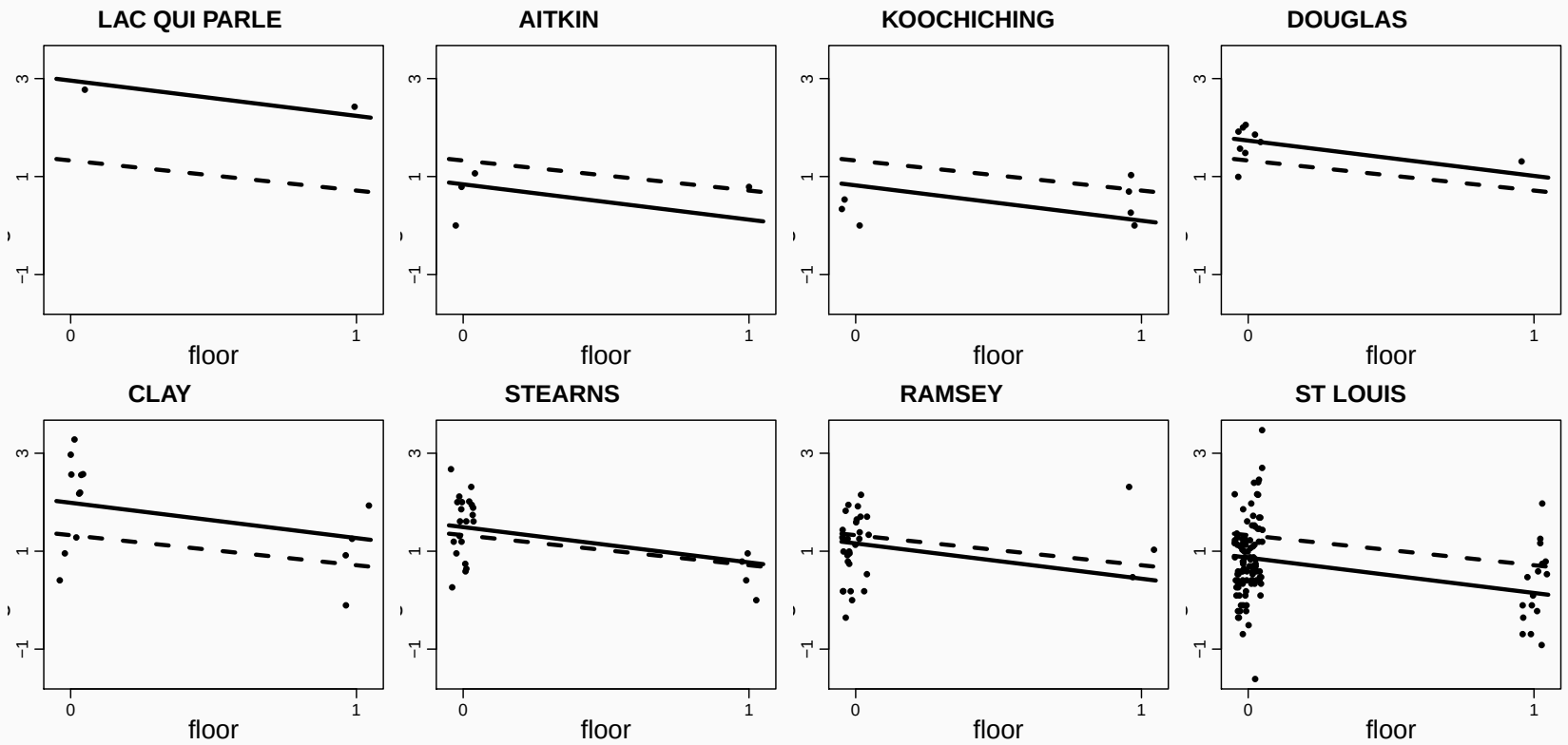


Figure 2: Complete pooling (dashed lines) and no pooling (solid lines) for 8 counties in Minnesota.

Partial pooling with predictors iii

Both these analysis have problems.

- The complete pooling analysis ignores any variation in average radon levels between counties.
- The no-pooling analysis has problems too, however, which we can see in Lac Qui Parle County, since the estimate is based on only two observations.

Let's fit now model (9) via the function `stan_lmer` of the `rstanarm` R package, and plot again the estimates.

Partial pooling with predictors iv

```
mlm.radon.pred <- stan_lmer(y ~ x+ (1|county))
print(mlm.radon.pred)
stan_lmer
 family:       gaussian [identity]
 formula:      y ~ x + (1 | county)
 observations: 919
------
            Median MAD_SD
(Intercept)  1.5    0.1
x           -0.7    0.1
```

Partial pooling with predictors v

```
Error terms:
 Groups   Name        Std.Dev.
 county   (Intercept) 0.33
 Residual             0.76
Num. levels: county 85
```

We obtain the following posterior estimates for the two sources of variation: $\hat{\tau} = 0.33, \hat{\sigma} = 0.76$.

Partial pooling with predictors vi

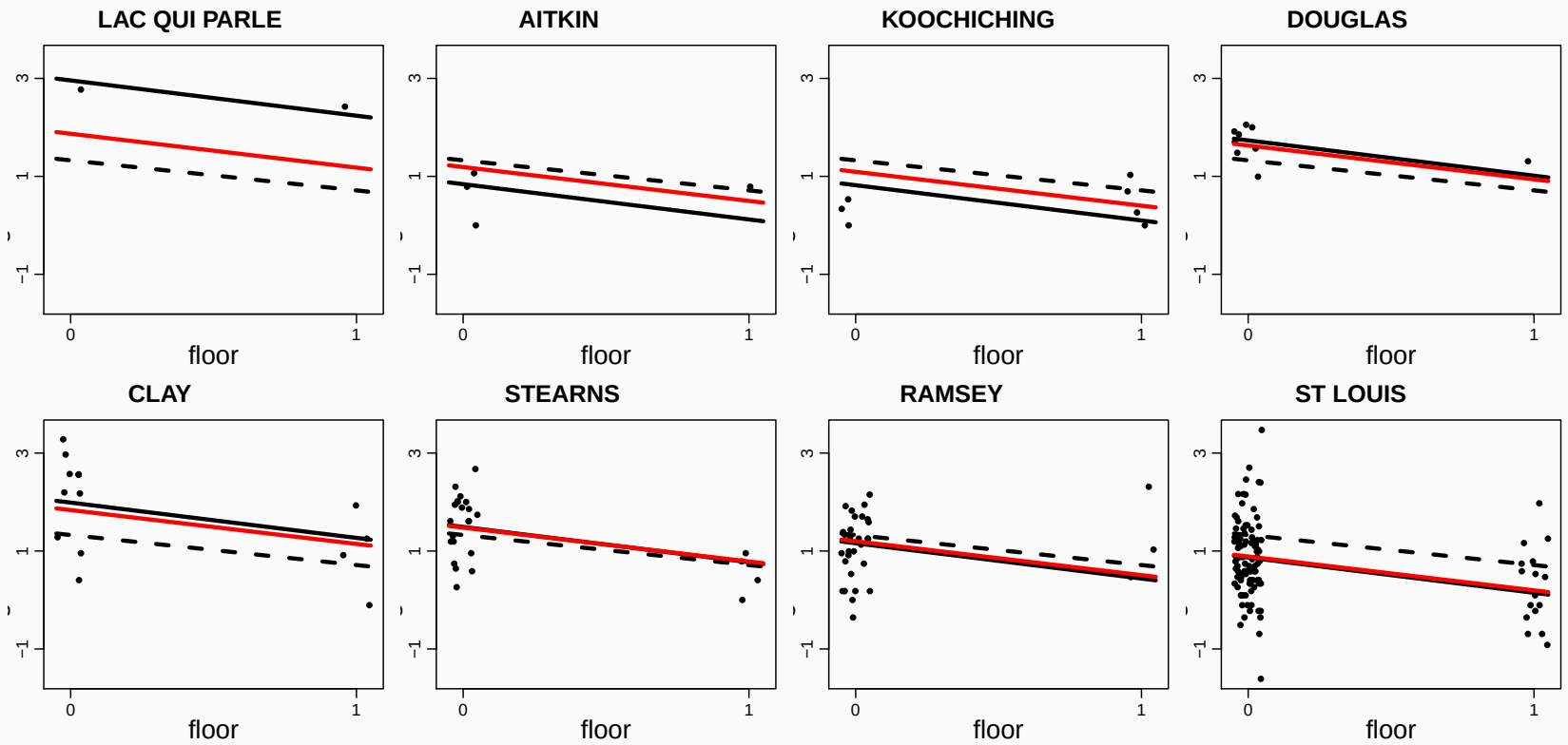


Figure 3: Complete pooling (dashed lines), no pooling (solid lines) and partial pooling (solid red lines).

Partial pooling with predictors vii

- The estimated line from the hierarchical model (9) in each county lies between the complete-pooling and no-pooling regression lines. There is strong pooling (solid red line closer to complete-pooling line) in counties with small sample sizes, and only weak pooling (solid red line close to no-pooling line) in counties containing many measurements.
- Classical regression models can be viewed as special cases of multilevel models. The limits $\tau \to 0$ (complete pooling) and $\tau \to \infty$ (no pooling) seem to be restrictive: given multilevel data, we can estimate τ, which acts as hyperparameter of a prior distribution on α.
- Note that the function `stan_lmer` works in the same way as the function `lmer` for classical inference. However, when the number of groups is small, it can be useful to switch to Bayesian inference, *to better account for uncertainty* in model fitting.

Partial pooling with predictors viii

We can generalize equation (8) as follows:

$$\hat{\alpha}_j \approx \frac{\frac{n_j}{\sigma^2}}{\frac{n_j}{\sigma^2} + \frac{1}{\tau_\alpha^2}}(\bar{y}_j - \beta\bar{x}_j) + \frac{\frac{1}{\tau_\alpha^2}}{\frac{n_j}{\sigma^2} + \frac{1}{\tau_\alpha^2}}\mu_\alpha, \tag{10}$$

a weighted average of the no-pooling estimate for its group $(\bar{y}_j - \beta\bar{x}_j)$ and the prior mean μ_α.

- Multilevel modeling partially pools the group-level parameters α_j toward their mean level, μ_α.
- There is more pooling when the group-level standard deviation τ is small.
- There is more smoothing for groups with fewer observations.

Partial pooling with predictors ix

We may disaggregate the information averaging over the counties, the *fixed* effects, and the county-level errors, the *random* effects, using the functions `fixef()` and `ranef()` of the `rstanarm` package:

```
fixef(mlm.radon.pred)
(Intercept)           x
  1.4623684  -0.6919822

ranef(mlm.radon.pred)
  $county
    (Intercept)
1  -0.264735142
2  -0.534511687
. . .
85 -0.073852110
```

The est. line for the first county is: $(1.46 - 0.26) - 0.69x = 1.20 - 0.69x$.

Hierarchical logistic regression

Multilevel modeling for GLMs

- Multilevel/hierarchical modeling is applied to logistic and probit regression and other generalized linear models (GLMs) in the same way as with linear regression: its coefficients are grouped into batches and a probability distribution is assigned to each batch.
- Also the computational tools to fit these models are basically the same as those used for multilevel linear regression.

1988 US polls i

1988 US polls (G&H book, chapter 14)
Dozens of national opinion polls are conducted by media organizations before every election, and it is desirable to estimate opinions at the levels of individual states as well as for the entire country. These polls are generally based on national random-digit dialing with corrections for non-response based on demographic factors such as sex, ethnicity, age, and education. We choose a single outcome—the probability that a respondent prefers the Republican candidate Bush against the democrat Dukakis for president—as estimated by a logistic regression model from a set of seven CBS News polls conducted during the week before the 1988 presidential election.

1988 US polls ii

- The aim is to fit a regression model for the individual response y given demographics and state. An average response θ_ℓ for each cross-classification ℓ of demographics and state is estimated. In this dataset we have sex (male or female), ethnicity (African American or other), age, education (4 categories each), and 51 states, for $\ell = 1, \ldots, L$=3264 categories.
- From the US census, we look up the adult population N_ℓ for each category ℓ. The estimated population average of the response y in any state j is then: $\theta_j = \sum_{\ell \in j} N_\ell \theta_l / \sum_{\ell \in j} N_\ell$, with each summation over the 64 demographic categories ℓ in the state. This weighting by population totals is called poststratification.
- We need many categories because (a) we are interested in estimates for individual states, and (b) non-response adjustments force us to include the demographics. As a result, any given survey will have few or no data in many categories. This is not a problem, however, if a multilevel model is fitted. Each factor or set of interactions in the model is automatically given a variance component.

1988 US polls. Varying-intercept model i

- We fit a simple model version by including two individual predictors, sex (`female`) and ethnicity (`black`):

$$\begin{aligned} \Pr(y_i = 1) &= \text{logit}^{-1}(\alpha_{j(i)} + \beta^{\text{female}}\text{female}_i + \beta^{\text{black}}\text{black}_i), \\ \alpha_j &\sim \mathcal{N}(\mu_\alpha, \tau^2_{\text{state}}),\ j = 1, \ldots, 51 \end{aligned} \tag{11}$$

  where $j(i)$ is the state index, and τ_α captures the between-state variability.

- We fit the model according to (a) a maximum likelihood approach through the function `glmer` of the `lme4` package, and (b) a Bayesian approach by using the `R` package `rstanarm` (function `stan_glmer`), relying on Hamiltonian Monte Carlo (HMC) sampling from the posterior distribution.

1988 US polls. Varying-intercept model ii

```
# frequentist fit

library(lme4)
M1 <- glmer (y ~ black + female + (1 | state),
                            family=binomial(link="logit"))
display(M1)

glmer(formula = y ~ black + female + (1 | state),
                family = binomial(link = "logit"))
            coef.est coef.se
(Intercept)  0.45     0.10
black       -1.74     0.21
female      -0.10     0.10

Error terms:
 Groups   Name        Std.Dev.
 state    (Intercept) 0.41
 Residual             1.00
---
number of obs: 2015, groups: state, 49
AIC = 2666.7, DIC = 2531.5
deviance = 2595.1
```

1988 US polls. Varying-intercept model iii

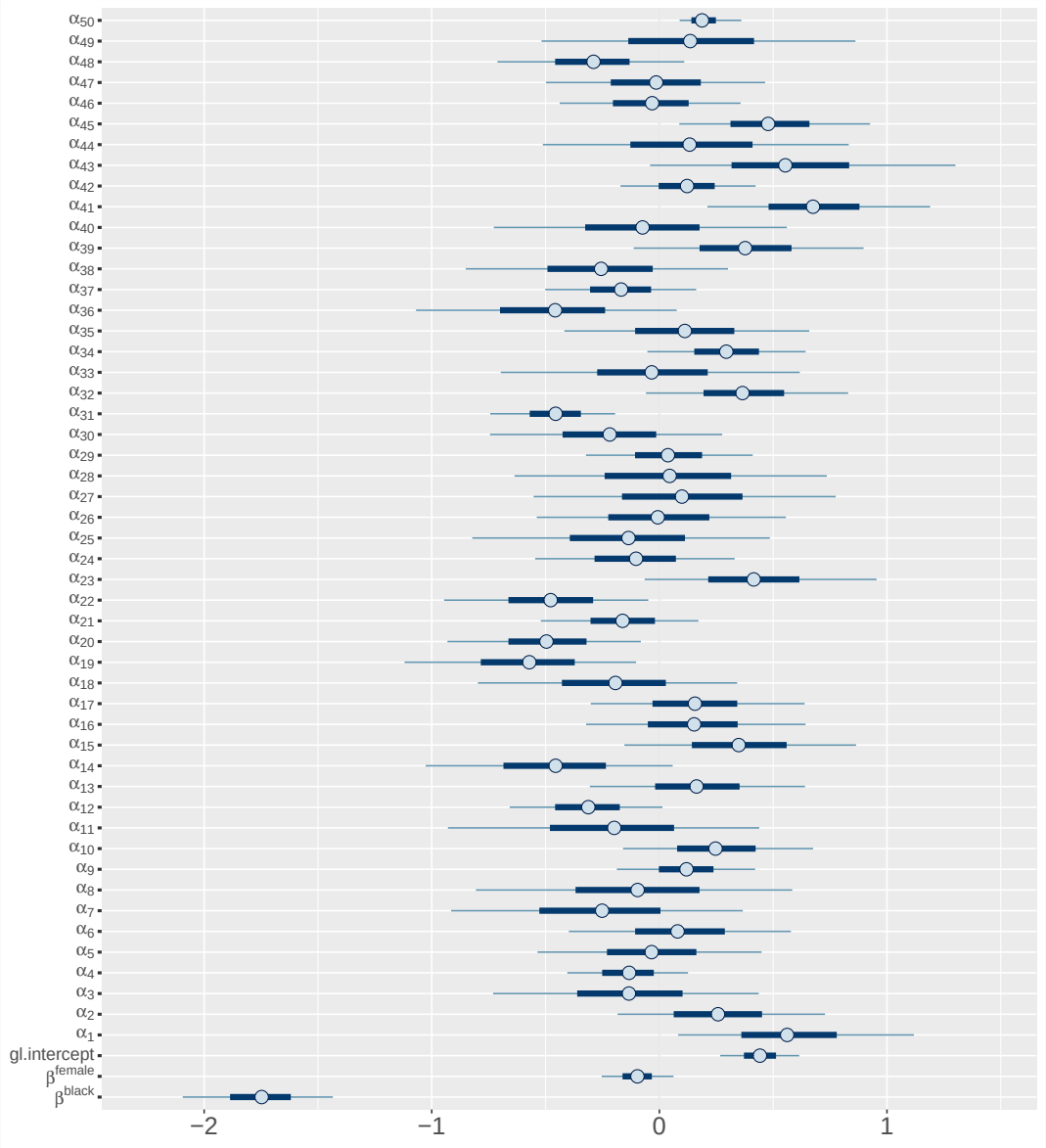
```
# Bayesian fit

library(rstanarm)
M1.rstanarm <- stan_glmer (y ~ black + female + (1 | state),
                           family=binomial(link="logit"))
print(M1.rstanarm)

stan_glmer
 family:       binomial [logit]
 formula:      y ~ black + female + (1 | state)
 observations: 2015
------
            Median MAD_SD
(Intercept)  0.4    0.1
black       -1.7    0.2
female      -0.1    0.1

Error terms:
 Groups Name        Std.Dev.
 state  (Intercept) 0.45
Num. levels: state 49
```

1988 US polls. Varying-intercept model iv

- The syntax `(1| state)` allows to include varying intercepts at the state level.
- The top part display gives the estimate of the average intercept (μ_α), the coefficients for `black` and `female`, and their standard errors.
- The between-state variation is estimated at $\hat{\tau}_{\text{state}} = 0.41$ under the frequentist approach, and 0.45 under the Bayesian approach. There is no residual standard deviation (which instead is given in the linear regression) because the logistic regression model does not have such a parameter. Finally, the model has an overdispersion of 1.0 (see `residual` in the first fit), because logistic regression with binary data cannot be overdispersed. The summary for the frequentist fit also reports the AIC, the DIC, and the model's deviance.
- The two procedures yield very similar results.

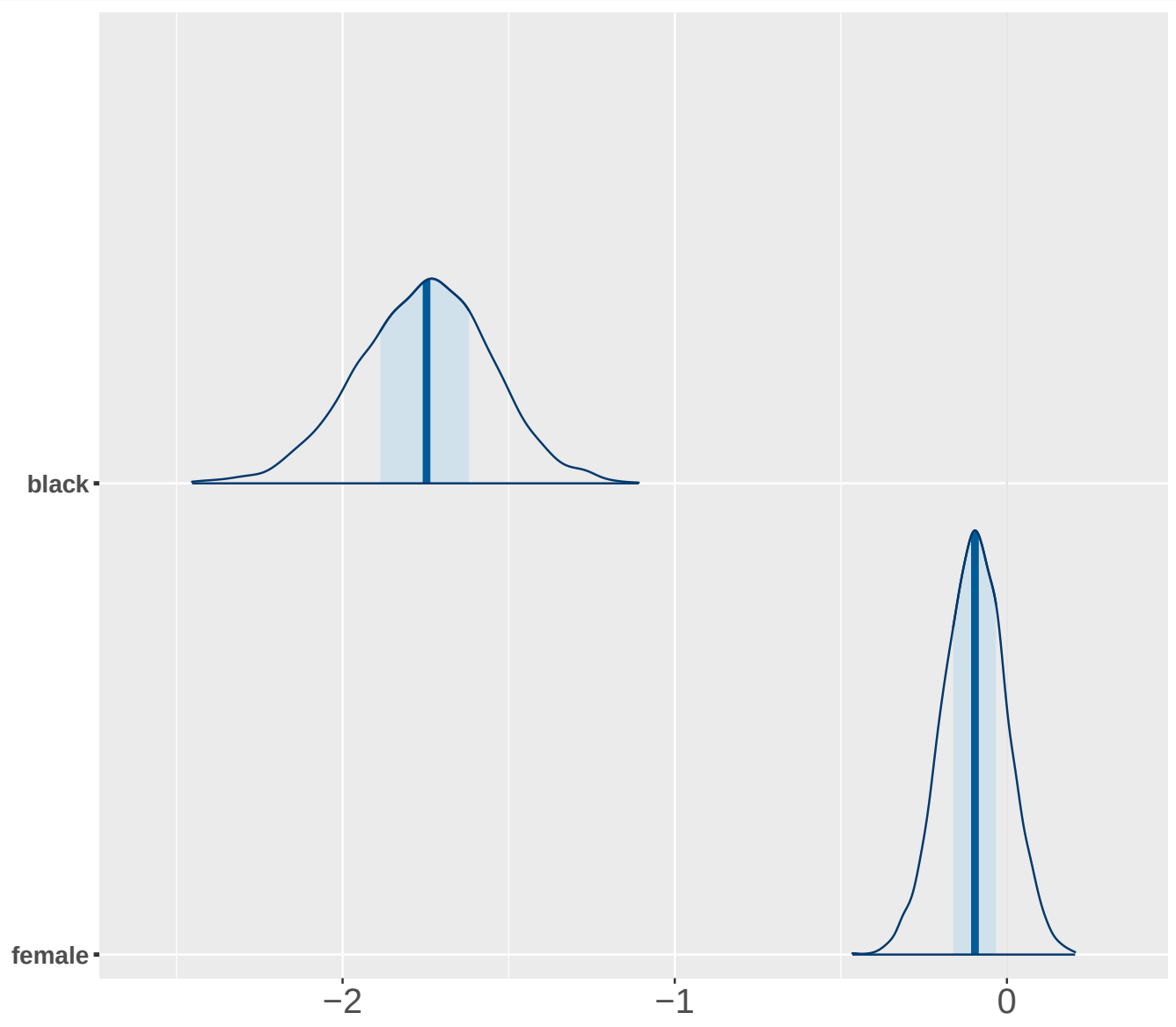
1988 US polls. Varying-intercept model v

From the Bayesan fit: 50% and 95% credible intervals for all the parameters.



1988 US polls. Varying-intercept model vi

From the Bayesian fit: posterior marginal densities along with 50% intervals for the 'fixed-effects' β^{black} and β^{female}.



1988 US polls. Varying-intercept model vii

Parameters' interpretation (for the Bayesian fit):

- The coefficient β^{black} reports a posterior estimate of -1.7: `black` is a categorical variable (coded as 1 for black people, 0 otherwise). A difference of 1 unit in this predictor has a linear effect of -1.7 on the logit probability of supporting Bush. In terms of odds ratios, being black gives an odds ratio of $\exp(-1.7) \approx 0.18$, causing a decrease in the odds of approximately 0.82 (82%).
- The coefficient β^{female} is estimated at -0.1. `female` is a categorical predictor (1 for women, 0 otherwise). Being a woman has an effect of -0.1 on the logit probability of supporting Bush. OR interpretation: $\exp(-0.1) \approx 0.9$, decrease in the odds of approx. 10%.

Be aware: understanding and interpreting model estimates is the first step! Ask, ask, ask yourself whether your estimates make sense...

1988 US polls. Varying-intercept model viii

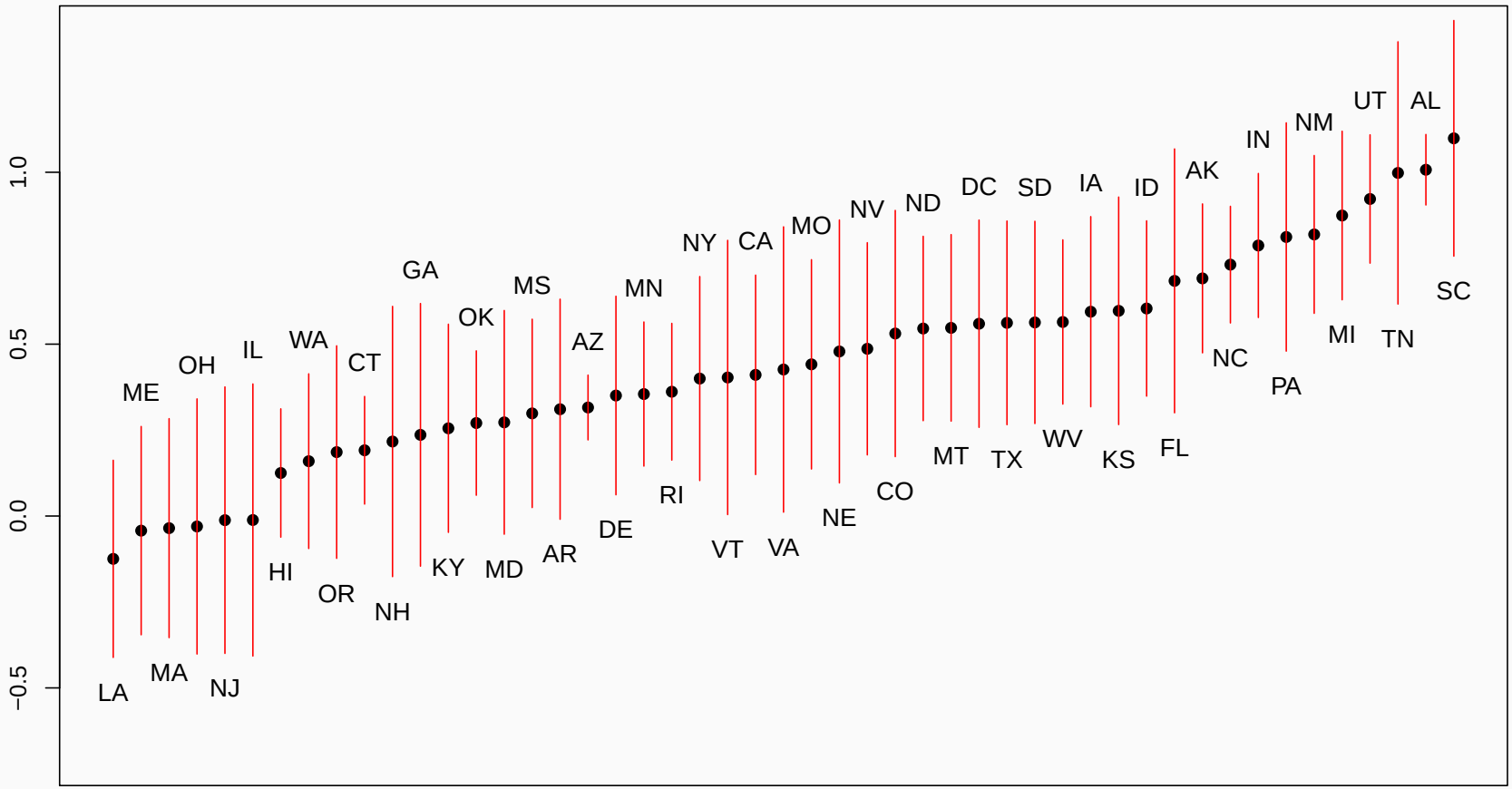
Many issues arise when you fit a model:

- Interpret your results. Do they make sense?
- Produce some plots for your estimates.
- Check your model. Is your model plausible, according to the data that you have? To be continued...
- Augment your model, if necessary: predictors, random effects,etc.
- Compare your model with other competing models. Is your model better than the others? Use AIC, DICC, LOOIC...To be continued...
- Use your model to make predictions.

Being a modeller represents a compromise between a mathematician and an artist. You can tremble between these two extremes.

1988 US polls. Varying-intercept model ix

'Random effects' α for the states: post. means $\pm$ s.e.



States

1988 US polls. Varying-intercept and slope model i

- We could ask ourself: is also the slope for the female varying in some states? Maybe, the women preference for Bush in Alabama is rather different than the same support in New Jersey...
- We propose a second model, a *varying-intercept and slope model*:

$$\Pr(y_i = 1) = \text{logit}^{-1}(\alpha_{j(i)} + \beta^{\text{female}}_{j(i)}\text{female}_i + \beta^{\text{black}}\text{black}_i), \quad i = 1, \ldots, n$$

$$\begin{pmatrix} \alpha_j \\ \beta^{\text{female}}_j \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_\alpha \\ \mu_\beta \end{pmatrix}, \begin{pmatrix} \tau^2_\alpha & \rho\tau_\alpha\tau_\beta \\ \rho\tau_\alpha\tau_\beta & \tau^2_\beta \end{pmatrix} \right), \quad j = 1, \ldots, 51, \tag{12}$$

where τ^2_α and τ^2_β are the variances for the intercepts and the slopes, repsectively, and ρ is the correlation coefficients between α and β.

1988 US polls. Varying-intercept and slope model ii

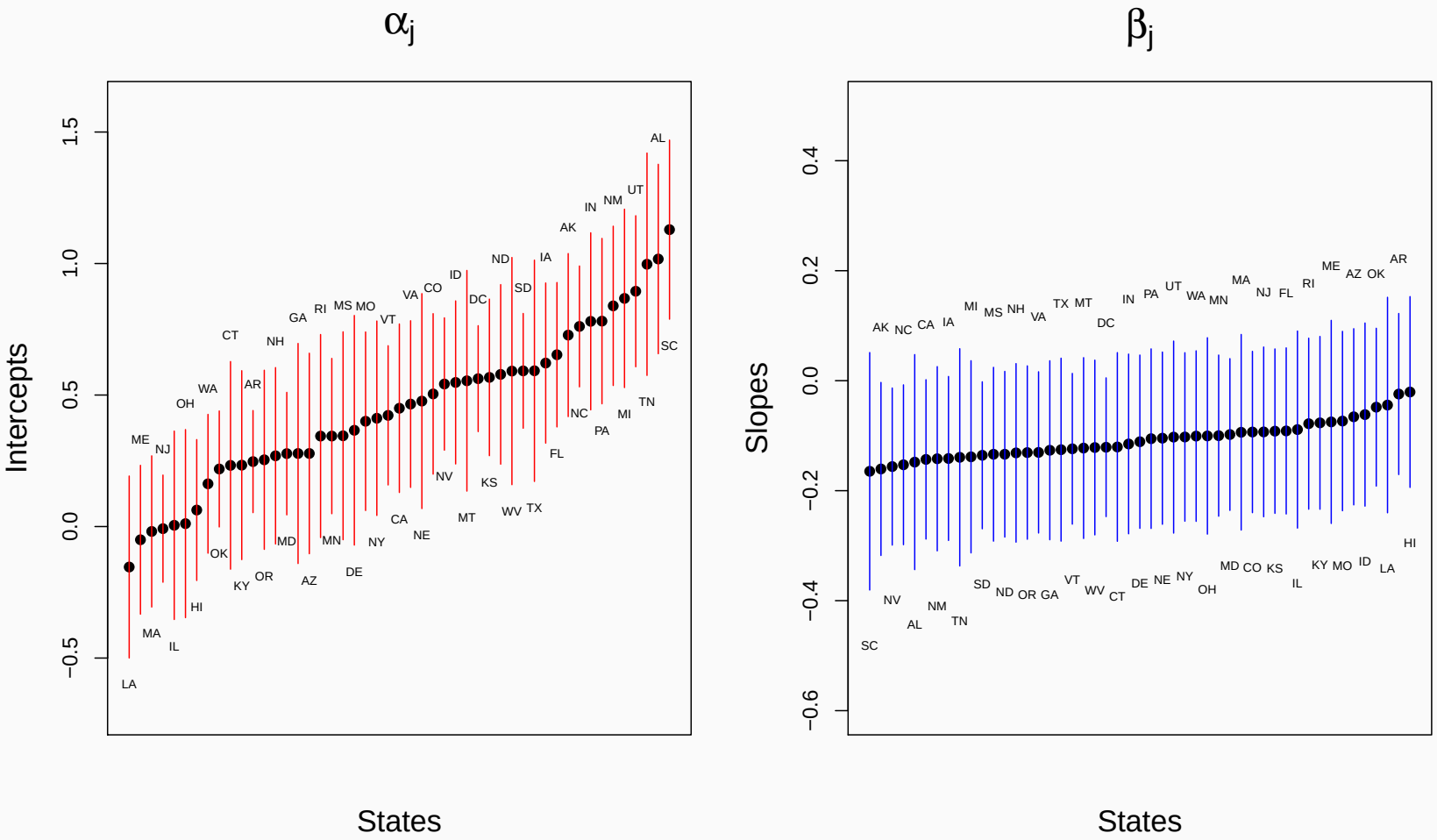
```
# Bayesian fit

M2.rstanarm <- stan_glmer (y ~ black + female + (1+ female | state),
             family=binomial(link="logit"))
print(M2.rstanarm)
stan_glmer
 family:       binomial [logit]
 formula:      y ~ black + female + (1 + female | state)
 observations: 2015
------
            Median MAD_SD
(Intercept)  0.5    0.1
black       -1.7    0.2
female      -0.1    0.1

Error terms:
 Groups Name        Std.Dev. Corr
 state  (Intercept) 0.47
        female      0.23     -0.40
```

1988 US polls. Varying-intercept and slope model iii

Parameters' interpretation:

- $\hat{\tau}_\alpha = 0.47$, the variation between the β^{female}, $\hat{\tau}_\beta$, is 0.23, whereas $\hat{\rho} = -0.4$. Thus, there is negative correlation between the states and the female effects.
- Other parameters are almost unchanged with respect to the varying-intercept model.

1988 US polls. Varying-intercept and slope model iv



Model comparison i

- We should start assessing the goodness of fit of our models. In Bayesian inference, the main tools to compare models are the penalized likelihood criteria: AIC, DIC, BIC,...
- We consider here also an extension of AIC based on cross validation, LOOIC, available via the `loo` package.
- The meaning is the same: the lower is the value of one among these criteria, and the better is the model fit.

Model comparison ii

```
# Bayesian fits
lpd1 <- log_lik(M1.rstanarm)
loo1 <- loo(lpd1)
lpd2 <- log_lik(M2.rstanarm)
loo2 <- loo(lpd2)
c(loo1$looic, loo2$looic)

[1] 2649.373 2651.668

# frequentist fits
d1 <- display(M1)
d2 <- diaplay(M2)
c(d1$AIC, d2$AIC)

[1] 2666.66 2668.721
```

- The varying-intercept and slope model fit is not better than the fit of the varying intercept model, in both the fitting procedures. The simpler the better (Occam rasor)!
- We could try to extend our model and, eventually, increase the goodness of fit (to be continued).

A fuller model including non-nested factors

- Finally, we expand the previous models to use all the demographic predictors in the CBS weighting, including the interactions sex $\times$ ethnicity and age $\times$ education. At the state level, we include indicators for the 5 regions (Northeast, Midwest, South, West, and District of Columbia, considered as a separate region because of its distinctive voting patterns) along with $\mathrm{v.prev}$, a measure of the previous Republican vote in the state. Then, a multilevel logistic regression including the four categorical predictors (sex, ethnicity, age, and education), along with the 51 states memberships and the 5 regions is provided:

$$
\begin{aligned}
\Pr(y_i = 1) = \mathrm{logit}^{-1}(&\beta_0 + \beta^{\mathrm{female}}\mathrm{female}_i + \beta^{\mathrm{black}}\mathrm{black}_i + \\
&+ \beta^{\mathrm{female.black}}\mathrm{female}_i \cdot \mathrm{black}_i + \alpha^{\mathrm{age}}_{k(i)} + \alpha^{\mathrm{edu}}_{l(i)} + \alpha^{\mathrm{age.edu}}_{k(i),l(i)} + \alpha^{\mathrm{state}}_{j(i)}), \\
\alpha^{\mathrm{state}}_j &\sim \mathcal{N}(\alpha^{\mathrm{region}}_{m(j)} + \beta^{\mathrm{v.prev}}\mathrm{v.prev}_j, \sigma^2_{\mathrm{state}}),\ j = 1, \ldots, 51 \\
\alpha^{\mathrm{age}}_k &\sim \mathcal{N}(0, \sigma^2_{\mathrm{age}}),\ k = 1, \ldots, 4 \\
\alpha^{\mathrm{edu}}_l &\sim \mathcal{N}(0, \sigma^2_{\mathrm{edu}}),\ l = 1, \ldots, 4 \\
\alpha^{\mathrm{age.edu}}_{k,l} &\sim \mathcal{N}(0, \sigma^2_{\mathrm{age.edu}}),\ k = 1, \ldots, 4,\ ł = 1, \ldots, 4 \\
\alpha^{\mathrm{region}}_m &\sim \mathcal{N}(0, \sigma^2_{\mathrm{region}}),\ m = 1, \ldots, 5.
\end{aligned}
\tag{13}
$$

Hierarchical Poisson regression

Hierarchical Poisson regression

- As with linear and logistic regression, generalized linear models can be fit to multilevel structures by including coefficients for group indicators and then adding group-level models.
- In modeling discrete data, such as counts, we need to take into account overdispersion and measures of exposures.

Hierarchical Poisson regression

- Data that are fit by a GLM are *overdispersed* if the data-level variance is higher than would be predicted by the model. Binomial and Poisson models are subject to overdispersion because they do not have variance parameters to capture the variance in the data.
- However, overdispersion can be directly modeled using a data-level variance component in a multilevel model. Consider a measure of exposure u_i, such that $\log(u_i)$ is the *offset*, then:

$$\begin{aligned} \text{Poisson regression}: \quad y_i &\sim \text{Poisson}(u_i e^{X_i\beta}), \\ \text{overdispersed Poisson regression}: \quad y_i &\sim \text{Poisson}(u_i e^{X_i\beta+\epsilon_i}) \\ \epsilon_i &\sim \mathcal{N}(0, \sigma_\epsilon^2). \end{aligned} \tag{14}$$

The new parameter σ_ϵ measures the amount of overdispersion, with $\sigma_\epsilon = 0$ corresponding to the classical Poisson regression.

Discrete data regression: cockroaches data i

Cockroaches data
A company that owns many residential buildings throughout New York City tells that they are concerned about the number of cockroach complaints that they receive from their 10 buildings in 12 months. They provide you some data collected in an entire year for each of the buildings and ask you to build a model for predicting the number of complaints over the next months and to understand which and how many of the available covariates could explain the number of complaints.

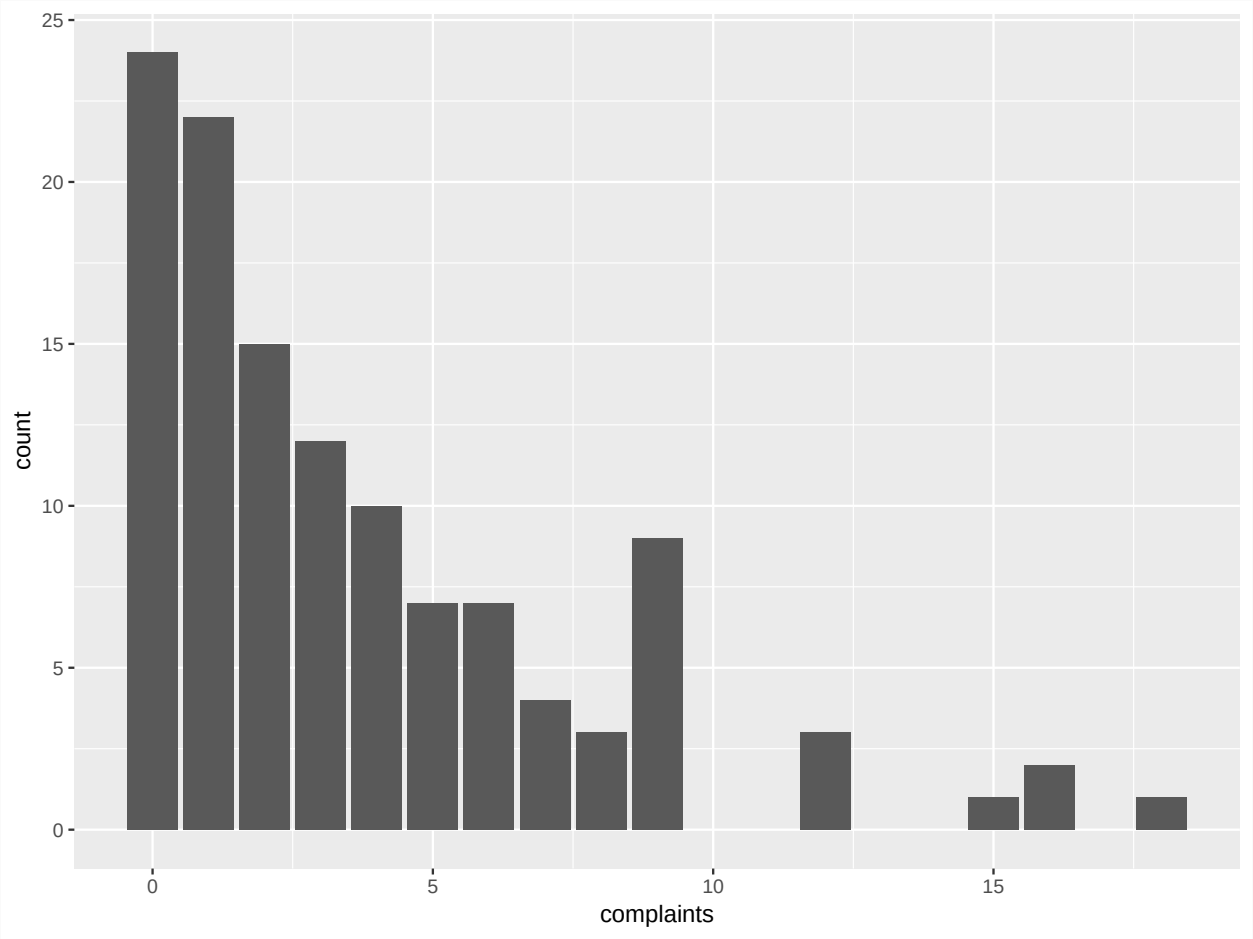
Discrete data regression: cockroaches data ii

We have access to the following fields (`pest_data.RDS`):

- `complaints`: Number of complaints per building in the current month
- `traps`: The number of traps used per month per building
- `live_in_super`: An indicator for whether the building has a live-in super
- `age_of_building`: The age of the building
- `total_sq_foot`: The total square footage of the building
- `average_tenant_age`: The average age of the tenants per building
- `monthly_average_rent`: The average monthly rent per building
- `floors`: The number of floors per building

Discrete data regression: cockroaches data iii

Let's make some plots of the raw data, such as the distribution of the complaints:



Poisson regression: cockroaches data i

- A common way of modeling this sort of skewed, single bounded count data is as a Poisson random variable. For simplicity, we will start assuming:
  - ungrouped data, with no building distinction
  - no time-trend structures
- We use the number bait stations placed in the building, denoted below as `traps`, as explanatory variable. This model assumes that the mean and variance of the outcome variable `complaints` (number of complaints) is the same. For the i-th complaint, $i = 1, \ldots, n$, we have

$$\begin{aligned} \text{complaints}_i &\sim \text{Poisson}(\lambda_i) \\ \lambda_i &= \exp(\eta_i) \\ \eta_i &= \alpha + \beta\, \text{traps}_i \end{aligned} \tag{15}$$

Poisson regression: cockroaches data ii

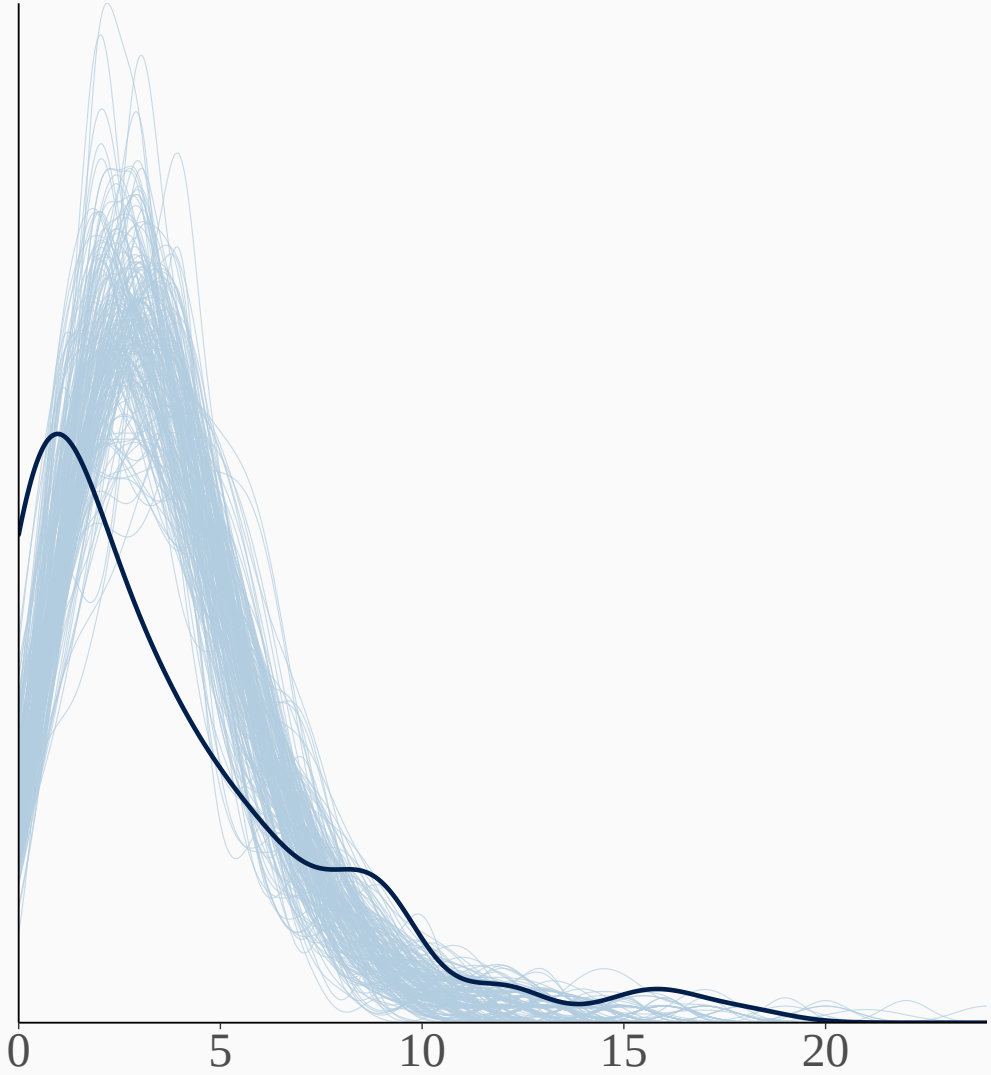
- We fit the model in Stan and we obtain the following posterior estimates (R output):

	mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
alpha	2.58	0.15	2.28	2.48	2.58	2.69	2.88	979	1
beta	-0.19	0.02	-0.24	-0.21	-0.19	-0.18	-0.15	997	1

- We could now check the model in terms of some graphical measures: for instance, in a Bayesian framework we may want to assess whether some replicated data under the model are close to the observed ones (this is the so-called posterior predictive checking approach).

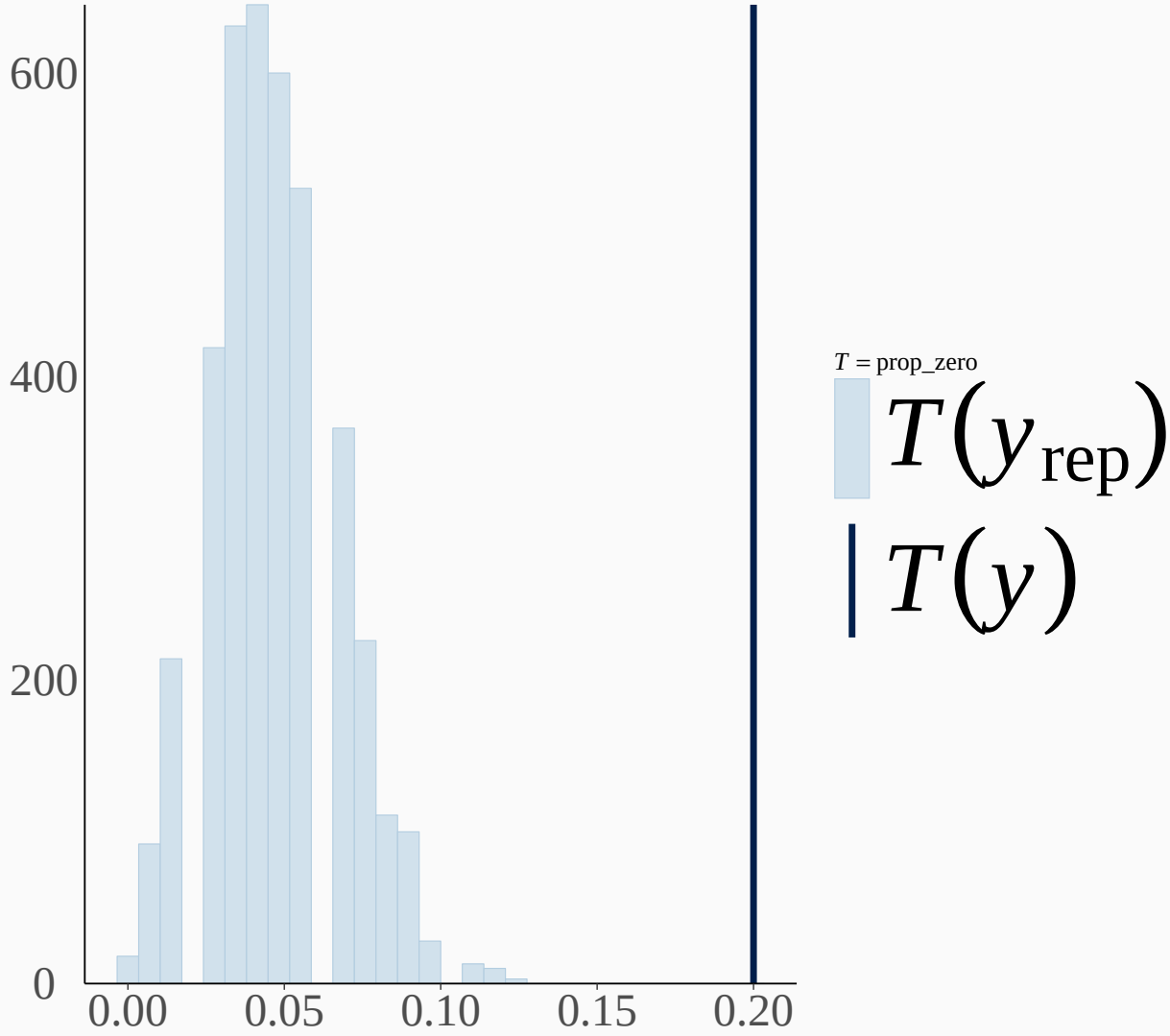
Poisson regression: cockroaches data iii

We check the model via some simulated data:



Poisson regression: cockroaches data iv

We check the proportion of zeros in the data and in the replications:



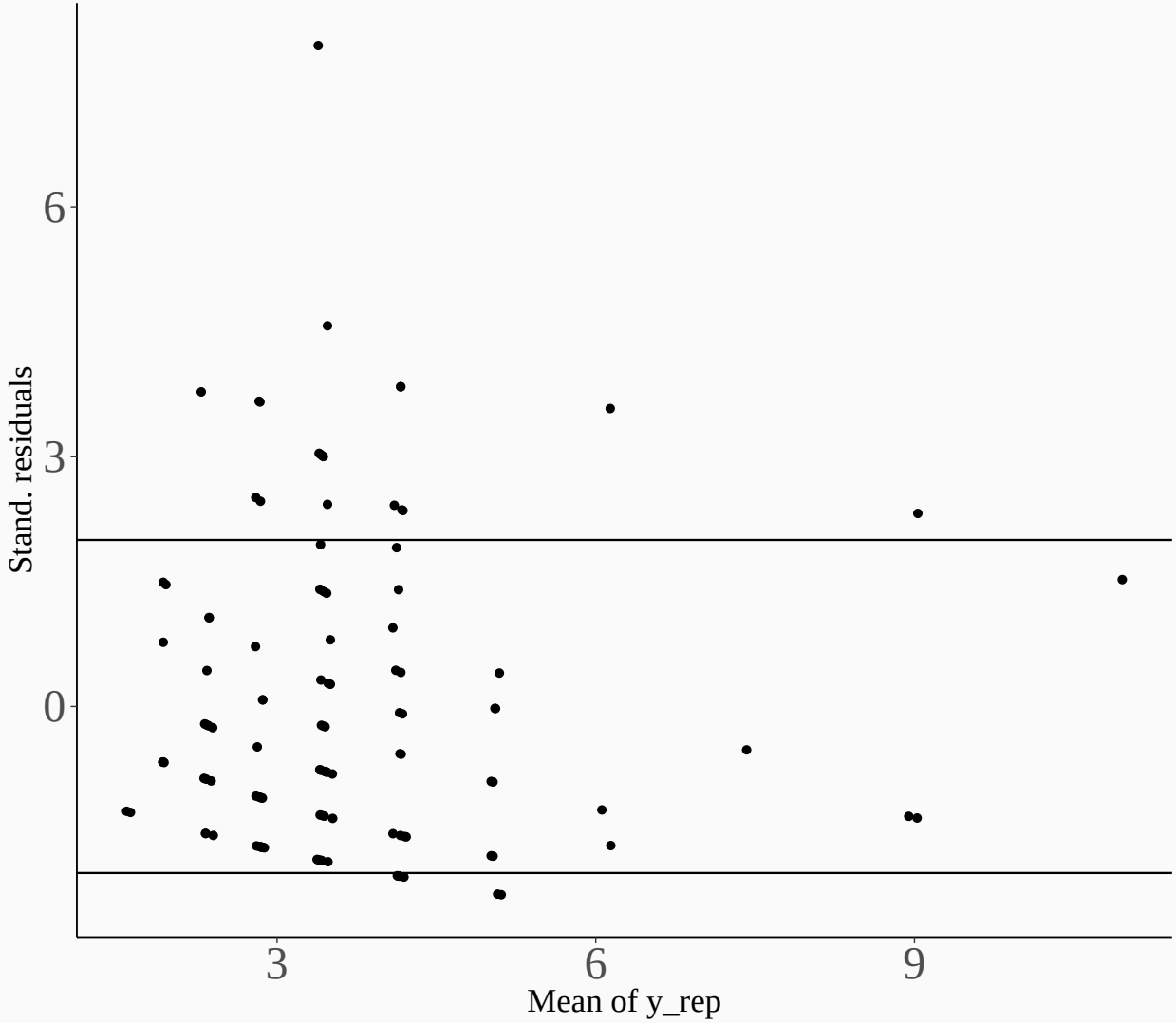
Poisson regression: cockroaches data v

Comments:

- We immediately realize that replicated distributions are far from the observed data distribution, and that the proportion of zero assumed by the Poisson model is quite underestimated...It is clear that the model does not capture this feature of the data well at all.
- Maybe the Poisson distribution distribution is not suited in this case...let's still explore the standardised residuals of the observed vs predicted number of complaints.
- We can also view how the predicted number of complaints varies with the number of traps.

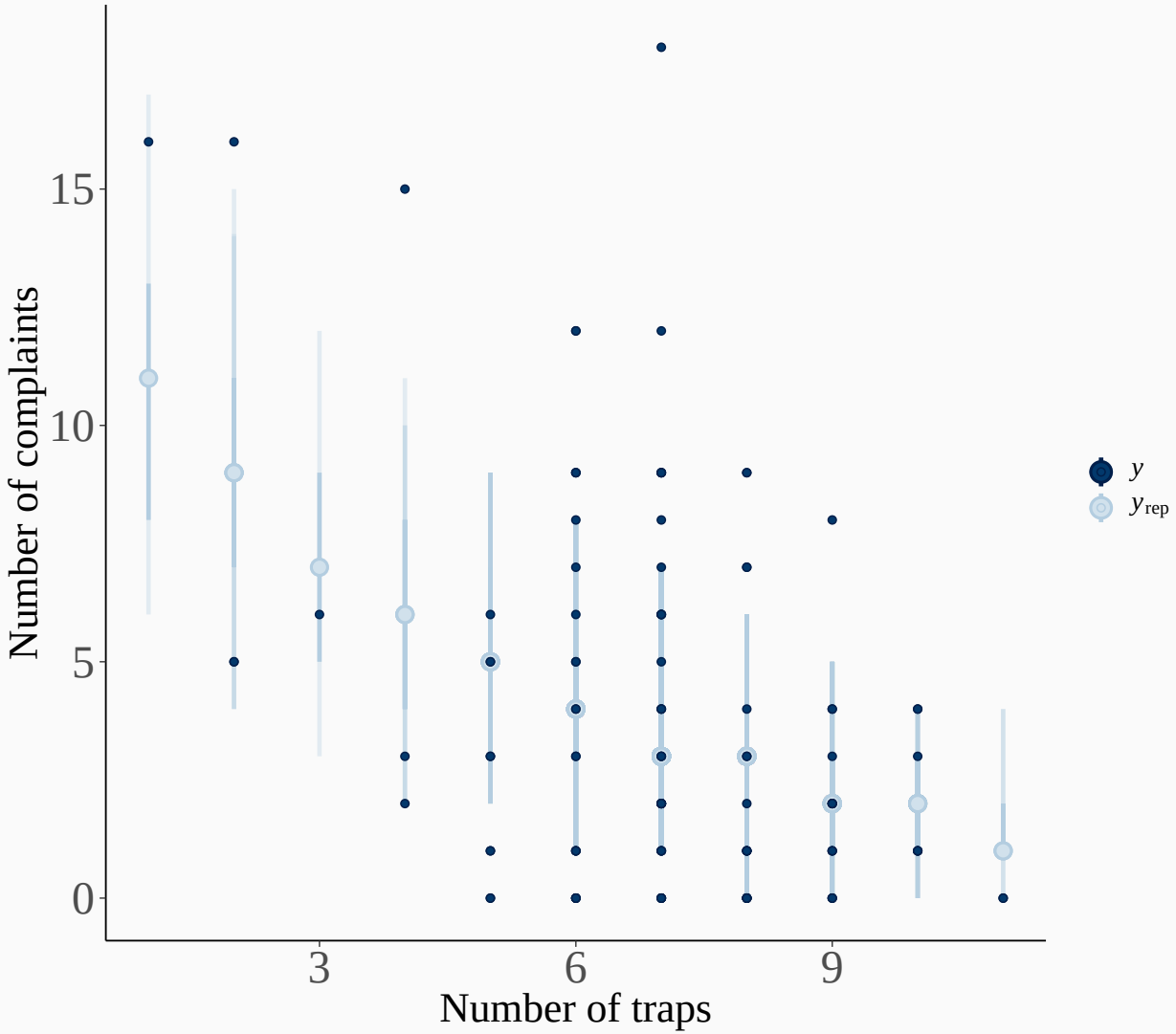
Poisson regression: cockroaches data vi

Standardized residuals:



Poisson regression: cockroaches data vii

Predictive intervals:



We can see that the model does not seem to fully capture the data.

Hierarchical NB regression i

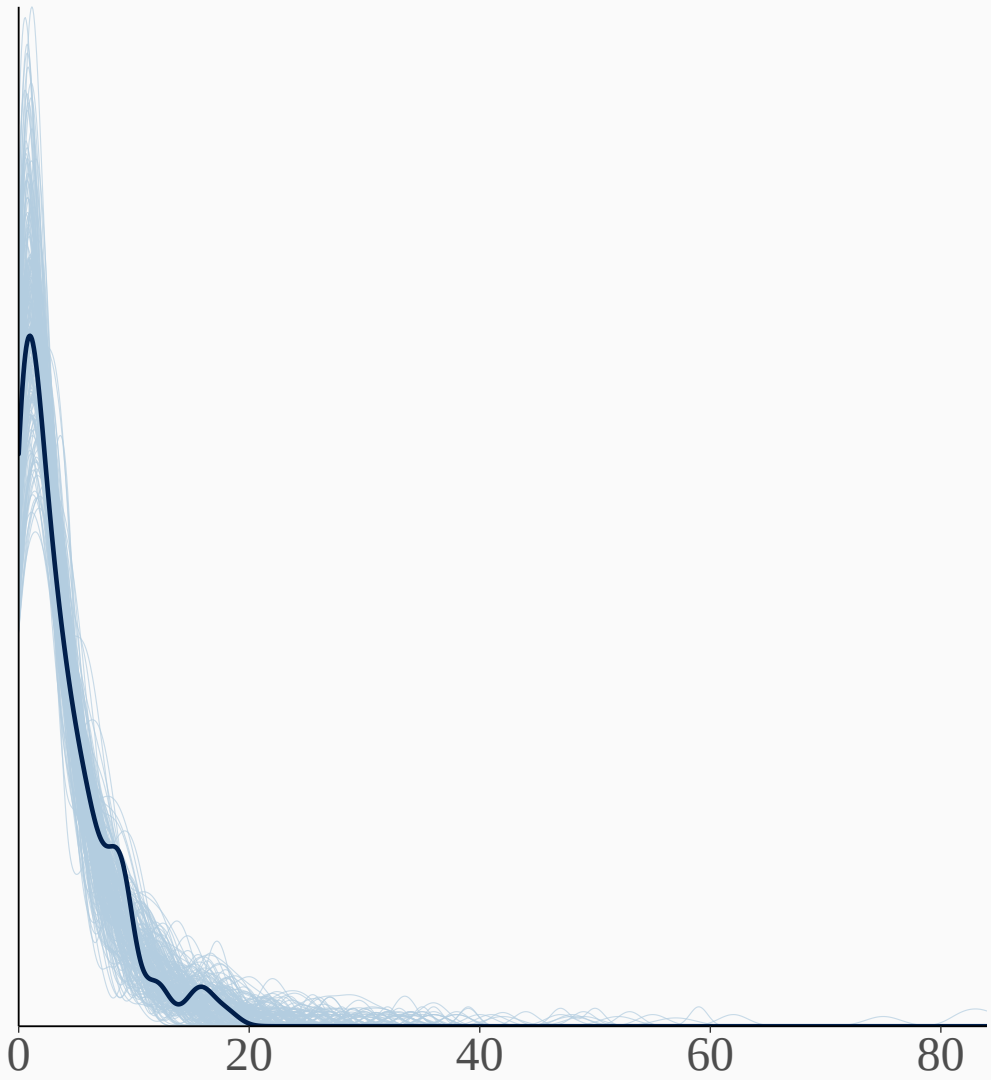
- A non-hierarchical model is not suited here...and we could swithc to the negative binomial distribution to capture overdispersion!
- We can extend the Poisson model (15) encoding hierarchical structure for the building and considering an offset term. Thus, for each complaint i we have:

$$
\begin{aligned}
\text{complaints}_{ib} &\sim \text{NegBin}(\lambda_{ib}, \phi) \\
\lambda_{ib} &= \exp(\eta_{ib}) \\
\eta_{ib} &= \alpha_{b(i)} + \beta\, \text{traps}_i + \beta_{\text{super}}\, \text{super}_i + \text{log\_sq\_foot}_i \qquad (16) \\
\alpha_b &\sim \mathcal{N}(\mu, \tau_\alpha^2), \\
\phi &\sim \mathcal{N}^+(0, 1)
\end{aligned}
$$

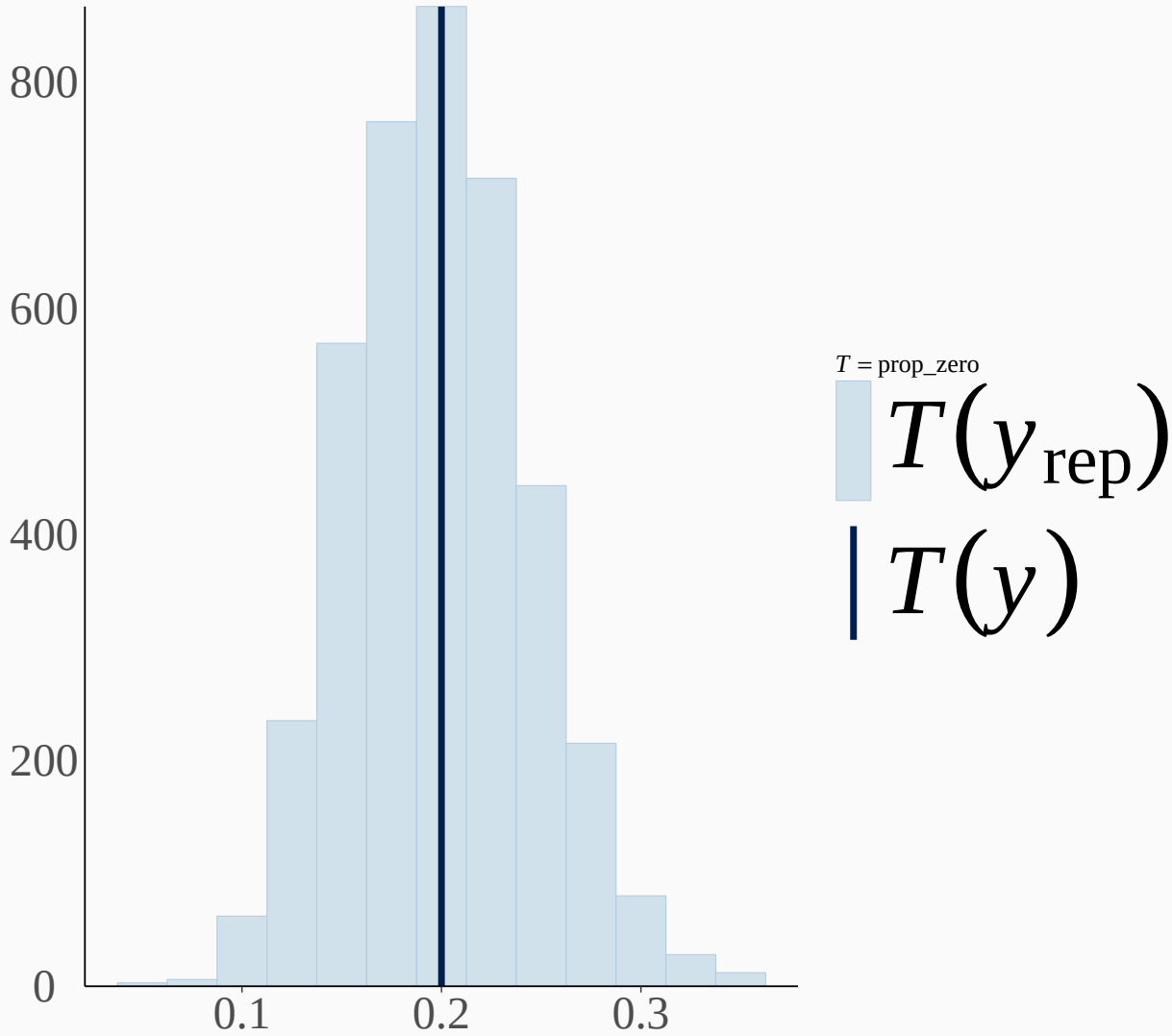
where $b(i)$ is the nested index for the building where the i-th complaint is registered.

- Using a hierarchical regression the model adequacy improves (see next slides).

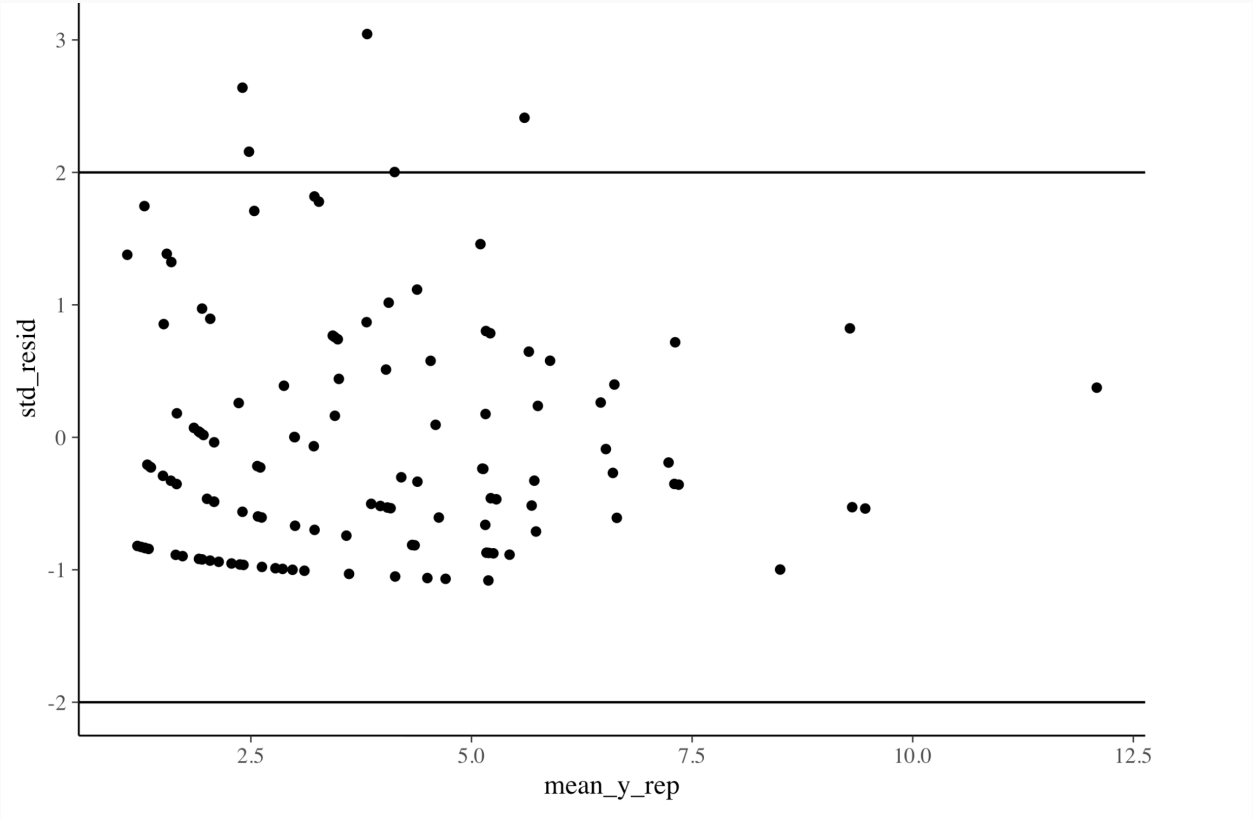
Hierarchical NB regression ii



Hierarchical NB regression iii



Hierarchical NB regression iv



Better!

Further reading

To properly capture the contents and the details about hierarchical/multilevel modeling, we strongly suggest the following further reading:

- Chapter 15 and 16 from *Bayesian Data Analysis*, by A.Gelman et al.
- Chapter 11, 12, 13, 14, 15 from *Data Analysis using Regression and Multilevel/Hierarchical models*, by A. Gelman and J. Hill.