Advanced statistical methods

Hierarchical models

Leonardo Egidi

2025/2026

Università di Trieste

Table of contents i

Towards multilevel/hierarchical models: an overview

- Grouped data
- Clustered data
- Repeated measurements
- Non-nested models

The hierarchical/multilevel framework

Hierarchical linear regression

- Radon data
- Eight schools data
- US elections data

Hierarchical logistic regression

- US polls data

Hierarchical Poisson regression

- Police stops data
- Cockroaches data

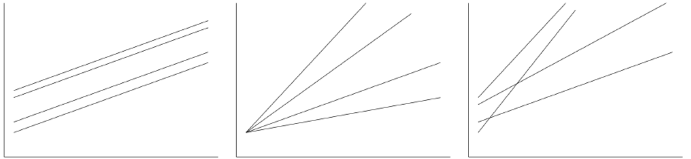
Towards multilevel/hierarchical models: an overview

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.

Varying-intercept and varying-slope models i

- With grouped data, a regression that includes indicators for groups is called a *varying-intercept model* because it can be interpreted as a model with a different intercept within each group



(a) Varying intercept

(b) Varying slope

(c) Varying intercept and slope

Varying-intercept and varying-slope models ii

- Model with one continuous predictor x and indicators for $J = 5$ groups. The model can be written as a regression with 6 predictors or, equivalently, as a regression with two predictors (x and the constant term), with the intercept varying by group (left figure panel):

$$y_i = \alpha_{j(i)} + \beta x_i + \epsilon_i, \quad \text{varying-intercept.}$$

- Another option (central panel) is to let the slope vary with constant intercept:

$$y_i = \alpha + \beta_{j(i)} x_i + \epsilon_i, \quad \text{varying-slope.}$$

Varying-intercept and varying-slope models iii

- Finally, the right panel shows a model in which both the intercept and the slope vary by group:

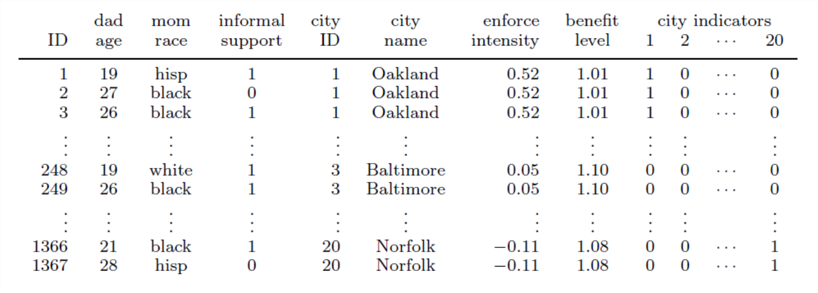
  $$y_i = \alpha_{j(i)} + \beta_{j(i)} x_i + \epsilon_i, \qquad \text{varying-intercept and slope.}$$

  The varying slopes are interactions between the continuous predictor x and the group indicators.
- It can be challenging to estimate all these α_j's and β_j's, especially when inputs are available at the group level.

Clustered data i

- With multilevel modeling we need to go beyond the classical setup of a data vector y and a matrix of predictors X. Each level of the model can have its own matrix of predictors.
- Observational study from Gelman and Hill, (2006): effect of city-level policies on enforcing child support payments from unmarried fathers.
- The treatment is at the group (city) level, but the outcome is measured on individual families.
- To estimate the effect of child support enforcement policies, the key “treatment” predictor is a measure of enforcement policies, which is available at the city level.
- Aim: estimate the probability that the mother received informal support, given the city-level enforcement measure and other city- and individual-level predictors.

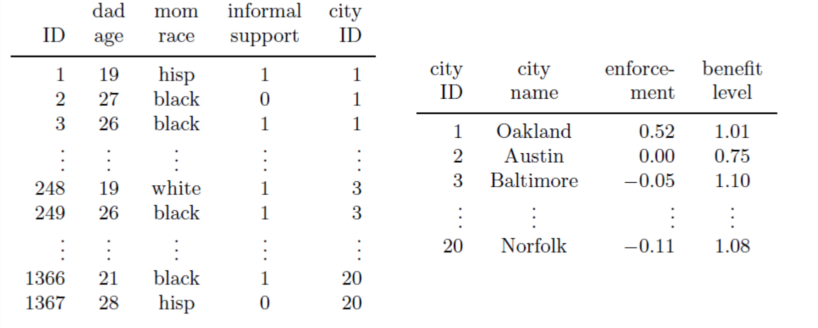
Clustered data ii



	dad	mom	informal	city	city	enforce	benefit	city indicators			
ID	age	race	support	ID	name	intensity	level	1	2	⋯	20
1	19	hisp	1	1	Oakland	0.52	1.01	1	0	⋯	0
2	27	black	0	1	Oakland	0.52	1.01	1	0	⋯	0
3	26	black	1	1	Oakland	0.52	1.01	1	0	⋯	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮		⋮
248	19	white	1	3	Baltimore	0.05	1.10	0	0	⋯	0
249	26	black	1	3	Baltimore	0.05	1.10	0	0	⋯	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮		⋮
1366	21	black	1	20	Norfolk	−0.11	1.08	0	0	⋯	1
1367	28	hisp	0	20	Norfolk	−0.11	1.08	0	0	⋯	1

Figure 1: Table 1: compact table for clustered data

Clustered data iii



ID	dad age	mom race	informal support	city ID
1	19	hisp	1	1
2	27	black	0	1
3	26	black	1	1
⋮	⋮	⋮	⋮	⋮
248	19	white	1	3
249	26	black	1	3
⋮	⋮	⋮	⋮	⋮
1366	21	black	1	20
1367	28	hisp	0	20

city ID	city name	enforce- ment	benefit level
1	Oakland	0.52	1.01
2	Austin	0.00	0.75
3	Baltimore	−0.05	1.10
⋮	⋮	⋮	⋮
20	Norfolk	−0.11	1.08

Figure 2: Table 2: two data-matrices for clustered data

- First table: data for the analysis as it might be stored in a computer package, with information on each of the 1367 mothers surveyed.

Clustered data iv

- Second table: to make use of the **multilevel structure** of the data, however, we need to construct two data matrices, one for each level of the model (city and mothers).
- Conceptually, the two-matrix, or multilevel, data structure has the advantage of clearly showing which information is available on individuals and which on cities.
- It also gives more flexibility in fitting models, allowing us to move beyond the classical regression framework.

Clustered data v

We briefly outline several possible ways of analyzing these data, as a motivation and lead-in to multilevel modeling.

- *Individual-level regression*: $\Pr(Y_i = 1) = \text{logit}^{-1}(X_i\beta)$ where X includes the constant term, the treatment (enforcement intensity), and the other predictors (father's age and indicators for mother's race at the individual level; and benefit level at the city level). X is thus constructed from the data matrix of Table 1.

  Problem: it ignores city-level variation beyond that explained by enforcement intensity and benefit level,which are the city-level predictors in the model.

- *Group-level regression on city averages*: perform a city-level analysis, with individual-level predictors included using their group-level averages. The outcome, y_j, would be the average total support among the respondents in city j, the enforcement indicator would be the treatment, and the other variables would also be included as predictors. Such a regression—in this case, with 20 data points—has the advantage that its errors are automatically at the city level.

  Problem: however, by aggregating, it removes the ability of individual predictors to predict individual outcomes.

Clustered data vi

- *Individual-level regression with city indicators, followed by group-level regression of the estimated city effects*: two-steps analysis, first fitting a logistic regression to the individual data y given individual predictors (in this example, father's age and indicators for mother's race) along with indicators for the 20 cities. Then, the next step is to perform a linear regression at the city level, considering the estimated coefficients of the city indicators (in the individual model that was just fit) as the "data" y_j . This city-level regression has 20 data points and uses, as predictors, the city-level data (in this case, enforcement intensity and benefit level).

  Problem: can run into problems when sample sizes are small in particular groups, or when there are interactions between individual- and group-level predictors.

Clustered data vii

Multilevel modeling is a more general approach that can include predictors at both levels at once.

- The multilevel model looks something like the two-step model we have described, except that both steps are fitted at once.
- Two components: a logistic regression with 1369 data points predicting the binary outcome given individual-level predictors and with an intercept that can vary by city, and a linear regression with 20 data points predicting the city intercepts from city-level predictors.

$$\Pr(Y_i = 1) = \text{logit}^{-1}(\alpha_{j(i)} + X_i\beta), \quad i = 1, \ldots, n,$$

where X is the matrix of individual-level predictors and $j(i)$ indexes the city where person i resides.

Clustered data viii

The second part of the model—what makes it "multilevel"—is the regression of the city coefficients:

$$\alpha_j \sim \mathcal{N}(U_j\gamma, \sigma_\alpha^2), \quad j = 1, \ldots, 20,$$

where U is the matrix of city-level predictors, γ is the vector of coefficients for the city-level regression, and σ_α is the standard deviation of the unexplained group-level errors.

- The key is the group-level variation parameter σ_α, which is estimated from the data (along with α, β).
- The model for the α allows us to include all 20 of them in the model *without having to worry about collinearity*.

Repeated measurements

- Another kind of multilevel data structure involves repeated measurements on persons (or other units)—thus, measurements are clustered within persons, and predictors can be available at the measurement or person level.

- Suppose a dataset where some people who bought an insurance policy are every year asked either to renew or to interrupt the policy. We basically have as many repeated measurements for each person as many years that person is observed/asked.

- A naive multilevel logistic regression could then be similar to the previous model, with each α_j defined here in terms of the *j*-th ensured for which the *i*-th policy was observed.

- Here also, we can work with a more rectangular-structured data matrix (similarly as Table 1) or with two-data matrices: the choice is done in terms of users' convenience.

Indicator variables and fixed or random effects i

- When including an input variable with J categories into a classical regression, standard practice is to choose one of the categories as a baseline and include indicators for the other $J - 1$ categories (in the child enforcement example, one could set city 1 (Oakland) as the baseline and include indicators for the other 19. The coefficient for each city then represents its comparison to Oakland.)
- In a multilevel model it is unnecessary to do this arbitrary step of picking one of the levels as a baseline. For example, in the child support study, one would include indicators for all 20 cities in the model. In a classical regression these could not all be included because they would be collinear with the constant term, but in a multilevel model this is not a problem because they are themselves modeled by a group-level distribution.

Indicator variables and fixed or random effects ii

- The varying coefficients (α_j's or β_j's) in a multilevel model are sometimes called **random effects**, a term that refers to the randomness in the probability model for the group-level coefficients.
- The term **fixed effects** is used in contrast to random effects—but not in a consistent way! Fixed effects are usually defined as varying coefficients that are not themselves modeled.
- As an interpretation issue, fixed effects are constant across individuals, and random effects vary.
- Varying slopes can be interpreted as *interactions* between an individual-level predictor and group indicators. As with classical regression models with interactions, the intercepts can often be more clearly interpreted if the continuous predictors are appropriately centered

Non-nested models i

- So far we have considered the simplest hierarchical structure of individuals i in groups j. We briefly discuss now more complicated grouping structures.
- Example: a psychological experiment with two potentially interacting factors. We collect success rates data on pilots of flight simulators, with $n = 40$ data points corresponding to $J = 5$ treatment conditions and $K = 8$ different airports, as shown in the next figure (from G&H book, sect. 13.5).

Non-nested models ii

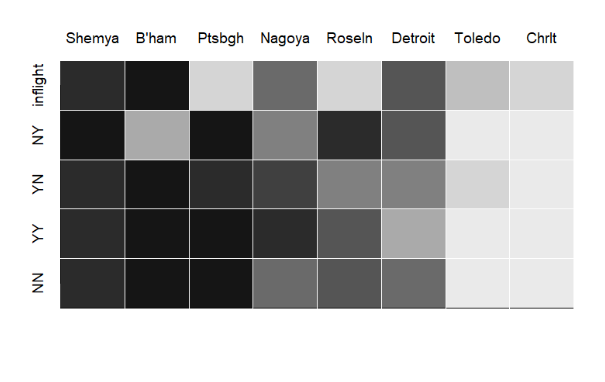
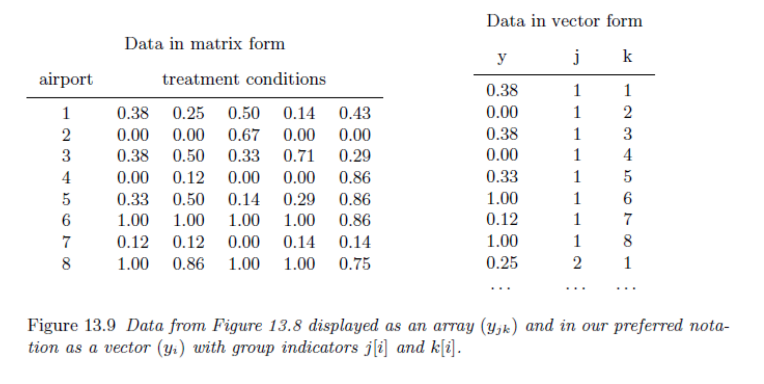


Figure 13.8 *Success rates of pilots training on a flight simulator with five different treatments and eight different airports. Shadings in the 40 cells i represent different success rates y_i, with black and white corresponding to 0 and 100%, respectively. For convenience in reading the display, the treatments and airports have each been sorted in increasing order of average success. These 40 data points have two groupings—treatments and airports—which are not nested.*

Non-nested models iii

The data stored as a matrix and as an array are displayed in the next figure (always from G&H book):



Data in matrix form

airport	treatment conditions				
1	0.38	0.25	0.50	0.14	0.43
2	0.00	0.00	0.67	0.00	0.00
3	0.38	0.50	0.33	0.71	0.29
4	0.00	0.12	0.00	0.00	0.86
5	0.33	0.50	0.14	0.29	0.86
6	1.00	1.00	1.00	1.00	0.86
7	0.12	0.12	0.00	0.14	0.14
8	1.00	0.86	1.00	1.00	0.75

Data in vector form

y	j	k
0.38	1	1
0.00	1	2
0.38	1	3
0.00	1	4
0.33	1	5
1.00	1	6
0.12	1	7
1.00	1	8
0.25	2	1
...	...	...

Figure 13.9 *Data from Figure 13.8 displayed as an array* (y_{jk}) *and in our preferred notation as a vector* (y_i) *with group indicators* $j[i]$ *and* $k[i]$.

Non-nested models iv

- The responses can be fit to a non-nested multilevel model of the form:

$$\begin{aligned} y_i &\sim \mathcal{N}(\mu + \gamma_{j(i)} + \delta_{k(i)}, \sigma_y^2),\ \ i = 1, \ldots, n \\ \gamma_j &\sim \mathcal{N}(0, \sigma_\gamma^2),\ \ j = 1, \ldots, J \\ \delta_k &\sim \mathcal{N}(0, \sigma_\delta^2),\ \ k = 1, \ldots, K, \end{aligned} \tag{1}$$

  where the parameters γ_j and δ_k represent treatment effects and airport effects. Their distributions are centered at zero because the regression model for y already has an intercept μ, and any nonzero mean for the γ and δ distributions could be folded into μ.

- When fit to the data in the figure, the estimated residual standard deviations at the individual, treatment and airport levels are $\hat{\sigma}_y = 0.23$, $\hat{\sigma}_\gamma = 0.04$ and $\hat{\sigma}_\delta = 0.32$. Thus, the variation among airports is huge—even larger than that among individual measurements—but the treatments vary almost not at all.

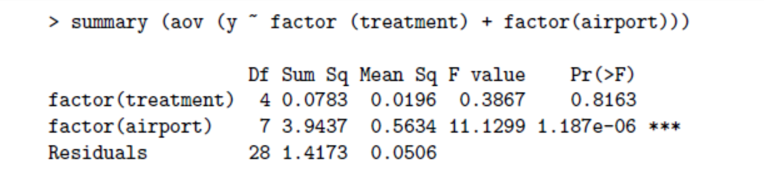
Non-nested models v

- Connection with **Analysis of Variance (ANOVA)**: as we know from classical statistics, ANOVA is typically used to learn the relative importance of different sources of variation in a dataset. In this example, how much of the variation in the data is explained by treatments, how much by airports, and how much remains after these factors have been included in a linear model? *If a multilevel model has already been fit, it can be summarized by the variation in each of its batches of coefficients*.
- In classical statistics, ANOVA refers either to a family of additive data decomposition, or to a method of testing the statistical significance of added predictors in a linear model. For the flight data simulator we can write:

$$y_i = \mu + \gamma_{j(i)} + \delta_{k(i)} + \epsilon_i, \tag{2}$$

and a classical two-way ANOVA can be obtained as follows:

Non-nested models vi



```
> summary (aov (y ~ factor (treatment) + factor(airport)))

                  Df Sum Sq Mean Sq F value    Pr(>F)
factor(treatment)  4 0.0783  0.0196  0.3867    0.8163
factor(airport)    7 3.9437  0.5634 11.1299 1.187e-06 ***
Residuals         28 1.4173  0.0506
```

which indicates that the variation among treatments is not statistically significant. Let's see the sources of variation and degrees of freedom:

- 5 treatment effects minus 1 constraint = 4 degrees of freedom
- 8 airports effects minus 1 constraint = 7 df
- 40 residuals minus 12 constraints (1 mean, 4 treatment effects, 7 airport effects) = 28 df

Non-nested models vii

- When comparing nested models, ANOVA is related to the classical test of the hypothesis that the smaller model is true, which is equivalent to the hypothesis that the additional predictors all have coefficients of zero when included in the larger model.
- When moving to multilevel modeling, the key idea we want to take from ANOVA is the estimation of the importance of different batches of predictors.
- A general solution to perform ANOVA here is to fit the model (2)—along with the random effects for γ, δ, and the error ϵ— and summarize the estimated variance components, $\hat{\sigma}_y, \hat{\sigma}_\gamma, \hat{\sigma}_\delta$.

Item-response and ideal-point models i

- Usually applied to data with multilevel structure, typically non-nested, for example with measurements associated with persons and test items, or judges and cases.
- A standard model for success or failure in testing situations is the logistic item-response model, also called the Rasch model. Suppose J persons are given a test with K items, with $y_{jk} = 1$ if the response is correct. Then the logistic model can be written as:

$$\Pr(y_{jk} = 1) = \text{logit}^{-1}(\alpha_j - \beta_k), \tag{3}$$

  with parameters:
  - α_j: the *ability* of person j,
  - β_k: the *difficulty* of item k.

Item-response and ideal-point models ii

In general, not every person is given every item, so it is convenient to index the individual responses as $i = 1, \ldots, n$, with each response i associated with a person $j(i)$ and item $k(i)$. Thus model (3) becomes:

$$\Pr(y_i = 1) = \mathrm{logit}^{-1}(\alpha_{j(i)} - \beta_{k(i)}). \tag{4}$$

- The model (4) is not identified, because a constant can be added to all the abilities α_j and all the difficulties β_k, and the predictions of the model will not change. From the standpoint of classical logistic regression, this nonidentifiability is a simple case of collinearity and can be resolved by constraining the estimated parameters in some way, for instance setting $\alpha_1 = \beta_1 = 0$, constraining the α_j's to sum to zero, or constraining the β_k's to sum to zero.

Item-response and ideal-point models iii

- In a multilevel model, such constraints are unnecessary. The natural multilevel model for (4) assigns some normal distributions to the ability and the difficulty parameters:

$$\begin{aligned} \alpha_j &\sim \mathcal{N}(\mu_\alpha, \sigma_\alpha^2),\ j = 1, \ldots, J, \\ \beta_k &\sim \mathcal{N}(\mu_\beta, \sigma_\beta^2),\ k = 1, \ldots, K. \end{aligned}$$

  Now it is μ_α and μ_β that are not identified, because a constant can be added to each without changing the predictions. The simplest way to identify the multilevel model is set $\mu_\alpha = 0$, or to set $\mu_\beta = 0$ (but not both).
- *Ideal-point* modeling is an application of item-response models to a settimg where what is being easured is not ability of individuals and difficulty of items, but rather positions of individuals and items on some scale of values.
- Example: Supreme Court voting (G&H book, sect. 14.3).

Non-nested NB model of structure in social networks i

- Understanding the structure of social networks, and the social processes that form them, is a central concern of sociology for both theoretical and practical reasons. Networks have been found to have important implications for social mobility, getting a job, the dynamics of fads and fashion, attitude formation, and the spread of infectious disease.
- Example (from book G&H, sect. 15.3): overdispersed Poisson regression model to learn about social structure. We fit the model to a random-sample survey of Americans who were asked, "How many X's do you know?" for a variety of characteristics X, defined by name (Michael, Christina, Nicole,...), occupation (postal worker, pilot, gun dealer,...), ethnicity (Native American), or experience (prisoner, auto accident victim,...).
- The original goals of the survey were (1) to estimate the distribution of individuals' network size, defined to be the number of acquaintances, in U.S. population and (2) to estimate the sizes of certain subpopulations, especially those that are hard to count using regular survey results.

Non-nested NB model of structure in social networks ii

- Modeling setup: for respondent $i = 1, \ldots, 1370$ and subpopulations $k = 1, \ldots, 32$, we use the notation y_{ik} for the number of persons in group k known by person i.
- We evaluate three possible models, assuming $y_{ik} \sim \text{Poisson}(\lambda_{ik})$:

$$\begin{aligned}\text{Erdos-Renyi model} : \lambda_{ik} &= ab_k \\ \text{null model} : \lambda_{ik} &= a_i b_k \\ \text{overdispersed model} : \lambda_{ik} &= a_i b_k g_{ik}.\end{aligned}$$

- *Null model*: in which individuals i have varying levels of gregariousness or popularity, so that the expected number of persons in group k known by person i will be proportional to this gregariousness parameter, which we label a_i. Departure from this model—patterns not simply explained by differing group sizes or individual popularities—can be viewed as evidence of structured social acquaintance networks.

Non-nested NB model of structure in social networks iii

- *Overdispersion* in these data can arise if the relative propensity for knowing someone in prison, for example, varies from respondent to respondent. We can write this in the generalized linear model framework as:

$$y_{ik} \sim \text{Poisson}(e^{a_i+b_k+\gamma_{ik}}),$$

where each $\gamma_{ik} = \log(g_{ik}) \equiv 0$ in the null model. For each subpopulation k, we let the multiplicative factors $g_{ik} = e^{\gamma_{ik}}$ follow a Gamma distribution with a value of 1 for the mean and a value of $1/(\omega_k - 1)$ for the shape parameter. In this way:

$$y_{ik} \sim \text{NegBin}(e^{a_i+b_k}, \omega_k),$$

Costs and benefits of multilevel modeling i

Before we go to the effort of learning multilevel modeling, it is helpful to briefly review what can be done with classical regression:

- Prediction for continuous or discrete outcomes,
- Fitting of nonlinear relations using transformations,
- Inclusion of categorical predictors using indicator variables,
- Modeling of interactions between inputs,
- Causal inference (under appropriate conditions).

Costs and benefits of multilevel modeling ii

Motivations for moving to multilevel models:

- Accounting for individual- and group-level variation in estimating group-level regression coefficients.
- Modeling variation among individual-level regression coefficients. In classical regression, one can do this using indicator variables, but multilevel modeling is convenient when we want to model the variation of these coefficients across groups, make predictions for new groups, or account for group-level variation in the uncertainty for individual-level coefficients.
- Estimating regression coefficients for particular groups
- A potential drawback to multilevel modeling is the additional complexity of coefficients varying by group.

Costs and benefits of multilevel modeling iii

- A multilevel model requires additional assumptions beyond those of classical regression—basically, each level of the model corresponds to its own regression with its own set of assumptions such as additivity, linearity, independence, equal variance, and normality.
- The usual alternative to multilevel modeling is classical regression—either ignoring group-level variation, or with varying coefficients that are estimated classically (and not themselves modeled)—or combinations of classical regressions.
- In various limiting cases, the classical and multilevel approaches coincide. When there is very little group-level variation, the multilevel model reduces to classical regression with no group indicators; conversely, when group-level coefficients vary greatly (compared to their standard errors of estimation), multilevel modeling reduces to classical regression with group indicators.

Costs and benefits of multilevel modeling iv

- When the number of groups is small (less than five, say), there is typically not enough information to accurately estimate group-level variation. As a result, multilevel models in this setting typically gain little beyond classical varying-coefficient models.
- Computational softwares: `lme4`, `WinBUGS`, `JAGS`, Stan (`rstan`, `rstanarm`).
- In this course, we will strongly rely on Stan and Bayesian methods. However, we will also cover classical procedures.

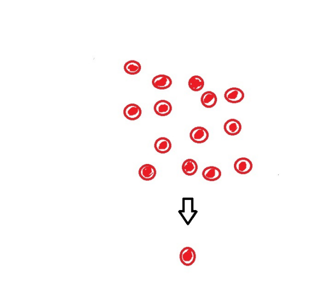
The hierarchical/multilevel framework

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

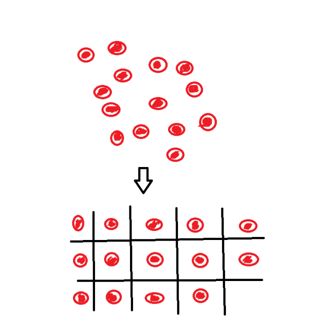
If we assume that every individual is equivalent then we can pool the data, but only at the expense of bias ⇔ Complete pooling.



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

Motivations iii

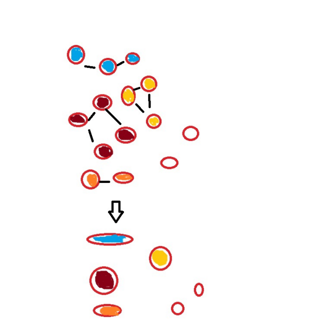
Conversely, modelling every individual separately avoids any bias, but then the data becomes very sparse and inferences weak ⇔ No pooling.



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

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:



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

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 $j = 1, \dots, J$, 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,

$$\pi(\theta|\phi) = \prod_{j=1}^{J} \pi(\theta_j|\phi). \tag{5}$$

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

$$\pi(\theta) = \int \left(\prod_{j=1}^{J} \pi(\theta_j|\phi) \right) \pi(\phi) d\phi. \tag{6}$$

The fundamental concept of exchangeability iv

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

$$p(\theta, y) = \prod_{i=1}^{n} p(y_{ij}|\theta_{j(i)})\pi(\theta_{j(i)}|\phi)\pi(\phi), \tag{7}$$

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

$$\pi(\theta, \phi|y) \propto \pi(\phi, \theta)p(y|\theta). \tag{8}$$

- 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.

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

- 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:

$$\begin{aligned} y_{ij} &\sim \mathcal{N}(\alpha_{j(i)} + x_{ij}\beta_{j(i)}, \sigma_y^2), \\ \begin{pmatrix}\alpha\\ \beta\end{pmatrix} &\sim \mathcal{N}\left(\begin{pmatrix}\mu_\alpha\\ \mu_\beta\end{pmatrix}, \begin{pmatrix}\sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta\\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta^2\end{pmatrix}\right), \end{aligned} \tag{9}$$

  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 (9) in vector-matrix form as:

  $$\begin{aligned} y_{ij} &\sim \mathcal{N}(X_j \boldsymbol{\beta}_{j(i)}, \sigma_y^2), \\ \boldsymbol{\beta}_j &\sim \mathcal{N}(\mu_\beta, \Sigma_\beta), \end{aligned} \tag{10}$$

  where Σ_β is a variance/covariance matrix for $\boldsymbol{\beta}_j$.

- 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:

$$\begin{aligned} y_{ij} \sim & \mathcal{N}(\alpha_{j(i)}, \sigma^2), \; i = 1, \ldots, n \quad \text{Individual level} \\ \alpha_j \sim & \mathcal{N}(\mu_\alpha, \tau^2), \; j = 1, \ldots, J \; \text{Group level} \end{aligned} \tag{11}$$

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 ± 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 ± standard errors are plotted against the number of observations for each county.

Partial pooling with no predictors iii

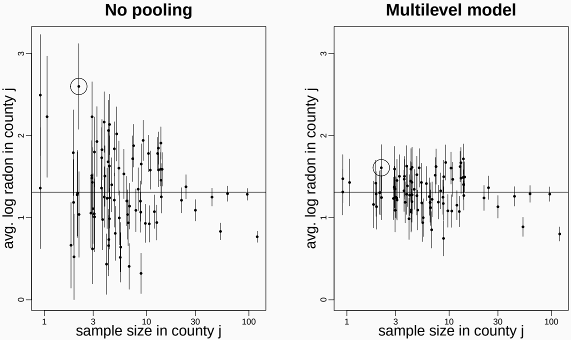


Figure 3: Estimates ± standard errors for the average log radon levels in Minnesota counties plotted versus the number of observations in the county.

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:

$$\boxed{\hat{\alpha}_j = \frac{\frac{n_j}{\sigma^2}\bar{y}_j + \frac{1}{\tau^2}\bar{y}_{\text{all}}}{\frac{n_j}{\sigma^2} + \frac{1}{\tau^2}} \qquad (12)}$$

  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 (12) 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:

$$\begin{aligned} y_{ij} \sim & \mathcal{N}(\alpha_{j(i)} + \beta x_i, \sigma^2), \ i = 1, \ldots, n \text{ Individual level} \\ \alpha_j \sim & \mathcal{N}(\mu_\alpha, \tau^2), \ j = 1, \ldots, J \qquad \text{Group level} \end{aligned} \tag{13}$$

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

Partial pooling with predictors ii

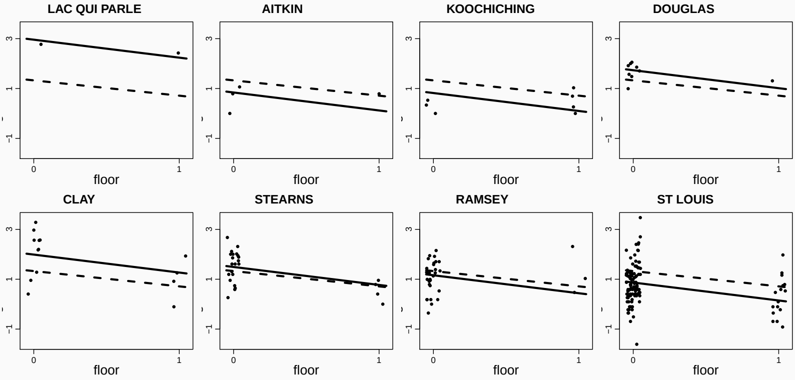


Figure 4: Complete pooling (dashed lines) and no pooling (solid lines) for 8 counties in Minnesota.

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 (13) 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

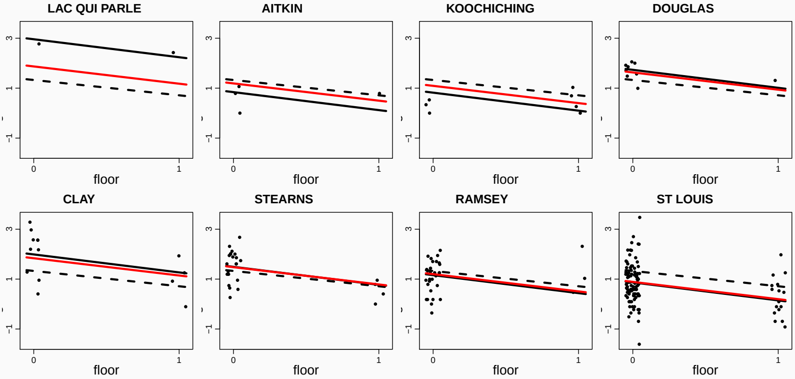


Figure 5: Complete pooling (dashed lines), no pooling (solid lines) and partial pooling (solid red lines).

Partial pooling with predictors vii

- The estimated line from the hierarchical model (13) 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 (12) as follows:

$$\hat{\alpha}_j \approx \frac{\frac{n_j}{\sigma^2}}{\frac{n_j}{\sigma^2} + \frac{1}{\tau_\alpha^2}}(\bar{y}_j - \beta\bar{x}_j) + \frac{\frac{1}{\tau_\alpha^2}}{\frac{n_j}{\sigma^2} + \frac{1}{\tau_\alpha^2}}\mu_\alpha, \tag{14}$$

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: $(1.46 - 0.26) - 0.69x = 1.20 - 0.69x$.

Sum-up about the Radon data example

- You find the data, some preliminary analysis shown in class, and some R code elaboration in the official Moodle course page.
- Please, check the code and repeat the analysis
- Some points for further open discussion in class:
  - Consider other modeling strategies, for instance change the likelihood, or the priors if you use a Bayesian approach. Feel free to use the function `stan_glmer` or the function `glmer`.
  - Plot the estimates by using the `bayesplot` package.
  - Divide the data in *training* and *test* and make some predictions.
  - Check the goodness-of-fit of your model, propose some measures for check.
  - Fit the model on other states, maybe one or two: are the analysis similar?
  - Simulate some fake-data (both the response variables and the explanatory variables) and repeat the multilevel analysis on these fake-data.

Eight schools example

Eight schools example (BDA book, 5.5)

We illustrate a normal model with a problem in which the hierarchical Bayesian analysis gives conclusions that differ in important respects from other methods.

A study was performed for the Educational Testing Service to analyze the effects of special coaching programs on test scores in each of eight high-schools.

The outcome variable in each study was a score, varying between 200 and 800, with mean about 500 and standard deviation about 100.

There is no prior reason to believe that any of the eight programs is more effective than any other.

As we'll see from a **Bayesian perspective**, the choice of the prior is of substantial importance here.

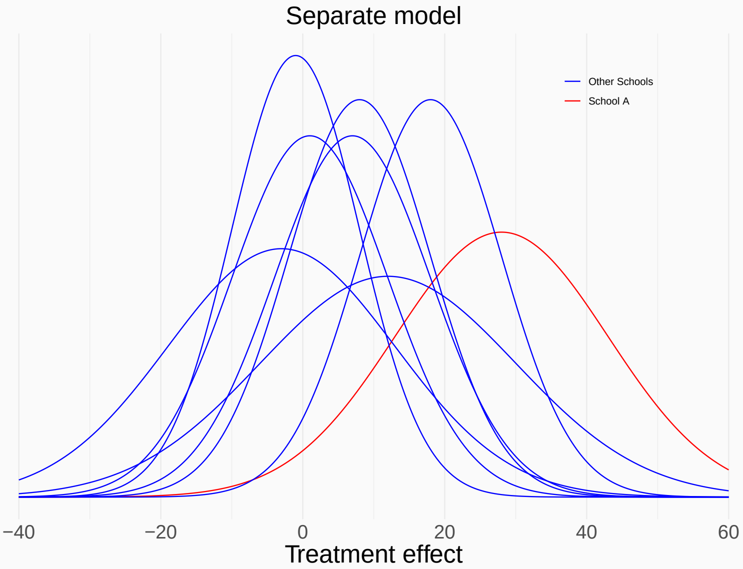
Eight schools

- We denote with y_{ij} the result of the i-th test in the j-th school. We assume the following model:

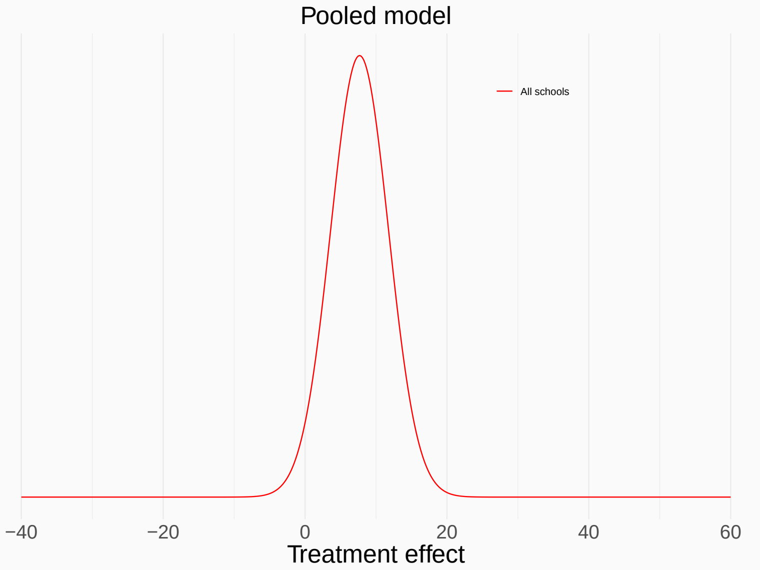
$$\begin{aligned} y_{ij} &\sim \mathcal{N}(\theta_j, \sigma_y^2) \\ \theta_j &\sim \mathcal{N}(\mu, \tau^2) \end{aligned} \tag{15}$$

- Do some schools perform better/worse according to these coaching effects?
- We will make three distinct analysis: separate analysis, pooled analysis and hierarchical modelling.
- Actually, for each school we have the estimated coaching effects y_j, $y = (28, 8, -3, 7, -1, 1, 18, 12)$, and a measure of standard deviation for them, $s = (15, 10, 16, 11, 9, 11, 10, 18)$.

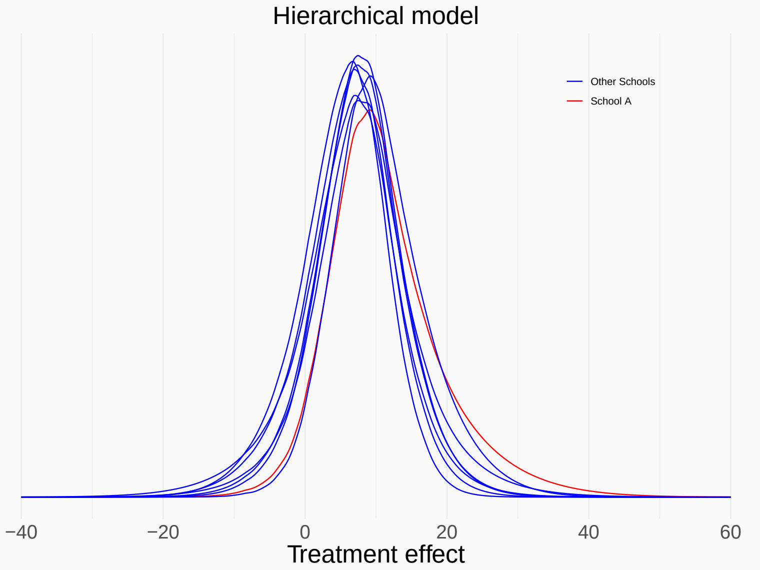
Eight schools: separate analysis



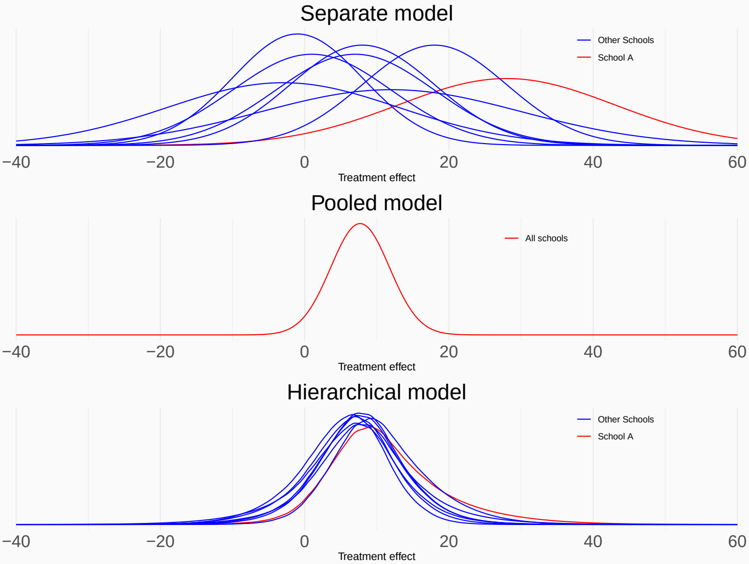
Eight schools: pooled analysis



Eight schools: hierarchical model



Eight schools: three models



Separate model

Pooled model

Hierarchical model

Eight schools: three models i

Comments:

- Separate analysis: the standard errors of these estimated effects make very difficult to distinguish between any of the experiments...treating each experiment separately and applying the simple normal analysis in each yields 95% posterior intervals that all overlap substantially.
- Pooled-analysis: under the hypothesis that all experiments have the same effect and produce independent estimates of this common effect, we could treat y as eight normally distributed observations with known variances. The pooled estimate is 7.7, and the posterior variance is 16.6.

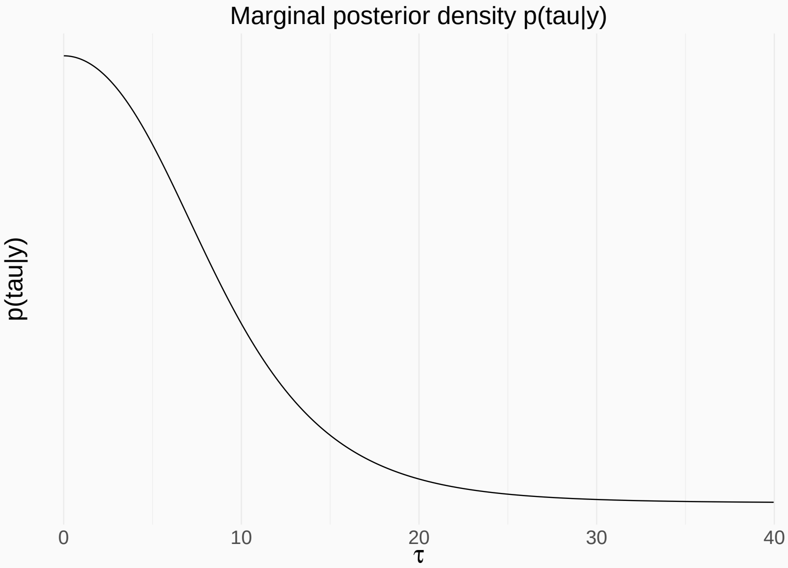
However, both the extreme analysis have difficulties.

Eight schools: three models ii

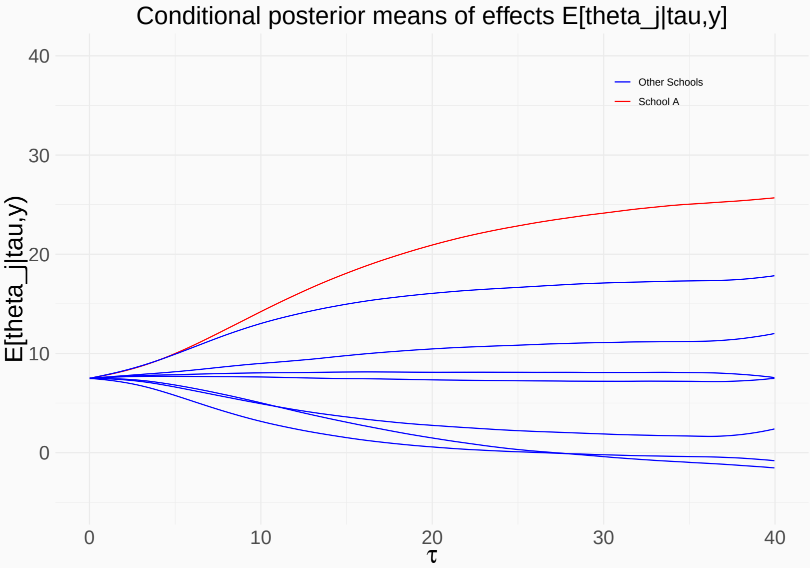
Other comments:

- Consider school A. The effect in school A is estimated as 28.4 with a standard error of 14.9 under the separate analysis, versus a pooled estimate of 7.7 with a standard error of 4.1. Mmh...should I flip a coin?
- We would like a compromise that combines information from all the eight experiments without assuming all the θ_j to be equal. The Bayesian analysis under the hierarchical model provides exactly that.
- As we may see from the third plot, the posterior distribution of $\theta_1, \ldots, \theta_8$ results to be closer to the complete analysis. Let's see now some other posterior analysis.

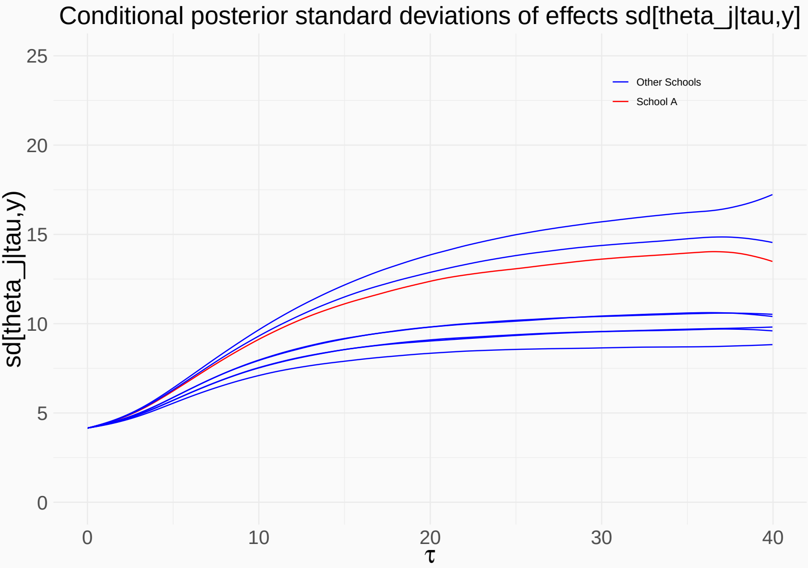
Eight schools: posterior summaries for hierarchical model i



Eight schools: posterior summaries for hierarchical model ii



Eight schools: posterior summaries for hierarchical model iii



Eight schools: posterior summaries for hierarchical model iv

- In the plot for the marginal posterior $\pi(\tau|y)$, $\tau = 0$ is the most likely value (no variation in θ, complete pooling).
- Conditional posterior means $\mathrm{E}(\theta_j|\tau, y)$ are displayed as functions of τ: for most of the likely values of τ, the estimated effects are relatively close together: as τ becomes larger (more variability among schools), the estimates approach the separate analysis results.
- Conditional standard deviations $\mathrm{sd}(\theta_j|\tau, y)$ become larger as τ increases.

Eight schools: discussion

Comments:

- The general conclusion from these posterior summaries is that an effect as large as 28.4 points (school A) in any school is unlikely. For the likely values of τ, the estimates in all schools are substantially less than 28 points.
- To sum up, the Bayesian analysis of this example not only allows straightforward inferences about many parameters, but provides posterior inferences that account for the partial pooling as well as the uncertainty in the hyperparameters.
- We have still to investigate the role of the prior for the population standard deviation τ.

Eight schools: priors for τ^2 i

- As we have already seen in other situations, assigning a prior may have a substantial effect on the final posterior inferences.
- In this example, τ^2 governs the extent of variation between the schools: which are some suitable priors?
- We review three choices:

$$\tau \sim \text{Uniform}(0, 100) \tag{16}$$

$$\tau^2 \sim \text{InvGamma}(0.01, 0.01) \tag{17}$$

$$\tau \sim \text{HalfCauchy}(0, 2.5) \tag{18}$$

Eight schools: priors for τ^2 ii

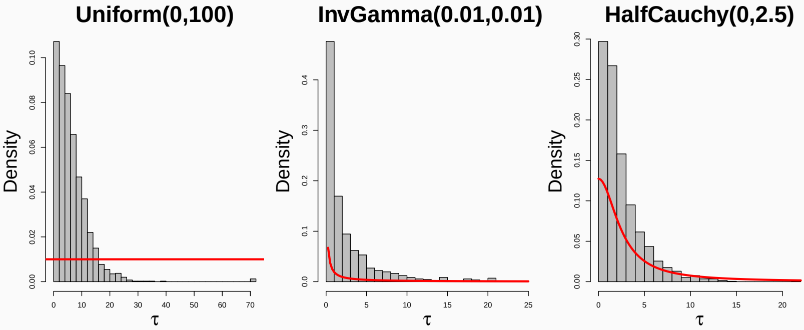


Figure 6: Marginal posterior (histograms) vs priors (solid red lines)

Eight schools: priors for τ^2

- Uniform The data show support for a range of values below $\tau = 20$, with a slight tail after that, reflecting the possibility of larger values, which are difficult to rule out given that the number of groups J is only 8 (that is, not much more than the $J = 3$ required to ensure a proper posterior density with finite mass in the right tail)
- Inverse gamma This prior distribution is sharply peaked near zero and further distorts posterior inferences, with the problem arising because the marginal likelihood for τ^2 remains high near zero. Moreover, the posterior is quite sensitive to the choices of the hyperparameters (try!)
- Half Cauchy less likely to dominate the inferences

Eight schools: priors for τ^2 iii

Comments:

- The InvGamma prior is not at all noninformative for this problem since the resulting posterior distribution remains highly sensitive to the choice of the hyperparameters.
- The Uniform prior distribution seems fine for the 8-school analysis, but problems arise if the number of groups J is much smaller, in which case the data supply little information about the group-level variance, and a noninformative prior distribution can lead to a posterior distribution that is improper or is proper but unrealistically broad.

Sum-up about the eight-schools example

- You find the data, some preliminary analysis shown in class, and some R code elaboration in the official Moodle course page.
- Please, check the code and repeat the analysis.
- Some points for further open discussion in class:
  - Fit the eight school example according to a frequentist approach and compare the results with those from the Bayesian analysis.
  - Repeat the analysis by changing the likelihood specification and/or the prior specification.
  - Compare the latter model with the Gaussian-Gaussian model in terms of predictive information criteria (AIC, DIC, WAIC, LOOIC...).

Extending linear models: predicting US elections i

Forecasting US presidential elections (BDA book, sect. 15.2)
Political scientists in the US have been interested in the idea that national elections are highly predictable, in the sense that one can accurately forecast election results using information publicly available several months before the election. We provide an example (see BDA, sect. 15.2 for further details) using a hierarchical linear model estimated from the elections through 1988 to forecast the 1992 elections.
The units of analysis are results in each state from each of the 11 presidential elections from 1948 to 1988. The response variable is the Democratic party candidate's share of the two party-vote for president in that state and year. In total, we have 511 observations.

Extending linear models: predicting US elections ii

- Let's have a preliminary graphical look at the data. The Figure below displays the democratic share of the two-party vote for president, for each state, in 1984 (x-axis) and 1988 (y-axis) in the left plot, and the same share for each state in 1972 (x-axis) and 1976 (y-axis) in the right plot.

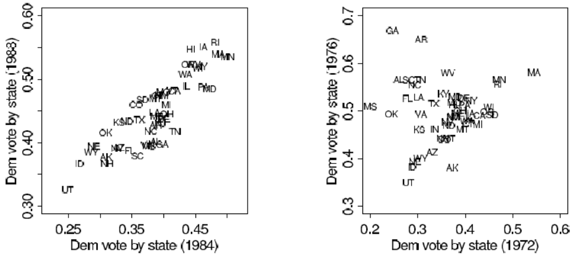


Figure 15.1 *(a) Democratic share of the two-party vote for president, for each state, in 1984 and 1988. (b) Democratic share of the two-party vote for president, for each state, in 1972 and 1976.*

Extending linear models: predicting US elections iii

- The left panel suggests that the presidential vote may be strongly predictable from one election to the next: the points fall close to a straight line! A linear model predicting y from x is reasonable and relatively precise.
- However, the pattern is not always so strong, see the right panel where the relation is not close to linear. Nevertheless, we can reveal some patterns:
    - the greatest outlying point, on the upper left, is Georgia (GA), the home state of Jimmy Carter, the Democratic candidate in 1976;
    - the other outlying points, all on the upper left side, are other states in the South, Carter's home region.

Extending linear models: predicting US elections iv

- Then, it appears that it may be possible to create a good linear fit by including other predictors in addition to the Democratic share of the vote in the previous election, such as indicator variables for the candidates' home states and home regions. (For political analysis, the US is typically divided into four regions: Northeast, South, Midwest, and West, with each region containing ten or more states.)
- We start by fitting a preliminary, non-hierarchical regression model trying to capture three levels of variations—at nation, regional and state level—with the following explanatory variables:

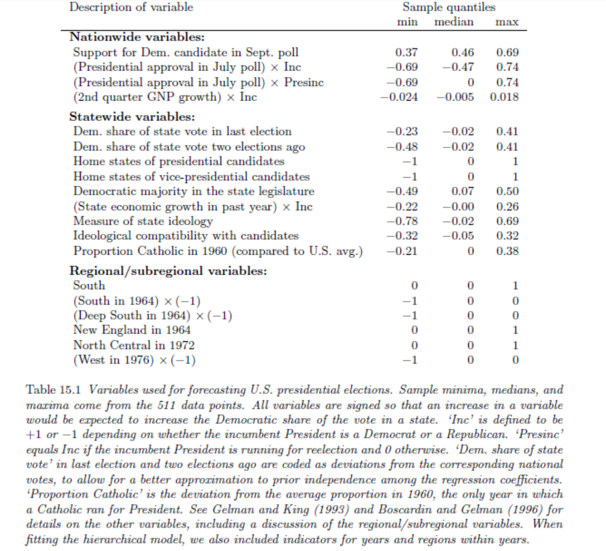
Extending linear models: predicting US elections v

  - *Nationwide*: nation measures of popularity of the candidates, popularity incumbent president, measures of condition of the economy in the past two years.
  - *Regional*: home-region indicators for the candidates and various adjustments for past elections in which regional voting had been important.
  - *Statewide*: Democratic's party share of the state's vote in recent presidential elections, measures of the state's economy and political ideology, and home-state indicators.

  See the next Table for a complete overview about the explanatory variables.

- We fit a classical regression including all the variables in the Table to the data up to 1988.

Extending linear models: predicting US elections vi



Description of variable	Sample quantiles		
	min	median	max
Nationwide variables:			
Support for Dem. candidate in Sept. poll	0.37	0.46	0.69
(Presidential approval in July poll) × Inc	−0.69	−0.47	0.74
(Presidential approval in July poll) × Presinc	−0.69	0	0.74
(2nd quarter GNP growth) × Inc	−0.024	−0.005	0.018
Statewide variables:			
Dem. share of state vote in last election	−0.23	−0.02	0.41
Dem. share of state vote two elections ago	−0.48	−0.02	0.41
Home states of presidential candidates	−1	0	1
Home states of vice-presidential candidates	−1	0	1
Democratic majority in the state legislature	−0.49	0.07	0.50
(State economic growth in past year) × Inc	−0.22	−0.00	0.26
Measure of state ideology	−0.78	−0.02	0.69
Ideological compatibility with candidates	−0.32	−0.05	0.32
Proportion Catholic in 1960 (compared to U.S. avg.)	−0.21	0	0.38
Regional/subregional variables:			
South	0	0	1
(South in 1964) × (−1)	−1	0	0
(Deep South in 1964) × (−1)	−1	0	0
New England in 1964	0	0	1
North Central in 1972	0	0	1
(West in 1976) × (−1)	−1	0	0

Table 15.1 *Variables used for forecasting U.S. presidential elections. Sample minima, medians, and maxima come from the 511 data points. All variables are signed so that an increase in a variable would be expected to increase the Democratic share of the vote in a state. 'Inc' is defined to be +1 or −1 depending on whether the incumbent President is a Democrat or a Republican. 'Presinc' equals Inc if the incumbent President is running for reelection and 0 otherwise. 'Dem. share of state vote' in last election and two elections ago are coded as deviations from the corresponding national votes, to allow for a better approximation to prior independence among the regression coefficients. 'Proportion Catholic' is the deviation from the average proportion in 1960, the only year in which a Catholic ran for President. See Gelman and King (1993) and Boscardin and Gelman (1996) for details on the other variables, including a discussion of the regional/subregional variables. When fitting the hierarchical model, we also included indicators for years and regions within years.*

Extending linear models: predicting US elections vii

- The ordinary linear regression model ignores the year-by-year structure of the data, treating them as 511 independent observations, rather than 11 sets of roughly 50 related observations each. The feature of these data that such a model misses is that partisan support across the states does not vary independently: in other words, because of the known grouping into years, the assumption of exchangeability among the 511 observations does not make sense, *even after controlling for the explanatory variables*.
- At this stage, one would need to *criticize* the model by assessing how and whether this fits the observed data. The crucial step of model checking (or also goodness-of-fit) will be addressed later on.
- An important use of the model is to forecast the nationwide outcome of the presidential election.

Extending linear models: predicting US elections viii

- Then, to check whether correlation of the observations from the same election has a substantial effect on nationwide forecasts, we could try to make some predictions to assess the model effectiveness: for instance, we could simulate a *test variable* that reflects the average precision of the model in predicting the national result, i.e. the square root of the average of the squared nationwide realized residuals for the 11 general elections in the dataset.
- The next Figure displays the values of this test variable (obtained from the posterior distribution of β) against the hypothetical replicated values under the model: the practical consequence of the failure of the model is that its forecasts of national election results are falsely precise.

Extending linear models: predicting US elections ix

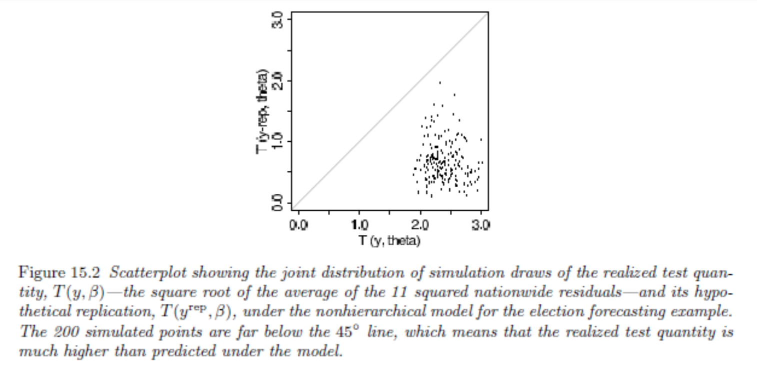


Figure 15.2 *Scatterplot showing the joint distribution of simulation draws of the realized test quantity, $T(y, \beta)$—the square root of the average of the 11 squared nationwide residuals—and its hypothetical replication, $T(y^{\rm rep}, \beta)$, under the nonhierarchical model for the election forecasting example. The 200 simulated points are far below the 45° line, which means that the realized test quantity is much higher than predicted under the model.*

Extending linear models: predicting US elections x

- We can improve the regression by:
  - adding an additional predictor for each year to serve as an indicator for nationwide partisan shifts unaccounted for by the other national variables; this adds 11 new components of β corresponding to the 11 election years in the data;
  - adding 44 region $\times$ year indicator variables to cover all regions in the elections and capture regional variability; because the South tends to act as a special region of the US politically, we give the 11 Southern regional variables their own common variance, and treat the remaining 33 regional variables as exchangeable with their own variance.

Extending linear models: predicting US elections xi

In total, we add 55 new β parameters and three new variance components to the model. We can then write the varying-coefficients model for data in state s, region $r(s)$, and year t as:

$$\begin{aligned} y_{st} &\sim \mathcal{N}(X_{st}\beta + \gamma_{r(s)t} + \delta_t, \sigma_y^2) \\ \gamma_{rt} &\sim \begin{cases} \mathcal{N}(0, \tau_{\gamma 1}^2) & \text{for } r = 1, 2, 3 \quad \text{(non-South)} \\ \mathcal{N}(0, \tau_{\gamma 2}^2) & \text{for } r = 4 \qquad\quad\ \text{(South)} \end{cases} \\ \delta_t &\sim \mathcal{N}(0, \tau_\gamma^2), \end{aligned} \tag{19}$$

where $\beta, \sigma_y, \tau_{\gamma 1}, \tau_{\gamma 2}, \tau_\gamma$ are hyperparameters.

- We repeat the same simulations done for the non-hierarchical model (see next Figure) and we realize that these hierarchical simulations fit the observed data much better.

Extending linear models: predicting US elections xii

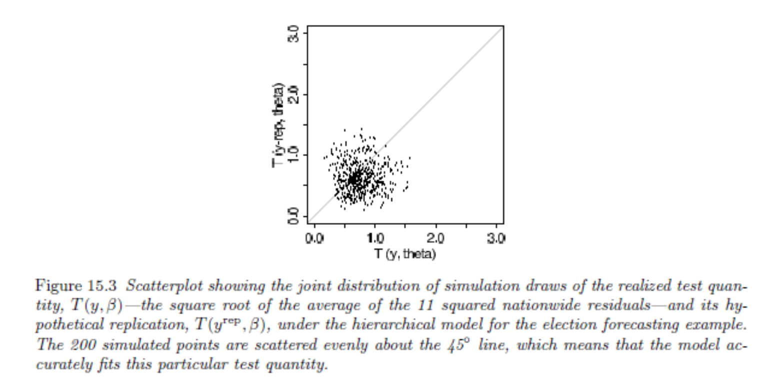


Figure 15.3 *Scatterplot showing the joint distribution of simulation draws of the realized test quantity, $T(y, \beta)$—the square root of the average of the 11 squared nationwide residuals—and its hypothetical replication, $T(y^{\text{rep}}, \beta)$, under the hierarchical model for the election forecasting example. The 200 simulated points are scattered evenly about the 45° line, which means that the model accurately fits this particular test quantity.*

Extending linear models: predicting US elections xiii

- In the next and final Figure we report the state-by-state predictions applied to data from 1992, with a forecasted 85% probability that the Democrats would win the national elecoral vote total. The forecasts for individual states have predictive standard errors between 5% and 6%.

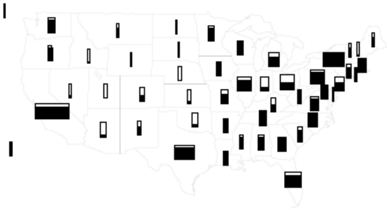


Figure 6.1 *Summary of a forecast of the 1992 U.S. presidential election performed one month before the election. For each state, the proportion of the box that is shaded represents the estimated probability of Clinton winning the state; the width of the box is proportional to the number of electoral votes for the state.*

Extending linear models: predicting US elections xiv

- To sum up, there are three main advantages of the hierarchical model here:
    - It allows the modeling of correlation within election years and regions.
    - Including the year and region × year terms without a hierarchical model, or not including these terms at all, corresponds to special cases of the hierarchical model with $\tau = \infty$ (no-pooling) or 0 (complete pooling), respectively. The more general model allows for a reasonable compromise between these extremes.
    - Predictions will have additional components of variability of regions and year and should therefore be more reliable.

Sum-up about the forecasting US election example

- You find the data in the official Moodle course page: the file name is `forecasting_us_elections.txt`.
- For further open discussion in class:
  - Load the data.
  - Reproduce the Figure 15.1 of BDA, Chapter 15 (see the slides).
  - Try to fit the simple non-hierarchical regression model and the multilevel model in (19).
  - Try to reproduce the main results (see Chapter 15 from BDA).

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`):

$$\begin{aligned}\Pr(y_i = 1) &= \text{logit}^{-1}(\alpha_{j(i)} + \beta^{\text{female}}\text{female}_i + \beta^{\text{black}}\text{black}_i),\\ \alpha_j &\sim \mathcal{N}(\mu_\alpha, \tau^2_{\text{state}}),\ j = 1, \dots, 51\end{aligned} \tag{20}$$

  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.

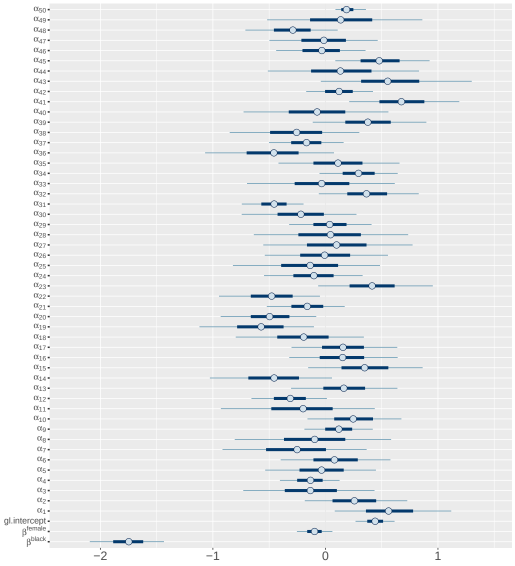
- The top part display gives the estimate of the average intercept (μ_α), the coefficients for `black` and `female`, and their standard errors.

- 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.

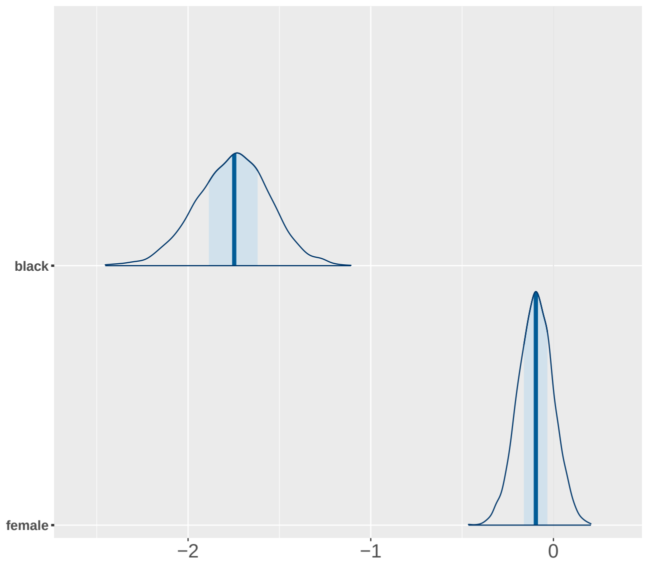
1988 US polls. Varying-intercept model v

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

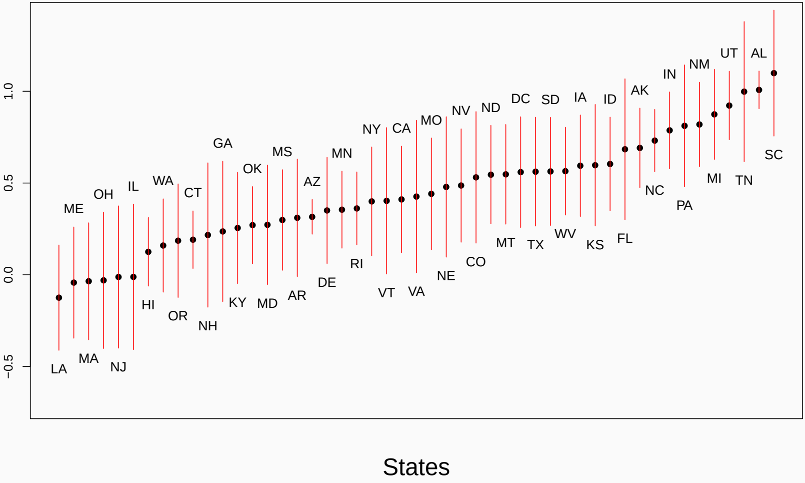
Many issues arise when you fit a model:

- 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 ± s.e.



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*:

$$
\begin{aligned}
&\mathrm{Pr}(y_i = 1) = \mathrm{logit}^{-1}(\alpha_{j(i)} + \beta_{j(i)}^{\mathrm{female}}\mathrm{female}_i + \beta^{\mathrm{black}}\mathrm{black}_i), \quad i = 1, \ldots, n \\
&\begin{pmatrix} \alpha_j \\ \beta_j^{\mathrm{female}} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_\alpha \\ \mu_\beta \end{pmatrix}, \begin{pmatrix} \tau_\alpha^2 & \rho\tau_\alpha\tau_\beta \\ \rho\tau_\alpha\tau_\beta & \tau_\beta^2 \end{pmatrix} \right), \quad j = 1, \ldots, 51,
\end{aligned}
\tag{21}
$$

where τ_α^2 and τ_β^2 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

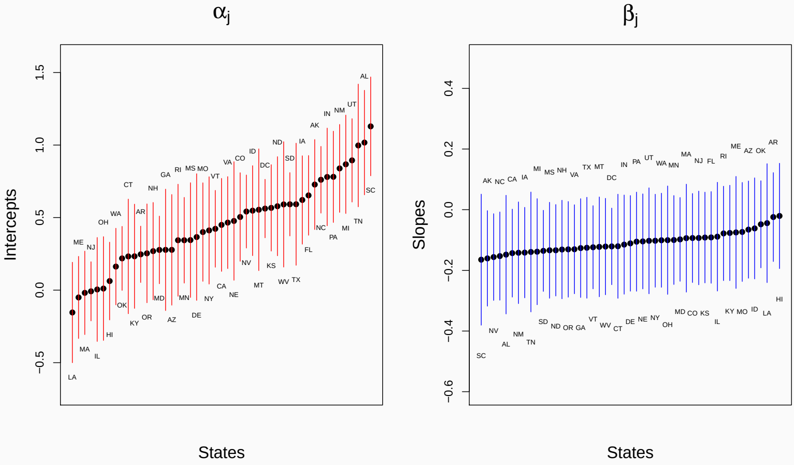
```
# Bayesian fit

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 × ethnicity and age × 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:

$$
\begin{aligned}
\Pr(y_i = 1) = \operatorname{logit}^{-1}(\beta_0 + \beta^{\mathrm{female}}\mathrm{female}_i + \beta^{\mathrm{black}}\mathrm{black}_i + \\
+ \beta^{\mathrm{female.black}}\mathrm{female}_i \cdot \mathrm{black}_i + \alpha^{\mathrm{age}}_{k(i)} + \alpha^{\mathrm{edu}}_{l(i)} + \alpha^{\mathrm{age.edu}}_{k(i),l(i)} + \alpha^{\mathrm{state}}_{j(i)}),\\
\alpha^{\mathrm{state}}_j &\sim \mathcal{N}(\alpha^{\mathrm{region}}_{m(j)} + \beta^{\mathrm{v.prev}}\mathrm{v.prev}_j, \sigma^2_{\mathrm{state}}),\ j = 1, \ldots, 51\\
\alpha^{\mathrm{age}}_k &\sim \mathcal{N}(0, \sigma^2_{\mathrm{age}}),\ k = 1, \ldots, 4\\
\alpha^{\mathrm{edu}}_l &\sim \mathcal{N}(0, \sigma^2_{\mathrm{edu}}),\ l = 1, \ldots, 4\\
\alpha^{\mathrm{age.edu}}_{k,l} &\sim \mathcal{N}(0, \sigma^2_{\mathrm{age.edu}}),\ k = 1, \ldots, 4,\ \text{ł} = 1, \ldots, 4\\
\alpha^{\mathrm{region}}_m &\sim \mathcal{N}(0, \sigma^2_{\mathrm{region}}),\ m = 1, \ldots, 5.
\end{aligned}
\tag{22}
$$

Sum-up about the US 1988 election example

- You find the data and the R code with some preliminary analysis and some model fits in the official Moodle course page: be careful, you need more than one file to load and process the data. For further details, check the G&H book, chapter 14. Check the code and repeat the analysis.

- For further open discussion in class:
  - Fit a complete-pooling and a no-pooling model and compare the estimates with one or more multilevel models.
  - Try to reproduce, or at least try to make a similar plot, the Figure 14.2 from the book G&H.
  - Fit the model (22) following both a Bayesian and a frequentist approach, and produce the estimates, also from a graphical perspective (hint: use the `bayesplot` package).
  - Divide the data in *training* and *test* sets and use the training data to predict the test data (hint: use measures of predictive accuracy such as accuracy, sensitivity and specificity).
  - Following (20), (21), and (22), write (not fit) a model where you want to further include the individual categorical predictor `income`, defined on a 1–5 scale (1 is poor, 5 is wealthy), and the regional continuous predictor `reg_income`, expressing an average of the citizens' income in the 5 regions in the past two years.

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:

$$\begin{aligned} \text{Poisson regression} : y_i &\sim \mathsf{Poisson}(u_i e^{X_i \beta}), \\ \text{overdispersed Poisson regression} : y_i &\sim \mathsf{Poisson}(u_i e^{X_i \beta + \epsilon_i}) \\ \epsilon_i &\sim \mathcal{N}(0, \sigma_\epsilon^2). \end{aligned} \tag{23}$$

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

Police stops in New York

Police stops data (G&H book, chapter 15)
There have been complaints in New York City and elsewhere that the police harass members of ethnic minority groups. The police has a policy of keeping records on every "stop and frisk", and this information was collated for all stops over a 15-month period in 1998-1999. One could analyse these data to see to what extent different ethnic groups were stopped by the police. Focus is on blacks (African Americans), hispanics (Latinos), and whites (European Americans). For each ethnic group $e = 1, 2, 3$ and precint of the New York City $p = 1, \dots, 75$, we model the number of stops using an overdispersed Poisson regression. The exposure u_{ep} is the number of arrests by people of ethnic group e in precint p in the previous year as recorded by the Department of Criminal Justices Services (the exposure is multiplied by 15/12 to scale to a 15-month period), so that $\log((15/12)u_{ep})$ is an offset.

Police stops: an overdispersed Poisson regression i

- For each ethnic group e and precint p:

$$\begin{aligned} y_{ep} &\sim \text{Poisson}\left(\frac{15}{12} u_{ep} e^{\mu+\alpha_e+\beta_p+\epsilon_{ep}}\right) \\ \alpha_e &\sim \mathcal{N}(0, \sigma_\alpha^2) \\ \beta_p &\sim \mathcal{N}(0, \sigma_\beta^2) \\ \epsilon_{ep} &\sim \mathcal{N}(0, \sigma_\epsilon^2), \end{aligned} \tag{24}$$

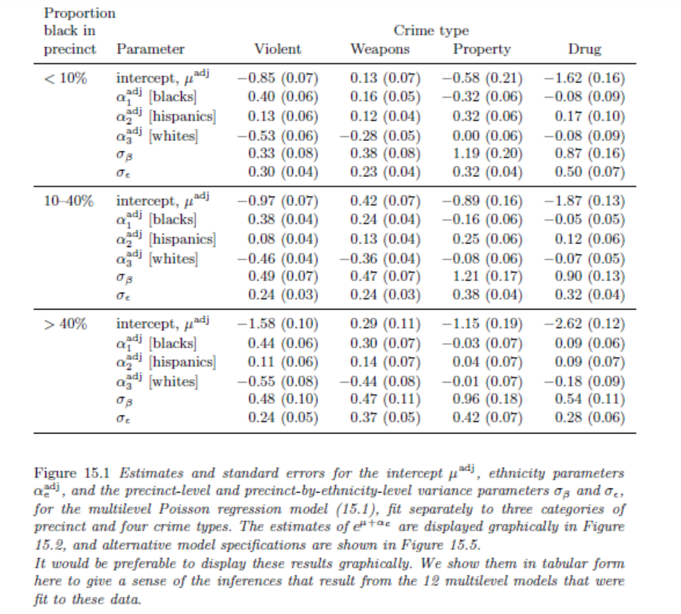
  where the coefficients α_e control the ethnic group, the β_p's adjust for variation among precints, and the ϵ_{ep}'s allow for overdispersion.

- Identifiability constraints: when comparing ethnic groups, we could look at the ethnicity coefficients relative to their mean:

$$\alpha_e^{\text{adj}} = \alpha_e - \bar{\alpha}, \quad e = 1, 2, 3.$$

- Having done this, we also adjust the intercept of the model accordingly: $\mu^{\text{adj}} = \mu + \bar{\alpha}$. Now $\mu^{\text{adj}} + \alpha_e^{\text{adj}} = \mu + \alpha_e$ for each ethnic group e, and so we can use μ^{adj} and α^{adj} in place of μ and α without changing the model for the data. See next Figures for the estimates.

Police stops: an overdispersed Poisson regression ii



		Crime type			
Proportion black in precinct	Parameter	Violent	Weapons	Property	Drug
$<$ 10%	intercept, μ^{adj}	−0.85 (0.07)	0.13 (0.07)	−0.58 (0.21)	−1.62 (0.16)
	α_1^{adj} [blacks]	0.40 (0.06)	0.16 (0.05)	−0.32 (0.06)	−0.08 (0.09)
	α_2^{adj} [hispanics]	0.13 (0.06)	0.12 (0.04)	0.32 (0.06)	0.17 (0.10)
	α_3^{adj} [whites]	−0.53 (0.06)	−0.28 (0.05)	0.00 (0.06)	−0.08 (0.09)
	σ_β	0.33 (0.08)	0.38 (0.08)	1.19 (0.20)	0.87 (0.16)
	σ_ϵ	0.30 (0.04)	0.23 (0.04)	0.32 (0.04)	0.50 (0.07)
10–40%	intercept, μ^{adj}	−0.97 (0.07)	0.42 (0.07)	−0.89 (0.16)	−1.87 (0.13)
	α_1^{adj} [blacks]	0.38 (0.04)	0.24 (0.04)	−0.16 (0.06)	−0.05 (0.05)
	α_2^{adj} [hispanics]	0.08 (0.04)	0.13 (0.04)	0.25 (0.06)	0.12 (0.06)
	α_3^{adj} [whites]	−0.46 (0.04)	−0.36 (0.04)	−0.08 (0.06)	−0.07 (0.05)
	σ_β	0.49 (0.07)	0.47 (0.07)	1.21 (0.17)	0.90 (0.13)
	σ_ϵ	0.24 (0.03)	0.24 (0.03)	0.38 (0.04)	0.32 (0.04)
$>$ 40%	intercept, μ^{adj}	−1.58 (0.10)	0.29 (0.11)	−1.15 (0.19)	−2.62 (0.12)
	α_1^{adj} [blacks]	0.44 (0.06)	0.30 (0.07)	−0.03 (0.07)	0.09 (0.06)
	α_2^{adj} [hispanics]	0.11 (0.06)	0.14 (0.07)	0.04 (0.07)	0.09 (0.07)
	α_3^{adj} [whites]	−0.55 (0.08)	−0.44 (0.08)	−0.01 (0.07)	−0.18 (0.09)
	σ_β	0.48 (0.10)	0.47 (0.11)	0.96 (0.18)	0.54 (0.11)
	σ_ϵ	0.24 (0.05)	0.37 (0.05)	0.42 (0.07)	0.28 (0.06)

Figure 15.1 *Estimates and standard errors for the intercept* μ^{adj}*, ethnicity parameters* α_e^{adj}*, and the precinct-level and precinct-by-ethnicity-level variance parameters* σ_β *and* σ_ϵ*, for the multilevel Poisson regression model (15.1), fit separately to three categories of precinct and four crime types. The estimates of* $e^{\mu+\alpha_e}$ *are displayed graphically in Figure 15.2, and alternative model specifications are shown in Figure 15.5.*
It would be preferable to display these results graphically. We show them in tabular form here to give a sense of the inferences that result from the 12 multilevel models that were fit to these data.

Police stops: an overdispersed Poisson regression iii

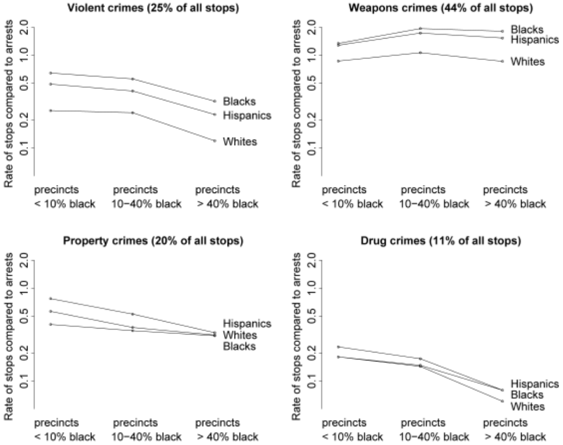


Figure 2. Estimated rates $e^{\mu+\alpha_e}$ at which people of different ethnic groups were stopped for different categories of crime, as estimated from hierarchical regressions (1) using previous year's arrests as a baseline and controlling for differences between precincts. Separate analyses were done for the precincts that had <10%, 10–40%, and >40% black population. For the most common stops—violent crimes and weapons offenses—blacks and Hispanics were stopped about twice as often as whites. Rates are plotted on a logarithmic scale. Numerical estimates and standard errors are given in Table 1.

Police stops: an overdispersed Poisson regression iv

- The previous Figure shows that, for the most frequent categories of stops (violent crimes and weapons offenses) blacks and hispanics were much more likely to be stopped than whites, in all categories of precints. For violent crimes, blacks and hispanics were stopped 2.5 times and 1.9 times as often as whites, respectively, and for weapons crimes, blacks and hispanics were stopped 1.8 times and 1.6 times as often as whites.
- We could extend the model (24), for instance by changing the batching of precints, or altering the role played by the previous year's arrests.
- Alternatively, one could also include some precint-level predictors:

$$y_{ep} \sim \text{ Poisson}\left(\frac{15}{12}u_{ep}e^{\mu+\alpha_e+\zeta_1 z_{1p}+\zeta_2 z_{2p}+\beta_p+\epsilon_{ep}}\right),$$

  where z_{1p} and z_{2p} represent the proportion of blacks and hispanics in precint p.
- For further modeling details, see the G&H Book, Chapter 15, and the paper 'An Analysis of the New York City Police Department's "Stop-and-Frisk" Policy in the Context of Claims of Racial Bias', by Gelman, Fagan and Kiss.

Sum-up about the NYC police stops data example

- You find the data in the official Moodle course page: the file name is `nyc_arrests.txt`.
- For further open discussion in class:
    - Load the data.
    - Fit the model (24) from a Bayesian and a frequentist point of view, provide the estimates (also from a graphical perspective) and comment the results.
    - Extend the model: write a modeling extension (fit is not required) where a further continuous covariate `income_ethnicity`, expressing an average income for the ethnicity e in New-York City, is available. (Hint: there is not a unique way to incorporate it). Finally, discuss the eventual sign of the estimated coefficient from a "socio-political" perspective.

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.

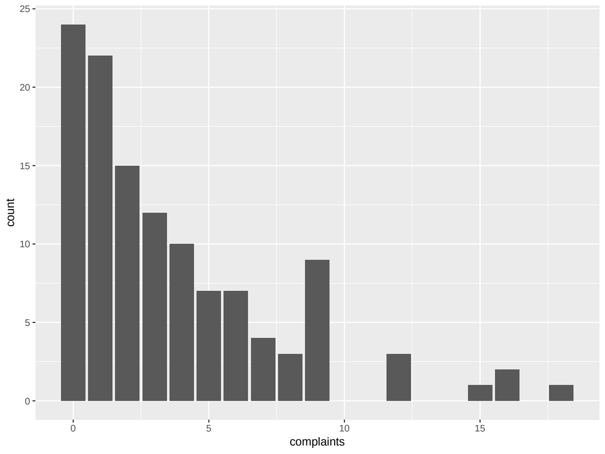
Discrete data regression: cockroaches data ii

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

$$\begin{aligned} \text{complaints}_i &\sim \text{Poisson}(\lambda_i) \\ \lambda_i &= \exp(\eta_i) \\ \eta_i &= \alpha + \beta\, \text{traps}_i \end{aligned} \tag{25}$$

Poisson regression: cockroaches data ii

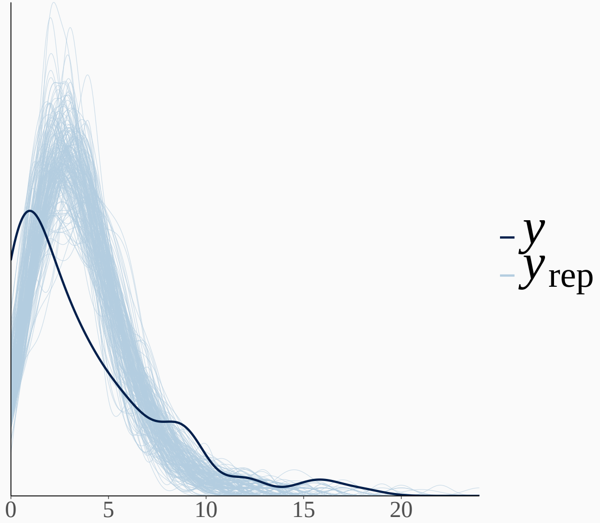
- We fit the model in Stan and we obtain the following posterior estimates (R output):

```
        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).

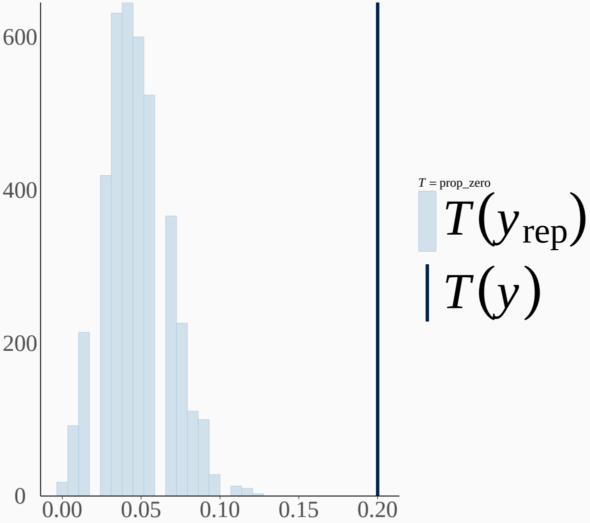
Poisson regression: cockroaches data iii

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:



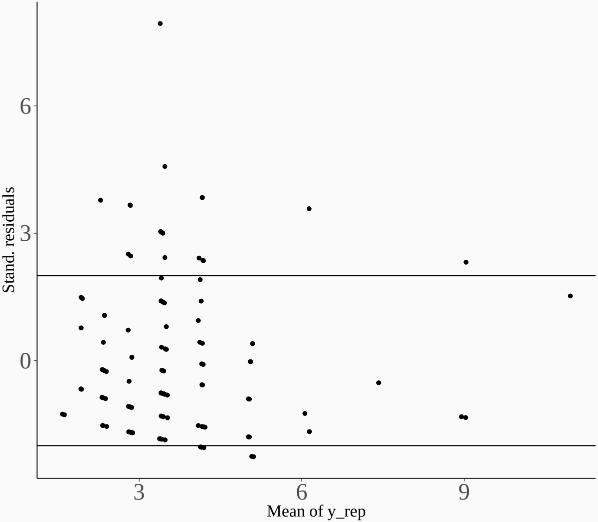
Poisson regression: cockroaches data v

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.

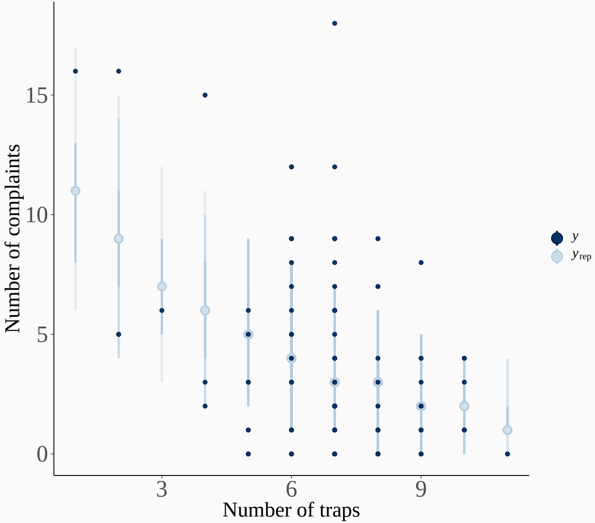
Poisson regression: cockroaches data vi

Standardized residuals:



Poisson regression: cockroaches data vii

Predictive intervals:



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

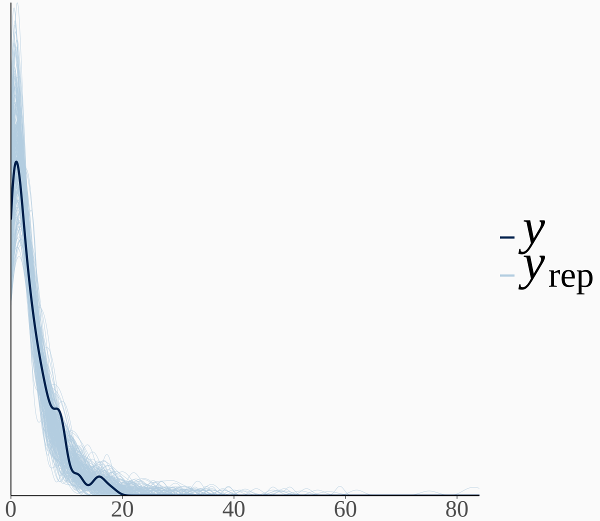
Hierarchical NB regression i

- 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 (25) encoding hierarchical structure for the building and considering an offset term. Thus, for each complaint i we have:

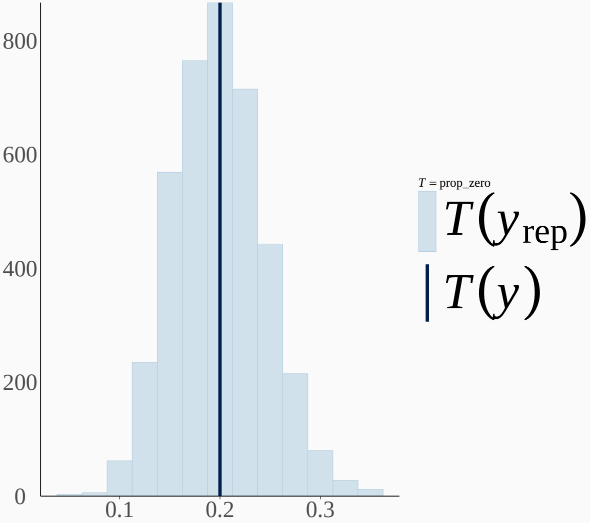
$$\begin{aligned} \text{complaints}_{ib} &\sim \mathrm{NegBin}(\lambda_{ib}, \phi) \\ \lambda_{ib} &= \exp(\eta_{ib}) \\ \eta_{ib} &= \alpha_{b(i)} + \beta\, \mathrm{traps}_i + \beta_{\mathrm{super}}\, \mathrm{super}_i + \mathrm{log\_sq\_foot}_i \\ \alpha_b &\sim \mathcal{N}(\mu, \tau_\alpha^2), \\ \phi &\sim \mathcal{N}^+(0, 1) \end{aligned} \tag{26}$$

  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).

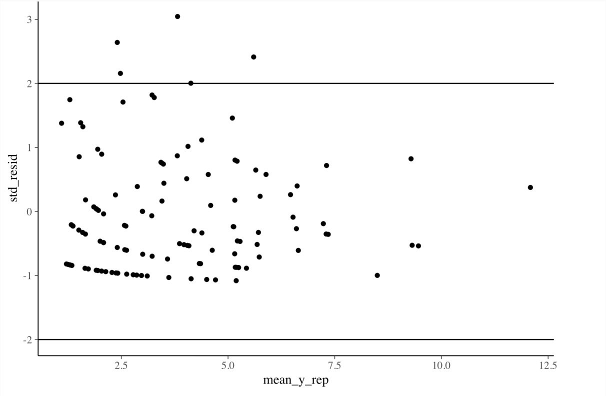
Hierarchical NB regression ii



Hierarchical NB regression iii



Hierarchical NB regression iv



Better!

Sum-up about the cockroaches data example

- You find the data and the relevant `R` code in the official Moodle course page.
- For further open discussion in class:
    - Check the code and repeat the analysis. Careful: all the analysis have been done through the canonical use of the Stan software, but feel free to use the functions `stan_glmer` and/or `glmer`.
    - Extend the model: write a modeling extension (fit is not required) where also the time structure is included in the model.

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.