UNIVERSITÀ DEGLI STUDI DI TRIESTE

Dipartimento di scienze economiche, aziendali, matematiche e statistiche "Bruno de Finetti"

Bayesian Statistics

Hierarchical models

Leonardo Egidi

A.A. 2018/19

Indice

Motivations

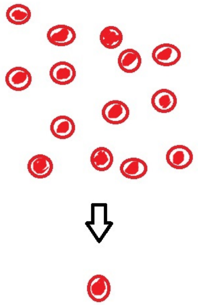
A common problem in applied statistics is modelling individuals/objects of a *population*.

Within this population, there may be some *subpopulations* sharing some common features. Thus, we should statistically acknowledge these different groups.

Multilevel/hierarchical models are extensions of regression in which data are structured in groups and coefficients can vary by group. We start with simple grouped data-person within cities, e.g.-where some information is available on individuals and some information is at the group level.

Motivations

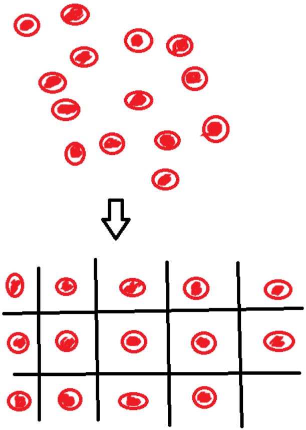
If we assume that every individual is equivalent then we can pool the data, but only at the expense of bias $\Leftrightarrow$ Complete pooling.



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

Motivations

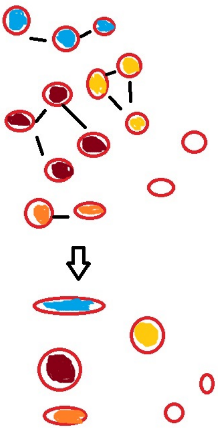
Conversely, modelling every individual separately avoids any bias, but then the data becomes very sparse and inferences weak $\Leftrightarrow$ No pooling.



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

Motivations

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



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

Motivations

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 modelled 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 - 1

In order to formalize this approach we need to consider exchangeability.

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 - 2

- This symmetry is represented probabilistically by exchangeability: the parameters $(\theta_1, \dots, \theta_J)$ are exchangeable in their joint prior distribution if $\pi(\theta_1, \dots, \theta_J)$ is invariant to permutations of the indexes $(1, \dots, J)$.
- 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 die and weigh the die 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 - 3

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

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

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

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

The fundamental concept of exchangeability - 4

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

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

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

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

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

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

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

Example: M5S voters. Role of exchangeability in inference.

M5S voters

Suppose you are an asian guy and let $\theta_1, \dots, \theta_5$ are the proportions of voters for MoVimento 5 Stelle (M5S) in five Italian regions from the last polls for the European Elections (May, 23). The regions, here in a random order, are: Piemonte, Liguria, Umbria, Puglia, Campania. What can you say about the M5S 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 proportion: 13.2, 14.3, 18.4, 21.5. Remember, you are asian, you do not know everything about M5S...what can you say about θ_5?

Example: M5S voters. Role of exchangeability in inference.

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.

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

You reconsider the four voters' proportions. You know that Luigi Di Maio is born in Pomigliano d'Arco, Campania, mmh...Giggino is loved for sure by his fellows, at least 30% of them will support him! Maybe the missing proportion θ_5 is Campania...You end up with a not exchangeable prior distribution.

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:

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

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}, $i = 1, \dots, n$, $j = 1, \dots J$;

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

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

$$\theta_j \sim \pi(\theta|\phi) \quad \text{prior}$$

- heterogeneity level: unobserved ϕ

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

Indice

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 obained by stratified or cluster sampling

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.

Extending linear models: radon data

Radon data

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

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:

$$
\begin{aligned}
y_{ij} &\sim \mathcal{N}(\alpha_{j(i)}, \sigma^2),\ i = 1, \ldots, n,\ \text{Individual level} \\
\alpha_j &\sim \mathcal{N}(\mu_\alpha, \tau^2),\ j = 1, \ldots, J, \qquad \text{Group level}
\end{aligned}
\tag{5}
$$

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

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

A third option is hierarchical modelling: estimates $\pm$ standard errors are plotted against the number of observations for each county.

Partial pooling with no predictors

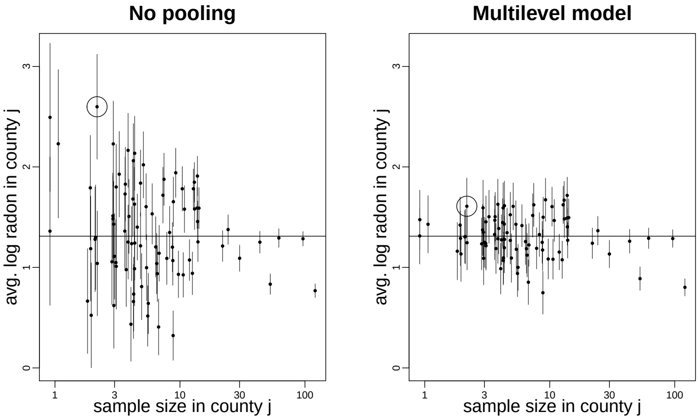


Figure: Estimates ± standard errors for the average log radon levels in Minnesota counties plotted versus the number of observations in the county.

Partial pooling with no predictors

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

$$\hat{\alpha}_j = \frac{\frac{n_j}{\sigma^2}\bar{y}_j + \frac{1}{\tau^2}\bar{y}_{\text{all}}}{\frac{n_j}{\sigma^2} + \frac{1}{\tau^2}} \tag{6}$$

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

The weighted average (6) 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

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:

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

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

Partial pooling with predictors

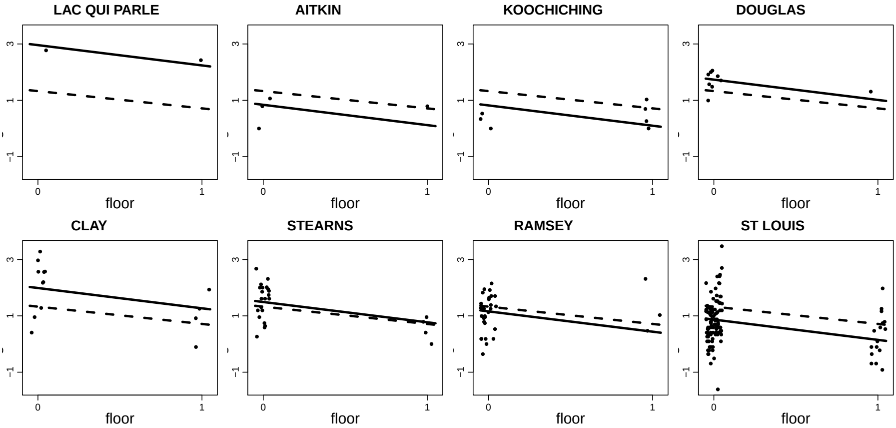


Figure: Complete poolimg (dashed lines) and no pooling (solid lines) for 8 counties in Minnesota.

Partial pooling with predictors

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 (7) via the function `stan_lmer` of the `rstanarm` R package, and plot again the estimates.

Partial pooling with predictors

```
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

```
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

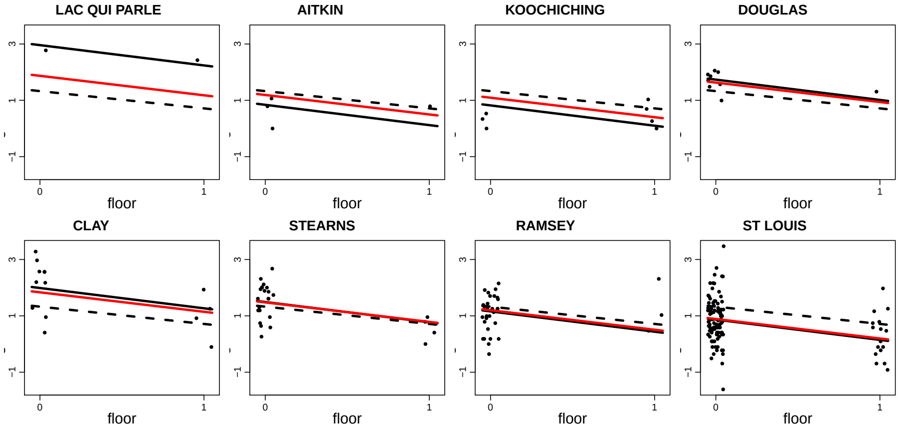


Figure: Complete pooling (dashed lines), no pooling (solid lines) and partial pooling (solid red lines).

Partial pooling with predictors

- The estimated line from the hierarchical model (7) 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

We can generalize equation (6) as follows:

$$\hat{\alpha}_j \approx \frac{\frac{n_j}{\sigma^2}}{\frac{n_j}{\sigma^2} + \frac{1}{\tau_\alpha^2}}(\bar{y}_j - \beta\bar{x}_j) + \frac{\frac{1}{\tau_\alpha^2}}{\frac{n_j}{\sigma^2} + \frac{1}{\tau_\alpha^2}}\mu_\alpha, \tag{8}$$

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

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

Eight schools example

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

Eight schools example (BDA, 5.5)

A study waw 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, the choice of the prior is of substantial importance here.

Eight schools

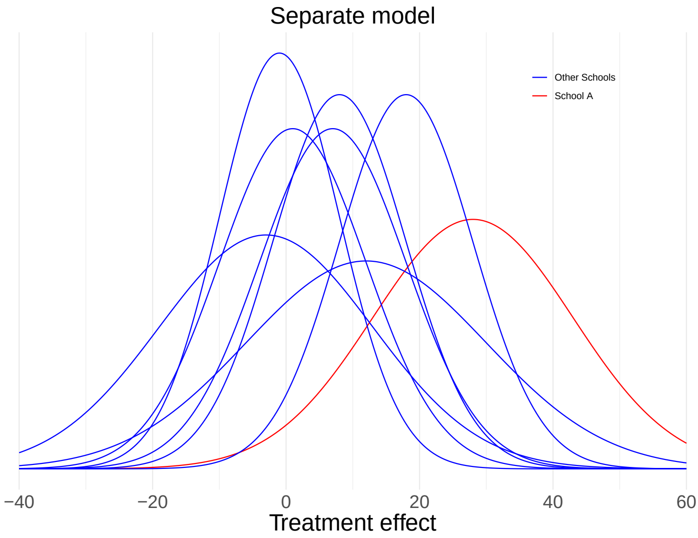
We denote with y_{ij} the result of the i-th test in the j-th school. We assume the following model:

$$\begin{aligned} y_{ij} &\sim \mathcal{N}(\theta_j, \sigma_y^2) \\ \theta_j &\sim \mathcal{N}(\mu, \tau^2) \end{aligned} \tag{9}$$

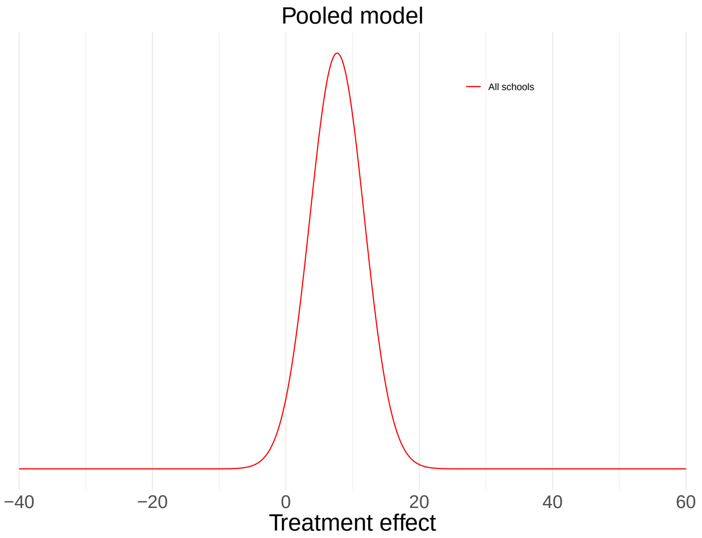
Do some schools perform better/worse according to these coaching effects?
We will make three distinct analysis: separate analysis, pooled analysis and hierarchical modelling.

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, $s = (15, 10, 16, 11, 9, 11, 10, 18)$.

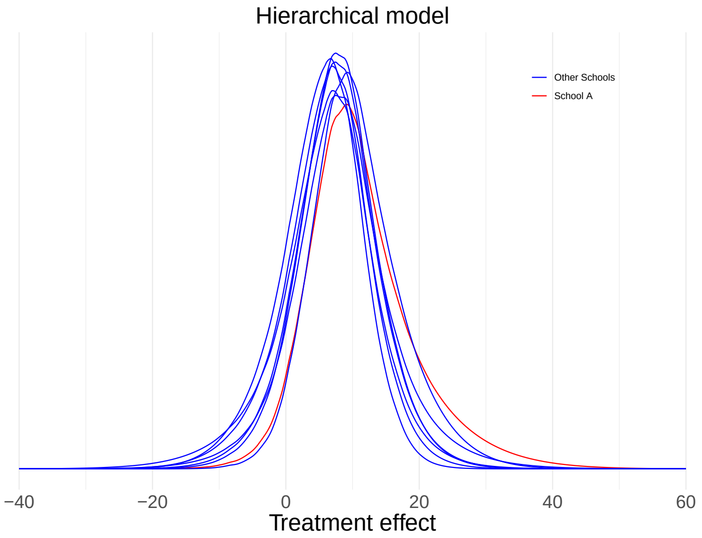
Eight schools: separate analysis



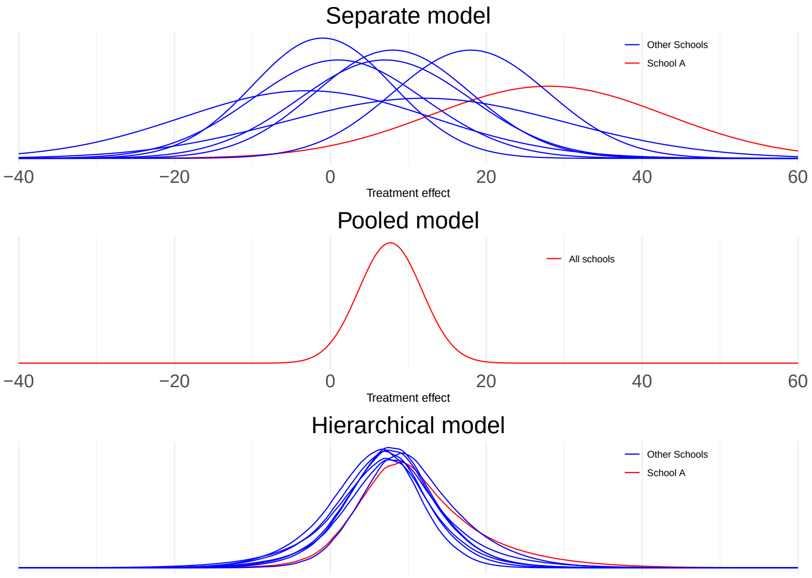
Eight schools: pooled analysis



Eight schools: hierarchical model



Eight schools: three models



Eight schools: three models

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.

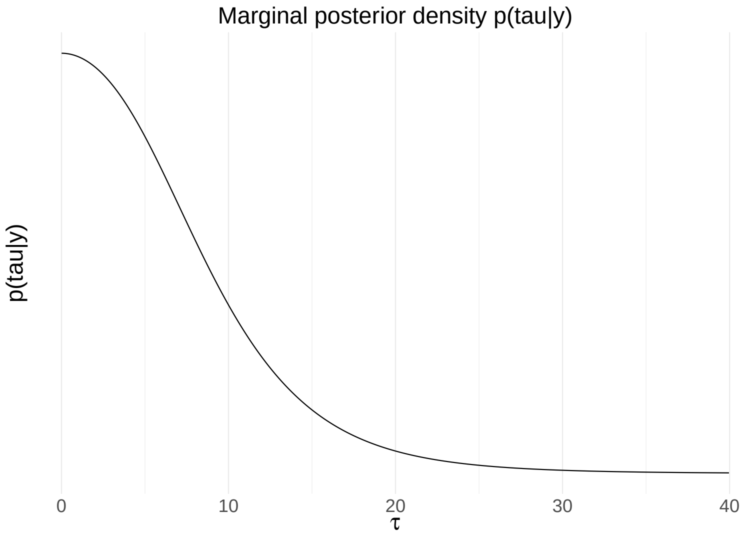
However, both the extreme analysis have difficulties.

Eight schools: three models

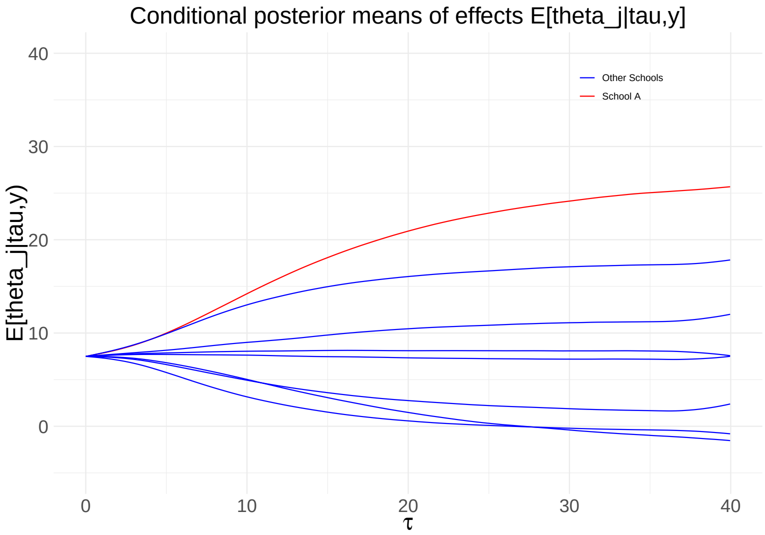
Other comments:

- 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.

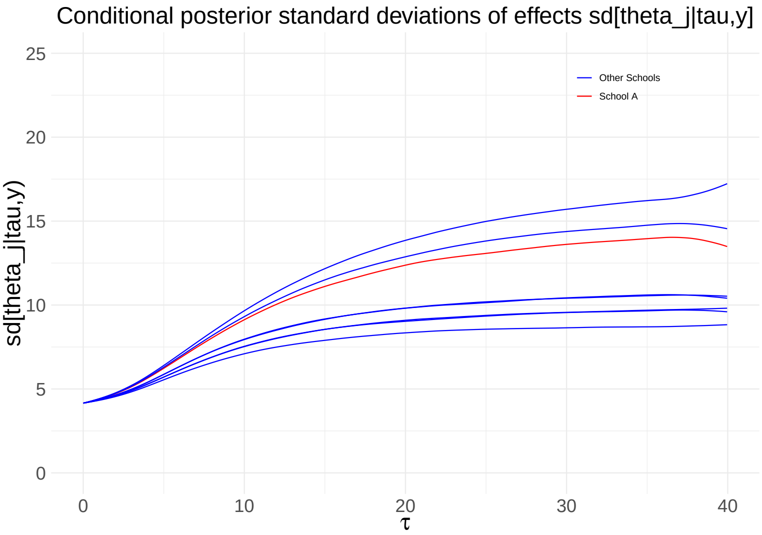
Eight schools: posterior summaries for hierarchical model



Eight schools: posterior summaries for hierarchical model



Eight schools: posterior summaries for hierarchical model



Eight schools: posterior summaries for hierarchical model

- In the plot for the marginal posterior $p(\tau|y)$, $\tau = 0$ is the most likely value (no variation in θ, complete pooling).
- Conditional posterior means $E(\theta_j|\tau, y)$ 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 sd τ.

Eight schools: priors for τ^2

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:

$$\tau \sim \text{Uniform}(0, 100) \tag{10}$$

$$\tau^2 \sim \text{InvGamma}(0.01, 0.01) \tag{11}$$

$$\tau \sim \text{HalfCauchy}(0, 2.5) \tag{12}$$

Eight schools: priors for τ^2

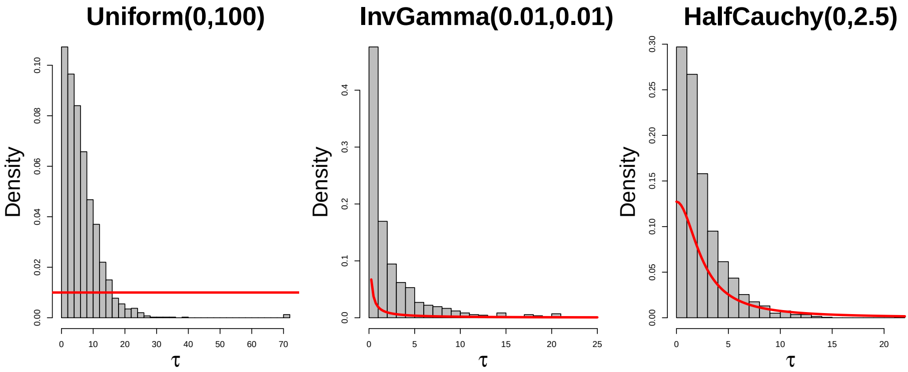


Figure: Marginal posterior (histograms) vs priors (solid red lines)

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

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.

Indice

Hierarchical logistic regression

1988 US polls

We choose a single outcome—the probability that a respondent prefers the Republican candidate 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.

We introduce multilevel logistic regression including two individual 0-1 predictors—`female` and `black`—and the 51 states:

$$
\begin{aligned}
\Pr(y_i = 1) =& \text{logit}^{-1}(\alpha_{j(i)} + \beta^{\text{female}}\text{female}_i + \beta^{\text{black}}\text{black}_i), \quad i = 1, \dots, n \\
\alpha_j \sim& \mathcal{N}(\mu_\alpha, \tau^2_{\text{state}}), \; j = 1, \dots, 51
\end{aligned}
\tag{13}
$$

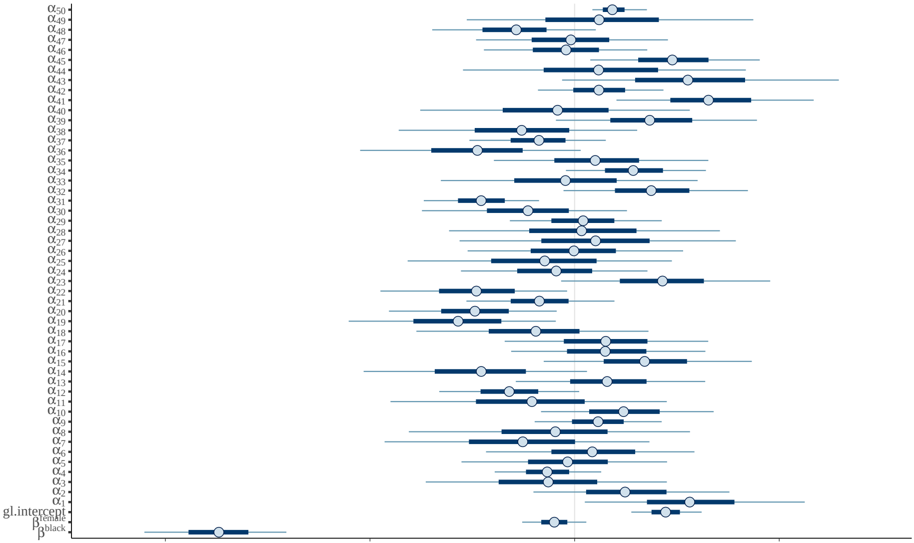
where $j(i)$ is the state index.

1988 US polls. Varying-intercept model

```
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
```

The state variation is estimated at $\hat{\tau}_{\text{state}} = 0.45$.

1988 US polls. Varying-intercept model



1988 US polls. Varying-intercept model

Parameters' interpretation

- 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.
- 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.

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

Hierarchical logistic regression: 1988 US polls

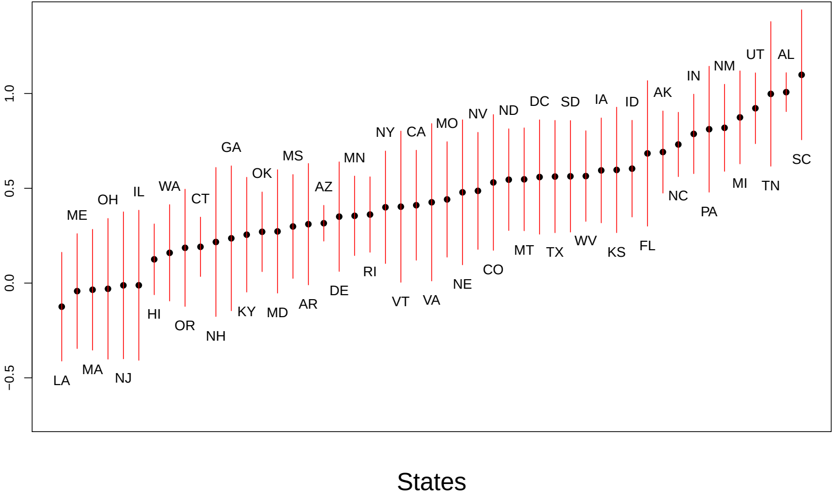
Many issues arise when you fit a model:

- 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, LOIIC...To be continued...
- Use your model to make predictions.

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

Hierarchical logistic regression: 1988 US polls

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



1988 US polls. Varying-intercept and slope

We could ask ourself: is also the slope for the female varying in some states? Maybe, the women Bush 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*:

$$\begin{aligned}\Pr(y_i = 1) =& \text{logit}^{-1}(\alpha_{j(i)} + \beta^{\text{female}}_{j(i)}\text{female}_i + \beta^{\text{black}}\text{black}_i), \quad i = 1, \ldots, n\\ \begin{pmatrix}\alpha_j\\ \beta_j\end{pmatrix} \sim& \mathcal{N}\left(\begin{pmatrix}\mu_\alpha\\ \mu_\beta\end{pmatrix}, \begin{pmatrix}\tau^2_\alpha & \rho\tau_\alpha\tau_\beta\\ \rho\tau_\alpha\tau_\beta & \tau^2_\beta\end{pmatrix}\right), \quad j = 1, \ldots, 51,\end{aligned} \tag{14}$$

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

1988 US polls. Varying-intercept and slope

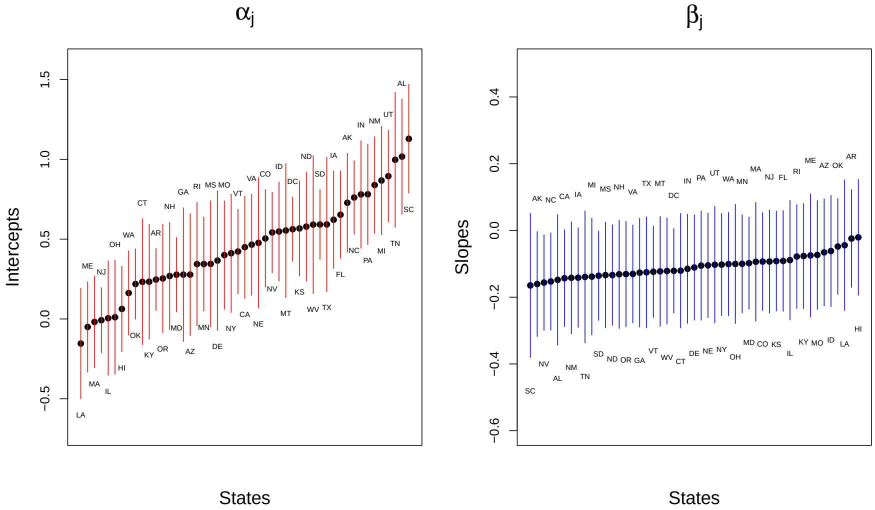
```
stan_glmer
 family:       binomial [logit]
 formula:      y ~ black + female + (1 + female | state)
 observations: 2015
------
            Median MAD_SD
(Intercept)  0.5    0.1
black       -1.7    0.2
female      -0.1    0.1

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

1988 US polls. Varying-intercept and slope

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' effects and the female effects.
- Other parameters are almost unchanged with respect to the varying-intercept model.

1988 US polls. Varying-intercept and slope



Model comparison

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

```
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
```

The varying-intercept and slope model does not improve over the fit of the varying intercept model. The simpler the better!

We could try to extend our model and, eventually, increase the goodness of fit (to be continued).

Indice

Hierarchical Poisson regression

We can extend Poisson models encoding hierarchical structure. Take again the cockroach regression, and consider now to include as many intercepts as many buildings. Thus, for each complaint i we have:

$$\begin{aligned}
\text{complaints}_{ib} \sim& \mathrm{Posson}(\lambda_{ib}) \\
\lambda_{ib} =& \exp{(\eta_{ib})} \\
\eta_{ib} =& \alpha_{b(i)} + \beta\,\mathrm{traps}_b + \beta_{\mathrm{super}}\ \mathrm{super}_i + \mathrm{log_sq_foot}_i \\
\alpha_b \sim& \mathcal{N}(\mu, \tau^2_{\alpha}),
\end{aligned} \tag{15}$$

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

Further reading

Further reading

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