

STA2201H Methods of Applied Statistics II

Monica Alexander

Week 9: More hierarchical models

Notes

- ▶ Midterm
- ▶ A3 out this week
- ▶ Also have to hand in research proposal and EDA

Research project

- ▶ Pick a dataset and research question and write a short paper
- ▶ Some examples from last year on Quercus (posted with permission)
- ▶ No real rules, but Stan/Bayesian (hierarchical) models **strongly** encouraged
 - ▶ GLMs
 - ▶ Hierarchical models
 - ▶ Temporal smoothing
 - ▶ Penalized splines
 - ▶ Measurement error/ data models
- ▶ Research proposal (DUE WITH A3):
 - ▶ What is your question and why is it interesting/important?
 - ▶ (brief) what is known? Hypotheses?
 - ▶ What's your dataset and main outcome/covariates of interest?
 - ▶ EDA: characteristics of dataset, any key patterns in outcome/relationships with other covariates
 - ▶ Should be done in the one Rmd/quarto file (hand in both Rmd/quarto and pdf)

Overview

- ▶ Funnels of hell and reparameterization
- ▶ GLMs in a hierarchical context

Reading

- ▶ Lesaffre and Lawson, 'Bayesian Biostatistics'. Lip cancer example is from here.
- ▶ GH Chapters 14-15
- ▶ BDA Chapters 15-16
- ▶ A nice overview of hierarchical funnels:
<https://arxiv.org/pdf/1312.0906.pdf>
- ▶ Xbox paper:
<https://www.microsoft.com/en-us/research/wp-content/uploads/2016/04/forecasting-with-nonrepresentative-polls.pdf>

Hierarchical models and funnels

Back to radon

$$y_i | \alpha_{j[i]}, \beta_{j[i]} \sim N(\alpha_{j[i]} + \beta_{j[i]} x_i, \sigma_y^2)$$

$$\alpha_j \sim N(\mu_\alpha, \sigma_\alpha^2)$$

$$\beta_j \sim N(\mu_\beta, \sigma_\beta^2)$$

- ▶ y_i is log radon level
- ▶ α_j is county-specific intercept
- ▶ β_j is county-specific slope
- ▶ x_i is floor 1/ basement dummy

Let's fit this in Stan

Just a moment

$$y_i | \alpha_{j[i]}, \beta_{j[i]} \sim N(\alpha_{j[i]} + \beta_{j[i]} x_i, \sigma_y^2)$$

$$\alpha_j \sim N(\mu_\alpha, \sigma_\alpha^2)$$

$$\beta_j \sim N(\mu_\beta, \sigma_\beta^2)$$

- ▶ What's the interpretation of σ_β^2 ?
- ▶ What happens to the β s when σ_β^2 is small?

Fit in Stan

```
transformed parameters {  
  vector[N] y_hat;  
  
  for (i in 1:N)  
    y_hat[i] = a1[county[i]] + a2[county[i]] * x[i];  
}  
model {  
  mu_a1 ~ normal(0, 1);  
  a1 ~ normal(mu_a1, sigma_a1);  
  mu_a2 ~ normal(0, 1);  
  a2 ~ normal(mu_a2, sigma_a2);  
  
  y ~ normal(y_hat, sigma_y);  
}
```

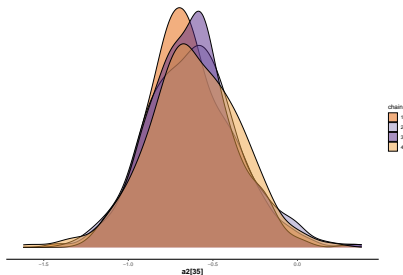
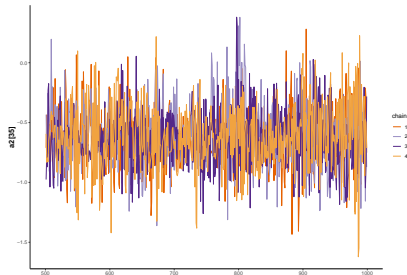
Fit in Stan

Fit to MN only. Here's some of the results:

##		mean	se_mean	n_eff	Rhat
##	a1[1]	1.1723465	0.005430858	2579.81006	1.0012350
##	a1[35]	1.0531867	0.005527846	1897.79097	0.9991013
##	a2[1]	-0.6296192	0.006593284	2069.29189	0.9993998
##	a2[35]	-0.6232028	0.006982481	1428.75164	1.0000627
##	sigma_a1	0.3374825	0.002053772	535.68104	1.0070451
##	sigma_a2	0.3116232	0.019996496	35.37614	1.0930507

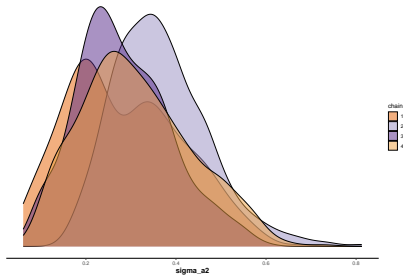
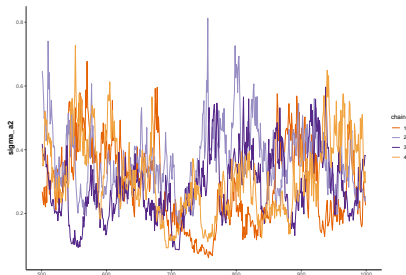
Let's look at some diagnostics

Pick a county (number 35) and plot the trace and density

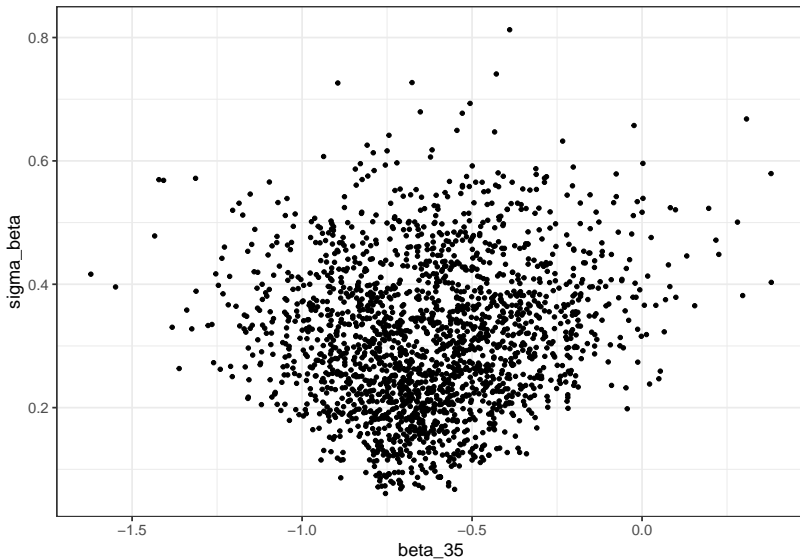


What about the variance parameter

What's happening here?



Scatterplot of β_{35} and σ_β



Funnels of hell

- ▶ The density of these models looks like a 'funnel', with a region of high density but low volume below a region of low density and high volume
- ▶ This property of the posterior is a characteristic of the model and not a problem in itself, but makes sampling hard
- ▶ Especially a problem for Gibbs, but still a problem for HMC
- ▶ The narrower the space, the smaller the step size has to be
- ▶ Larger step sizes more likely to get rejected, so the sampler can get stuck

We fit a centered parameterization

$$y_i | \alpha_{j[i]}, \beta_{j[i]} \sim N(\alpha_{j[i]} + \beta_{j[i]} x_i, \sigma_y^2)$$

$$\alpha_j \sim N(\mu_\alpha, \sigma_\alpha^2)$$

$$\beta_j \sim N(\mu_\beta, \sigma_\beta^2)$$

- e.g. β_j is directly dependent on hyperparameters μ_β and σ_β^2

Non-centered parameterization

$$y_i | \alpha_{j[i]}, \beta_{j[i]} \sim N(\alpha_{j[i]} + \beta_{j[i]} x_i, \sigma_y^2)$$

...

$$\beta_j = \mu_\beta + \eta_j \cdot \sigma_\beta$$

$$\eta_j \sim N(0, 1)$$

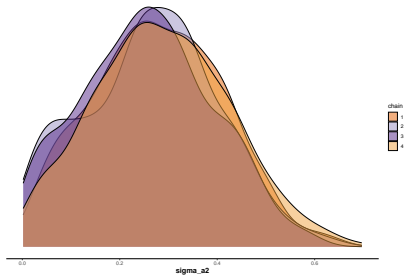
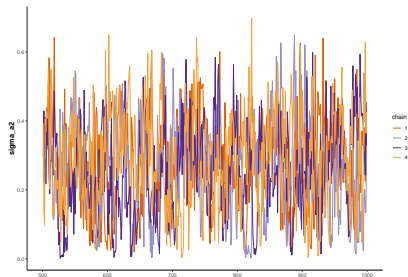
...

- ▶ β_j is now a global mean + some offset, scaled by the β -specific standard deviation.
- ▶ In a non-centered parameterization we do not try to fit the group-level parameters directly, rather we fit a latent Gaussian variable from which we can recover the group-level parameters with a scaling and a translation
- ▶ The variables we are actually sampling are uncorrelated

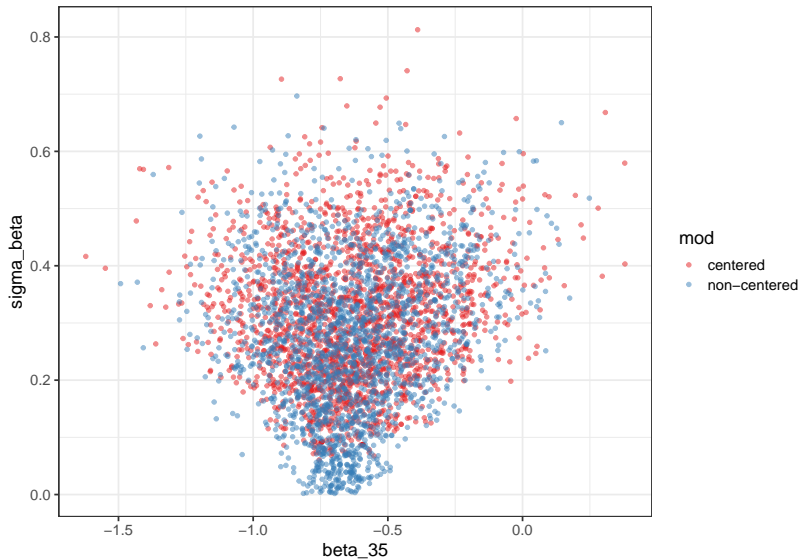
Non-centered Stan model

```
transformed parameters {  
  vector[85] a1;  
  vector[85] a2;  
  vector[N] y_hat;  
  
  a1 = mu_a1 + sigma_a1 * eta1;  
  a2 = mu_a2 + sigma_a2 * eta2;  
  
  for (i in 1:N)  
    y_hat[i] = a1[county[i]] + a2[county[i]] * x[i];  
}  
model {  
  mu_a1 ~ normal(0, 1);  
  mu_a2 ~ normal(0, 1);  
  eta1 ~ normal(0, 1);  
  eta2 ~ normal(0, 1);  
  y ~ normal(y_hat, sigma_y);  
}
```

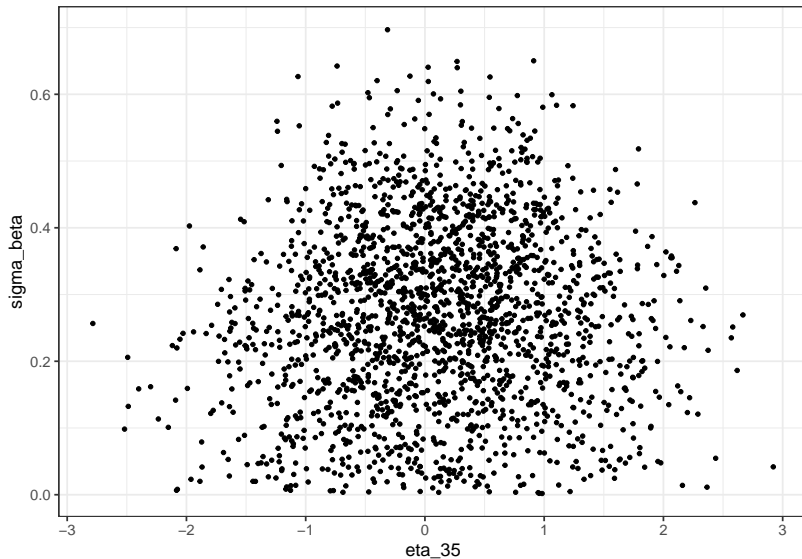
Diagnostics for σ_β



Better exploration of the funnel



Less dependence between sampled parameters



Summary

- ▶ Funnel structure of posterior density is a consequence of hierarchical model structure
- ▶ Not just the shape but the mass of density
- ▶ Mostly a problem when you don't have much data! More shrinkage = more mass in narrow bit of funnel

GLMs in a hierarchical context

GLMs in a hierarchical context

- ▶ Last week we introduced hierarchical models in the setting where the data are assumed to be normally distributed
- ▶ But can easily extended to model other types of non-normal data hierarchically

Let's revisit the Poisson case:

$$y_i | \lambda_i \sim \text{Poisson}(\lambda_i)$$

where $\lambda_i > 0$, may depend on covariates, and vary by group membership $j[i]$.

Lip cancer

- ▶ Let's look at an example of estimating the risk of lip cancer by region in the former German Democratic Republic (Lesaffre and Lawson, chapter 9)
- ▶ In 1989, 2,342 deaths were recorded from lip cancer among males in 195 regions of GDR.
- ▶ For each region i we observed the number of deaths y_i
- ▶ We also know the expected count e_i , based on
 - ▶ age-specific mortality rates for whole country
 - ▶ age distribution of each region
- ▶ We also know the percentage of the population working outside
- ▶ Goal: estimate **relative risk** for each region θ_i

Lip cancer

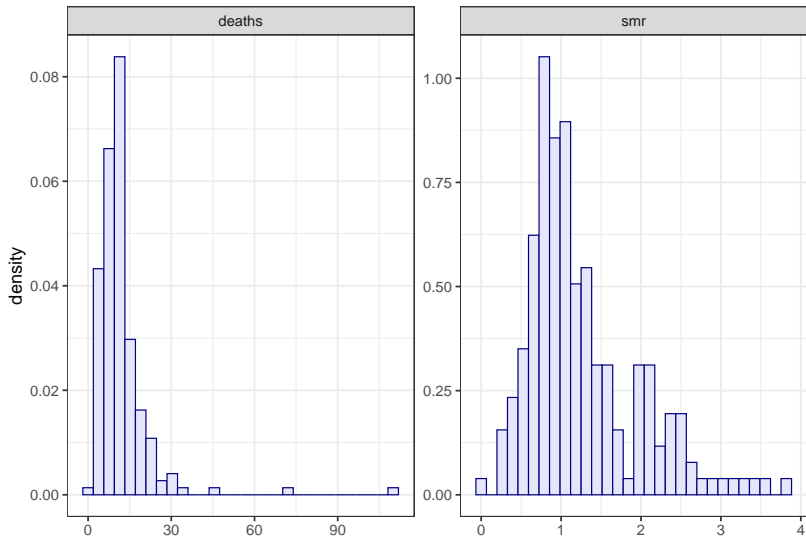
- ▶ Define the standardized mortality rate $SMR_i = y_i / e_i$
- ▶ This is an estimate of the underlying relative risk θ_i , which captures the relative difference in risk of dying for region i as compared to the reference population.
- ▶ Problem of small populations, some of the SMRs are based on very low counts, so are very uncertain
 - ▶ 15% based on counts < 5
 - ▶ 62% based on counts < 10

Set-up

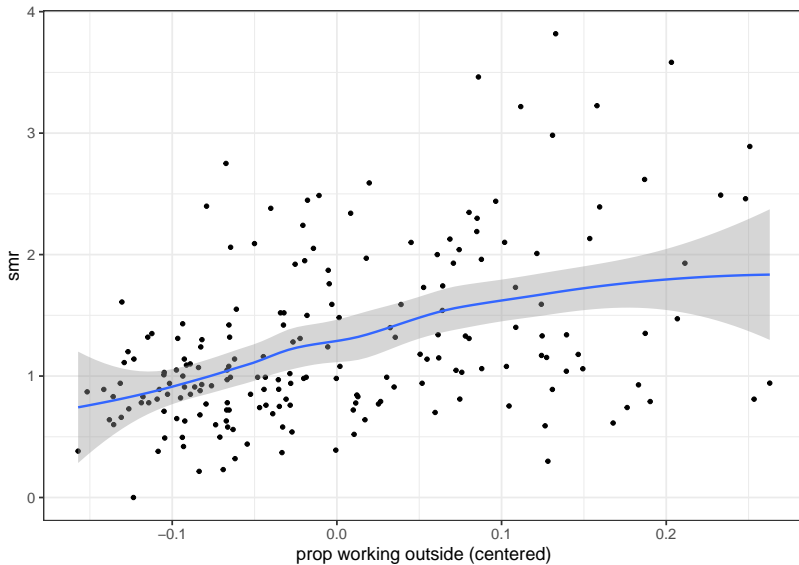
$$\begin{aligned} y_i &\sim \text{Poisson } (\theta_i \cdot e_i) \\ \theta_i &= ?? \end{aligned}$$

- ▶ From last week, in a broad sense, what are our three options for θ_i ?

What the data look like



What the data look like



Lip cancer

1. Model θ_i s separately
2. Find one θ for all regions
3. Use hierarchical model to exchange information about θ_i s across regions

$$\begin{aligned}y_i &\sim \text{Poisson}(\theta_i \cdot e_i) \\ \log \theta_i &= \alpha_i + \beta(x_i^c) \\ \alpha_i &\sim N(\mu, \sigma_\mu^2)\end{aligned}$$

Where x_i^c is the (centered) percent of male population engaged in agriculture/forestry and fisheries in region i

Lip cancer

Full model

$$\begin{aligned}y_i &\sim \text{Poisson}(\theta_i \cdot e_i) \\ \log \theta_i &= \alpha_i + \beta(x_i^c) \\ \alpha_i &\sim N(\mu, \sigma_\mu^2) \\ \mu, \beta &\sim N(0, 1) \\ \sigma_\mu &\sim N_+(0, 1)\end{aligned}$$

Fitting in Stan

```
data {  
  int<lower=1> N;  
  vector[N] x;  
  vector[N] offset;  
  int<lower=0> deaths[N];  
  int<lower=0> region[N];  
}  
parameters {  
  vector[N] alpha;  
  real mu;  
  real beta;  
  real<lower=0> sigma_mu;  
}  
model {  
  vector[N] log_lambda;  
  for (i in 1:N){  
    log_lambda[i] = alpha[i] + beta*x[i] + offset[i];  
  }  
  
  alpha ~ normal(mu, sigma_mu);  
  
  mu ~ normal(0, 1);  
  beta ~ normal(0,1);  
  sigma_mu ~ normal(0, 1);  
  
  deaths ~ poisson_log(log_lambda);  
}
```

Stan interlude

Note in the previous slide I wrote

```
deaths ~ poisson_log(log_lambda)
```

for the likelihood.

There are a million (well, at least two) other options, see here:

https://mc-stan.org/docs/2_18/functions-reference/poisson.html

- ▶ `target += poisson_lpmf(deaths | lambda)`
i.e. “Increment target log probability density with...”
- ▶ `deaths ~ poisson(lambda)` this is shorthand for the above, to make people used to BUGS/JAGS happier.

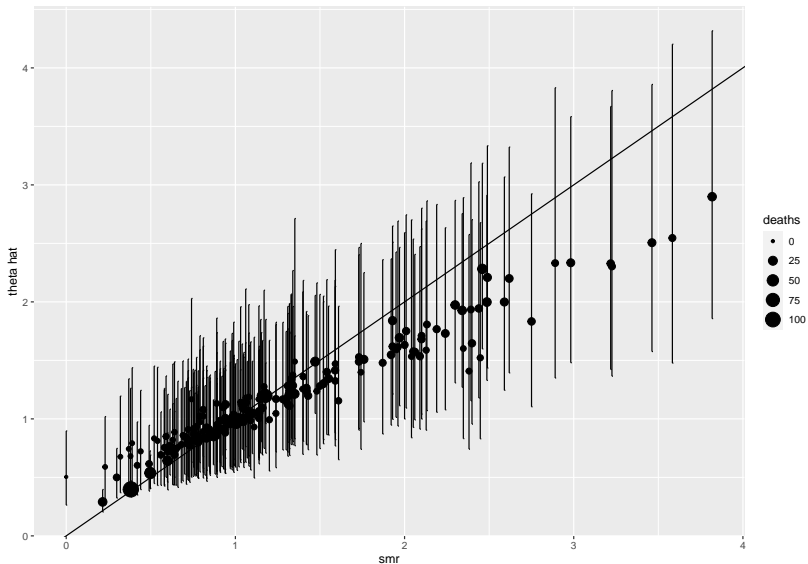
In the above example, I used `poisson_log` so I didn't have to exponentiate my expression, but you do you.

Also note that I used an ‘offset’ in the same way that we did in glm, but you could equally write in terms of the expected deaths e.g. `lambda[i] = theta[i]*expected[i]`, etc etc

Interpretation of coefficient on proportion working outside?

##		mean	se_mean	n_eff	Rhat
## mu		0.08693161	0.0005012467	5014.020	0.9996318
## beta		1.96371501	0.0058524931	3275.073	1.0002089
## sigma_mu		0.38669402	0.0006168194	2494.219	0.9993429

Observed SMR v estimated θ_i



Overdispersion

- ▶ Recall that in many applications, counts are likely to be overdispersed
- ▶ Actually not bad in lip cancer case (how do you tell?)
- ▶ Going back k weeks, what were our two main options for dealing with overdispersion?

Overdispersion with a quasi-Poisson set-up

In the lip cancer case we had

$$y_i \sim \text{Poisson} (\exp(\alpha_i + \beta x_i^c) \cdot e_i)$$

...

For the overdispersed case we would have

$$y_i \sim \text{Poisson} (\exp(\alpha_i + \beta x_i^c + \varepsilon_i) \cdot e_i)$$

...

$$\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$$

...

How to compare?

Logistic regression

Logistic regression in a hierarchical context

Can easily extend these hierarchical models to model binary outcomes and the probability of an event happening by various groups of interest, e.g. in the simplest form we would have

$$y_i | \pi_i \sim \text{Bern}(\pi_i), \text{ OR}$$

$$y_i | \pi_i \sim \text{Bin}(n_i, \pi_i), \text{ if total number of trials is } n_i$$

and then

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right)$$

= some function of covariates, e.g.

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_i$$

and then could put some hierarchical structure on the β 's for instance.

Fertility intentions

- ▶ Using data from the 2015-2017 National Survey of Family Growth
- ▶ Gathers information on family life, marriage and divorce, pregnancy, infertility, use of contraception, and men's and women's health in the US.
- ▶ Some microdata are public:
<https://www.cdc.gov/nchs/nsfg/index.htm>

Fertility intentions

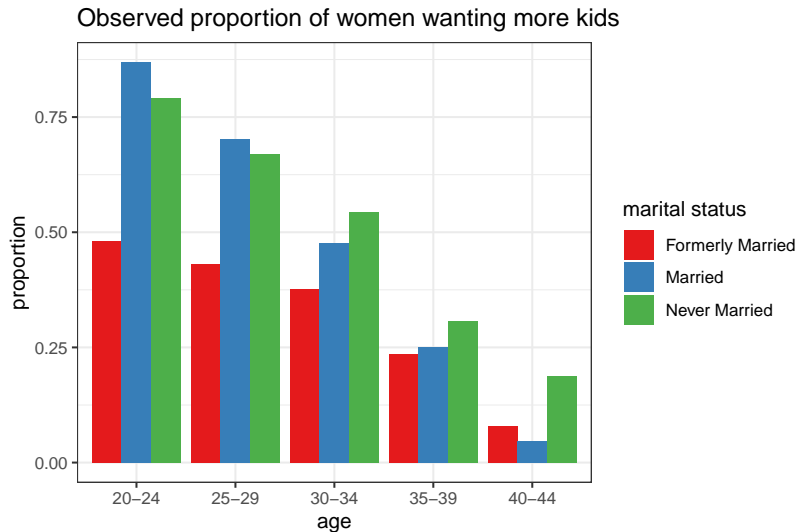
Interested in looking at fertility intentions, i.e. do women intend to have more children (and if so, how many)

- ▶ predictor of childbearing
- ▶ gaps tell you something about unmet need, due to social, economic conditions, cultural changes

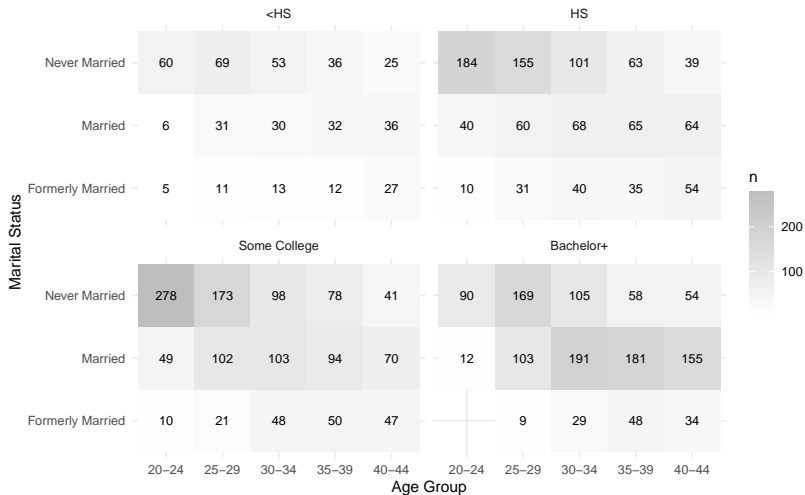
We are interested in the probability that a women wants more children.

- ▶ Individual covariates: age, education, marital status
- ▶ Possible extensions: state/region, parity (current number of children)

Some clear patterns



Sample size differences



Non-nested hierarchical model

- ▶ So far we have considered the simplest hierarchical structure of individuals i in groups j
- ▶ In the polls example, individuals have different group memberships (based on age and education) that we may want to pool across.

E.g. a reasonable model set-up would be

- ▶ to have hierarchies for age and education
- ▶ an extended model that includes geography may also have county nested within state, for example

Just one hierarchical model

A hierarchical model with just age would be

$$y_i | \pi_i \sim \text{Bern}(\pi_i)$$

$$\Pr(y_i = 1) = \pi_i = \text{logit}^{-1}(\alpha_{j[i]})$$

$$\alpha_j \sim \text{N}(\mu_\alpha, \sigma_\alpha^2), \text{ for } j = 1, \dots, 5$$

Add in other effects

$$y_i | \pi_i \sim \text{Bern}(\pi_i)$$

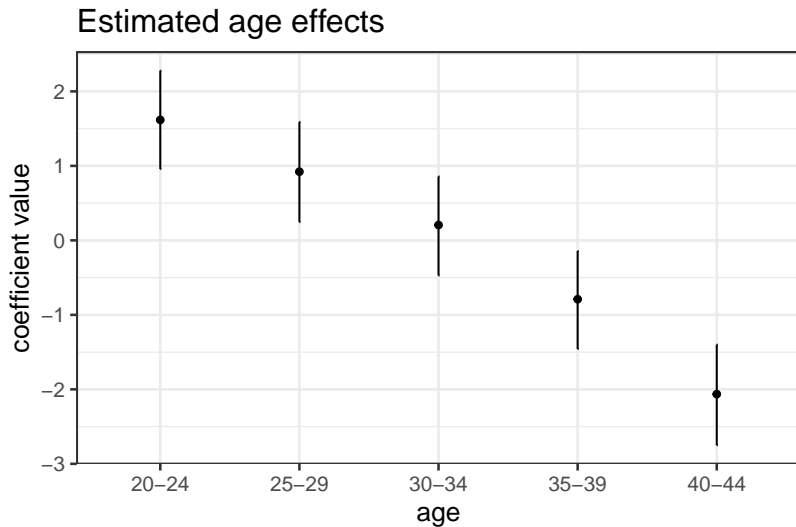
$$\pi_i = \text{logit}^{-1} \left(\beta_0 + \beta_1 \text{formerly married}_i + \beta_2 \text{married}_i + \alpha_{j[i]}^{\text{age}} + \alpha_{k[i]}^{\text{edu}} \right)$$

$$\alpha_j^{\text{age}} \sim \text{N}(0, \sigma_{\text{age}}^2), \text{ for } j = 1, \dots, 5$$

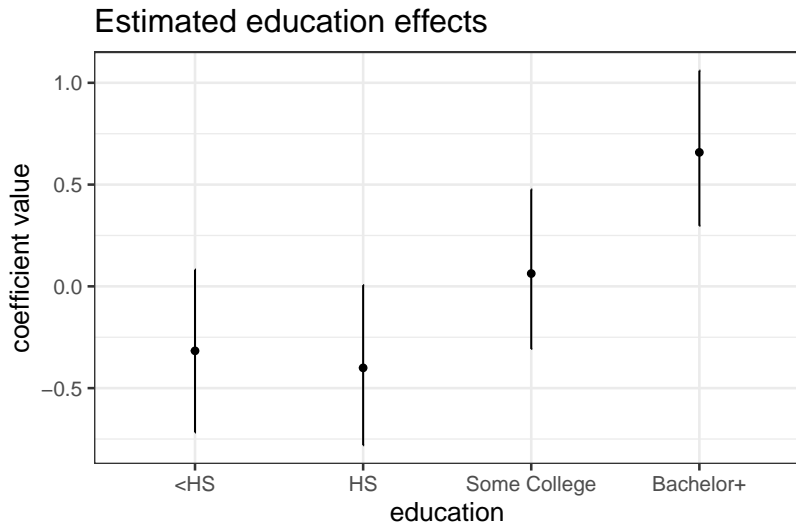
$$\alpha_k^{\text{edu}} \sim \text{N}(0, \sigma_{\text{edu}}^2), \text{ for } k = 1, \dots, 4$$

- ▶ Indexes: individual i , age j , education k
- ▶ Notice that the effects for age, educ, are centered around zero and now there is a 'global' intercept β_0
 - ▶ any non-zero means for the α s could be folded into the global intercept
- ▶ The baseline category is never married

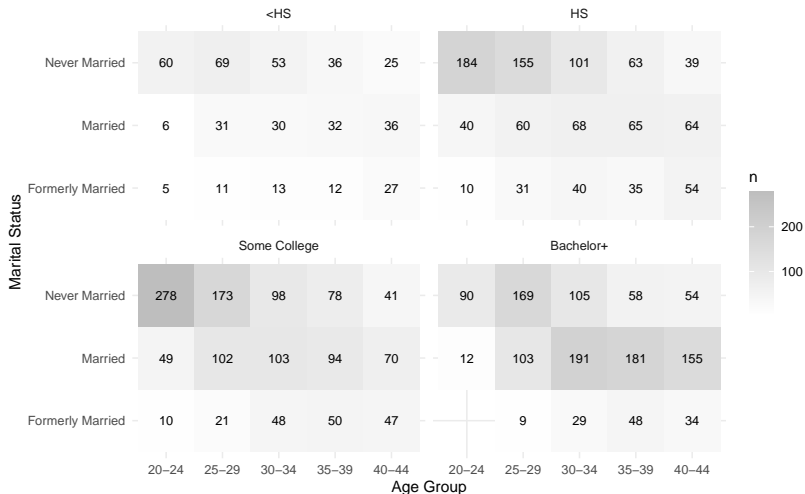
Some results



Some results

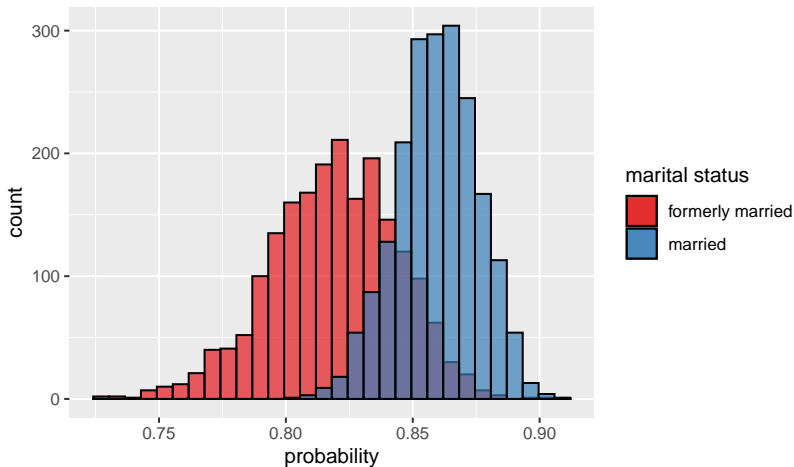


Remember we had no observations from one cell



Estimates from model

Estimated probabilities of wanting more children
20–24 year olds with Bachelor+



What about getting estimates for the 45-49 year olds?

How do we generate these?

$$y_i | \pi_i \sim \text{Bern}(\pi_i)$$

$$\pi_i = \text{logit}^{-1} \left(\beta_0 + \beta_1 \text{formerly married}_i + \beta_2 \text{married}_i + \alpha_{j[i]}^{\text{age}} + \alpha_{k[i]}^{\text{edu}} \right)$$

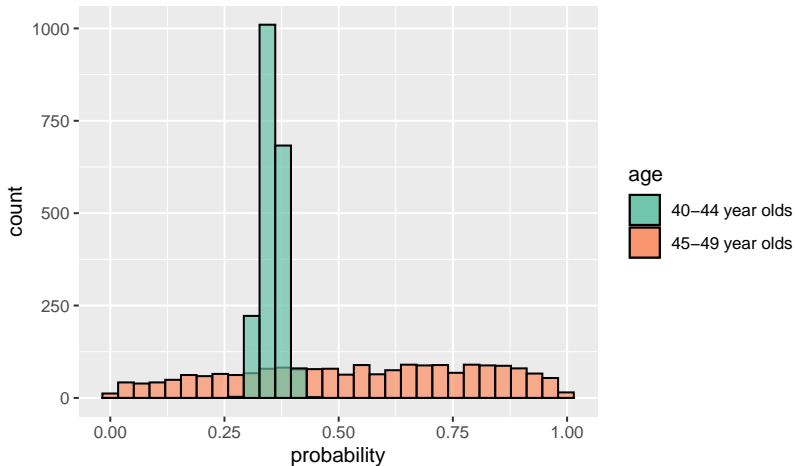
$$\alpha_j^{\text{age}} \sim \text{N} \left(0, \sigma_{\text{age}}^2 \right), \text{ for } j = 1, \dots, 5$$

$$\alpha_k^{\text{edu}} \sim \text{N} \left(0, \sigma_{\text{edu}}^2 \right), \text{ for } k = 1, \dots, 4$$

Results, compared to 40-44 year olds

Estimated probabilities of wanting more children

Married with Bachelor+



Issues

We draw a new $\tilde{\alpha}_j^{\text{age}}$ based on

$$\tilde{\alpha}_j^{\text{age}} \sim \text{N}\left(0, \sigma_{\text{age}}^2\right)$$

This assumes the age effects are conditionally independent of each other and centered at zero.

But we can do better!

An alternative specification on age

Consider the model as before but with one change:

$$y_i | \pi_i \sim \text{Bern}(\pi_i)$$

$$\pi_i = \text{logit}^{-1} \left(\beta_0 + \beta_1 \text{formerly married}_i + \beta_2 \text{married}_i + \alpha_{j[i]}^{\text{age}} + \alpha_{k[i]}^{\text{edu}} \right)$$

$$\alpha_j^{\text{age}} \sim \text{N} \left(\alpha_{j-1}^{\text{age}}, \sigma_{\text{age}}^2 \right), \text{ for } j = 2, \dots, 5$$

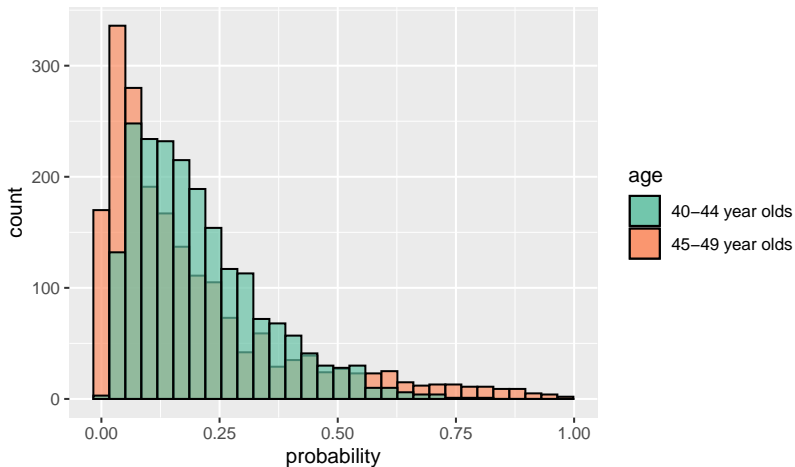
$$\alpha_k^{\text{edu}} \sim \text{N} \left(0, \sigma_{\text{edu}}^2 \right), \text{ for } k = 1, \dots, 4$$

- ▶ This is assuming that the age effect in group j is similar to that in the previous age group (plus some noise)
- ▶ We are placing a random walk on the age effects; assuming some structure over age
- ▶ Note: need to put a prior on the first age group here

Now look at the estimated probabilities for 45-49 year olds

Estimated probabilities of wanting more children

Married with Bachelor+, alternative model



Summary

- ▶ Can model non-normal data in hierarchical setting
- ▶ Can have different hierarchies going on at the same time
- ▶ In the fertility intentions case, hierarchical models makes sense here because cell counts get very small: want to stabilize estimates.
- ▶ We can think about structured priors on the hierarchical effects (e.g. over age)
 - ▶ this idea extends to other dimensions, e.g. particularly thinking about geographic space

Multilevel regression and poststratification

Background context

- ▶ Often want to calculate some population-level estimate of outcome of interest
- ▶ We have estimates of the outcome of interest for particular groups
- ▶ E.g. in fertility intentions example, want a national estimate of the proportion of women who would like more children, $\hat{p}_{\text{national}}$,
- ▶ Assume we can calculate estimates for each age, education, and marital status group, \hat{p}_g for $g = 1, 2, \dots, G$ where $G = 3 \times 4 \times 5$ by just taking the proportions from the survey

How to generate population-level estimates?

Possibilities:

1. Simple average

$$\hat{p}_{\text{national}} = \frac{1}{G} \sum_{g=1}^G \hat{p}_g$$

2. Use survey weights (if available)

$$\hat{p}_{\text{national}} = \sum_{g=1}^G w_g \cdot \hat{p}_g$$

3. Post-stratification

- If we have another dataset that had counts of women in each group g , call them N_g , then we could calculate post-stratified estimates

$$\hat{p}_{\text{national}} = \frac{1}{\sum_g N_g} \sum_{g=1}^G N_g \cdot \hat{p}_g$$

Another option: MRP

4. Multilevel regression and post-stratification (MRP)

- ▶ Instead of using raw survey estimates of the proportion in each group (\hat{p}_g), we could use our modelled estimates from the multilevel (hierarchical) regression, call them \hat{p}_g^{MR} . The national MRP estimate is then

$$\hat{p}_{\text{national}}^{MRP} = \frac{1}{\sum_g N_g} \sum_{g=1}^G N_g \cdot \hat{p}_g^{MR}$$

- ▶ What are the advantages of the MRP estimate?

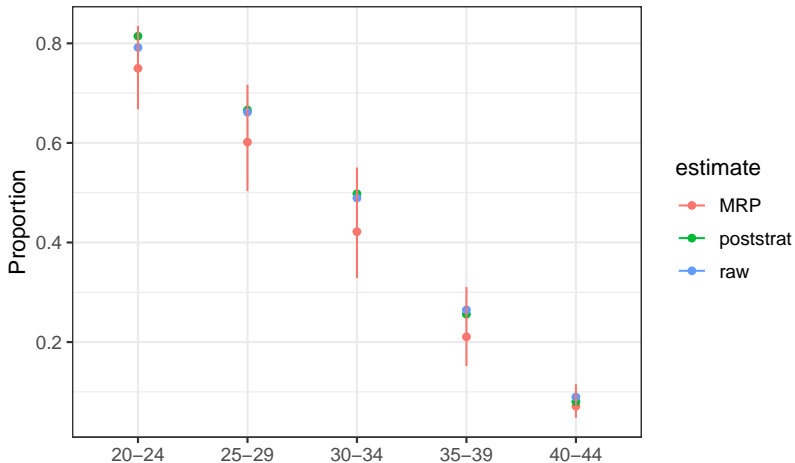
MRP

Note that we can also calculate MRP estimates across groups, for example, by age group a :

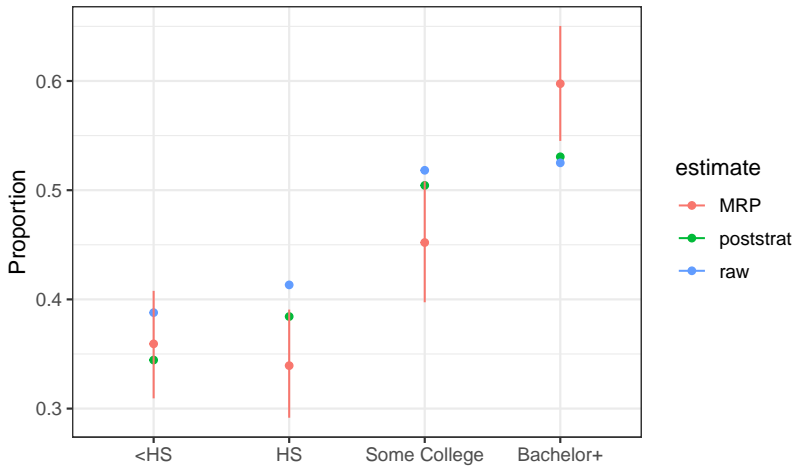
$$\hat{p}_a^{MRP} = \frac{1}{N_a} \sum_{k=1}^K N_{k,a} \cdot \hat{p}_{k,a}^{MR}$$

- ▶ N_a is number of people in age group a based on post-stratification dataset
- ▶ The k index refers to the number of education/marital status groups $k = 1, 2, \dots, K$, with $K = 3 \times 4$

Proportion of women wanting more kids by age



Proportion of women wanting more kids by education



MRP

- ▶ MRP very common in political polling (YouGov)
- ▶ Increasingly common as obtaining representative samples of populations becomes more and more difficult
- ▶ But think about model assumptions (both in terms of hierarchical structure (exchangeability) and post-strat assumptions)
- ▶ Big limitation: need post-strat info