STA2201H Methods of Applied Statistics II

Monica Alexander

Week 9: Hierarchical models II

Overview

- GLMs in a hierarchical context
- Why do we center predictors
- Funnels of hell and reparameterization

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

Poisson regression

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

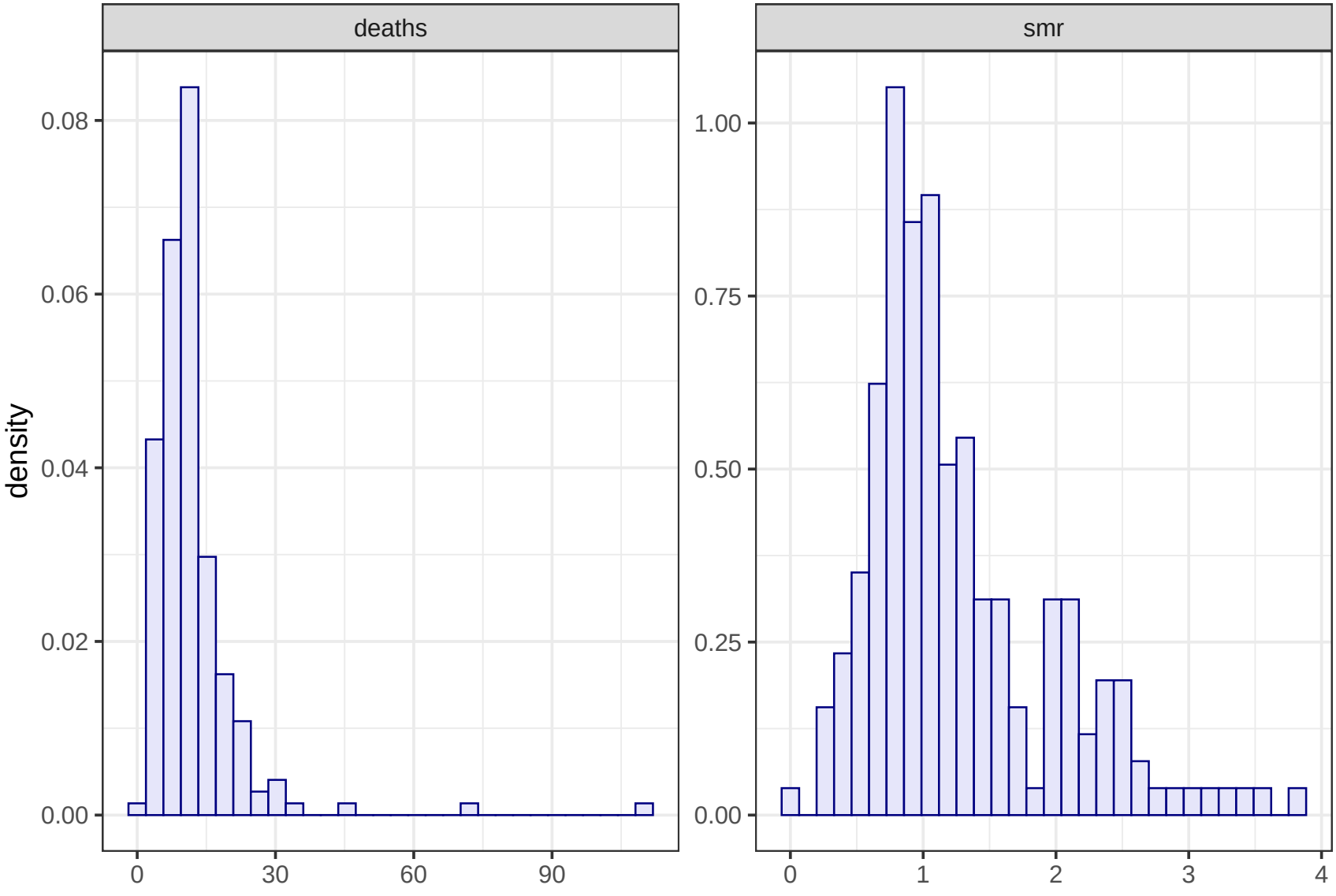
- Define the standardized mortality rate $SMR_i = y_i/e_i$
- This is an estimate of the underlying relative risk θ_i, which captures the relative difference in risk of dying for region i as compared to the reference population.
- Problem of small populations, some of the SMRs are based on very low counts,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

Set-up

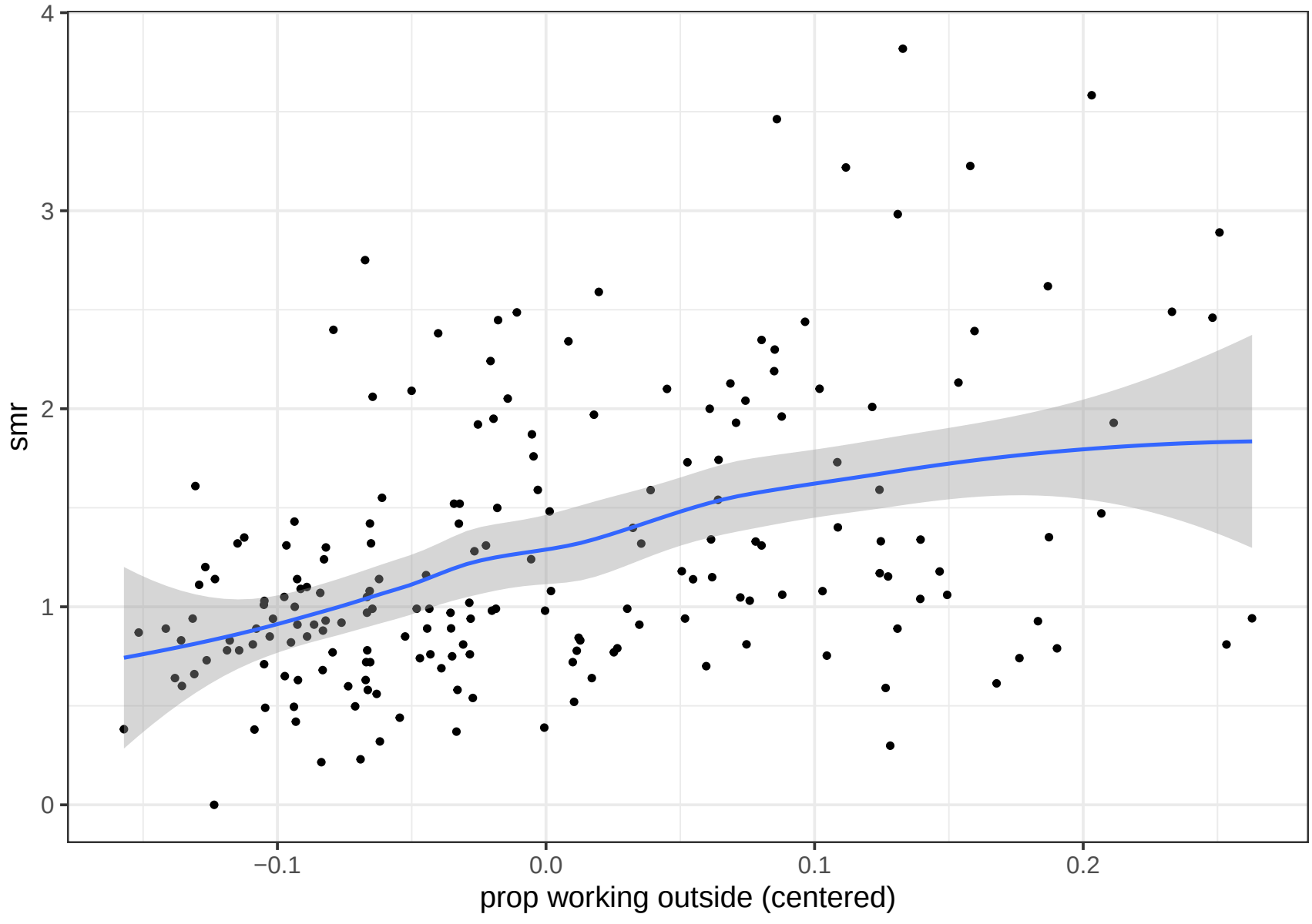
$$\begin{aligned} y_i &\sim \text{Poisson } (\theta_i \cdot e_i) \\ \theta_i &= ?? \end{aligned}$$

- 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_is separately
2. Find one θ for all regions
3. Use hierarchical model to exchange information about θ_is 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

- Relatively straightforward extension of normal models
- Be careful of types

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.

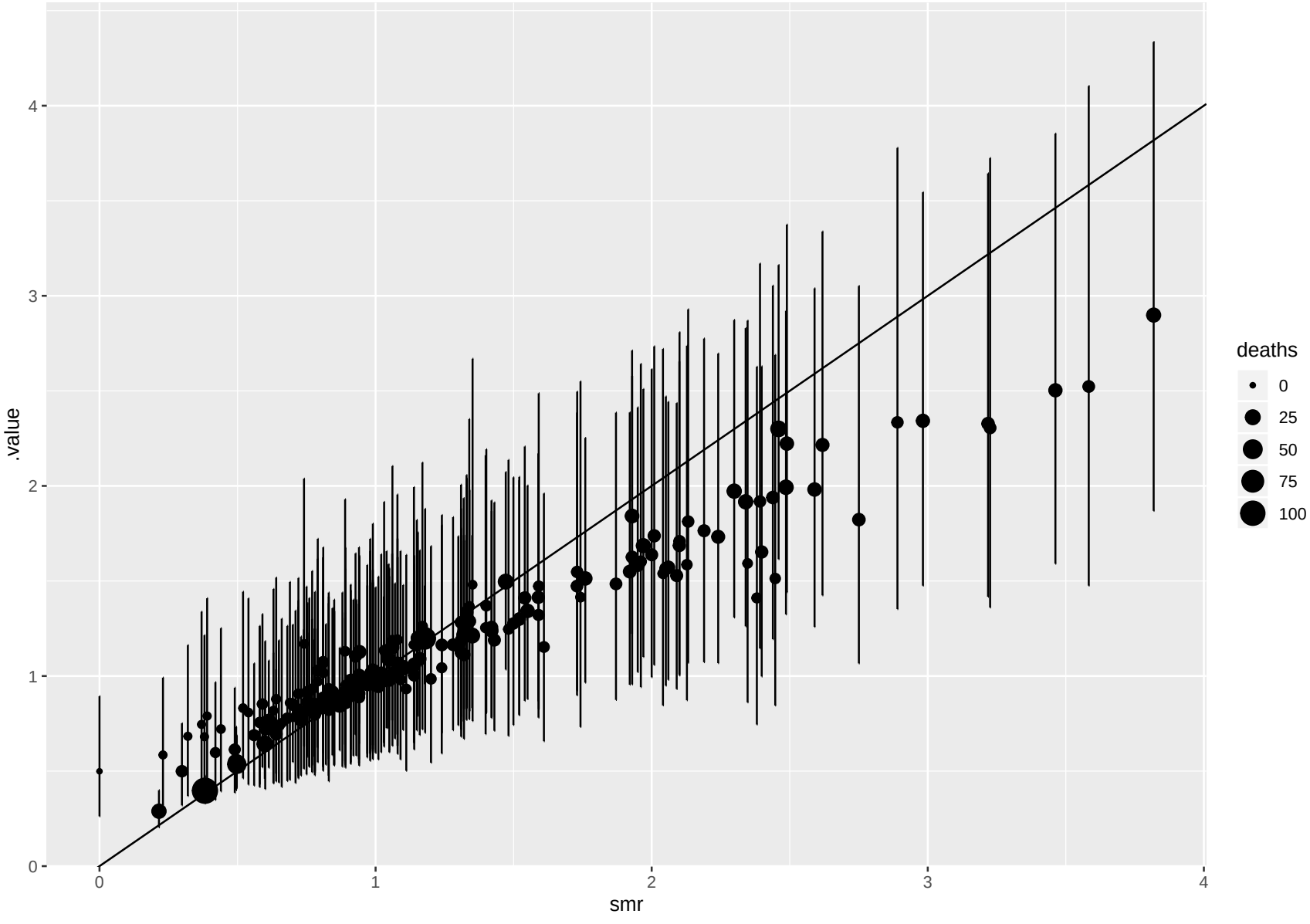
In the above example, I used `poisson_log` so I didn't have to exponentiate my expression, but you do you.

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         0.08589533 0.0005404030 4688.323 0.9997637
## beta       1.98307145 0.0061931908 2943.957 1.0005834
## sigma_mu   0.38702354 0.0006514144 2225.356 1.0002862
```

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

$$\begin{aligned} y_i &\sim \text{Poisson } (exp(\alpha_i + \beta x_i^c) \cdot e_i) \\ &... \end{aligned}$$

For the overdispersed case we would have

$$\begin{aligned} y_i &\sim \text{Poisson } (exp(\alpha_i + \beta x_i^c + \varepsilon_i) \cdot e_i) \\ &... \\ \varepsilon_i &\sim N(0, \sigma_\varepsilon^2) \\ &... \end{aligned}$$

How to compare?

Can we do hierarchical survival models?

Yes! Particularly wrt to parametric or PCH models. e.g. recall example from week four, we were looking at time to second birth:

$$\lambda_i(t) = \lambda_k \cdot \exp(\beta X) \text{ for } t \text{ in } [\tau_{k-1}, \tau_k)$$

λ_k is the baseline hazard in interval k, and the effect of the covariates is proportional (but could be relaxed by letting β vary by interval).

What is an example covariate that would (probably) benefit from a hierarchical structure?

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

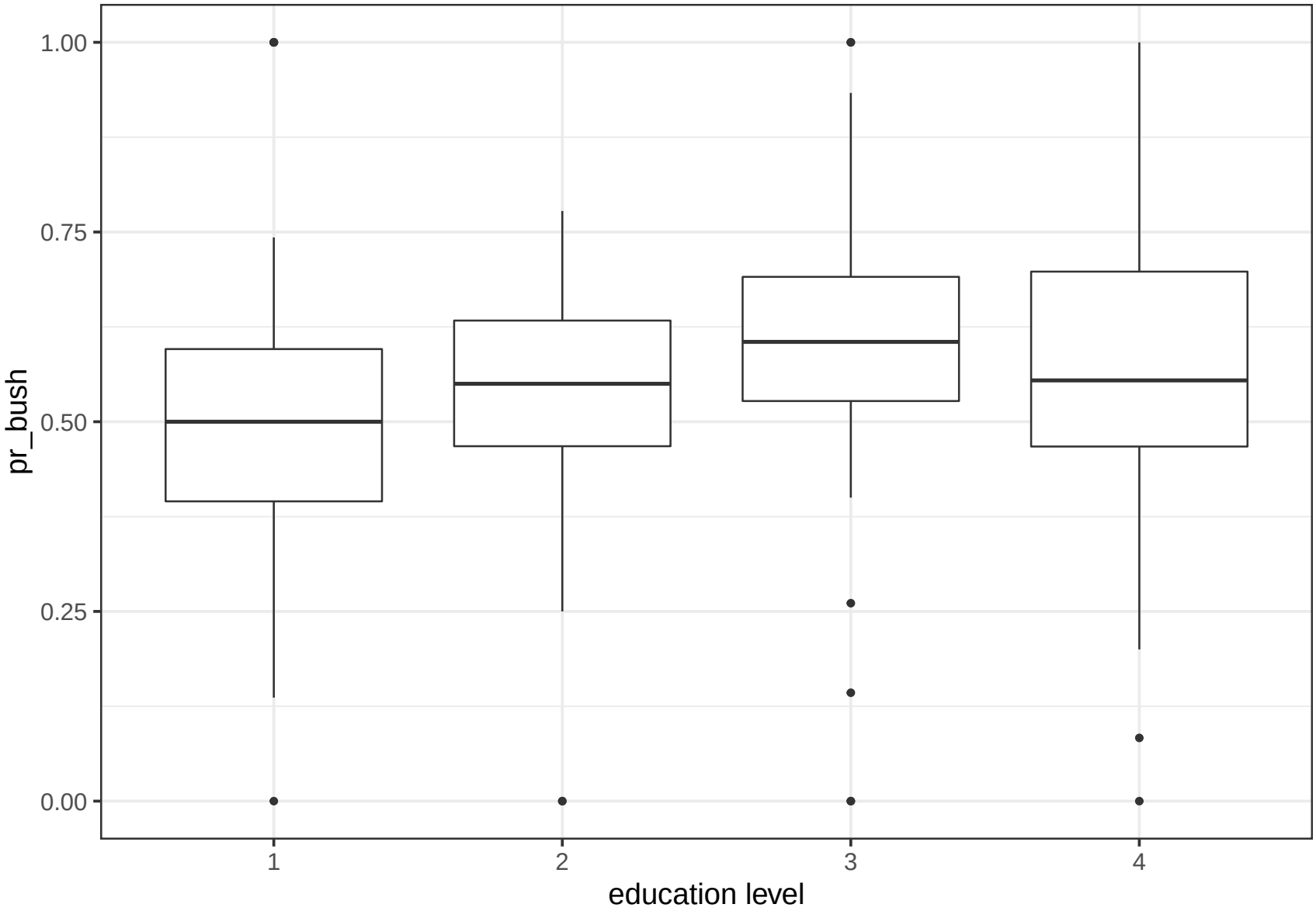
$$\begin{aligned}\text{logit}(\pi_i) &= \log\left(\frac{\pi_i}{1-\pi_i}\right) \\ &= \text{ some function of covariates, e.g.} \\ \text{logit}(\pi_i) &= \beta_0 + \beta_1 x_i\end{aligned}$$

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

Political polling - the classic example

- Dataset from GH (chapter 14): relates to the US case in 2004. (Bush v Kerry)
- We are interested in Pr(prefers Bush)
- Individual covariates: state, education, age, sex, race/ethnicity
- Region/state level covariates: results from previous elections

E.g. variation in Bush support across states by education



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, educ, state) that we may want to pool across.

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

- to have hierarchies for age, education and state
- to have state pooling would have another level that pools information within regions (within the country)

Let's build this up.

Step 1: effects for sex, (simplified) race, pooled effect for state

What we've seen before

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

$$\Pr\left(y_i = 1\right) = \pi_i = \text{logit}^{-1}\left(\alpha_{\text{j[i]}} + \beta^{\text{female}} \cdot \text{female}_i + \beta^{\text{NHB}} \cdot \text{NHB}_i\right)$$

$$\alpha_j \sim \text{N}\left(\mu_\alpha, \sigma^2_{\text{state}}\right), \text{ for } j = 1, \ldots, 51$$

Step 2: add in group-level predictor

Based on past election results. Still nothing new

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

$$\Pr\left(y_i = 1\right) = \pi_i = \text{logit}^{-1}\left(\alpha_{\text{j[i]}} + \beta^{\text{female}} \cdot \text{female}_i + \beta^{\text{NHB}} \cdot \text{NHB}_i\right)$$

$$\alpha_j \sim \text{N}\left(\beta^{\text{prev}} \cdot \text{prev}_j, \sigma^2_{\text{state}}\right), \text{ for } j = 1, \ldots, 51$$

Step 3: add in pooled effects by age and education

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

$$\pi_i = \text{logit}^{-1}\left(\beta_0 + \beta^{\text{female}}\ \text{female}_{\ i} + \beta^{\text{NHB}}\text{NHB}_i + \alpha^{\text{age}}_{k[i]} + \alpha^{\text{edu}}_{l[i]} + \alpha^{\text{state}}_{j[i]}\right)$$

$$\alpha^{\text{age}}_k \sim \text{N}\left(0, \sigma^2_{\text{age}}\right), \text{ for } k = 1, \ldots, 4$$

$$\alpha^{\text{edu}}_l \sim \text{N}\left(0, \sigma^2_{\text{edu}}\right), \text{ for } l = 1, \ldots, 4$$

$$\alpha^{\text{state}}_j \sim \text{N}\left(\beta^{\text{prev}} \cdot \text{prev}_j, \sigma^2_{\text{state}}\right), \text{ for } j = 1, \ldots, 51$$

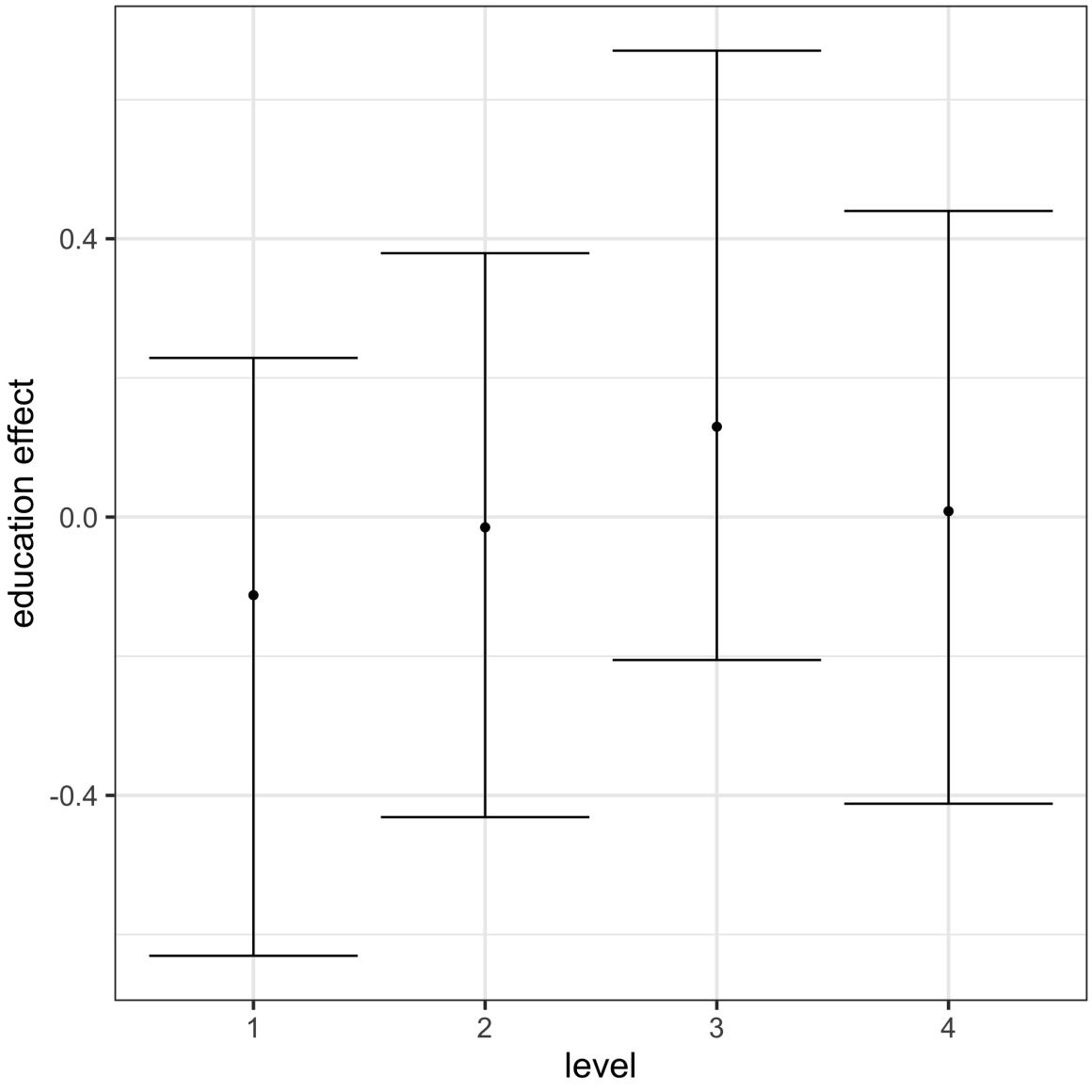
- Indexes: individual i, state j, age k, education l
- Notice that the effects for age, educ, state are centered around zero and now there is a 'global' intercept β_0
  - any non-zero means for the αs could be folded into the global intercept

Step 4: add another geographic hierarchy

$$
\begin{aligned}
y_i|\pi_i &\sim \text{Bern}\left(\pi_i\right) \\
\pi_i &= \text{logit}^{-1}\left(\beta_0 + \beta^{\text{female}}\ \text{female}_{\ i} + \beta^{\text{NHB}}\text{NHB}_i + \alpha^{\text{age}}_{k[i]} + \alpha^{\text{edu}}_{l[i]} + \alpha^{\text{state}}_{j[i]}\right) \\
\alpha^{\text{state}}_j &\sim \text{N}\left(\alpha^{\text{region}}_{m[j]} + \beta^{\text{prev}}\ \cdot \text{prev}_j, \sigma^2_{\text{state}}\right),\ \text{for } j = 1, \ldots, 51 \\
\alpha^{\text{age}}_k &\sim \text{N}\left(0, \sigma^2_{\text{age}}\right),\ \text{for } k = 1, \ldots, 4 \\
\alpha^{\text{edu}}_l &\sim \text{N}\left(0, \sigma^2_{\text{edu}}\right),\ \text{for } l = 1, \ldots, 4 \\
\alpha^{\text{region}}_m &\sim \text{N}\left(0, \sigma^2_{\text{region}}\right),\ \text{for } m = 1, \ldots, 5
\end{aligned}
$$

Yet another index m.

Education effect



Summary

- Can model non-normal data in hierarchical setting
- Can have different hierarchies going on at the same time
- In the polling case, hierarchical models makes sense here because cell counts get very small: want to stabilize estimates.
- An open question on what the hierarchical structure should be
  - e.g. in the polling case, we are assuming that age effects are iid draw from a common distribution
  - but likely to be correlation over age!

Why should we center/centre predictors

A simple example

$$
\begin{aligned}
y_i|\mu_i &\sim N\left(\mu_i, \sigma_y^2\right) \\
\mu_i &= \alpha_{j[i]} + \beta_{j[i]} x_i \\
\begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} &\sim N_2\left(\begin{pmatrix} \mu_\alpha \\ \mu_\beta \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta^2 \end{pmatrix}\right)
\end{aligned}
$$

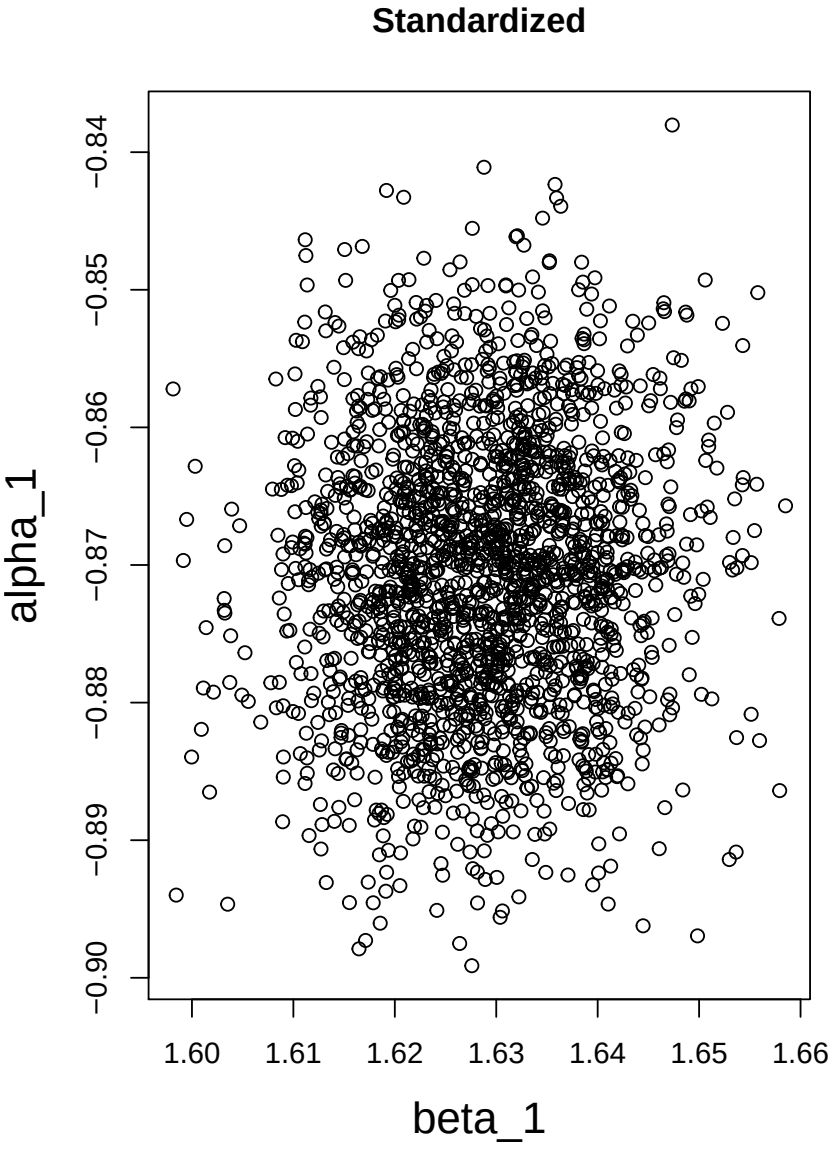
- Simulate some data
- Let's run model with x_i centered and non-centered and see the difference

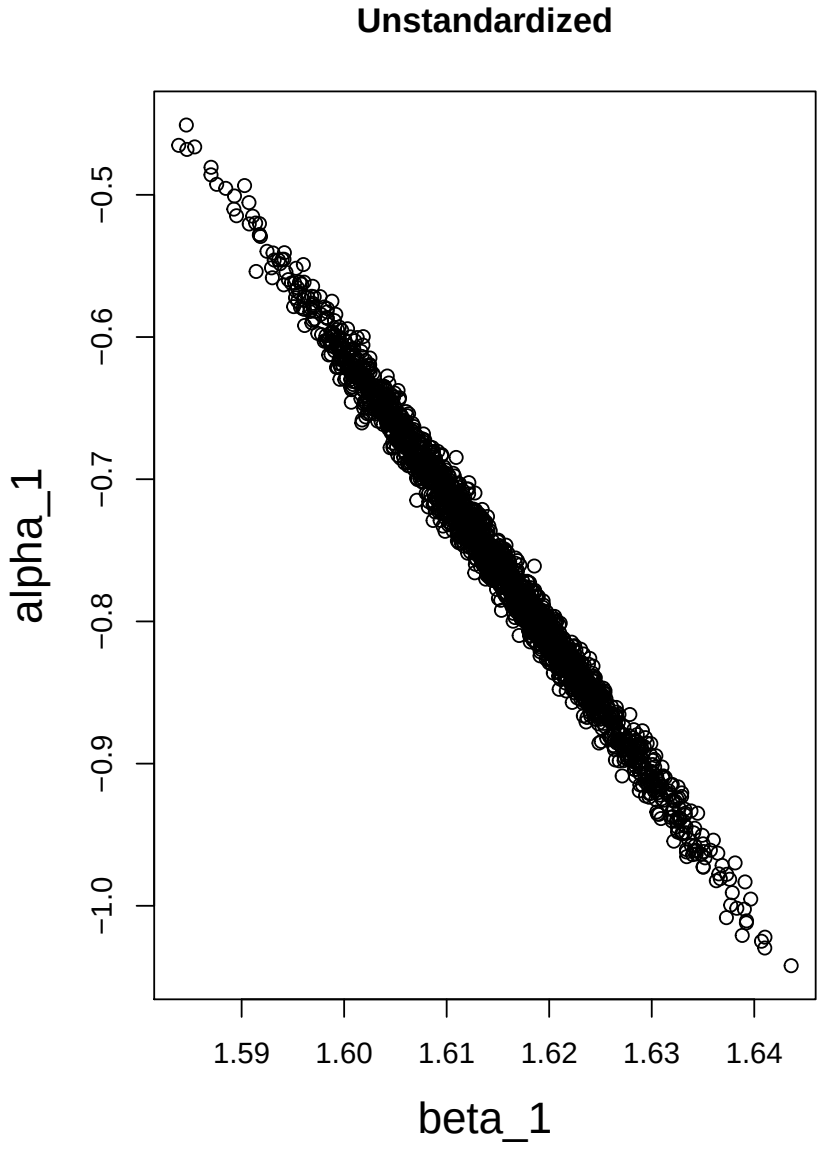
Simulation set-up

- $J = 100$
- $n = 100$
- $\mu_\alpha = 0$
- $\mu_\beta = 1$
- $\sigma_\alpha = \sigma_\beta = 1$
- $\sigma_y = 0.1$
- $\rho = -0.5$

Simulate y's and x's based on these settings, set unstandardized x's to have a mean of 10.

Results


Standardized
alpha_1
−0.84
−0.85
−0.86
−0.87
−0.88
−0.89
−0.90
1.60
1.61
1.62
1.63
1.64
1.65
1.66
beta_1


Unstandardized
alpha_1
−0.5
−0.6
−0.7
−0.8
−0.9
−1.0
1.59
1.60
1.61
1.62
1.63
1.64
beta_1

Centering predictors

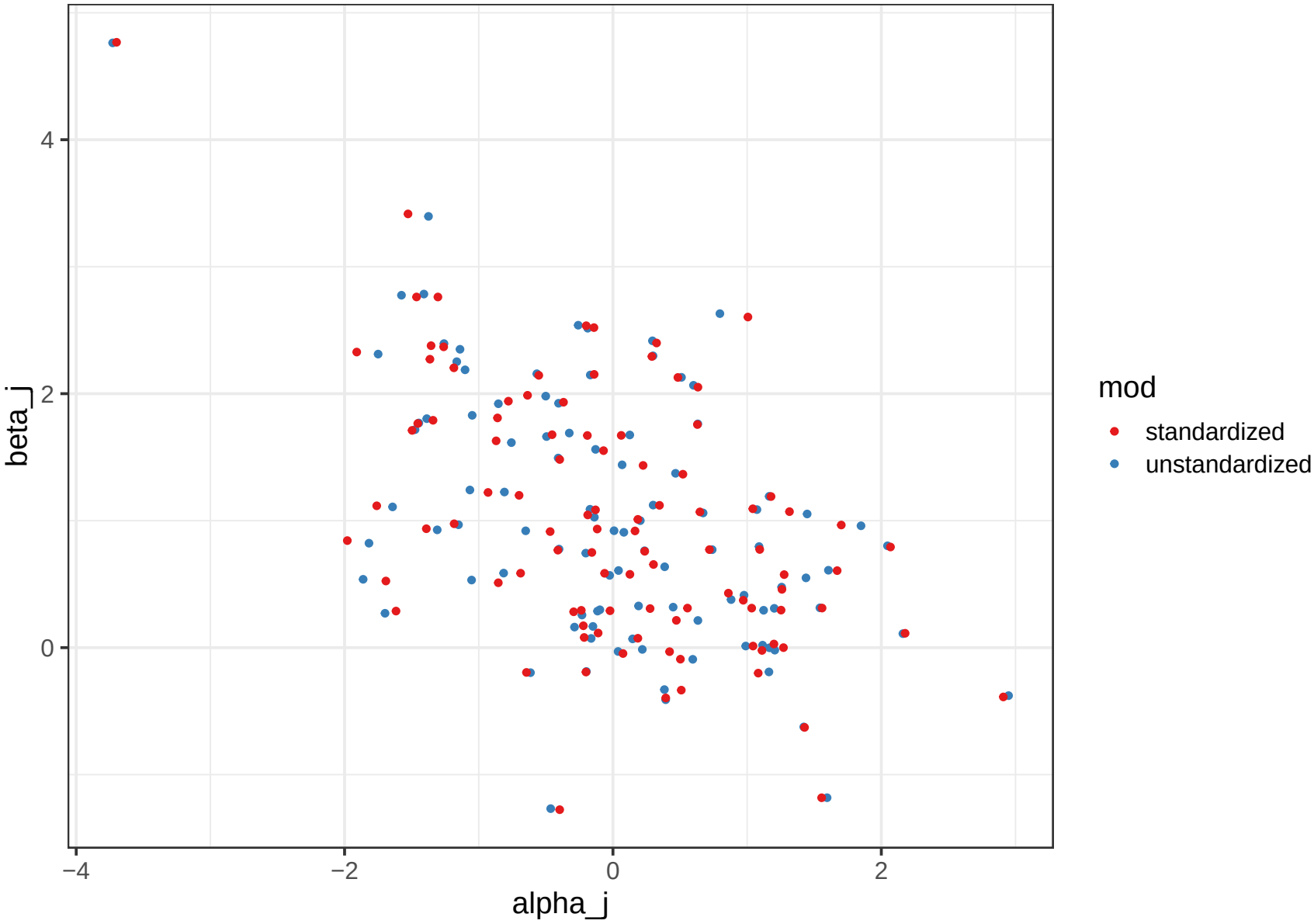
- When the mean of the predictors is far away from zero, changes in the slope induce the opposite change in the intercept
- Hard to interpret what intercepts mean
- Harder to sample: reducing correlation may speed up convergence

Is this ρ?

$$
\begin{aligned}
y_i|\mu_i &\sim N\left(\mu_i, \sigma_y^2\right) \\
\mu_i &= \alpha_{j[i]} + \beta_{j[i]} x_i \\
\begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} &\sim N_2\left(\begin{pmatrix} \mu_\alpha \\ \mu_\beta \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta^2 \end{pmatrix}\right)
\end{aligned}
$$

Across-group correlation

Correlation in both models is close to -0.5



Summary

- The across-samples correlation between the posterior samples $\alpha_j^{(s)}$ and $\beta_j^{(s)}$ depends on $\bar{x}$
- The across-groups correlation in parameters α_j and β_j is ρ
- centering the (continuous) predictor variable reduces correlation between the posterior samples of group-specific intercepts and slopes for each group
- centering predictors also improves interpretation

Hierarchical models and funnels

Back to radon

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

- 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

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

- What's the interpretation of σ_β^2?
- What happens to the βs 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);
}
```

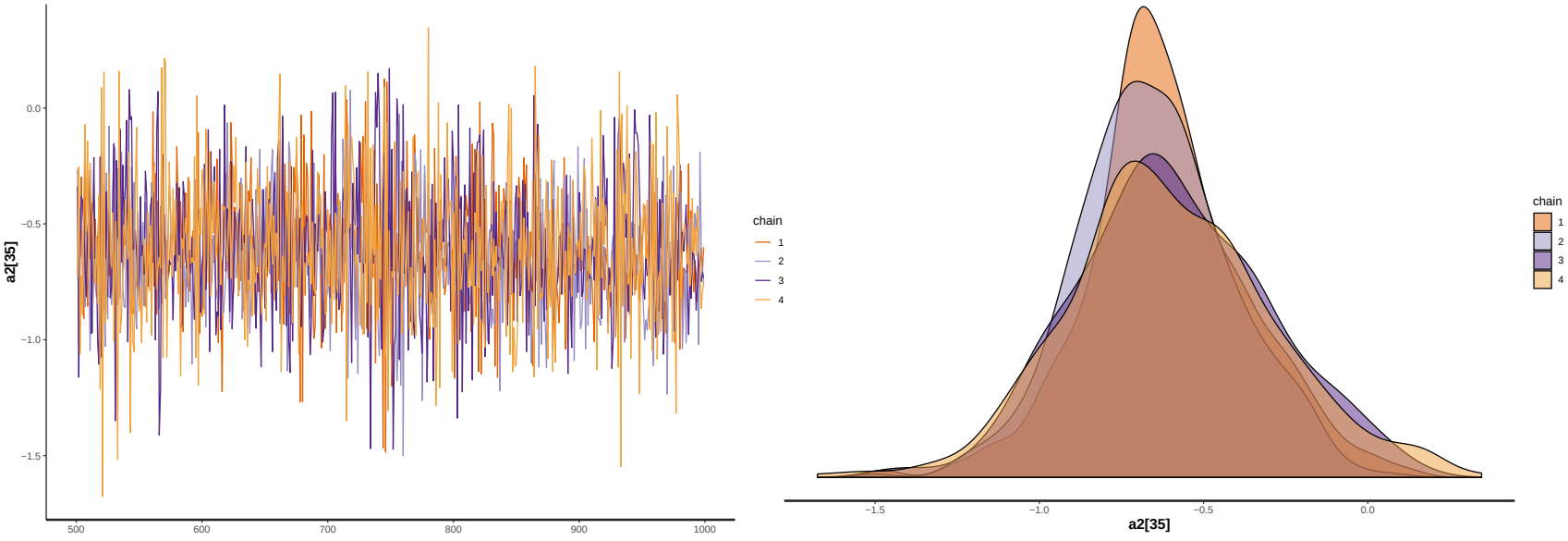
Fit in Stan

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

```
##                mean     se_mean      n_eff      Rhat
## a1[1]     1.1730539 0.005867669 2041.95833 0.9987677
## a1[35]    1.0405942 0.004815826 2467.79909 1.0000419
## a2[1]    -0.6146912 0.007621215 1784.00065 1.0022707
## a2[35]   -0.6148191 0.005761053 2236.87171 1.0047579
## sigma_a1  0.3438139 0.001861216  639.31369 1.0104995
## sigma_a2  0.3271075 0.023754571   25.81252 1.1325311
```

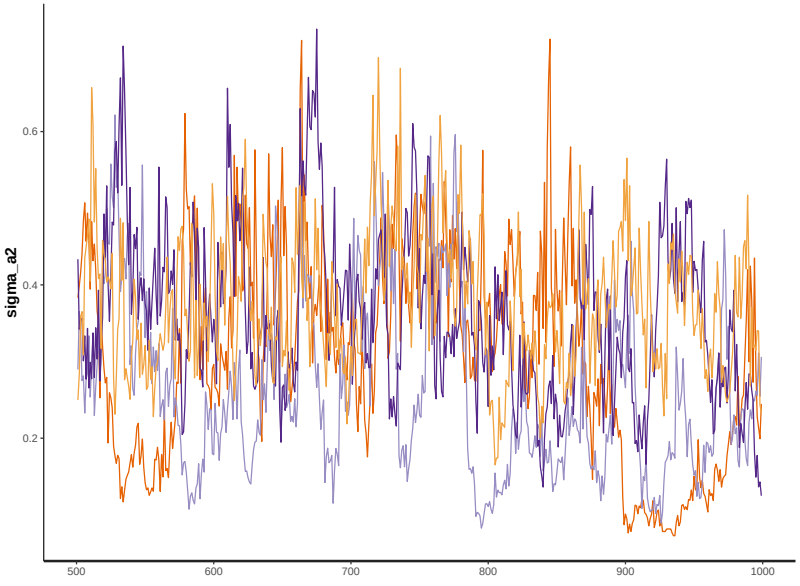
Let's look at some diagnostics

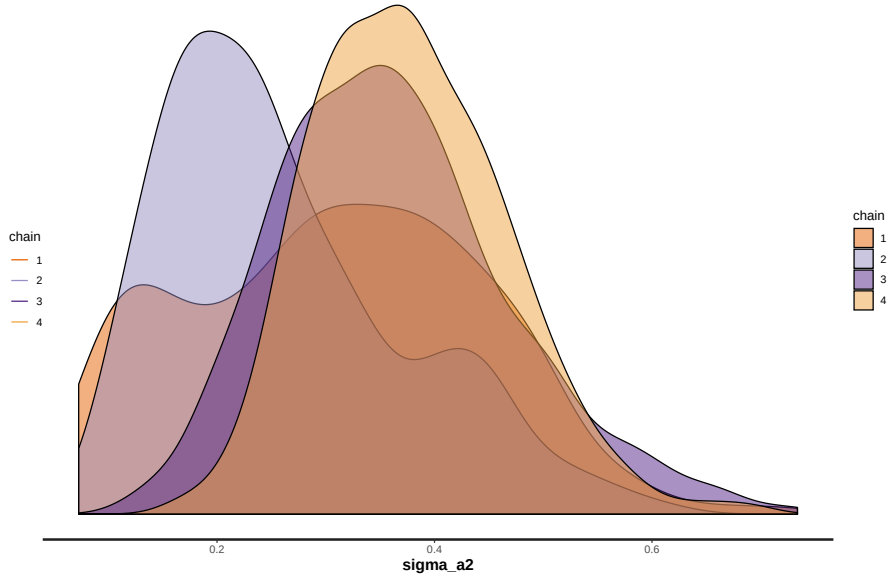
Pick a county (number 35) and plot the trace and density



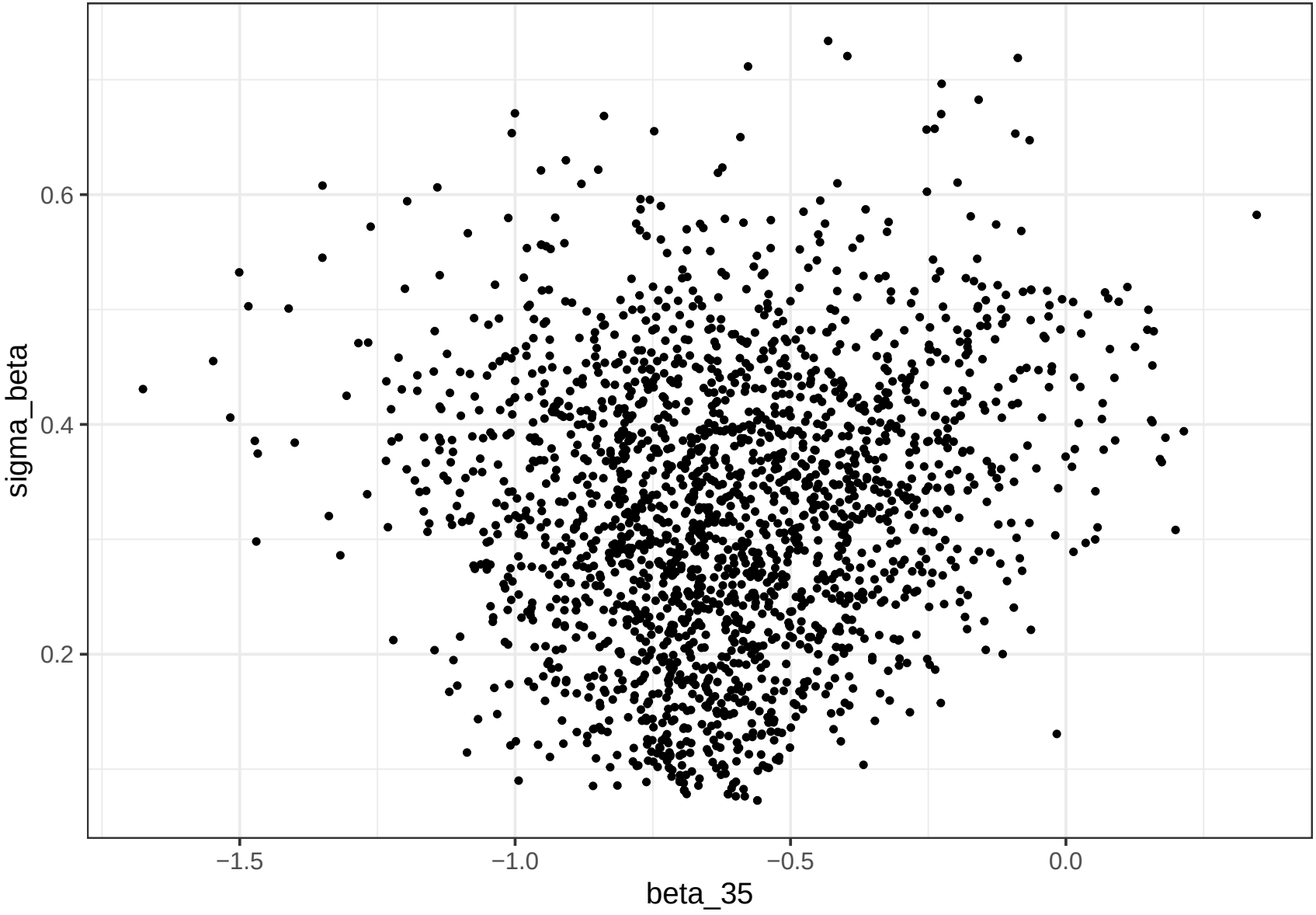
What about the variance parameter

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

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

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

Non-centered parameterization

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

$$\ldots$$

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

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

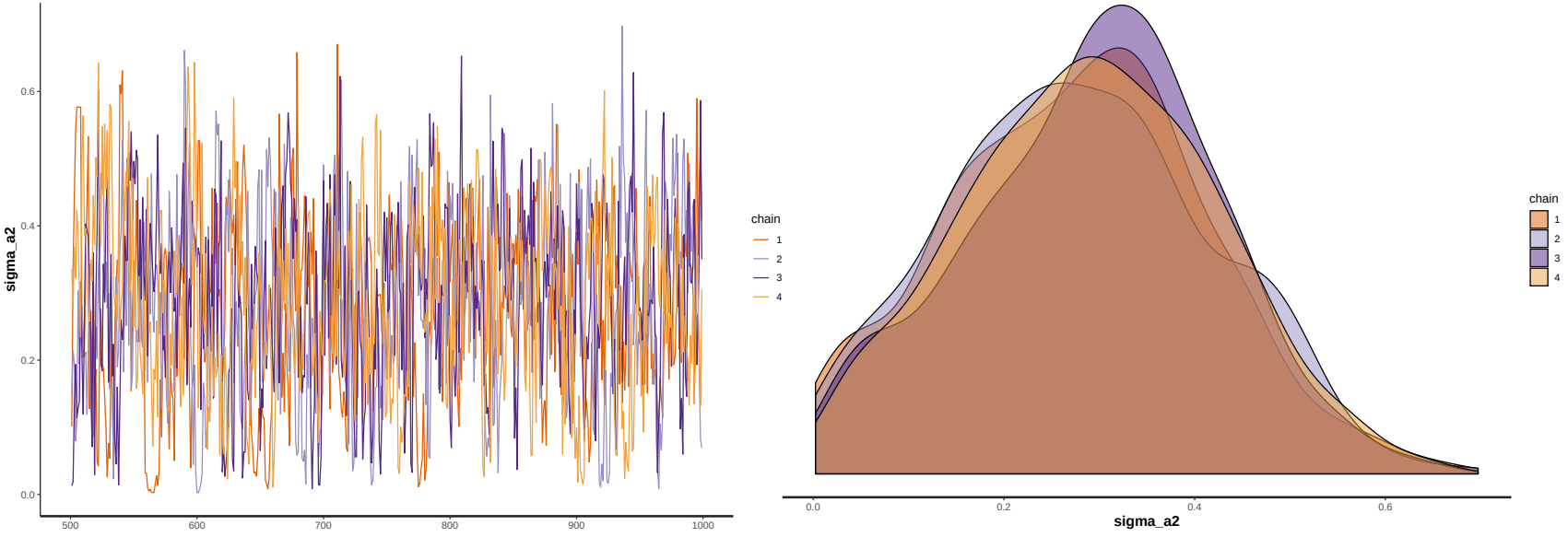
$$\ldots$$

- 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

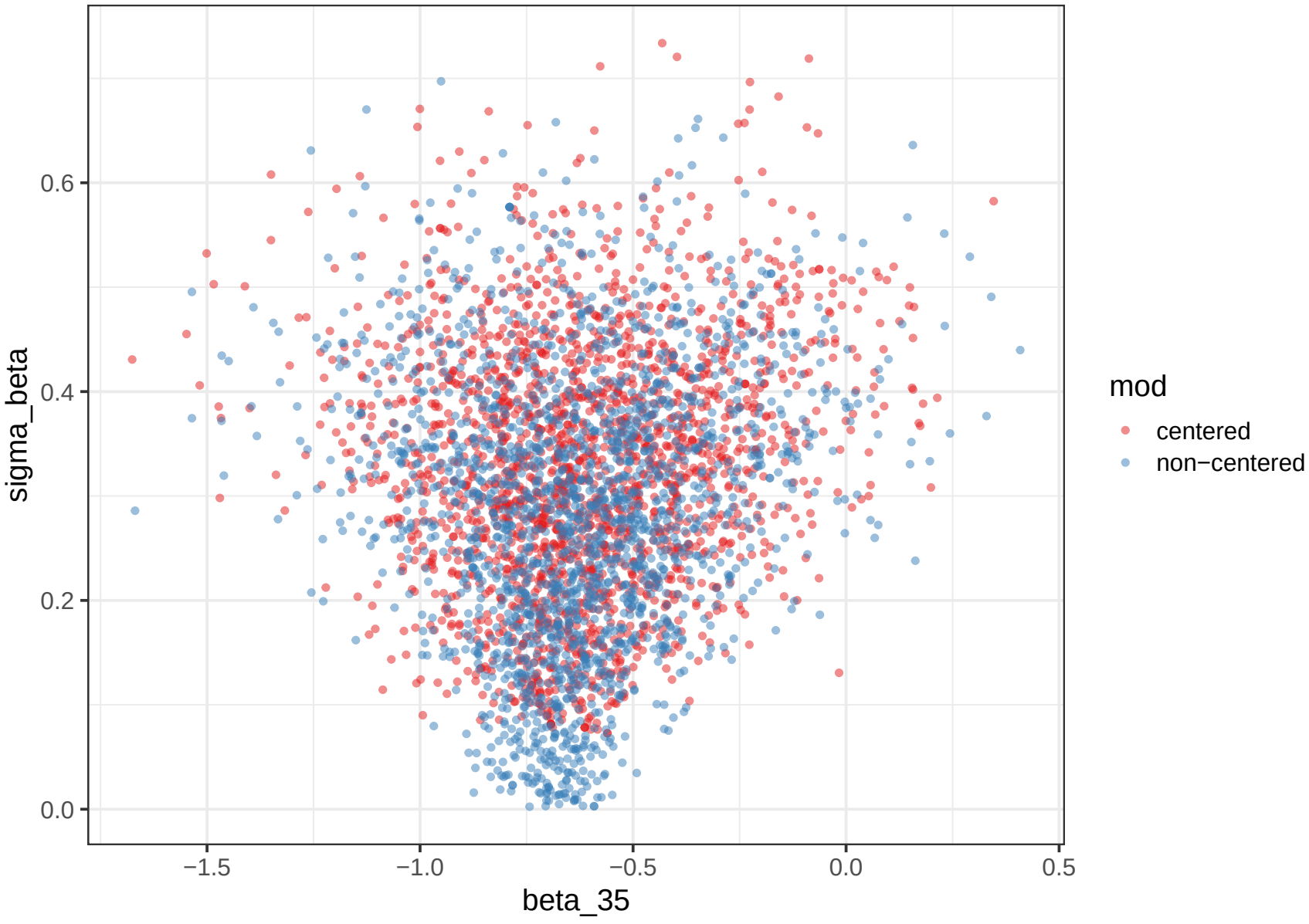
Non-centered Stan model

```
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);
}
```

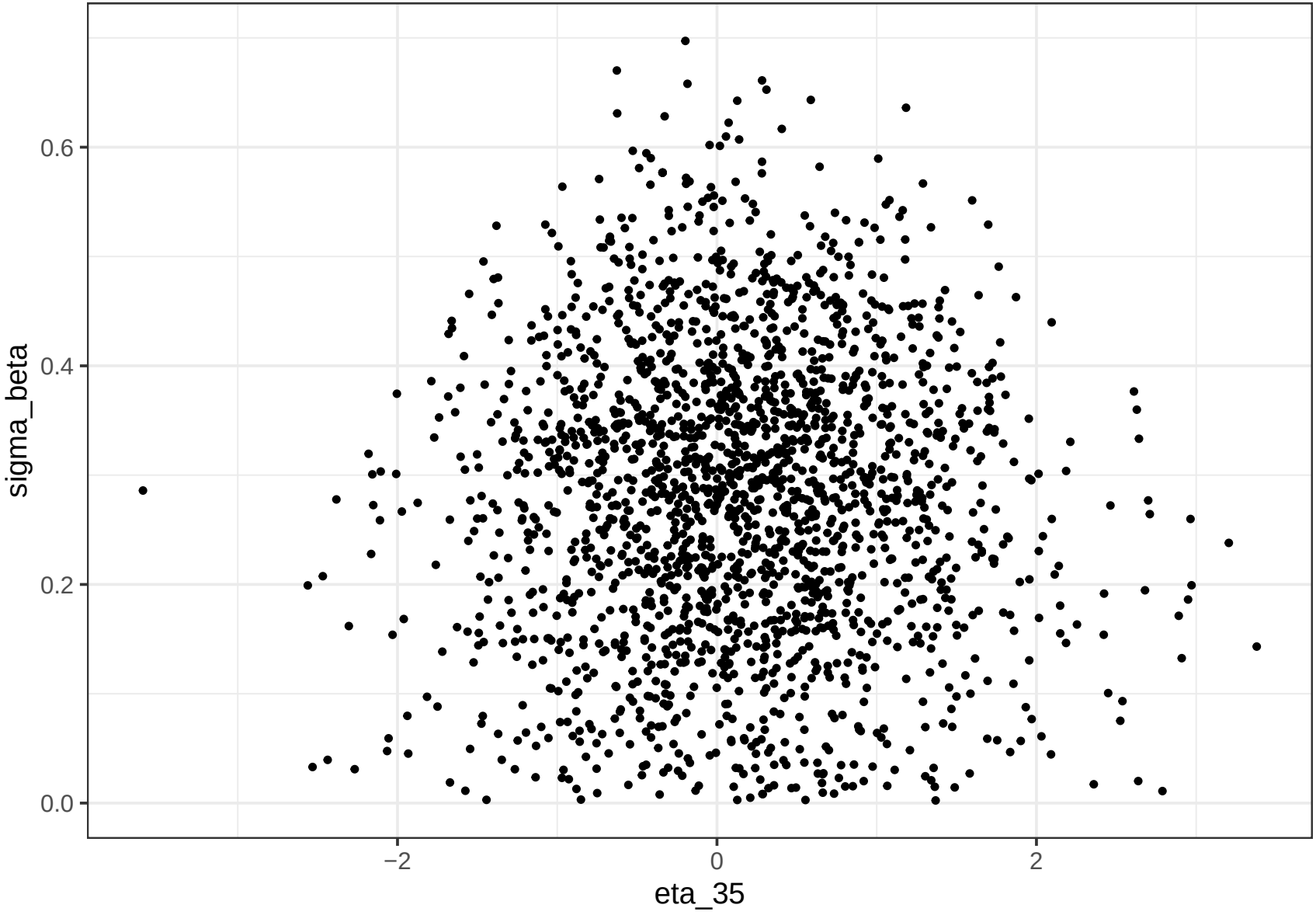
Diagnostics for σ_β



Better exploration of the funnel

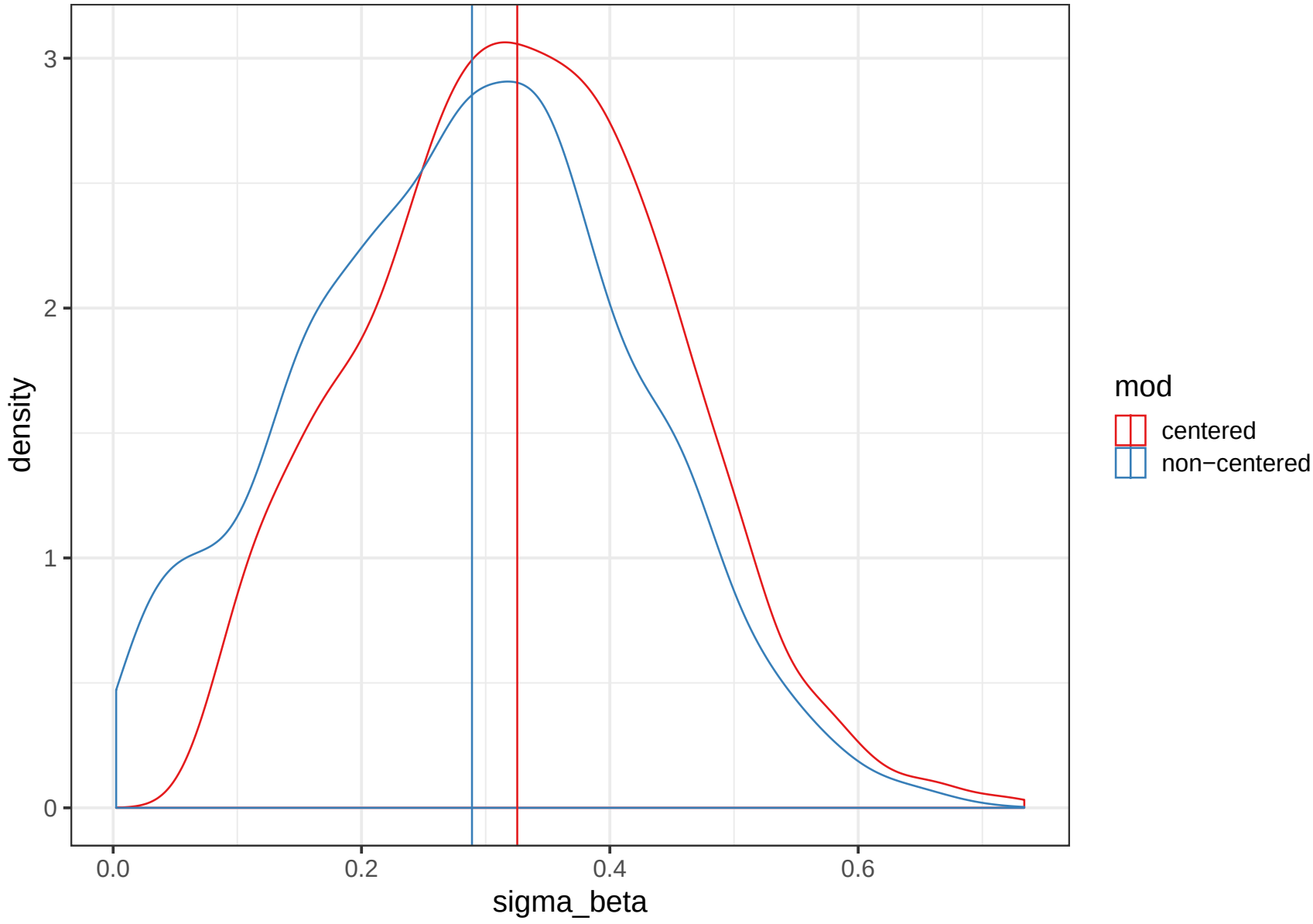


Less dependence between sampled parameters



Does it matter?

4% change in estimate for σ_β



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