SOC6707 Intermediate Data Analysis

Monica Alexander

Week 10: Hierarchical models II

Recap

Last week we started talking about hierarchical models

- Account for hierarchical structure in data (e.g. houses within counties)
- ‘Happy medium’ between treating all groups separately versus all groups the same
- Group-level effects are treated as coming from a common distribution, which allows information to be pooled across groups
- Particularly useful when some groups have small sample sizes

This week

- Varying slopes
- Hierarchical GLM (logistic regression)

Last week

For radon, we got up to

$$
\begin{aligned}
y_i &\sim N\left(\alpha_{j[i]} + \beta x_i, \sigma_y^2\right), \text{ for } i = 1, 2, \ldots, n \\
\alpha_j &\sim N\left(\gamma_0 + \gamma_1 u_j, \sigma_\alpha^2\right), \text{ for } j = 1, 2, \ldots, J
\end{aligned}
$$

Varying slopes

What about letting the effect of x_i vary by county?

- In last model, we assume that the difference between basement and first floor measurement is the same across houses, no matter which county the house is in.
- What if that difference varies by county?

$$\begin{aligned} y_i &\sim N\left(\alpha_{j[i]} + \beta_{j[i]} x_i, \sigma_y^2\right), \text{ for } i = 1, 2, \ldots, n \\ \alpha_j &\sim N\left(\mu_\alpha, \sigma_\alpha^2\right), \text{ for } j = 1, 2, \ldots, J \\ \beta_j &\sim N\left(\mu_\beta, \sigma_\beta^2\right), \text{ for } j = 1, 2, \ldots, J \end{aligned}$$

Allowing for varying slopes.

In R

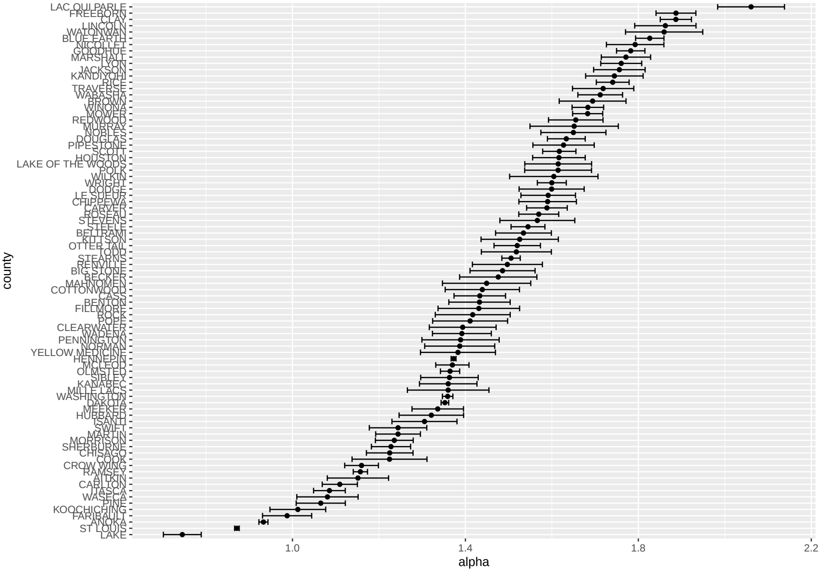
```
mod_hier_slopes <- lmer(log_activity ~ floor + (1+floor|county),
                        data = d_mn)
```

Allowing for varying slopes at unit level

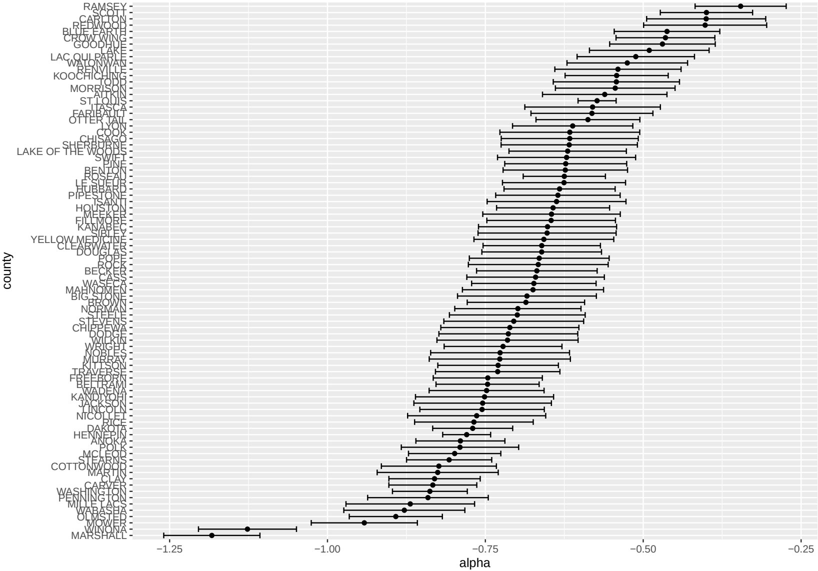
$$
\begin{aligned}
y_i &\sim N\left(\alpha_{j[i]} + \beta_{j[i]} x_i, \sigma_y^2\right), \text{ for } i = 1, 2, \ldots, n \\
\alpha_j &\sim N\left(\mu_\alpha, \sigma_\alpha^2\right), \text{ for } j = 1, 2, \ldots, J \\
\beta_j &\sim N\left(\mu_\beta, \sigma_\beta^2\right), \text{ for } j = 1, 2, \ldots, J
\end{aligned}
$$

- Estimate of μ_α is 1.46
- Estimate of μ_β is -0.679

County-specific intercepts



County-specific slopes



Hierarchical logistic regression

Hierarchical logistic regression

We can easily extend the idea of modeling hierarchical data to cases where our outcome of interest is a binary variable and we want to use logistic regression.

- Recall for binary data, we have observations of our outcome $y_1, y_2, \ldots, y_n$ where y_i is equal to 1 if the outcome of interest occurred for observation i and 0 otherwise.
- We are interested in estimating the probability that the outcome occurs associated with one or more covariates X_i, i.e. $\Pr(Y_i = 1|X_i)$
- In usual logistic regression we model this as

$$\text{logit}\ \Pr(Y_i = 1|X_{i1}, \ldots X_{ik}) = \beta_0 + \beta_1 X_{i1} \cdots + \beta_k X_{ik}$$

Hierarchical logistic regression

Changing the notation slightly, this is the same as

$$Y_i \sim \text{Bernoulli}(p_i)$$

with

$$\text{logit } p_i = \beta_0 + \dots$$

So for hierarchical logistic regression we can model the probabilities of interest (p_i's) with a hierarchical set-up

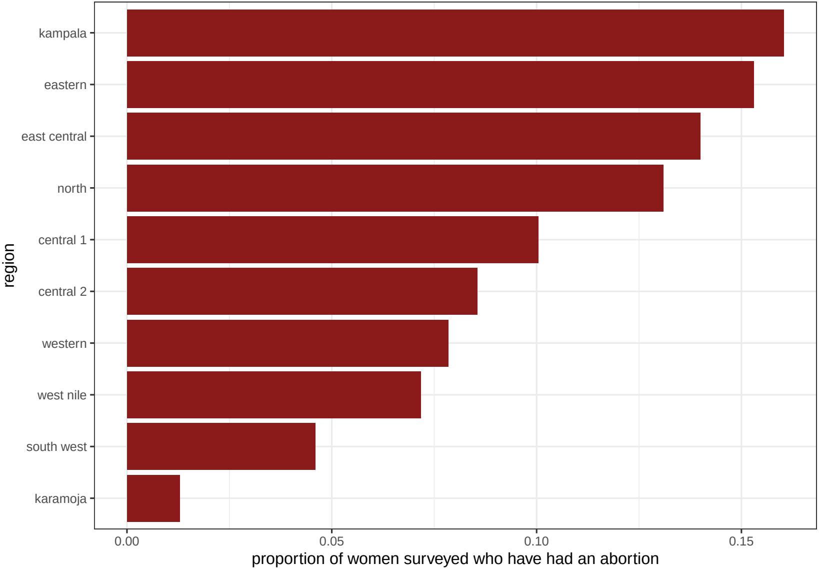
Motivating example

Abortion outcomes in Uganda

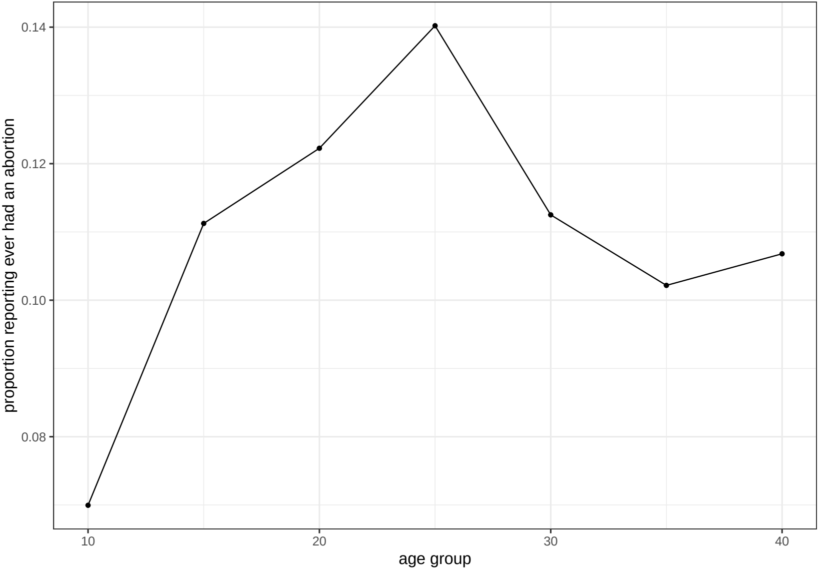
- Data from 2018 PMA survey (via IPUMS)
- Interested in factors associated with women ever having an abortion
- Outcome of interest: ‘ever had abortion (yes/no)’
- Notes: dropping don’t knows, including ‘unsuccessful abortions’ in ‘yes’.
- More notes: Self-reported abortion is very likely to be under-reported

Some graphs

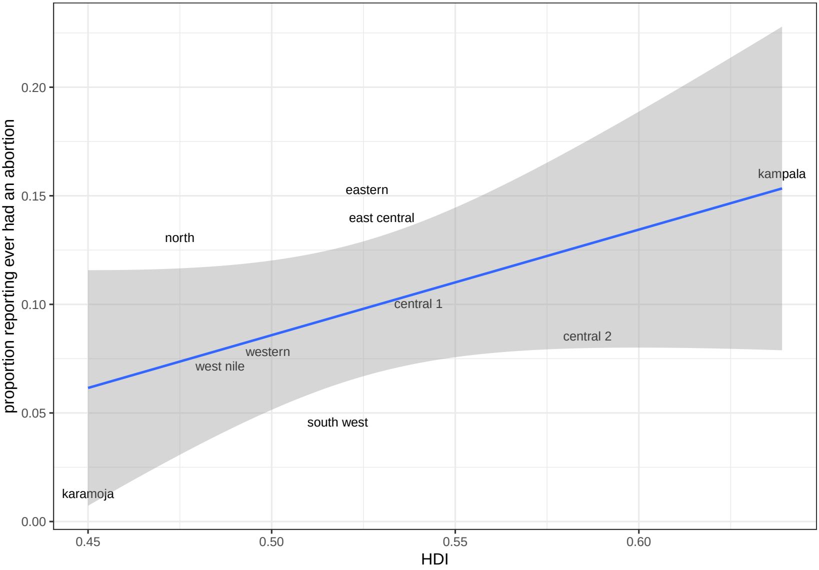
Proportion by region



Proportion by age group



Relationship with HDI



Model 1

Let's fit

$$Y_i \sim \text{Bernoulli}(p_i)$$

$$\text{logit} p_i = \alpha_{j[i]} + \beta x_i$$

$$\alpha_j \sim N(\mu, \sigma^2)$$

where

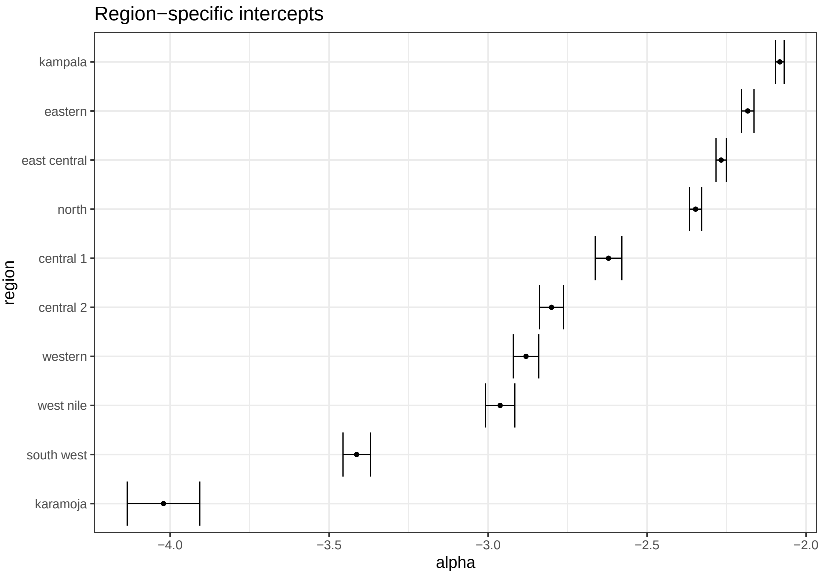
- Y_i refers to whether or not individual i has had an abortion
- The index j refers to region
- x_i refers to age

Fit in R

```r
mod <- glmer(abortion ~ (1|region)+age, data = d, family = "binomial")
summary(mod)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: abortion ~ (1 | region) + age
##    Data: d
##
##      AIC      BIC   logLik deviance df.resid
##   2412.1   2430.7  -1203.0   2406.1     3645
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -0.5223 -0.3959 -0.3222 -0.2417  6.3466
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  region (Intercept) 0.3788   0.6154
## Number of obs: 3648, groups:  region, 10
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.771946   0.251597 -11.017  < 2e-16 ***
## age          0.019118   0.006163   3.102  0.00192 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Correlation of Fixed Effects:
##     (Intr)
## age -0.568
```

Region-specific intercepts



Converting to probabilities

What is probability of abortion for a woman aged 30 in north region? The info we need:

```
alphas %>% filter(region == "north")
```

```
##      alpha region         se
## 1 -2.34757  north 0.01921455
```

```
coef(summary(mod))
```

```
##                Estimate  Std. Error    z value     Pr(>|z|)
## (Intercept) -2.77194614 0.251597408 -11.017387 3.150710e-28
## age          0.01911841 0.006162747   3.102255 1.920527e-03
```

Model 2

$$Y_i \sim \text{Bernoulli}(p_i)$$

$$\text{logit} p_i = \alpha_{j[i]} + \beta x_i$$

$$\alpha_j \sim N(\gamma_0 + \gamma_1 z_j, \sigma^2)$$

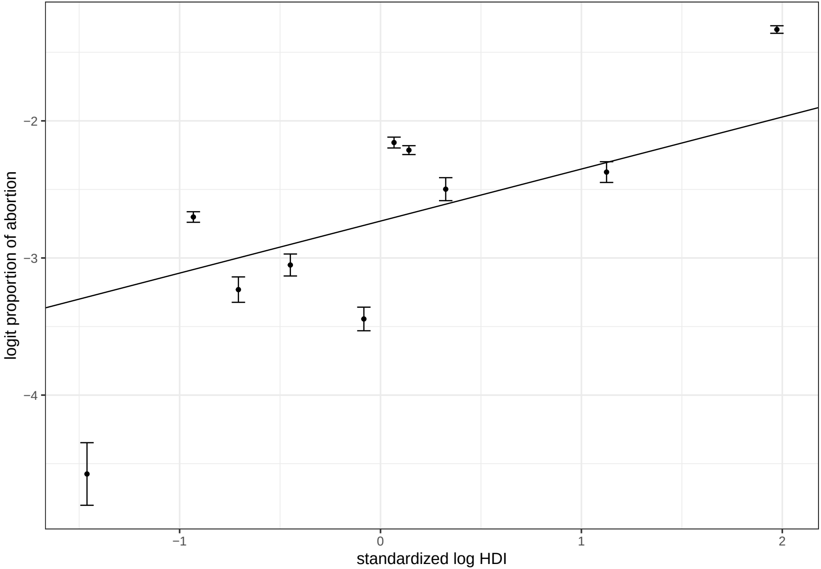
where everything is as before and z_j is the standardized log HDI at the region level

Fit in R

```
d <- d %>% mutate(log_hdi_c = (log(hdi)-mean(log(hdi)))/sd(log(hdi)))
mod2 <- glmer(abortion ~ (1|region)+age+log_hdi_c, data = d, family = "binomial")
summary(mod2)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: abortion ~ (1 | region) + age + log_hdi_c
##    Data: d
##
##      AIC      BIC   logLik deviance df.resid
##   2410.1   2435.0  -1201.1   2402.1     3644
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -0.5302 -0.3959 -0.3244 -0.2453  6.5454
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  region (Intercept) 0.253    0.503
## Number of obs: 3648, groups:  region, 10
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.730772   0.224626 -12.157  < 2e-16 ***
## age          0.019211   0.006164   3.117  0.00183 **
## log_hdi_c    0.379540   0.182386   2.081  0.03744 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Correlation of Fixed Effects:
##           (Intr) age
## age       -0.634
## log_hdi_c  0.019  0.018
```

Region-specific intercepts versus HDI



Age as factor

```
d <- d %>% mutate(age_group = factor(age_group)) %>%
  mutate(age_group = fct_relevel(age_group, "20", after = 0))
mod3 <- glmer(abortion ~ (1|region)+age_group+log_hdi_c, data = d, family = "binomial")
coef(summary(mod3))
```

```
##                Estimate Std. Error     z value     Pr(>|z|)
## (Intercept) -2.12953815  0.2077000 -10.2529503 1.147854e-24
## age_group10 -0.69216405  0.1809605  -3.8249462 1.308007e-04
## age_group15 -0.13498316  0.1656174  -0.8150301 4.150551e-01
## age_group25  0.17967992  0.1802938   0.9965952 3.189610e-01
## age_group30 -0.08585389  0.2011052  -0.4269103 6.694446e-01
## age_group35 -0.19036099  0.2220876  -0.8571439 3.913654e-01
## age_group40 -0.17131168  0.3442972  -0.4975692 6.187877e-01
## log_hdi_c    0.38649272  0.1857092   2.0811719 3.741817e-02
```

Non-nested hierarchies

Non-nested hierarchies

We can extend hierarchical structure to be on more than one variable, e.g. by region and age group.

$$Y_i \sim \text{Bernoulli}(p_i)$$

$$\text{logit} p_i = \beta_0 + \alpha_{j[i]} + \eta_{k[i]}$$

$$\alpha_j \sim N(0, \sigma_j^2)$$

and

$$\eta_k \sim N(0, \sigma_k^2)$$

where k refers to age group of individual i

- Notice now everthing is centered at zero and we have a global intercept

Rewrite model

Adding non-nested hierarchies in R

```
mod4 <- glmer(abortion ~ (1|region)+(1|age_group), data = d, family = "binomial")
summary(mod4)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: abortion ~ (1 | region) + (1 | age_group)
##    Data: d
##
##      AIC      BIC   logLik deviance df.resid
##   2408.7   2427.3  -1201.4   2402.7     3645
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -0.5001 -0.4089 -0.3172 -0.2301  6.1034
##
## Random effects:
##  Groups    Name        Variance Std.Dev.
##  region    (Intercept) 0.39825  0.6311
##  age_group (Intercept) 0.05989  0.2447
## Number of obs: 3648, groups:  region, 10; age_group, 7
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.3333     0.2328  -10.02   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Summary

- Can extend hierarchical models to be varying slopes as well as intercepts
- Can use hierarchical models in a logistic regression context where the outcome of interest is binary
- In general, useful when you have small sample sizes in some cells (i.e. some population subgroups)
- Standardizing covariates (minusing meand and dividing by standard deviation) is useful for interpretation (sometimes) but also to fit models

Lab - R Markdown