Multilevel models, part 1

Frank Edwards

Multilevel data

Many natural and social processes have inherent groupings

- People live in neighborhoods (within cities)
- Students learn in classrooms (within schools, within districts, …)
- Cities are subject to state laws

These groups often lead to patterns in our data called *clustering*

- Variables are often correlated within clusters
- These correlations can be due to observed variables
- But they can also be due to unobserved (or unobservable) features of each cluster
- If we don't account for clustering, our inferences will be misleading

Data for today (available at `https://github.com/f-edwards/intermediate_stats/tree/master/data`)

```
dat<-read_csv("./data/violent_arrest_data.csv")
```

The Distribution of poverty and crime across counties
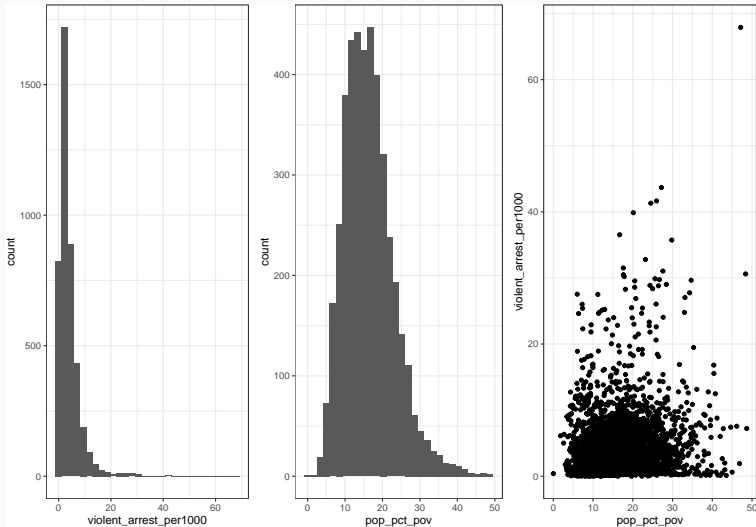
$$y_c = \beta_0 + \beta_1 x_c + \varepsilon_c$$

$$\varepsilon \sim N(0, \sigma^2)$$

This model, with no included parameters for group differences, is called a *complete pooling* model. Data from all groups is pooled to estimate a single intercept and single slope.

Where violent crime rates are *y*, poverty rates are *x*, counties are *c*, states are *s*

Estimating the complete-pooling model

```
m_complete_pooling <- lm(violent_arrest_per1000 ~ pop_pct_pov,
                         data = dat)

summary(m_complete_pooling)


##
## Call:
## lm(formula = violent_arrest_per1000 ~ pop_pct_pov, data = dat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.816 -2.370 -1.086  1.002 60.566
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.207479   0.170868   12.92   <2e-16 ***
## pop_pct_pov 0.108641   0.009632   11.28   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.21 on 4284 degrees of freedom
## Multiple R-squared:  0.02884,    Adjusted R-squared:  0.02861
## F-statistic: 127.2 on 1 and 4284 DF,  p-value: < 2.2e-16
```
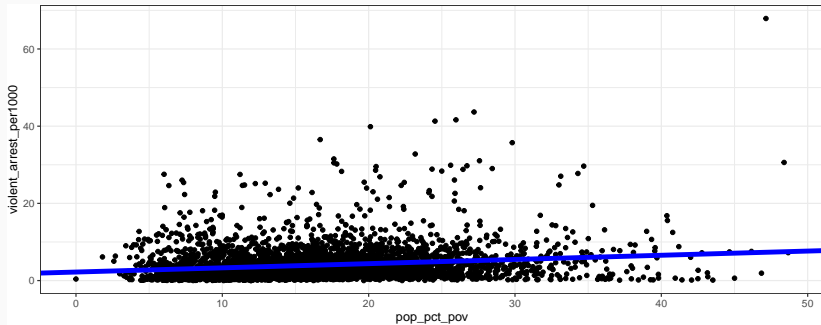
Sometimes units in the data have features that make them similar to other units.

- Repeated observations of counties over time (panel data)
- Observations of many counties from the same time-period (cross-sectional data)
- Counties nested within larger geographies (states, regions)
- We can adjust our models to learn from these features and improve inference

One source of clustering: geography (state)
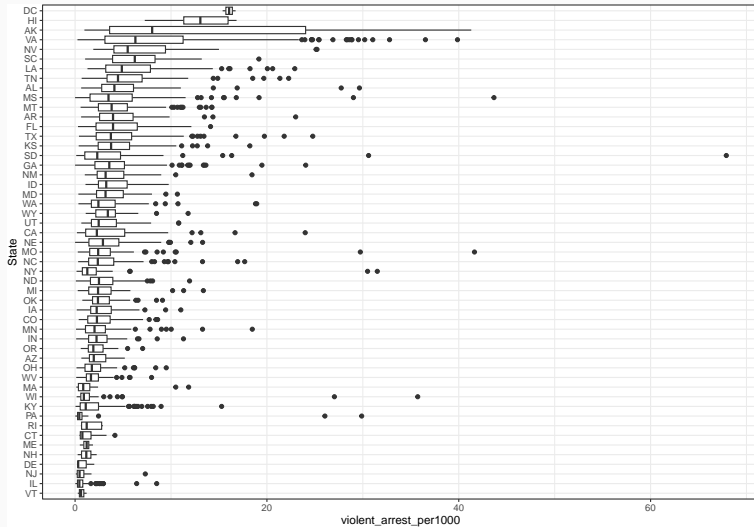
Intercepts as regression parameters

Let's say that we think that states may have different characteristics that may be associated with variation in violent crime arrest rates.

- This could include things like law, economics, politics, history, or population features.

Let's estimate a model where NO information across states is included in our estimation of intercepts.

- Every state gets their own intercept (computed as the exact within-state violent crime rate mean).

Where violent crime rates *y* are a function of poverty rates *x* in state *s*, and county *c*

Complete pooling shares information across states for all parameters, but makes no distinctions across states

$$E(y_c) = \beta_0 + \beta_1 x_c$$

No pooling does shares no information across states for computing intercepts. There is one (pooled) slope for all counties in this model

$$E(y_c) = \beta_s + \beta_1 x_c$$

Estimating the state fixed effects model

```
m_state_fe<-lm(violent_arrest_per1000 ~
                 pop_pct_pov +
                 factor(stusps),
              data = dat)
broom::tidy(m_state_fe)
```

```
## # A tibble: 52 x 5
##    term              estimate std.error statistic  p.value
##    <chr>                <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)           11.6     0.808      14.3 2.78e-45
##  2 pop_pct_pov          0.134    0.0103      12.9 1.24e-37
##  3 factor(stusps)AL     -9.07     0.909     -9.98 3.35e-23
##  4 factor(stusps)AR     -9.62     0.903     -10.7 3.58e-26
##  5 factor(stusps)AZ     -11.9      1.11     -10.8 7.64e-27
##  6 factor(stusps)CA     -10.2     0.888     -11.4 7.76e-30
##  7 factor(stusps)CO     -10.5     0.898     -11.7 3.56e-31
##  8 factor(stusps)CT     -11.2      1.55     -7.22 6.30e-13
##  9 factor(stusps)DC      2.09      2.78     0.751 4.53e- 1
## 10 factor(stusps)DE     -12.3      2.32     -5.29 1.30e- 7
## # i 42 more rows
```

Intercepts are computed as the average violent arrest rate in each state, when poverty is fixed at zero - Fixed effects adjust for differences in mean violent arrest rates *across states*

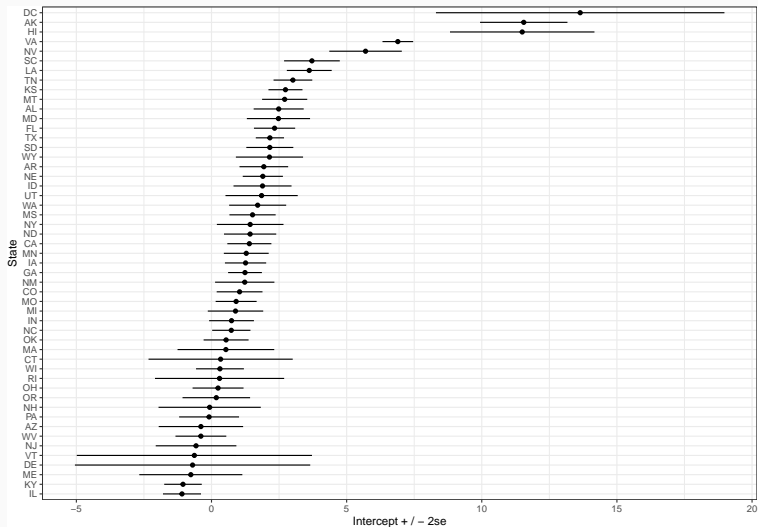# Visualizing the fixed effects (state intercepts) model
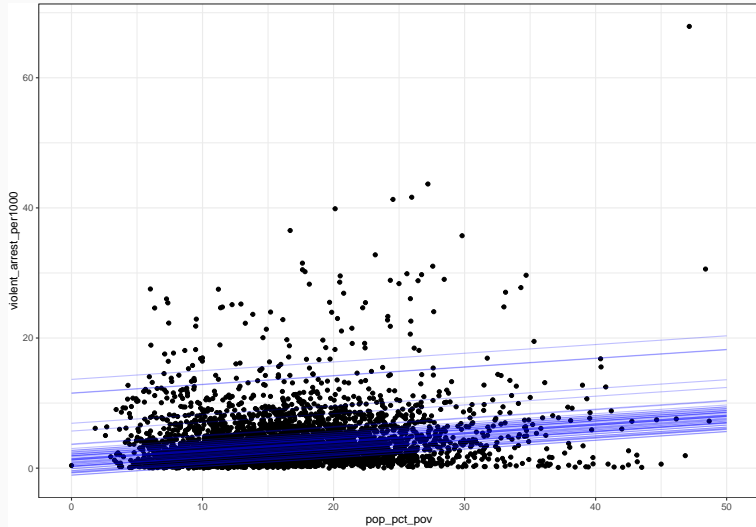
A *big* improvement!

```
BIC(m_complete_pooling, m_state_fe)
```

```
##                       df      BIC
## m_complete_pooling    3   24507.84
## m_state_fe           53   23917.00
```

Let's look at the distribution of intercepts. What do they mean?

After adjusting for all unmeasured (time-stable) confounding, we expect a county with zero poverty in state s to have β_s average arrests for violent crime per 1,000 persons per year when poverty is zero.

Problems with a fixed effects approach

- We *only* use information on counties within a state to compute intercepts
- States with few counties have *much* more error than states with many counties. This may cause too much certainty in some states, too little in others.
- To make a prediction, we *must* specify which state the prediction comes from

Introducing the multilevel model

- Let's assume that violent crime rates in counties *across* states differ from each other.
- We will estimate a *distribution* of intercepts that lets us think about differences both across and within states

- Under a complete-pooling model, we estimate a single intercept for the full data based on the average violent crime rate *across all counties*
- Under a no-pooling model, we estimate an intercept for every state based *only* on the data from that state
- Under a partial-pooling model, we estimate an intercept for every state based based on a model that *assumes that state intercepts are generated from an underlying probability distribution*, using data from all states

A basic multilevel model

Where violent crime rates *y* are a function of poverty rates *x* in state *s*, and county *c*

$$y_c = \alpha_s + \beta_1 x_c + \varepsilon_c$$

$$\varepsilon \sim N(0, \sigma^2)$$

$$\alpha \sim N(\bar{x}, \sigma_\alpha^2)$$

Let's compare the fixed effects and multilevel specifications

Fixed effects

$$y_c = \beta_s + \beta_1 x_c + \varepsilon_c$$
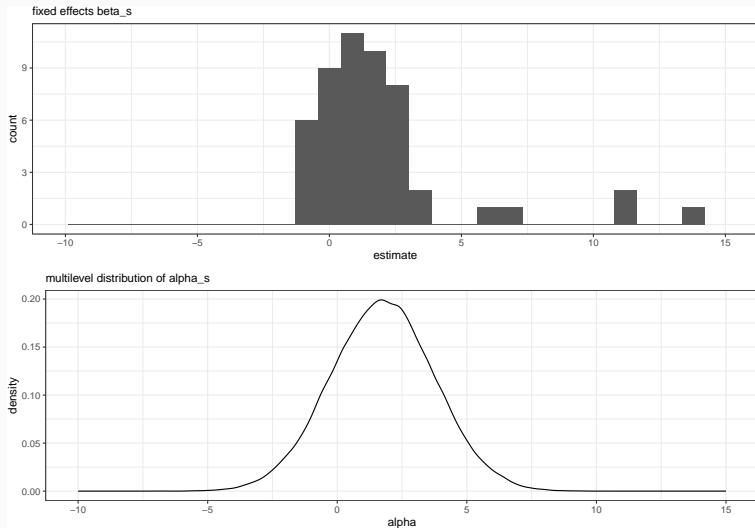$$\varepsilon \sim N(0, \sigma^2)$$

Multilevel

$$y_c = \alpha_s + \beta_1 x_c + \varepsilon_c$$
$$\varepsilon \sim N(0, \sigma^2)$$
$$\alpha \sim N(\bar{x}, \sigma_\alpha^2)$$

Our model *regularizes* (or shrinks) intercept estimates by nudging them toward the global mean

This shift is larger when:

- The group mean (state here) is far away from the global (national) mean
- We have few observations for a unit (small N of counties here)

This shift is smaller when:

- The group mean is close to the global mean
- We have many observations for a unit (large N of counties)

We use the `lmer` function from the `lme4` package for multilevel modeling in R, `stan_glmer()` and `brm()` follow the same syntax.

```r
library(lme4)

m_state_ml<-lmer(violent_arrest_per1000 ~
                    pop_pct_pov +
                    (1|stusps),
                 data = dat)
```

The (1|Variable) notation is used to specify the categorical variables that will have intercepts estimated.

Model output (some familiar, some new!)

```
summary(m_state_ml)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: violent_arrest_per1000 ~ pop_pct_pov + (1 | stusps)
##    Data: dat
##
## REML criterion at convergence: 23701.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.8999 -0.4908 -0.1751  0.2522 15.7779
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  stusps   (Intercept)  5.942   2.438
##  Residual             14.196   3.768
## Number of obs: 4286, groups:  stusps, 51
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  1.86112    0.38769   4.801
## pop_pct_pov  0.13340    0.01029  12.968
##
## Correlation of Fixed Effects:
##             (Intr)
## pop_pct_pov -0.401
```

We now have *two* variance components of our model. The residual ε and the intercept α.

Our model results tell us that $\sigma = 3.8$ and $\sigma_\alpha = 2.4$

What does that mean?

We have effectively divided our residual into a *state* component (α), and a *county-year* component ε

Residual variance is the unexplained difference between observed violent crime and the regression line based on poverty.

So how much variation is between states, and how much is between counties after we adjust for state differences?

ICC measures the proportion of total unexplained variation captured by your random effects

For our model, we can compute ICC as $\frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma^2} = \frac{2.4^2}{2.4^2 + 3.8^2} = 0.40$

This tells us that, after accounting for poverty in the county, 40 percent of the residual variation in violent crime rates across counties is explained by time-stable features of the state that they are in.
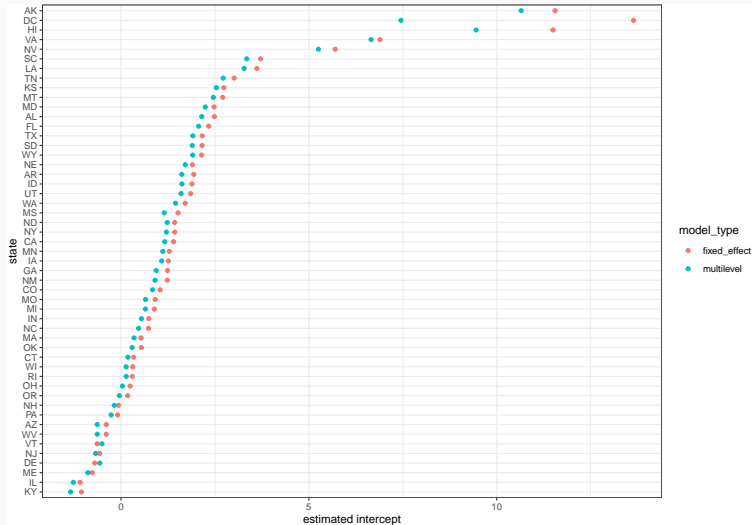
Let's compare results on the console between these two models

```
m_state_fe<-lm(violent_arrest_per1000 ~ pop_pct_pov +
                  factor(stusps),
               data = dat)

m_state_re<-lmer(violent_arrest_per1000 ~ pop_pct_pov +
                  (1|stusps),
               data = dat)
```

Let's check out the regularization we obtained from the multilevel model

Summary

- Estimated intercepts are equal to empirical group means
- This approach typically *overfits* the data, and makes for poor prediction, unless we have a lot of data on each unit

- The complete pooling approach assumes all units are the same
- The no-pooling approach assumes all units are different
- The *partial pooling* approach assumes that all units are different, but come from the same underlying distribution

Multilevel models are a good default

Multilevel models are very flexible and useful for modeling:

- Repeat observations of units over time
- Data with clusters

Multilevel models

- Shrink estimates toward group means
- Provide variance estimates across and within clusters
- Allow for flexible inference (especially for new units)

- Using multilevel models to estimate both varying intercepts and varying slopes across units
- Inference using multilevel models