# WELCOME TO THE 2021 NDACAN SUMMER TRAINING SERIES!

- The session will begin at 12pm EST.
- Please submit questions to the Q&A box.
- This session is being recorded.

# NDACAN SUMMER TRAINING SERIES

National Data Archive on Child Abuse and Neglect

Cornell University & Duke University

Children's Bureau
An Office of the Administration for Children & Families

NDACAN

# DATA STRATEGIES FOR THE STUDY OF CHILD WELFARE

# NDACAN SUMMER TRAINING SERIES SCHEDULE

- July 7, 2021 - Introduction to NDACAN
- July 14, 2021 - Survey Based Data
- July 21, 2021 - Administrative Data and Linking
- July 28, 2021 - VCIS Data and Special Populations
- **August 4, 2021 - Multilevel Modeling Workshop**
- August 11, 2021 - Latent Class Analysis Workshop

# SESSION AGENDA

- Describe the basic components of a multilevel model
- Describe the multi-level structure of AFCARS
- Estimate simple multilevel models in R using AFCARS

# UNDERSTANDING MULTILEVEL DATA

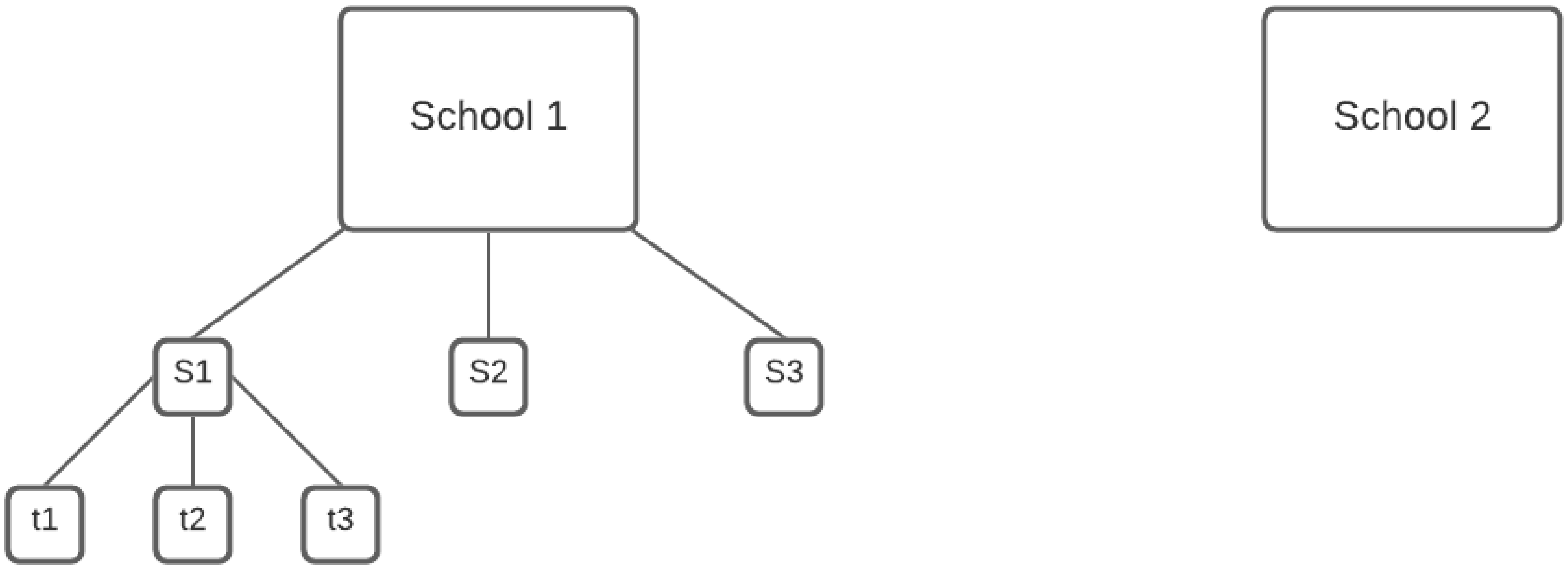
# THINKING ABOUT LEVELS

- We typically model data with the *iid* (independent and identically distributed) assumption.
  - This is a key assumption of most regression models (i.e. OLS)
- What does *iid* really mean?
  - Observations are *independent of each other*
  - Observations come from the *same data generating process* (or probability distribution)

# DOES THE IID ASSUMPTION HOLD?

- Students take a reading test 3 times during the school year.
  - All students in third grade take the test
  - The test is administered at 3 schools
  - We obtain 4 variables:
    - (1) student test score, (2) test wave, (3) student school, (4) student family's adjusted gross income.
- We'd want to know if students from high-income families score higher on the test than their low-income peers.

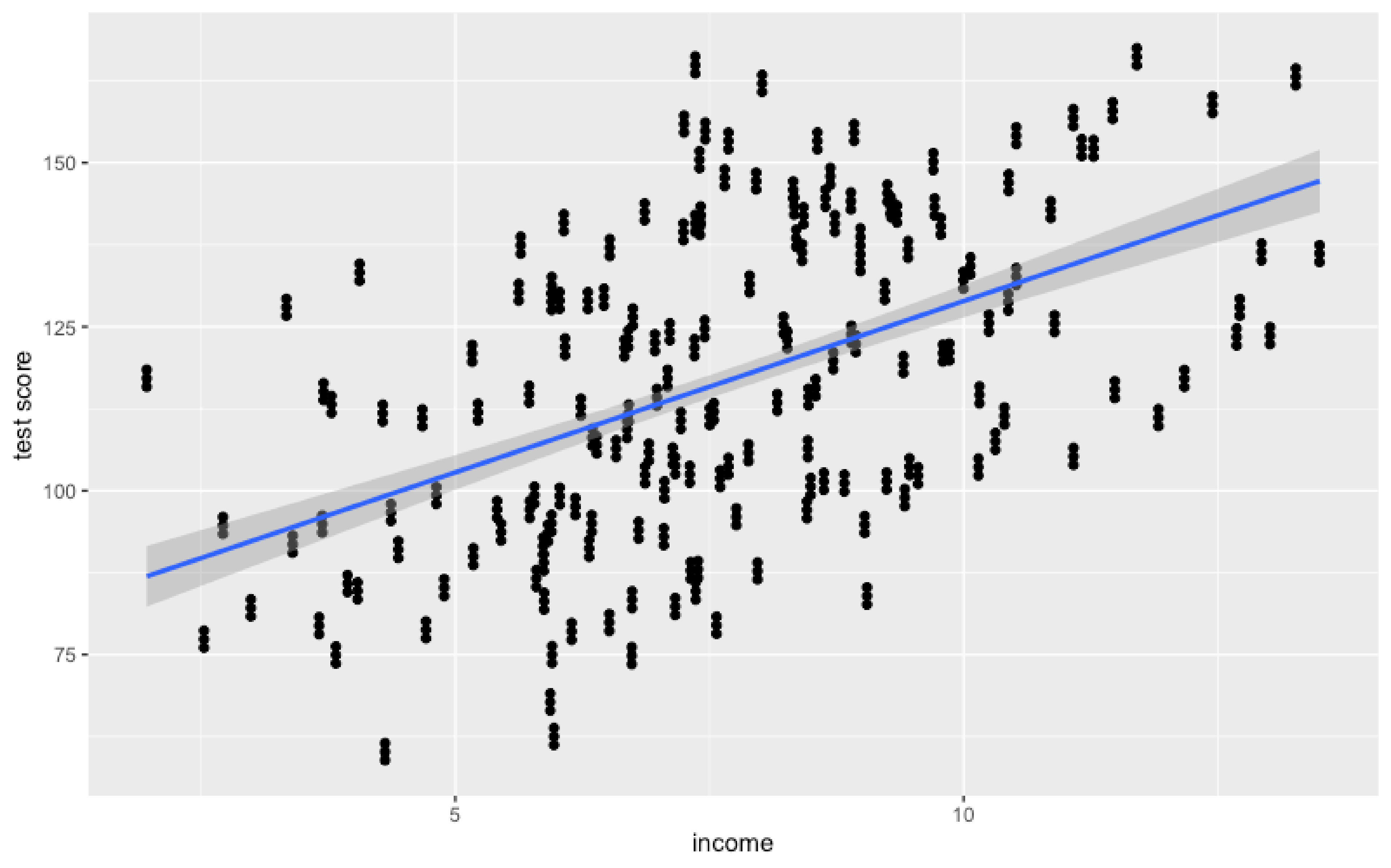
# THE STRUCTURE OF THE DATA



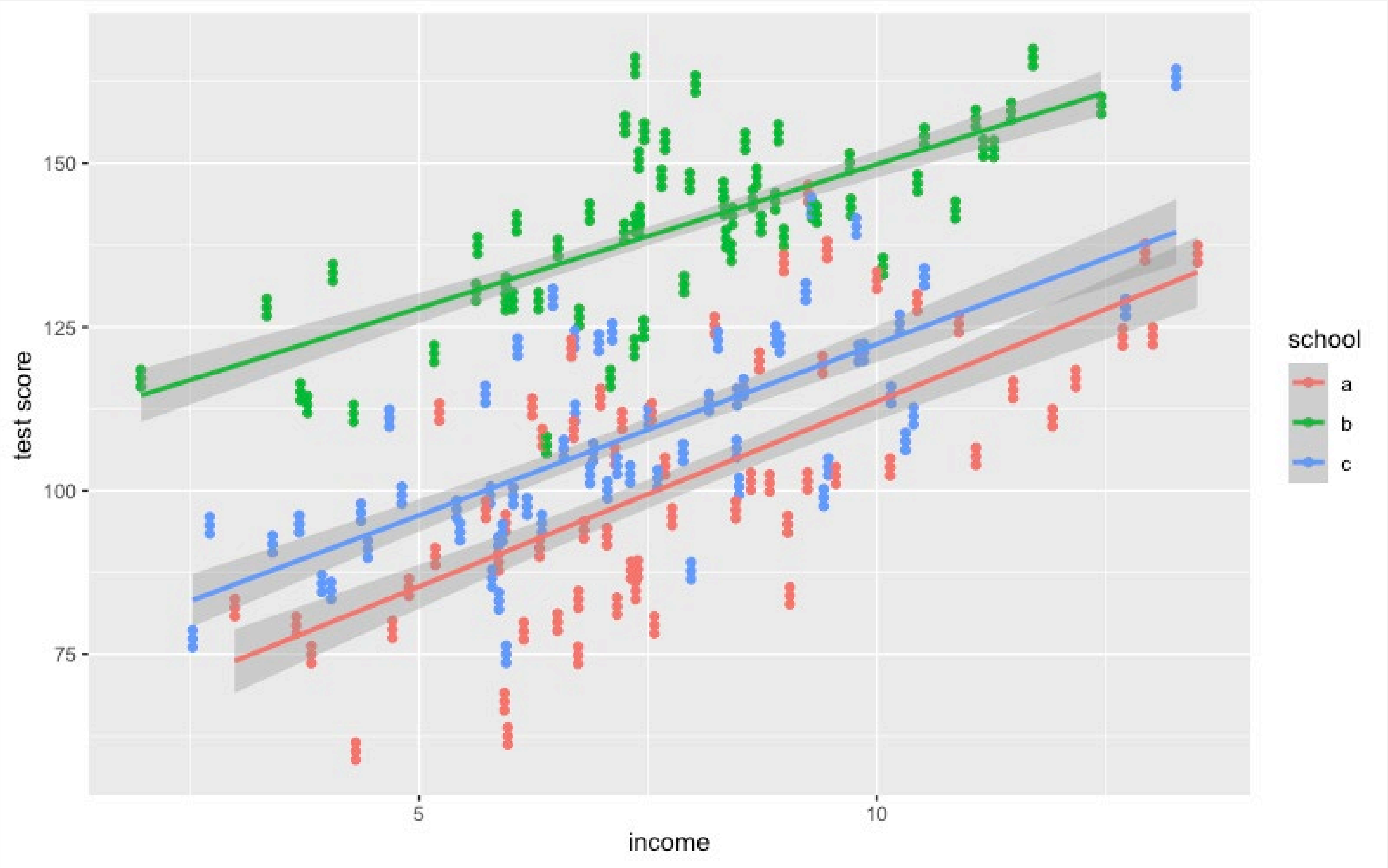
# THE STRUCTURE OF THE DATA

- Each test score is produced by student *i*, taught by teacher *t*, within school *s*.
  - A student's test scores at wave *w* are likely correlated with a student's test score at wave $w + 1$
  - Students within schools likely have correlated test scores
- We can identify a student's test score *y* as $y_{iwst}$
- Let's call student family income $x_i$

# IGNORING THE MULTILEVEL STRUCTURE



# INCORPORATING SCHOOL INTO THE MODEL



# DESCRIBE THE MULTILEVEL STRUCTURE OF AFCARS

# AFCARS FOSTER CARE ANNUAL FILE DATA STRUCTURE

- AFCARS provides annual case-level information on all children in foster care in the US
- It provides a single entry per child-year
- It includes an anonymized unique identifier for each child that is (generally) valid within states over time
- It includes data on the state with custody of the child, the county of the agency that has responsibility for the case, and many other case characteristics

# THE MULTILEVEL STRUCTURE OF AFCARS

- Case-years are nested within children
- Children are nested within counties
- Counties are nested within states
- States are subject to national trends and policy (time)

**Ignoring this structure can produce misleading inferences!**

# ESTIMATE A SIMPLE MULTILEVEL MODEL

# THE BASIC MULTILEVEL MODEL

- We can describe the simple regression model for individual *i*:

$$\mathrm{y}_i \sim \mathrm{Normal}(\mu_i, \sigma^2)$$
$$\mu = \beta_0 + \beta_1 x_i$$

- The basic multilevel model extends this, by adding a group-level intercept for group *j*

$$\mathrm{y}_{ij} \sim \mathrm{Normal}(\mu_{ij}, \sigma^2)$$
$$\mu = \beta_0 + \beta_1 x_i + \delta_j$$
$$\delta \sim \mathrm{Normal}(0, \sigma_\delta^2)$$

# ESTIMATING A MULTILEVEL MODEL USING AFCARS

- This demonstration uses R , the lme4 package, and the 2019 AFCARS child file.
  - In Stata, you can use meglm, or PROC MIXED in SAS

# GETTING SET UP

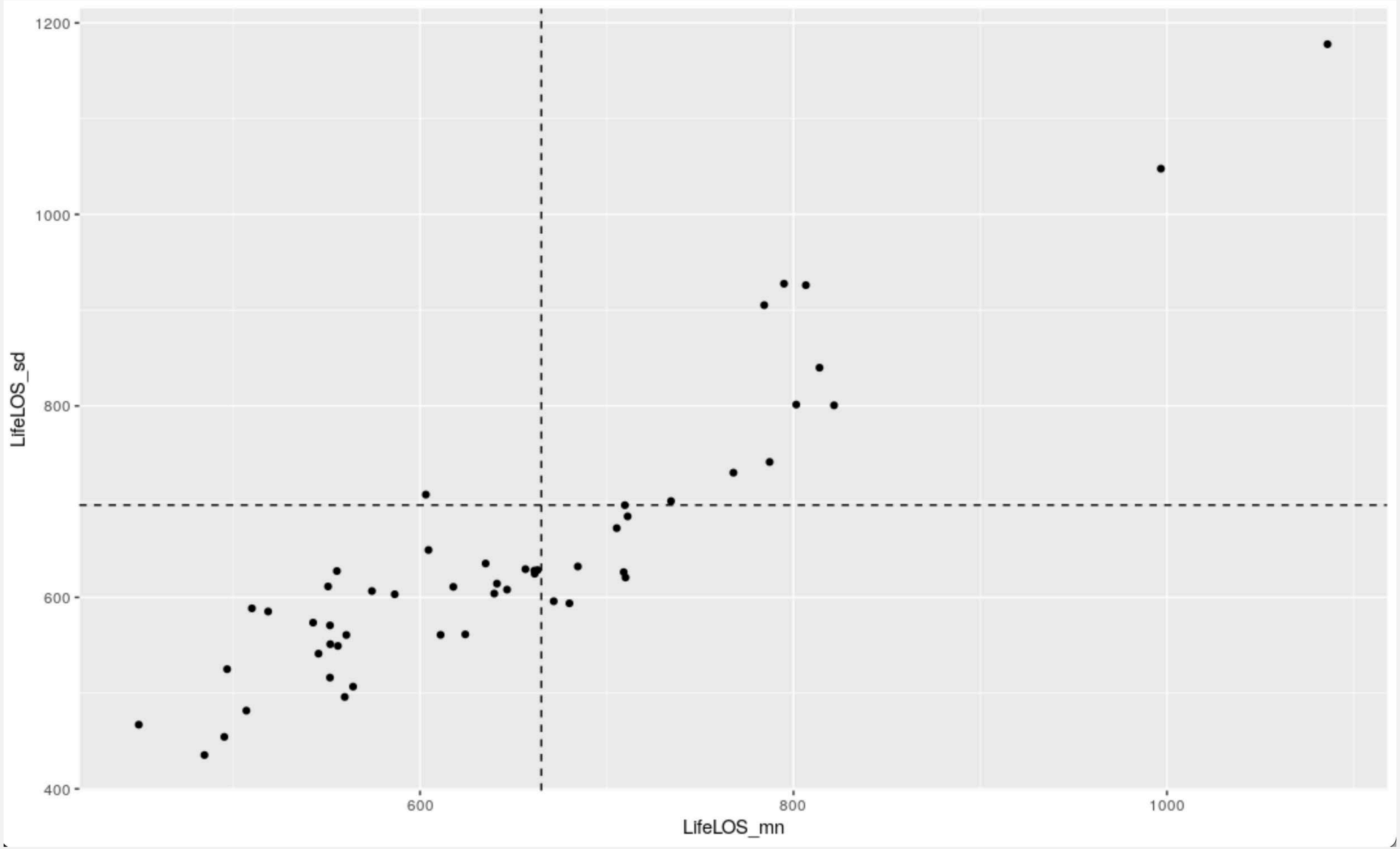
```
#### install the lme4 package if needed
#install.packages("lme4")
### load the package
library(lme4)

### read in afcars 2019 FC child file
afcars19<-read.delim("./afcars/FC2019v1.tab")
```

# IS THERE A MULTILEVEL STRUCTURE HERE?

```
> afcars19 %>% sample_n(5)
  STATE SEX AgeAtStart LifeLOS
1    55   2          3    1212
2    45   2          8      10
3    39   1          7     519
4    53   2          1     210
5    26   1          8    1452
>
```

# STATE LEVEL MEANS AND STANDARD DEVIATIONS (LIFELOS)



# ESTIMATING A NAIVE MODEL

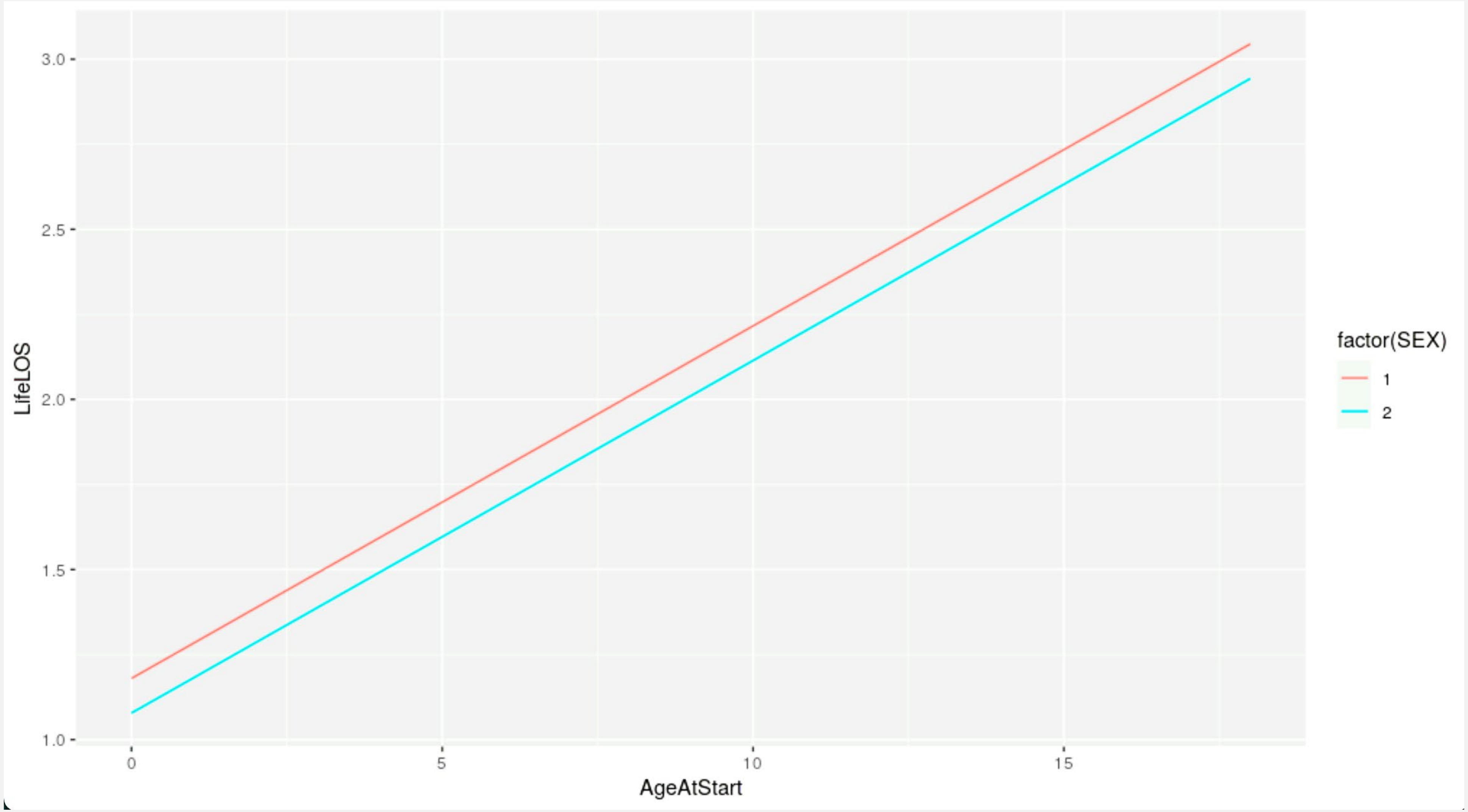
- First we'll estimate a linear model for lifetime length of stay (in years) as a function of child age and child sex

- $\widehat{\mathrm{LifeLOS}}_{\mathrm{i}} = \beta_0 + \beta_1 \times \mathrm{Age} + \beta_2 \mathrm{Sex}$

```
> m1<-lm(LifeLOS ~ AgeAtStart +
+            factor(SEX),
+          data = afcars19)
>
> tidy(m1)
# A tibble: 3 x 5
  term          estimate std.error statistic   p.value
  <chr>            <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept)     1.08    0.00414      262.  0
2 AgeAtStart      0.112   0.000393     285.  0
3 factor(SEX)2   -0.0974  0.00448      -21.8 7.27e-105
```

# INTERPRETING THE MODEL

- Let's estimate the expected lifetime length of stay for boys and girls at all ages



# ESTIMATING A MULTILEVEL MODEL TO ACCOUNT FOR STATE VARIATION

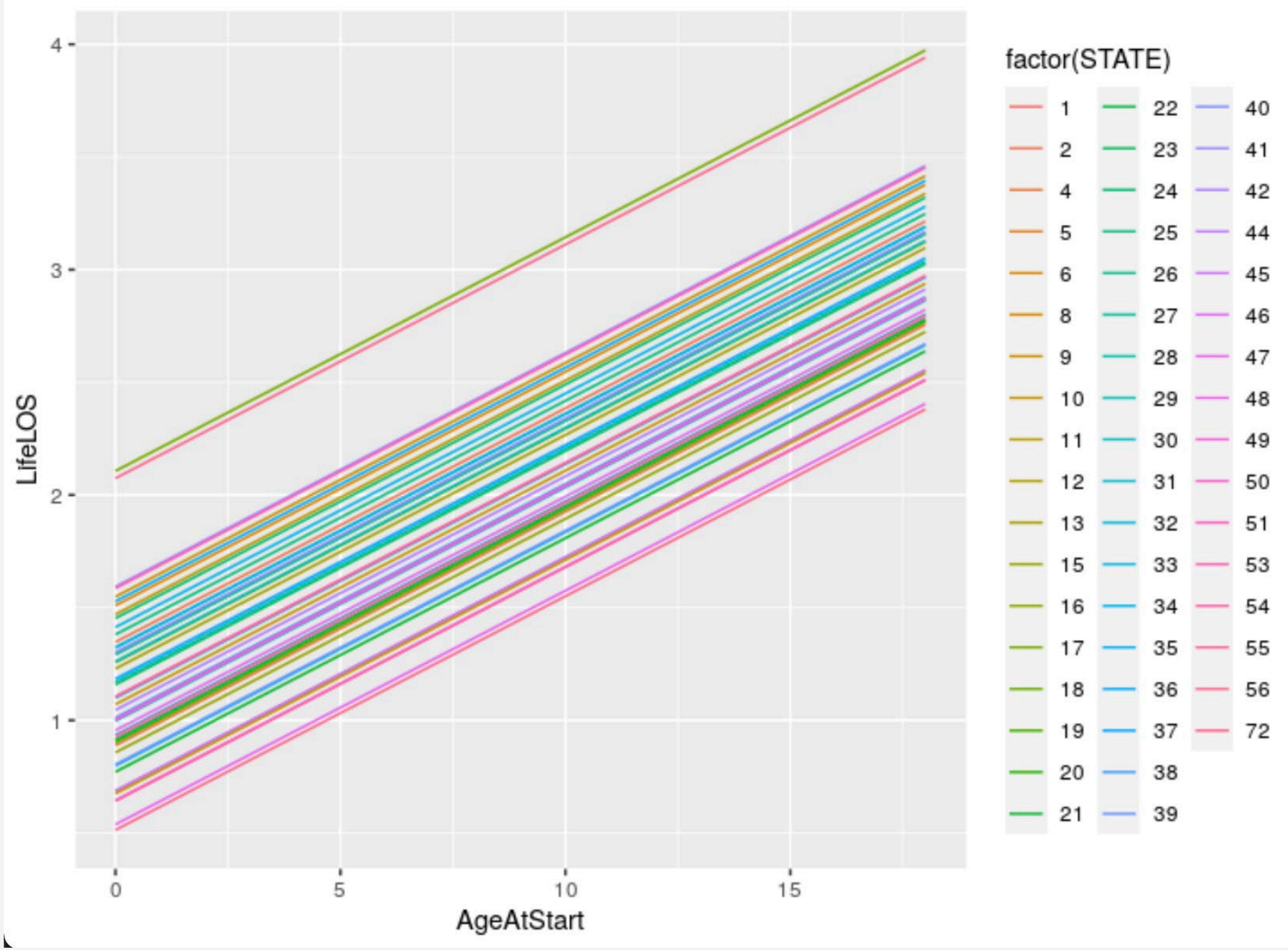
- Next, we'll add a *random effect* for each state. This estimates a separate intercept for each state.
- $\widehat{\text{LifeLOS}}_{\text{is}} = \beta_0 + \beta_1 \times Age + \sigma_s$

```
### estimating a multilevel model for states
m1_ml<-lmer(LifeLOS ~ AgeAtStart +
                factor(SEX) +
              (1|STATE),
              data = afcars19)
```

```
Random effects:
 Groups   Name        Variance Std.Dev.
 STATE    (Intercept) 0.1191   0.345
 Residual             3.2819   1.812
Number of obs: 610595, groups:  STATE, 52

Fixed effects:
              Estimate Std. Error t value
(Intercept)   1.1194244  0.0481413   23.25
AgeAtStart    0.1037294  0.0004245  244.34
factor(SEX)2 -0.1090684  0.0046410  -23.50
```

# INTERPRETING THE MODEL



# NEXT STEPS

- Additional possible levels for multiple years of data: individual-level, year, county
- Here we've estimated intercepts, but we can also estimate random slopes
- Random slopes and intercepts allow us to flexibly fit models with varying underlying relationships between variables
- These models are very effective ways to model differences in units across places and over time.

# QUESTIONS?

FRANK EDWARDS
ASSISTANT PROFESSOR, RUTGERS
FRANK.EDWARDS@RUTGERS.EDU

SARAH SERNAKER
STATISTICIAN, DUKE
SARAH.SERNAKER@DUKE.EDU

# NEXT WEEK…

**Date:**
**August 11th**

**Presenter:**
**Sarah Sernaker**

**Topic:**
**Latent Class Analysis**