S-043/Stat-151
Analysis for Clustered and Longitudinal Data
(Multilevel & Longitudinal Models)

Lecture 1.2
Clustered data as many small worlds

Instructor: Prof. Luke W. Miratrix

lmiratrix@g.harvard.edu

Larsen 603

Course Details
(including section times, office hours)

See “Shopping Slides” on Canvas site
https://canvas.harvard.edu/courses/67934
for details and overview.

Also,
YOU MUST
READ THE SYLLABUS

Problem Set 0:
Intro to R & Regression Refresher

Problem Set 0 has been posted on Canvas.
It is short.

It is due to Friday (midnight).

TFs & Instructor available up to 5pm

Weekly Sections

Wednesday 2pm-3pm

Thursday 12:30pm-1:30pm

All in Larsen G06

All followed by 1/2 hour of R help

This week:

All of it is R support

Come and just do homework 0 with a TF around to give support.

Miratrix Office Hours

Luke Miratrix's hours

Tuesday 10-11am

and 1/2 hr directly after all classes

No need for appointment; just drop in.

If private conversation needed, please email

Some "encouragement" to go to these:

If you don't come to office hours at least once this semester you will not get your full "participation grade."

No big questions needed. See syllabus.

Poll Everywhere

Our poll everywhere website is:

www.pollev.com/yayMLM

How to figure out what to read

The Canvas site has the following document:

Class by Class Reading List and Overview


Onwards to
some real
content!

Todays Goals

★ Introduce the concept of clustered data being a collection of tiny datasets that can each be analyzed for local trends.

★ Talk about how “fixed effect regression” is one possible road for dealing with such data.

★ Introduce the types of substantive questions one might ask of multilevel data.

★ Show how to aggregate data in R

★ Start talking about how individual trends can be tied together (pooled) via “random effects”

Planting some seeds of thought: fixed effects regression

The High School and Beyond Data Set

- ★ Nationally representative sample of US public and Catholic high schools from 1982
- ★ We have: 160 schools (90 public, 70 Catholic) with 7,185 students
- ★ We are interested in math achievement and SES.
- ★ See Raudenbush & Bryk text

The HSIS data (students, our special set of 10 schools)

```
> head(dat_ten) # shows the first six observations
       id minority female    ses mathach
1098 2526        0      1 -0.528  19.112
1099 2526        0      1  0.522  16.900
1100 2526        0      1  0.692  19.698
1101 2526        0      1  0.412  17.776
1102 2526        0      1  0.852  20.733
1103 2526        0      1  0.752  10.941

> tail(dat_ten, n = 3)
       id minority female    ses mathach
7124 9550        0      1  1.212  19.599
7125 9550        0      1 -0.198  15.995
7126 9550        0      1 -0.858  10.249
```

See Lecture 1.1 for how we took our subset of 10 schools.

A simple research question (RQ)

RQ1: What is the association between math achievement and SES?

A model for achievement on SES with "school fixed effects"

$$MATHACH_i \sim \beta_0 + \beta_1 SES_i + \beta_2 I_{ID_i=2} + \ldots + \beta_k I_{ID_i=k} + \epsilon_i$$

The $I_{ID=k}$ terms are indicator variables (dummy variables) for whether student *i* is in school *k*

Indicators ideally adjust for school-level differences. These are called **fixed effects**.

For students in school 1, the model is

$$MATHACH_i \sim \beta_0 + \beta_1 SES_i + \epsilon_i$$

For students in school 2, the model is

$$MATHACH_i \sim \left(\beta_0 + \beta_2\right) + \beta_1 SES_i + \epsilon_i$$

What does this model mean?

$$MATHACH_i \sim \beta_0 + \beta_1 SES_i + \beta_2 I_{ID_i=2} + \ldots + \beta_k I_{ID_i=k} + \epsilon_i$$

We're assuming that SES has the same relationship with achievement across schools, but in some schools students are systematically higher/lower achieving than in others.

The β_k, $k > 1$, capture the differences between the schools (in relation to the de-facto baseline school 1).

OLS in R: Easy

```
> M1 = lm( mathach ~ ses + id, data=dat.ten )
> summary( M1 )

Call:
lm(formula = mathach ~ ses + id, data = dat.ten)

Residuals:
    Min      1Q  Median      3Q     Max
-17.191  -5.024   0.236   5.385  15.981

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   13.352      1.297   10.30  < 2e-16 ***
ses            1.305      0.456    2.86   0.0044 **
id2917        -4.331      1.682   -2.57   0.0104 *
id3610         1.846      1.528    1.21   0.2276
id4292         0.147      1.549    0.09   0.9245
id4420         0.815      1.736    0.47
id6443        -3.416      1.767   -1.93
id6484        -0.198      1.702   -0.12
id8854        -8.124      1.775   -4.58
id9225         0.985      1.687    0.58
id9359         1.456      1.575    0.92
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '

Residual standard error: 6.48 on 404 degrees of freedom
Multiple R-squared:  0.223, Adjusted R-squared:  0.204
F-statistic: 11.6 on 10 and 404 DF,  p-value: <2e-16
```

To get this, we had to tell R to treat school ID as a *factor variable*. That's why it create indicator variables for id, but treated SES as numeric.

R automatically makes indicator variables for you. This is AWESOME.

TRUE means record what our final covariates are.

```
> M1 = lm( mathach ~ ses + id, data=dat.ten, x=TRUE )
```

See all the dummy variables we automatically made?

This is slicing: we are grabbing rows out of our model matrix.

```
> M1$x[ c(1, 51, 101, 151, 201 ), ]
```

	(Intercept)	ses	id2917	id3610	id4292	id4420	id6443
48	1	-0.788	0	0	0	0	0
1621	1	-0.578	1	0	0	0	0
2299	1	0.132	0	1	0	0	0
2986	1	-1.208	0	0	1	0	0
3188	1	-0.558	0	0	0	1	0

	id6484	id8854	id9225	id9359
48	0	0	0	0
1621	0	0	0	0
2299	0	0	0	0
2986	0	0	0	0
3188	0	0	0	0

This is a "Design Matrix" or "Model Matrix" - the covariates of OLS.

Once you fit a model, you can ask it questions

Here we ask "what are your estimated coefficients?"

```
> coef( M1 )
(Intercept)         ses      id2917      id3610      id4292
    13.3521      1.3053     -4.3311      1.8461      0.1469
     id4420      id6443      id6484      id8854      id9225
     0.8148     -3.4158     -0.1980     -8.1245      0.9845
     id9359
     1.4564
```

We can also remove our intercept to escape having a baseline

```
> M3 = lm( mathach ~ ses + id - 1, data=dat.ten)

> coef( M3 )
   ses id1288 id2917 id3610 id4292 id4420 id6443 id6484 id8854
 1.305 13.352  9.021 15.198 13.499 14.167  9.936 13.154  5.228
id9225 id9359
14.337 14.808
```

This is a different way of saying "no intercept"

```
> M3 = lm( mathach ~ 0 + ses + id, data=dat.ten)
```

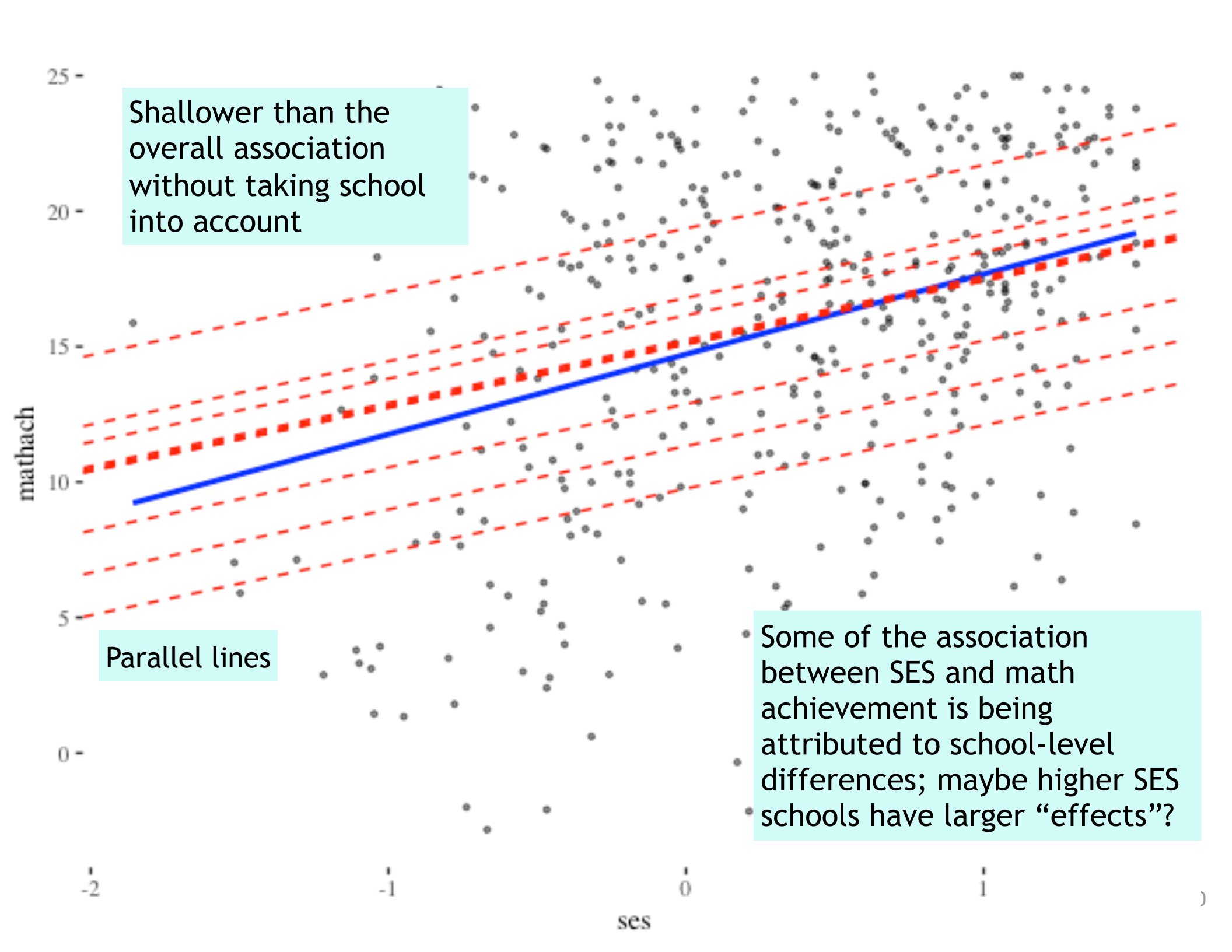
Plotting our fit lines

```
# Fit a no intercept, fixed effects model
M1 = lm( mathach ~ 0 + ses + id, data=dat.ten )


# make a table of our best fit lines
lines = data.frame( inter = coef(M1)[2:11],
                    slope = coef(M1)[[1]] )

# plot our data and our 10 best fit lines on top of it
ggplot( dat.ten, aes( ses, mathach ) ) +
  geom_point( size=0.75, alpha=0.5 ) +
  geom_smooth( method="lm", se=FALSE, col="blue" ) +
  geom_abline( data=lines, aes( slope=slope,
                                intercept=inter ),
               col="red", lty=2 )
```

This code will make an overall best fit line ignoring the school information. It will then add in the individual school lines from our model (which made all the slopes the same).


Shallower than the overall association without taking school into account
Parallel lines
Some of the association between SES and math achievement is being attributed to school-level differences; maybe higher SES schools have larger “effects”?
mathach
ses
25
20
15
10
5
0
-2
-1
0
1

What have we learned so far?

We just saw that regression in R is relatively straightforward.

We also saw that the “fixed intercept” model is saying that our 10 schools have different intercepts but have the same slope.

We have *not pooled* the intercept but have *completely pooled* the slope across our ten schools.

The plot shows us an estimate of the relationship of math achievement and SES for each of our ten schools.

Something to be upset about

Our model says the relationship between SES and math achievement in each school is **the same**

We might not think that is the case

Key concept:

Models impose structure

If the structure is not true, then our estimates can be difficult to interpret… or wrong!

Completely un-pooled: Fitting interactions is easy

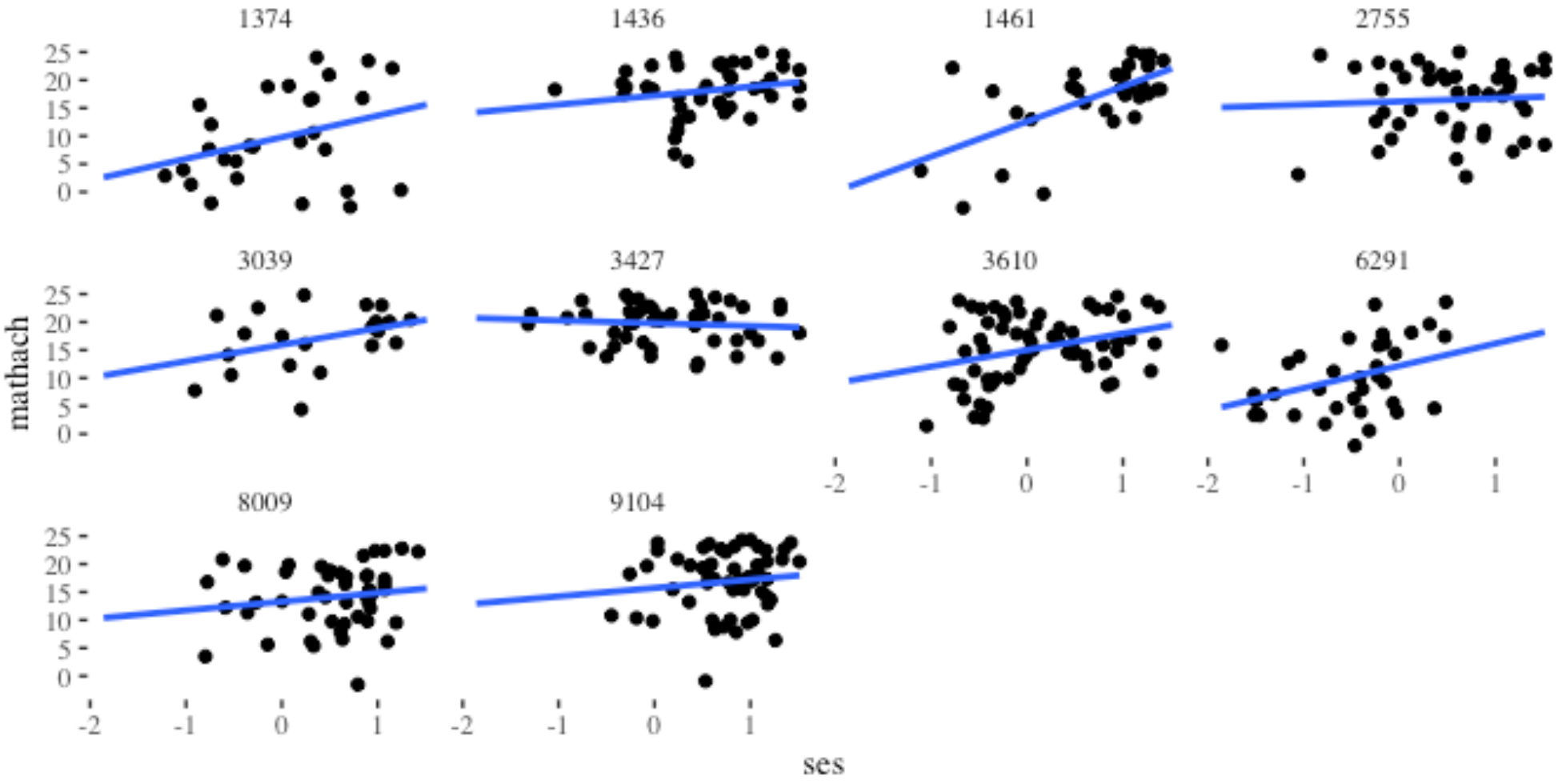
```
> M4 = lm( mathach ~ ses * id - 1, data=dat.ten )
> coef( M4 )
       ses     id1288     id2917     id3610     id4292
    3.2554    13.1149     8.8856    14.9999    12.7863
    id4420     id6443     id6484     id8854     id9225
   14.5376     9.2131    13.0245     5.7070    13.9361
    id9359 ses:id2917 ses:id3610 ses:id4292 ses:id4420
   15.5657    -2.1196    -0.2996    -3.4161    -0.2968
ses:id6443 ses:id6484 ses:id8854 ses:id9225 ses:id9359
   -3.9988    -2.6498    -1.3166    -0.3696    -4.0889
```

What did we do here? What does this model mean?

Why does everything have to be so easy?

10 worlds, completely separate

```
ggplot( dat.ten, aes( ses, mathach, group=id ) ) +
  facet_wrap( ~ id ) +
  geom_point() +
  geom_smooth( method="lm", se=FALSE, fullrange=TRUE )
```



Some conceptual questions

★ What do the intercepts of any of the lines mean?

★ What differences, if any, are there between running a new linear model on each school vs. running the interacted model on the set of 10 schools?

★ Do we trust the blue lines on the plot? Why or why not?

★ What about the variability in the slopes and intercepts of the blue lines?

Two Research Questions to Compare

1. What is the association between student SES and student math achievement?

   (This is about how a level-1 variable is associated with another level-1 variable.)

2. What is the association between school mean SES and school mean math achievement?

   (This is about how a level-2 variable is associated with another level-2 variable.)

Aggregating data to look at how schools vary

In the following we look at several different types of research question, and how we might try to answer them using simple approaches.

A simple idea:
Make a school-level data set

Collapse your individuals into groups: calculate values of interest for each school, and create a new dataset in which the *schools* are the observations.

Pros:

- Easy methods of OLS to analyze
- Easy to communicate
- Flexible

Cons:

- Doesn't take different precision of groups into account.
- Disallows some cross-level questions
- Individuals' trends get lost: the question has changed.

Group aggregation: Making 2nd level variables from 1st

```
> col.dat = dat %>% group_by( id ) %>%
+     summarize( per.fem = mean(female),
+                per.min = mean(minority),
+                mean.ses = mean(ses),
+                mean.ach = mean(mathach),
+                n.stud = n() )
> head(col.dat)
 id    per.fem per.min mean.ses mean.ach n.stud
1 1224    0.596  0.0851  -0.434      9.72     47
2 1288    0.44   0.12     0.122    13.5       25
3 1296    0.646  0.979   -0.426      7.64     48
```

We have generated new “level 2” variables. Specifically, the percent of students who are female, the percent who are non-white, mean student SES, and school size.

But these are not perfect due to, essentially, *measurement error* or *sampling error.*

What else could we aggregate at the school level?

Almost anything!

Anything that is a property of a school could count as a school characteristic

E.g. indices of racial diversity, indices of socio-economic diversity, geographical measures, standard deviation of achievement scores, and whatever else you can think of!

Connecting your new variables to other level-2 variables

```
> sdat = merge( sdat, col.dat, by="id", all=TRUE )


> head( sdat )
    id size sector pracad disclim himinty meanses   per.fem     mean.ses n.stud
1 1224  842      0   0.35   1.597        0  -0.428 0.5957447  -0.43438298     47
2 1288 1855      0   0.27   0.174        0   0.128 0.4400000   0.12160000     25
3 1296 1719      0   0.32  -0.137        1  -0.420 0.6458333  -0.42550000     48
4 1308  716      1   0.96  -0.622        0   0.534 0.0000000   0.52800000     20
```

Some of these variables were pre-existing

And some we just created

Aside: Notice anything that seems off?

RQs on How US high schools vary in mean math achievement

RQ 2: What is the average school mean achievement?

We can estimate this using our school-level data. How?

RQ 3: How much ***variation*** is there in school mean achievement?

We cannot quite estimate this. Why not?

see math derivation posted online.

! Problem: uncertainty in individual school means will inflate variability of the school means

Simple summary statistics tell an imperfect story

```
> library( skimr )
> skim( sdat$mean.ach )
Skim summary statistics

── Variable type:numeric ──────────────────────────────────────────────────────
      variable missing complete   n  mean   sd   p0   p25
 sdat$mean.ach       0      160 160 12.62 3.12 4.24 10.47

> skim( dat$mathach )

Skim summary statistics

── Variable type:numeric ──────────────────────────────────────────────────────
     variable missing complete    n  mean   sd    p0  p25
 dat$mathach       0     7185 7185 12.75 6.88 -2.83 7.28
>
```

This number is suspect. Too large.

This number doesn't answer our question
It is across all students and schools

More RQs we might ask at the school level

RQ 4: Does the mean SES of a school predict mean math achievement of that school?

RQ 5: Is there is variation in mean math achievement *beyond what is explained* by mean SES?

Problem: uncertainty in individual school means make this second question difficult, and dilutes our relationship in the first.

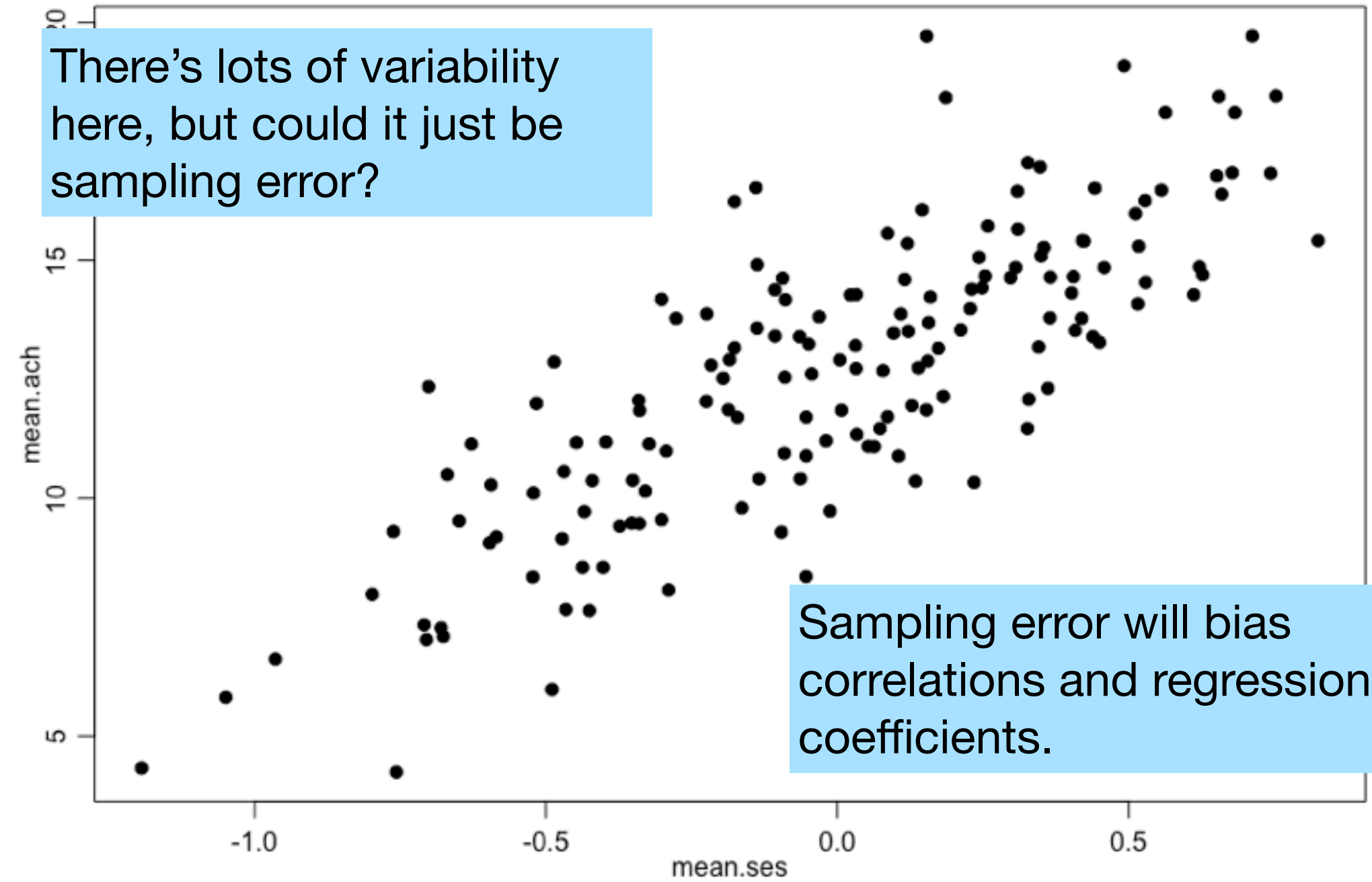
Analysis plan for RQ 4

1. For each school, estimate the mean math achievement and mean SES
2. Regress mean achievement on mean SES!

This could answer our first question.

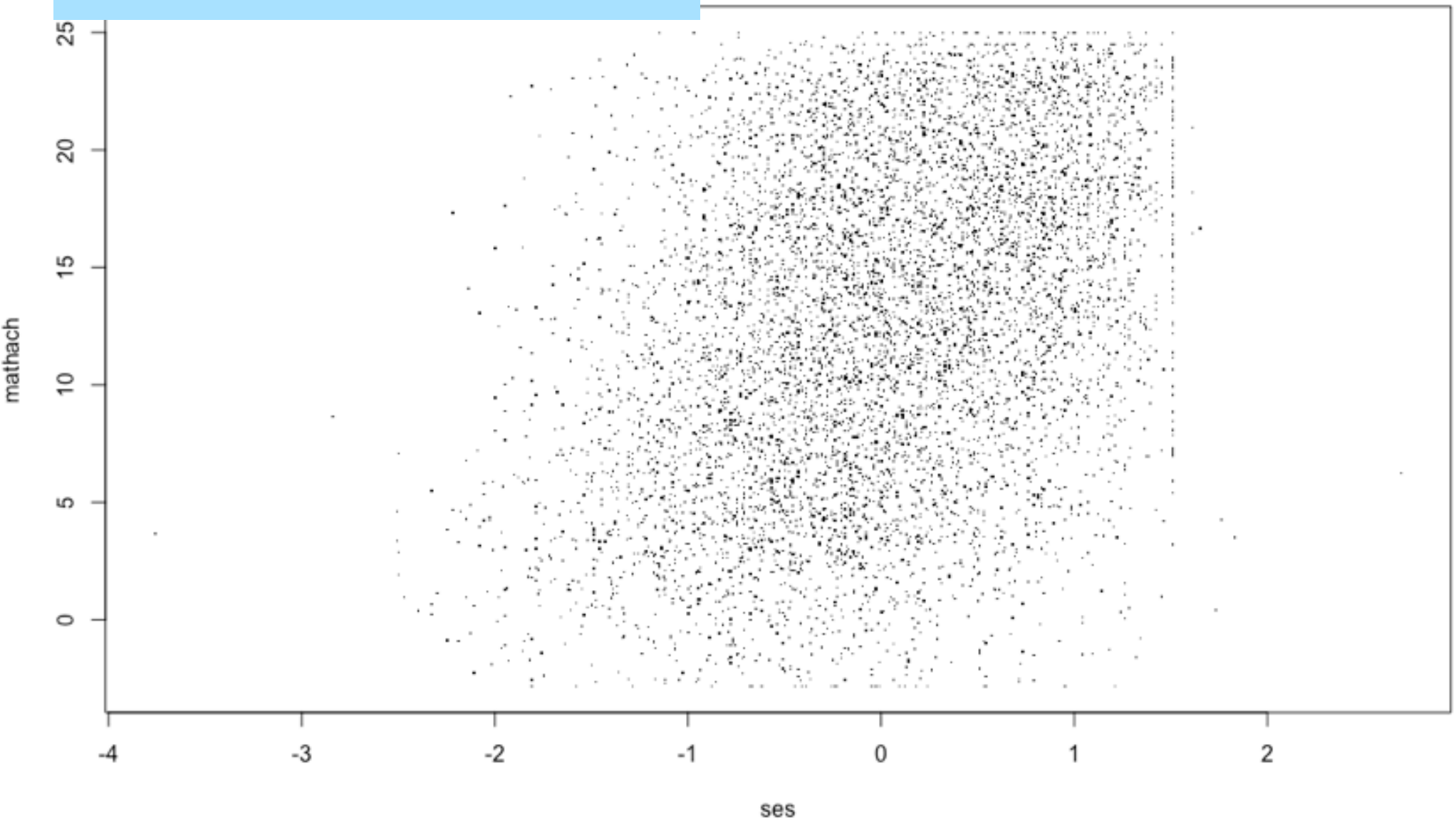
(Sort of)

Alas, this won't answer our second question!



General rule: these approaches will give smaller estimates than the truth, in general

The relationship is much less clear at the student level; maybe within schools there's no association?



Generalizing RQs 4 and 5: Why do schools vary?

RQ 6: Why do schools have different mean levels of achievement?
(or, which variables are associated with school mean achievement?)

RQ 7: Why do schools vary in their relationship between SES and achievement?
(or, which variables are associated with school associations between SES and achievement?)

Now we are looking to explain the individual school models with school-level variables.

Analysis plan (for RQ 6)

- ★ Regress our mean estimates on a bunch of school-level variables
- ★ Try to shoehorn our findings into some sort of theoretical framework
- ★ Publish
- ★ ???
- ★ Profit

A school-level regression

```
> ll.exp = lm( mean.ach ~ sector + size + pracad +
                    disclim + himinty, data=sdat )
> summary( ll.exp )

Residuals:
    Min      1Q  Median      3Q     Max
-6.087 -1.142   0.025   1.350   5.214

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.178672   0.572409   14.29  < 2e-16 ***
sector       0.523456   0.528555    0.99  0.32355
size         0.001015   0.000284    3.57  0.00047 ***
pracad       7.328571   0.868332    8.44  2.2e-14 ***
disclim     -0.318911   0.238575   -1.34  0.18328
himinty     -2.446769   0.361405   -6.77  2.5e-10 ***

Residual standard error: 1.99 on 154 degrees of freedom
Multiple R-squared:  0.606,  Adjusted R-squared:  0.594
F-statistic: 47.4 on 5 and 154 DF,  p-value: <2e-16
```

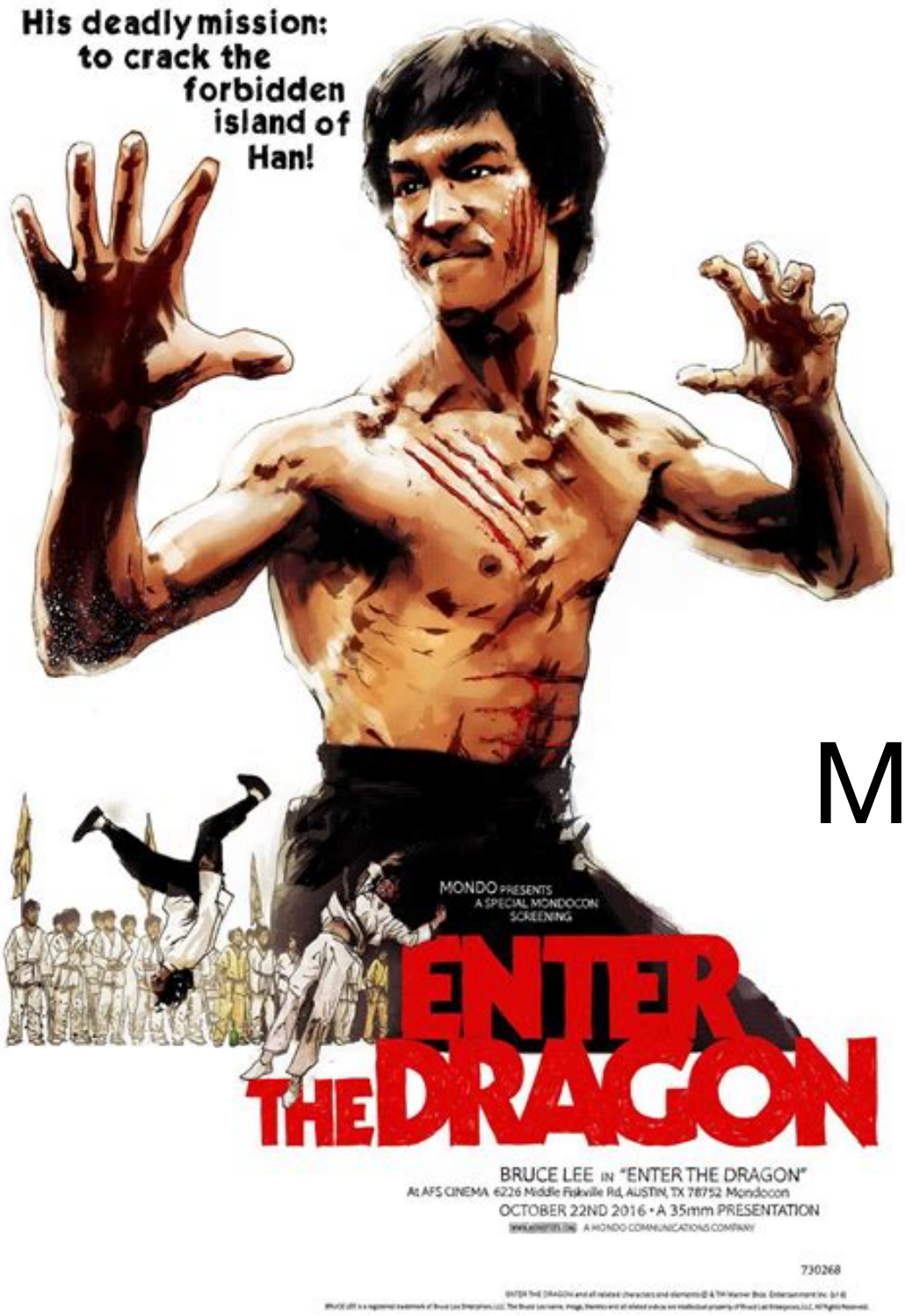
P-values and SEs are suspect!! (Why?)

This is unsatisfying

We need a better way to analyze multi-level data. This aggregation is cool, but it

1. Is underpowered (school level sampling error gives us problems)
2. Is biased (sampling error in the predictors introduces bias)
3. Makes implausible assumptions (homoscedasticity is almost never going to be correct)

Admittidly, we can fix 3 with econometrics (robust SEs)



Enter the Multilevel Model

Multilevel Models are really lots of little models

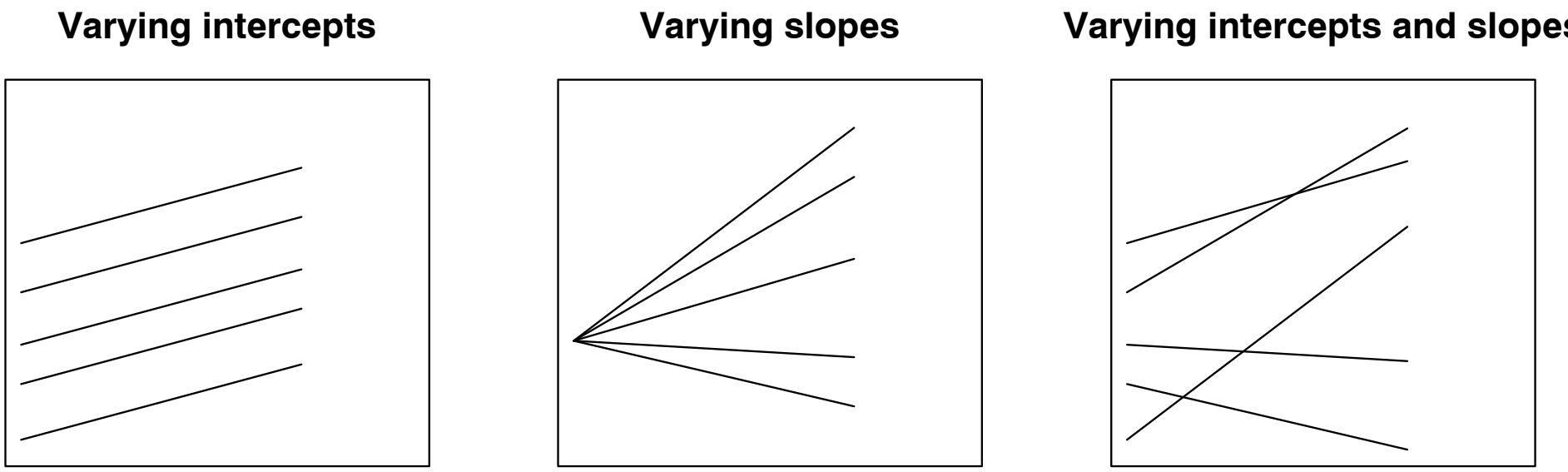


Figure 11.1 *Linear regression models with (a) varying intercepts ($y = \alpha_j + \beta x$), (b) varying slopes ($y = \alpha + \beta_j x$), and (c) both ($y = \alpha_j + \beta_j x$). The varying intercepts correspond to group indicators as regression predictors, and the varying slopes represent interactions between x and the group indicators.*

We use them to answer all the different kinds of questions we were thinking about, and more!

- ★ How does SES correlate with math achievement?
- ★ How does a school's mean SES relate to mean math achievement?
- ★ How do US high schools vary in mean math achievement?
- ★ Is the association between SES and math different across schools?
- ★ How do public vs. Catholic schools compare on math achievement or math achievement vs. SES, *controlling for mean SES?*

What are the levels of these questions?

For HS&B: Many schools, many models

Each school *j* has *its own model*:

$$Math_i = \alpha_{j[i]} + \beta_{j[i]} SES_i + \epsilon_i$$

$$\epsilon_i \sim N(0, \sigma_y^2)$$

Notes:

- The j[i] notation means the school that student *i* is attending.
- Each school has an intercept and slope. But how do we connect the schools?
  (Earlier in the lecture each school was on its own.)

What just happened to me?

In summation

Multilevel data are everywhere, even if we don't always realize it.

Fixed effects approaches can somewhat alleviate problems when focused on individual data.

Recognizing data as multilevel permits us to ask new and interesting questions.

There are hacky ways of dealing with these data, but the hacks don't work as well as we might like (they conflate sampling error with true variance)

The multilevel model may help us and perhaps that's why there is a course on them.

Appendix: Other examples of multilevel data

These slides just give a few more examples of multilevel data to give a sense of where we are going.

Survey responses (similar to factor analysis/SEM)

Survey respondents are presented with a scenario (two people meet at a party and go home together).

They are asked how likely it is that various things will happen (e.g., “Nothing physical will happen”, “They will make out”, “They take each other’s clothes off”, etc….)

(I am stopping just before the good/bad stuff)

From research on behavior expectations (ask Joe McIntyre)

How can we use the responses?

Option 1: form a composite score by aggregating the responses within survey respondent (simple mean or PCA)

This is the idea we just discussed where survey responses are nested within individuals

Option 2: Treat the individual responses as observations nested within a respondent

(This is what factor analysis/SEM does)

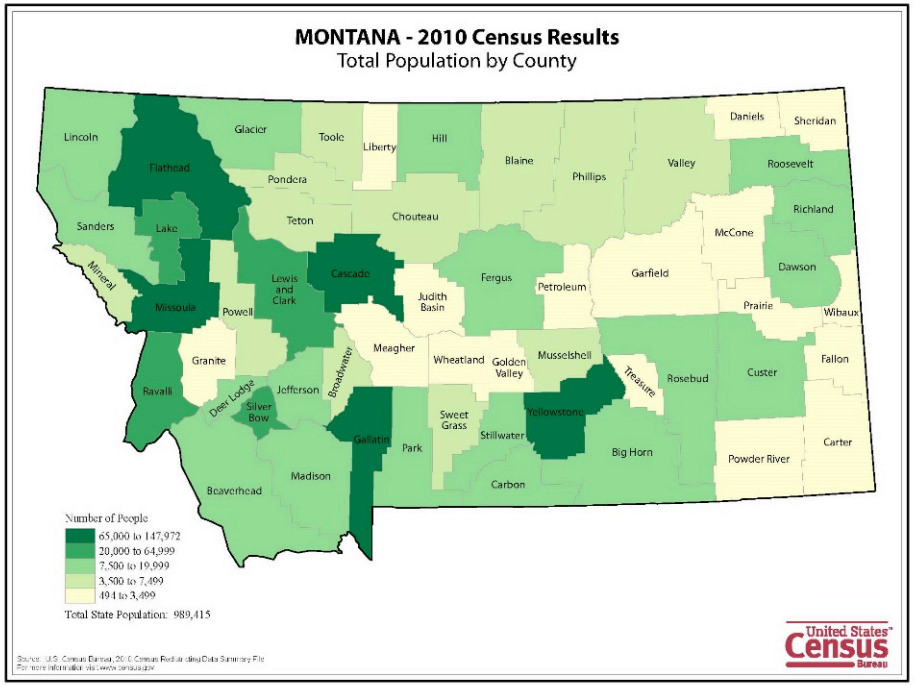
Small area analysis (important for spatial statistics, and for the Census)

Want to measure certain characteristics of an area (say presence of toxins). This can help us predict the risk of toxins in different areas

Take a large sample of houses and measure toxins

But some areas will have very, very few houses.

(See Radon in G&H reading.)



How can we use the responses?

Option 1: form a composite score by aggregating the values up to some larger area, say county or city level

Here estimates will be very imprecise for areas with few houses sampled

Option 2: Treat the individual homes as observations nested within an area

(Spatial statistics do something a little more complex, but ultimately similar; we'll see how this can improve our estimates for small areas)

Predicting batter success

Teams want to know how good their batters are.

Can base this on the proportion of times they bat in which they get a hit (batting average).

But batters who only get a couple of chances will tend to have extreme batting

How can we use the responses?

Option 1: form a composite score by averaging chances to the batter level

Here estimates will be very imprecise for batters who almost never play

Option 2: Treat the individual chances as observations nested within batters, and shrink averages calculated from a few hits in towards the mean.

(This is related to Bayes theorem)

Lung capacity measurement devices

We have patients breathing into a device two times to measure their lung capacity:

$$Y_{it} = \beta_i + \epsilon_{it} \text{ with } t = 1, 2$$

$$\epsilon_{it} \sim N(0, \sigma_y^2)$$

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

What are the two variances and what do they mean?

Chapter 2 of RH&S text

Australian Adolescents and their smoking

★ 2000 kids watched for 3 years

★ *Repeated Measures*

★ Track kids across time. Ascertain whether they smoke or not at each time point.

See page 241 (chap 11) in G&H
Warning: R code in book somewhat nonsensical to me. Ignore.

Format #1

person ID	sex	parents smoke? mom	dad	wave 1 age	smokes?	wave 2 age	smokes?	⋯
1	f	Y	Y	15:0	N	15:6	N	⋯
2	f	N	N	14:7	N	15:1	N	⋯
3	m	Y	N	15:1	N	15:7	Y	⋯
4	f	N	N	15:3	N	15:9	N	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱

age	smokes?	person ID	wave
15:0	N	1	1
14.7	N	2	1
15:1	N	3	1
15:3	N	4	1
⋮	⋮	⋮	⋮
15:6	N	1	2
15:1	N	2	2
15:7	Y	3	2
15:9	N	4	2

person ID	sex	parents smoke? mom	dad
1	f	Y	Y
2	f	N	N
3	m	Y	N
4	f	N	N
⋮	⋮	⋮	⋮

Format #2

Two Models for Same Thing

★ Model 1 (Individual *j* at time *t)*:

$$\Pr(y_{jt}=1) = \text{logit}^{-1}(\beta_0 + \beta_1 \text{psmoke}_j + \beta_2 \text{female}_j + \\ + \beta_3(1-\text{female}_j)\cdot t + \beta_4 \text{female}_j \cdot t + \alpha_j)$$

Individual effect

★ Model 2 (Observation *i* for *j[i]* at time *t[i])*:

$$\Pr(y_i=1) = \text{logit}^{-1}(\beta_0 + \beta_1 \text{psmoke}_{j[i]} + \beta_2 \text{female}_{j[i]} + \\ + \beta_3(1-\text{female}_{j[i]})\cdot t[i] + \beta_4 \text{female}_{j[i]} \cdot t[i] + \alpha_{j[i]})$$

Being able to shift from model to model is part of living in the real world. We will practice.

"Fixed" vs. "Random" effects

Three different possible models

$$Y_i = \alpha_{j[i]} + \beta_{j[i]} \cdot age_i$$

$$Y_i = \alpha_{j[i]} + \beta \cdot age_i$$

$$Y_i = \alpha + \beta \cdot age_i$$

Putting a distribution on alpha is not the same as forcing it to be the same for all schools.

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

vs.

The alpha are coefficients for fixed dummy variables S_j

QUESTION: How many potential models do we have using stuff on this slide?

Non-nested, but still multilevel

- ★ Study of earnings with variables of state of residence and occupation
- ★ E.g., 1500 people in 50 states and 40 job categories

$$y_i = X_i\beta + \alpha_{j[i]} + \gamma_{k[i]} + \epsilon_i, \text{ for } i = 1, \ldots, n,$$

$$\alpha_j \sim \mathrm{N}(U_j a, \sigma_\alpha^2), \text{ for } j = 1, \ldots, 40,$$

$$\gamma_k \sim \mathrm{N}(V_k g, \sigma_\gamma^2) \text{ for } k = 1, \ldots, 50$$

See pg 244 in