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

A Brief Welcome and Unit 1, Lecture 1 What is clustered data? How can you wrangle data in R?

Instructor: Prof. Luke W. Miratrix

lmiratrix@g.harvard.edu

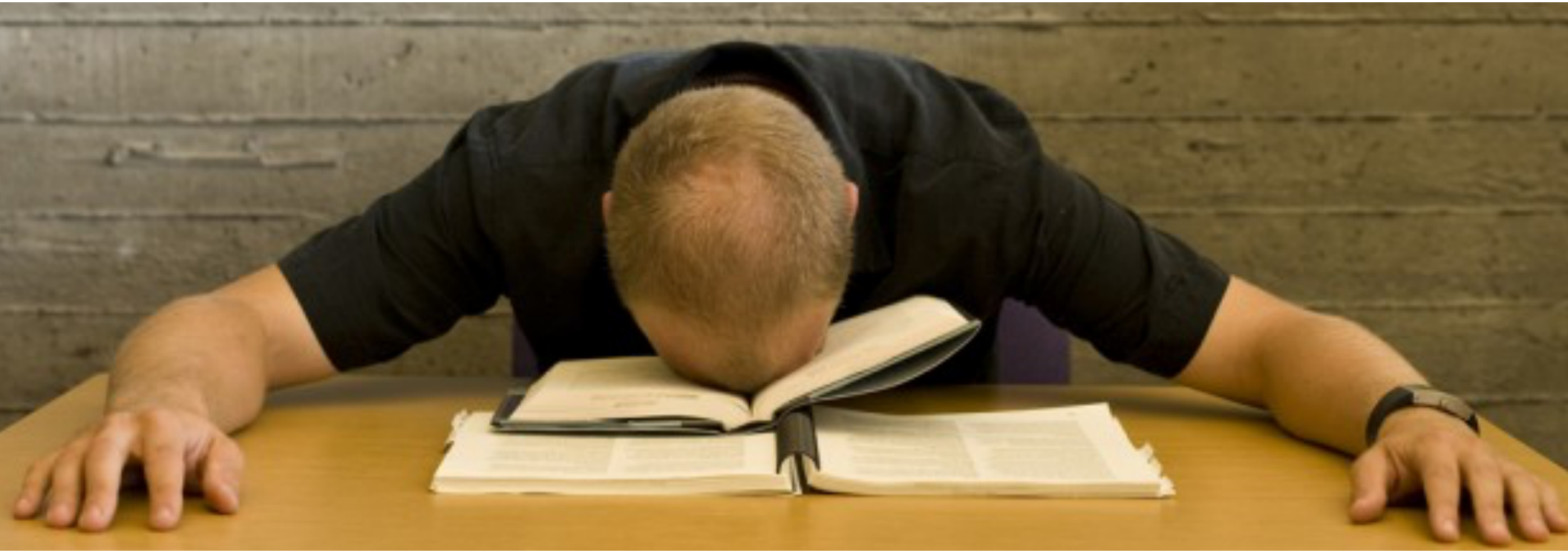
Larsen 603

Goals for today

- ★ Go through some of the course bureaucracy
- ★ Introduce a running example dataset and briefly discuss clustered data
- ★ Refresh your memory of multiple regression
- ★ Start to introduce R
- ★ Start to familiarize you with some potentially new mathematical notation

http://madan.org.il/sites/default/files/archive/8388548Boredom_TS_94788575_111913-617x416.jpg

Course Bureaucracy



Who is taking this course?

This course probably has a mix of the following:

- ★ PhD students in the School of Education
- ★ Masters students in statistics and education
- ★ PhD students from the School of Public Health, psychology, sociology, environmental science, ...
- ★ Masters of Public Health students,
- ★ Undergraduate statistics majors in FAS

Different folks, different goals, different skills ⇒ interesting opportunities

A brief FAQ

Is this course hard? Yes.

Is data analysis and open-ended work expected? Yes.

Can this be, at times, confusing and unsettling? Yes.

Is this on purpose? Yes.

Do we sometimes provide too much information and let you follow up on the parts that interest you? Yes.

Will you have the chance to do real work, real research? Yes.

Is this course primarily geared towards those actively engaged in doing research? Yes.

Are the other types of students in the class more prepared than I am? Yes and No.

Class Meetings, Attendance, Behavior

- ★ Lectures will be interactive. Questions are encouraged.
- ★ Lectures will be adapted to the needs of the students.
- ★ We will sometimes use polling systems to foster immediate feedback and discussion.
- ★ You are expected to come to every class meeting.
- ★ You are expected to behave professionally in lecture (no web surfing, phones are silent, no IM, etc) especially because it negatively impacts others.

Microagressions

Examples (modified quotations)

- ★ Hearing “those kids” in class in reference to poor children of color.
- ★ Being told it was unprofessional to share your disability/ invisible identity
- ★ Being called “our little lady” in a roomful of men.
- ★ Being asked “How is your English so good?”
- ★ Turning people’s names into nicknames without invitation when they are hard to pronounce.

How to increase awareness? Knowing we all do it (hopefully inadvertently) and raising awareness of it when it happens.

When I stumble, I invite you to email me or talk to me about it.

Resources of S-043

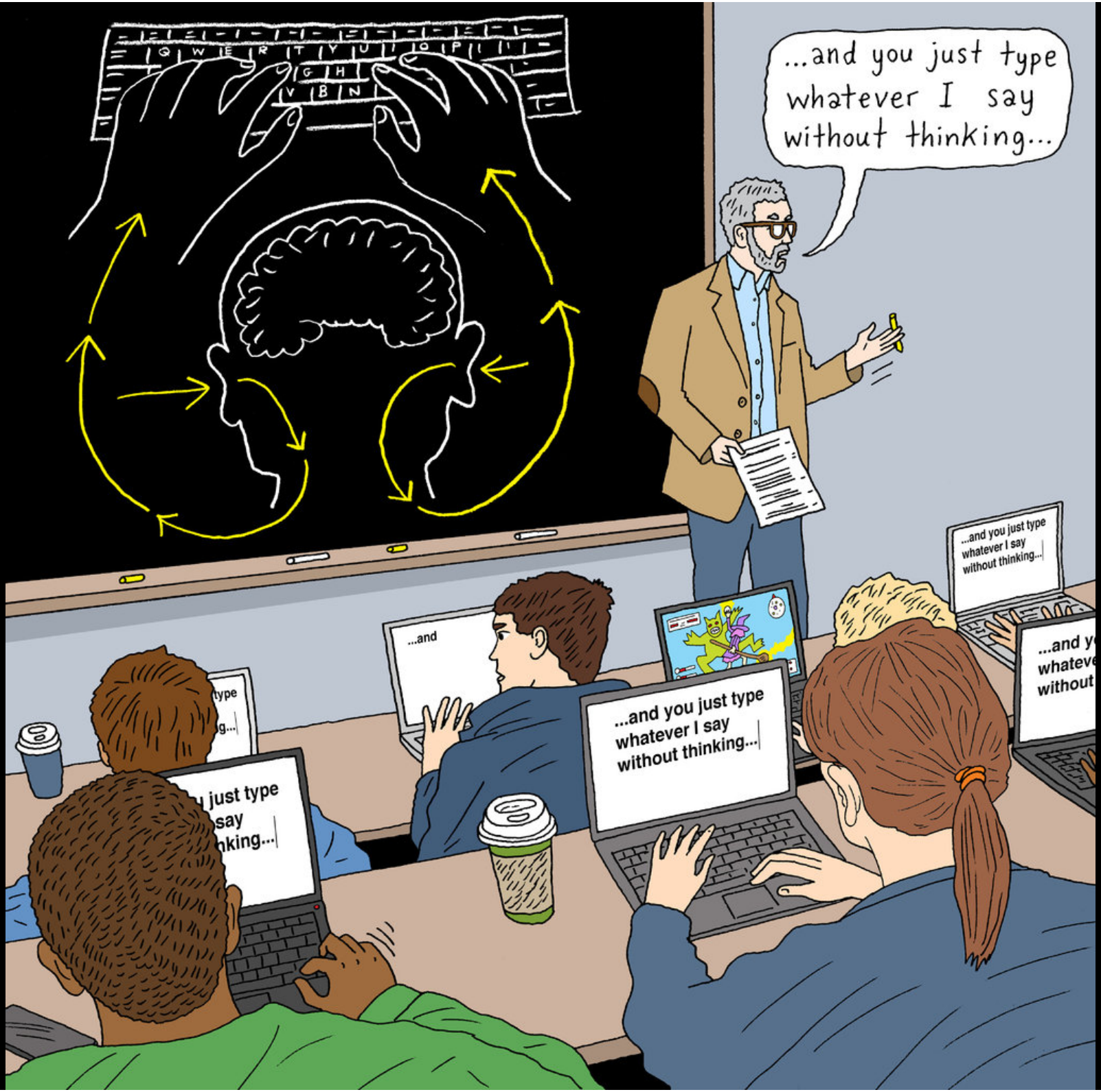
★ Canvas Site
- Springboard to everything else
- Announcements and updates
- Assignment due dates
- You submit work here
- Discussion forum for asking questions of TFs and classmates
- Warehouse of scanned reading materials
- Lecture slides, code, data

★ Quiet Questions
- For in-class silent participation and getting further support in lectures

★ RStudio and R
- The statistical software we use

★ Raudenbush and Bryk textbook
- The textbook for the course

★ Weekly Sections
- TFs will alternate. 2 different hour slots. Extra help, extra depth, R skills.


...and you just type whatever I say without thinking...
...and
...and you just type whatever I say without thinking...
...and you just type whatever I say without thinking...

Learning R

UHF (1989)

STATA vs. R

This class is taught in R

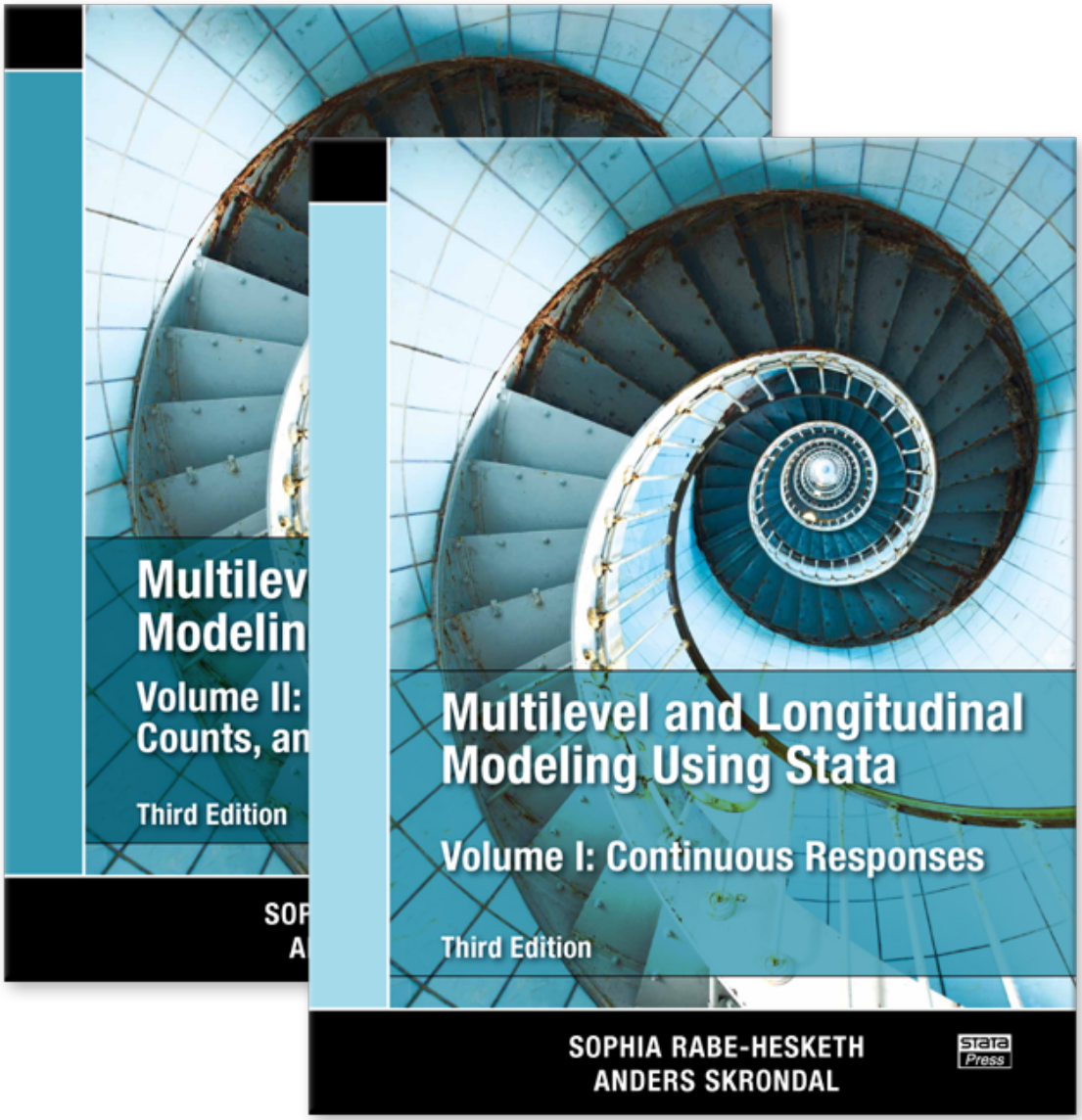
And yet, STATA is often used at HGSE.

These excellent books give example code for everything we do in STATA.

We will also provide some translation of STATA code provided in S-052 and S-040 to R.

I think R is better (see syllabus)

Statistics is not the tool used to analyze data.



Getting Started with R and RStudio on Your Computer

★ R is free and easy to install on your computer.

http://lib.stat.cmu.edu/R/CRAN/

★ Once you install R, download and install RStudio.

http://www.rstudio.com

We will help you with this during office hours and section if you get stuck. It's important to learn how to support yourself when using R, but it's also important to use the resources we provide!

Feel free to execute the R code which we demo in lecture on your own machine in real time. Extensively commented scripts will be made available.

A Focus on Communication and Critique

★ Communication and the ability to digest statistical argument are important skills we will focus on.

★ To tackle this we will try a few experimental things

★ For example, we may try some (low-stakes) peer review of parts of problem sets.

You are waving some rights, but not all

★ We reserve the right to try out different approaches to running the course and teaching you.

★ We will **preserve** and **encourage** your right to squawk about what is not working.

★ We commit to listening to such feedback and adjusting accordingly.

Sections, Office Hours, and Extra Work

★ Weekly sessions, focused on effectively using R and core conceptual concerns, each week.

★ Office hours will exist, and will be scheduled soon as the dust settles. (For now email to set up appointments.)

★ We will provide resources, such as pointers to the math behind the stats, for those interested.

Tools Table Add-ons Help See new changes

nal text Arial 11

Yes once we start standardizing variables, we'll look at point clouds centered at the origin. Mathematically, the actual center and spread of the point cloud is less important than its shape, at least for many of the statistics we use. Pragmatically, the exact location of the cloud is quite important.

4) Is the slope of the line 1 if r is 1? Or does r of 1 just mean it is a straight line, not necessarily a slope of 1?

Not exactly; mathematically, slope will be determined by the scale of your variables, whereas r is independent of this.

Only if X and Y have the same standard deviation (or, equivalently, variance). This will come up in unit 4, but notice that we can always change the slope of the line by changing the metric by which X or Y is measured, but this won't change the correlation.

(thanks!)

5) I'm a bit confused by what is meant by "scale." Does that mean plotting student/teacher ratio

Quiet questions (from Andrew Ho): a place to ask questions in class if you don't want to ask out loud. TFs will answer in real time, or alert the instructor if there are consistent questions.

Finding Course Bureaucracy Details

Please see posted lecture slides on the website.

https://canvas.harvard.edu/courses/67934

Also,

YOU MUST

READ THE SYLLABUS

Problem Set 0:
Intro to R & Regression Refresher

Problem Set 0 has been posted on the shared dropbox. It is short.

It is technically due next Wednesday, but due to shopping season the following Friday (5pm) is fine.

Intro to R Workshops

taught by

Eddie Kim

(stathelp@gse.harvard.edu)

Intro to R: 9/6 (Friday) 10:00-11:30am

Intro to R (repeat): 9/9 (Monday) 2:00-3:30pm

Plotting in R: 9/13 (Friday) 10:00-11:30am

Plotting in R (repeat): 9/16 (Monday) 2:00-3:30pm

All in Gutman 302

Please try to install R and RStudio beforehand.

Ok,
let’s get this course started!

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

Unit 1, Lecture 1

What is clustered data?
How can you wrangle data in R?

Instructor: Prof. Luke W. Miratrix

lmiratrix@g.harvard.edu

Larsen 603

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

★ What makes these data multilevel?

- We have a *sample of schools* with each school having a *sample of students* inside it.
- School is “Level 2”
- Student is “Level 1”

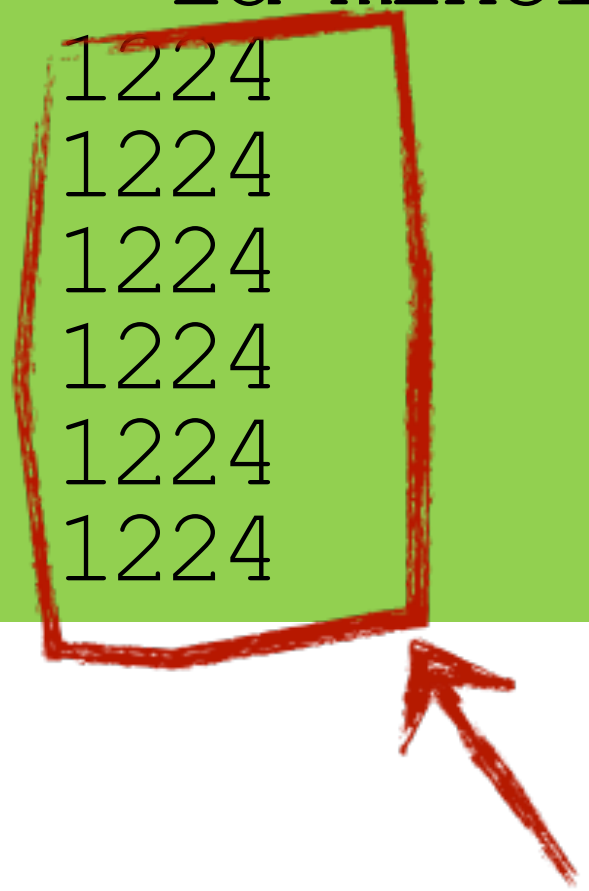
Loading data in R

```
> library( foreign )
> library( tidyverse )

> # read student data
> dat = read.spss( "data sets/High School and Beyond/hsb1.sav",
                                to.data.frame=TRUE )
> str( dat )
'data.frame':  7185 obs. of  5 variables:
 $ id       : Factor w/ 160 levels "1224","1288",..: 1 1 1 1 ...
 $ minority: num  0 0 0 0 0 0 0 0 0 0 ...
 $ female   : num  1 1 0 0 0 0 1 0 1 0 ...
 $ ses      : num  -1.528 -0.588 -0.528 -0.668 -0.158 ...
 $ mathach : num  5.88 19.71 20.35 8.78 17.9 …

> head( dat )
    id minority female    ses mathach
1 1224        0      1 -1.528   5.876
2 1224        0      1 -0.588  19.708
3 1224        0      0 -0.528  20.349

> nrow( dat )
[1] 7185
```

Note: The code we give you has more comments and looks a little different from the code we display. But this can be found in the sample lecture code.

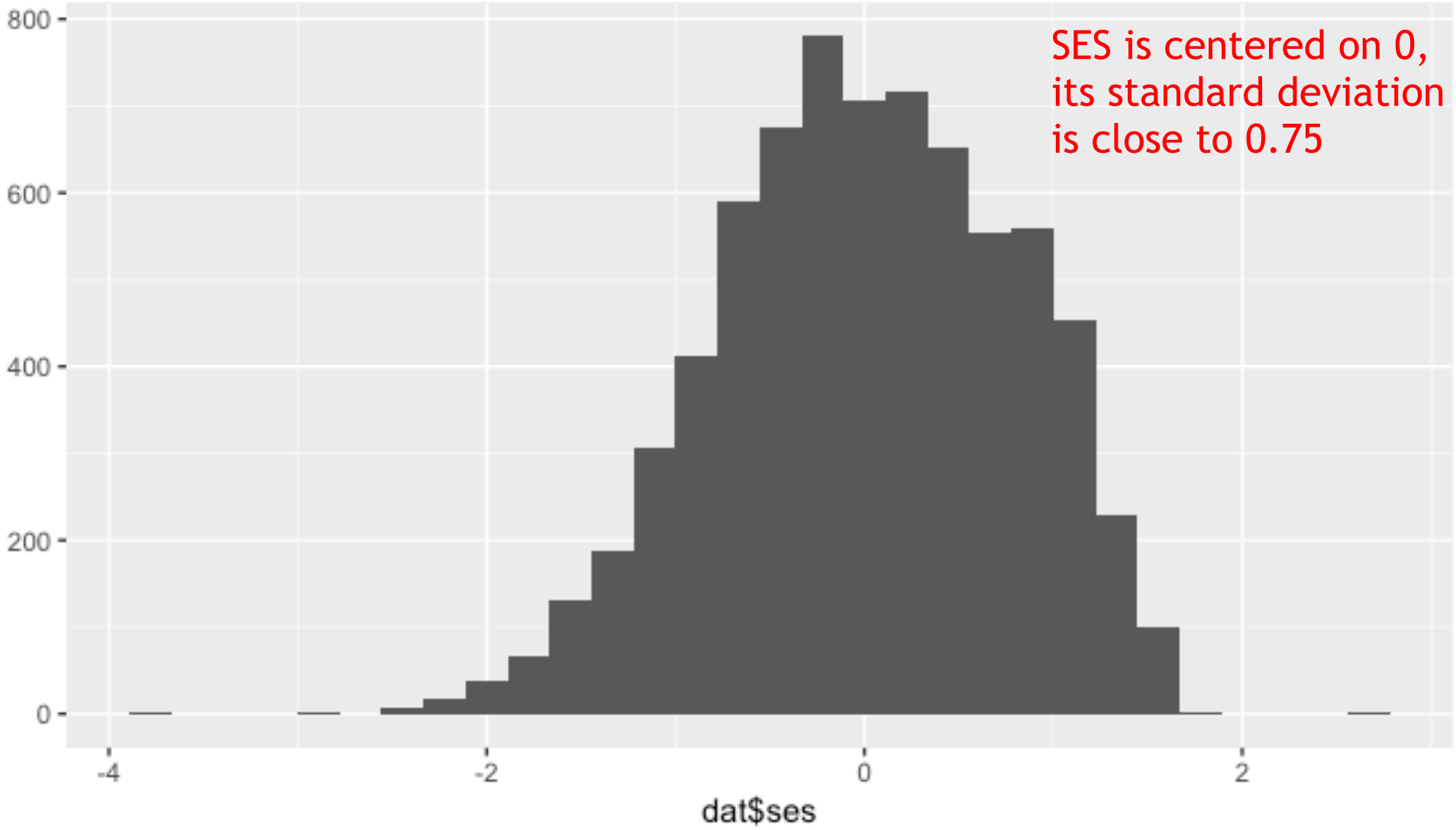
Individual cases (the students)

```
head(dat)
     id minority female    ses mathach
1  1224        0      1 -1.528   5.876
2  1224        0      1 -0.588  19.708
3  1224        0      0 -0.528  20.349
4  1224        0      0 -0.668   8.781
5  1224        0      0 -0.158  17.898
6  1224        0      0  0.022   4.583
```



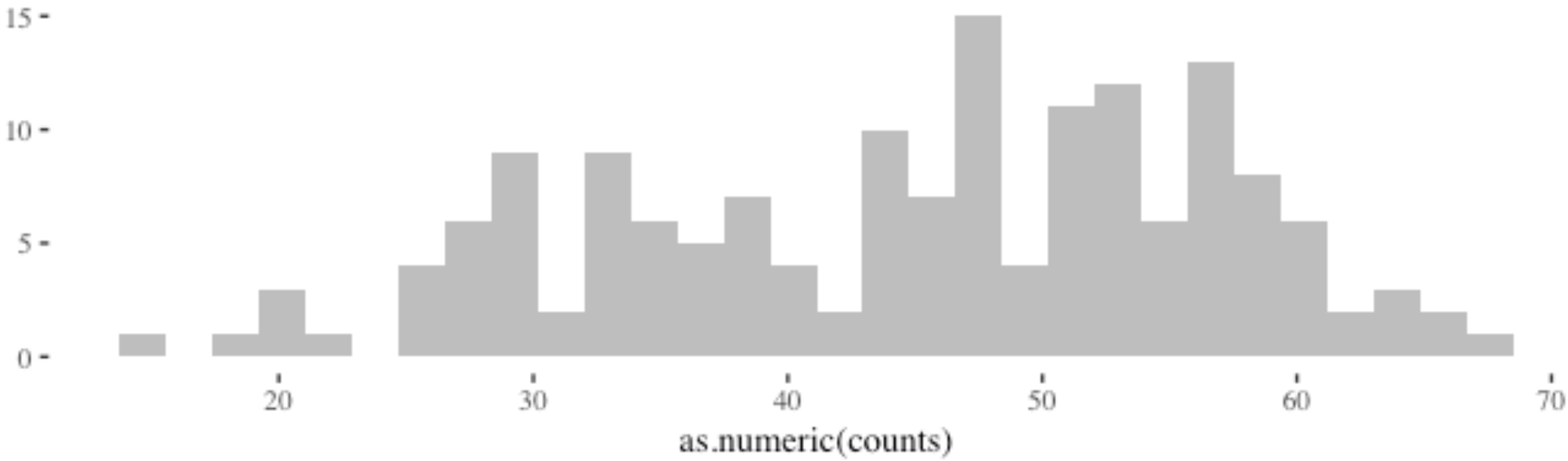
The school ids. We are seeing students from a single school in the first 6 rows.

Examining SES across students




```
> qplot( dat$ses, breaks=30, col="grey" )
```

Looking at sizes of samples across our sample of schools



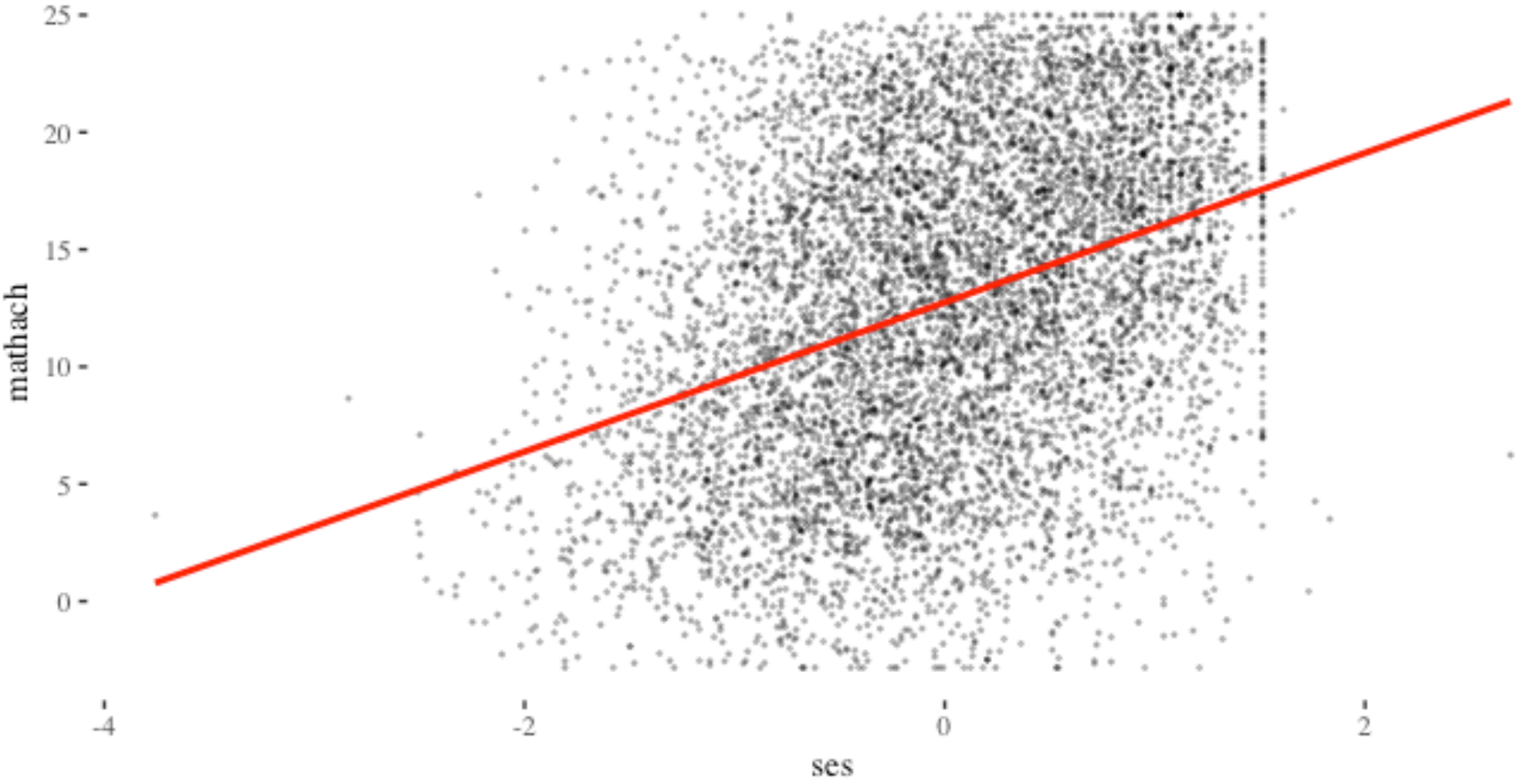

```
> counts = table( dat$id )

> head( counts )
1224 1288 1296 1308 1317 1358
  47   25   48   20   48   30
```

Variables in datasets, use the "$" notation

School IDs at top

Each school has number of students

Is SES connected to math achievement?




```
ggplot( dat, aes( ses, mathach ) ) +
  geom_point( alpha=0.25, size=0.3 ) +
  geom_smooth( method="lm", se=FALSE, col="red" )
```

Two Representations of Data: #1: People and Schools, separate

```
> head( dat )
   id minority female    ses mathach
1 1224        0      1 -1.528   5.876
2 1224        0      1 -0.588  19.708
3 1224        0      0 -0.528  20.349
```

Individual variables "Level 1" (7185 rows)

```
> head( sdat )
   id size sector pracad disclim himinty meanses
1 1224  842      0   0.35   1.597       0  -0.428
2 1288 1855      0   0.27   0.174       0   0.128
3 1296 1719      0   0.32  -0.137       1  -0.420
```

Link with school IDs

School variables "Level 2" (160 rows)

Two Representations of Data: #2: Single file (AKA STATA)

```
> dat = merge( dat, sdat, by="id", all.x=TRUE )
> head( dat )
     id minority female    ses mathach size sector p
1224           0      1 -1.528   5.876  842      0   0
1224           0      1 -0.588  19.708  842      0   0
1224           0      0 -0.528  20.349  842      0   0
> tail( dat )
9586           0      1  1.332  19.641  262      1
9586           0      1 -0.008  16.241  262      1
9586           0      1  0.792  22.733  262      1
> nrow( dat )
[1] 7185
```

Individual variables "Level 1"

School variables "Level 2"

Taste of R

Going from two datasets (first representation) to one dataset (second representation)

Merging and linking

Level 1 Data: Students

Level 2 Data: Schools

Each student in Level 1 has a school ID that references a particular school at Level 2.

To *link* data we find, for each student, the school that corresponds to that student.

This process is called *merging.*

Anatomy of merge()

```
merge(set1, set2,
      by = SOMETHING,
      all.x = TRUE,
      all.y = FALSE)
```

SOMETHING is, in the simplest case a VARIABLE NAME in quotes.

- ★ E.g., “id”

all.x and all.y means keep each row in the first or second set, even if no match is found.

If you have multiple matches, you get one row for each possible match.

POP QUIZ

```
A = data.frame( ID = c( 1, 2, 3 ),
                X = c( 10, 20, 30 ) )
B = data.frame( ID = c( 2, 2, 4 ),
                Y = c( 0.2, 0.3, 0.1 ) )

C = merge( A, B, by="ID", all.x=TRUE, all.y=FALSE )
C
```

What do you get?

(you could literally do this at your computer right now and see what happens)

Regression in R

Making a toy example: Taking a subset of schools

```
> set.seed( 12345 )

> sids = unique( dat$id )
> length( sids )
[1] 160

> winners = sample( sids, 10 )
> winners
 [1] 7332 8707 7635 9550 4458 2526 3533 4931 7011 9158
160 Levels: 1224 1288 1296 1308 1317 1358 1374 1433
1436 1461 1462 1477 1499 1637 1906 1909 1942 1946 ...
9586

> nrow( dat )
[1] 7185

> dat.ten = subset( dat, id %in% winners )
> nrow( dat.ten )
[1] 473
```

R can only generate pseudo-random numbers, so if we all start at the same point, we'll all get the same results.

Only keep students from these ten schools

Diversion: R uses “factors” (school ids are factors)

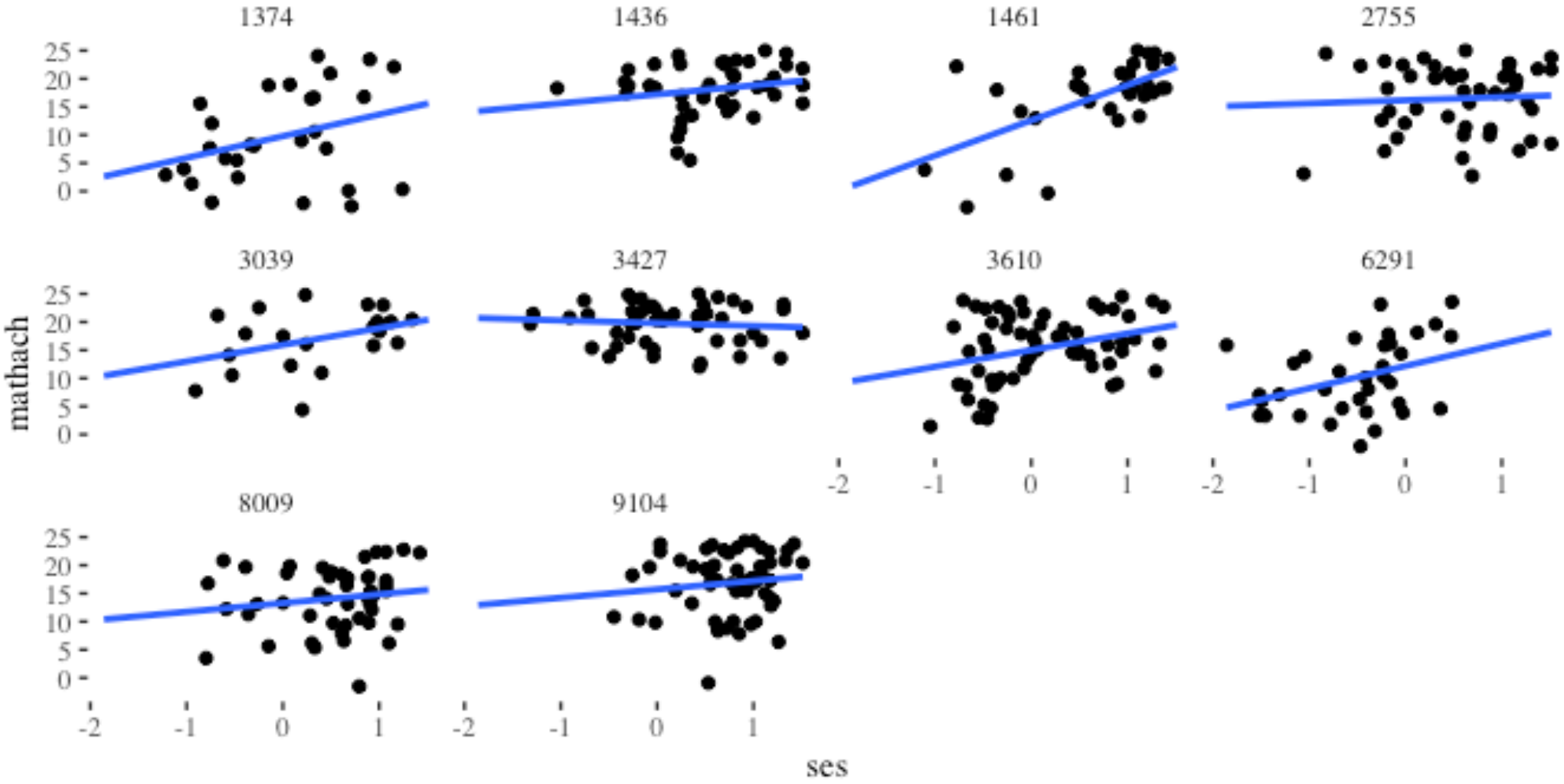
```
> dat.ten = droplevels( dat.ten )
> table( dat.ten$id )

2526 3533 4458 4931 7011 7332 7635 8707 9158 9550
  57   48   48   58   33   48   51   48   53   29
```

Aside: If we didn’t call droplevels() we would have gotten this:

```
> table( dat.ten$id )

1224 1288 1296 1308 1317 1358 1374 1433 1436 1461 1462
1477 1499 1637 1906 1909 1942 1946 2030 2208 2277 2305
   0   25    0    0    0    0    0    0    0    0    0
0    0    0    0    0    0    0    0    0    0    0
2336 2458 2467 2526 2626 2629 2639 2651 2655 2658 2755
2768 2771 2818 2917 2990 2995 3013 3020 3039 3088 3152
   0    0    0    0    0    0    0    0    0    0    0
0    0    0   43    0    0    0    0    0    0    0
3332 3351 3377 3427 3498 3499 3533 3610 3657 3688 3705
```

10 little worlds

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



OLS in R: Easy as pie

```
> M0 = lm( mathach ~ 1 + ses, data = dat.ten )
> summary( M0 )

Call:
lm(formula = mathach ~ 1 + ses, data = dat.ten)

Residuals:
     Min
-17.9611

Coefficien

                                          t|)
(Intercept)                               -16 ***
ses                                       -16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 6.114 on 471 degrees of freedom
Multiple R-squared:  0.189,  Adjusted R-squared:  0.1873
F-statistic: 109.8 on 1 and 471 DF,  p-value: < 2.2e-16
```

BUT THIS REGRESSION IS SUPER BAD!

(WHY?)

```
> 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
```

A taste of lecture 1.2: "Fixed Effect" regression.

Note, still easy to get R to do!

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 makes indicator variables for you

TRUE means record what our final covariates are.

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

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
```

See all the dummy variables we automatically made?

What just happened?

Recap of Lecture

★ Bureaucracy: there was a bunch of stuff that I can look up later

★ Clustered data is *hierarchical* in that it has *levels.*

★ R is a statistical programming language that can do stuff. The instructor says it's cool.

★ *Regression* is a simple thing to run, but with clustering simple regression is invalid.

Get started now!

There is a **Problem Set 0** posted on the Dropbox

This will get you involved in using R right away.

Next steps:

- ★ Get the textbook
- ★ Install R and RStudio
- ★ Attend the “Gentle Intro to R” workshop
- ★ Start on Homework 0