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

An overview of what this course is about, and possibly even why you should take it.

Instructor: Prof. Luke W. Miratrix

lmiratrix@g.harvard.edu

Larsen 603

S-043: multilevel and longitudinal models

- Who am I?
  - Luke Miratrix
  - Assistant professor at HGSE. Was assistant professor at Harvard Stat before this.
- Who are you?
  - HGSE PhD students
  - HGSE master’s students
  - PhD/master’s students from other schools (Harvard and others)
  - Undergrads?

What is this course for?

This course is to provide you with the tools you need to do research using data with a multilevel or longitudinal structure (i.e., clustered or nested data)

Multilevel data? What’s that?

In reality: Data where you have groups of things

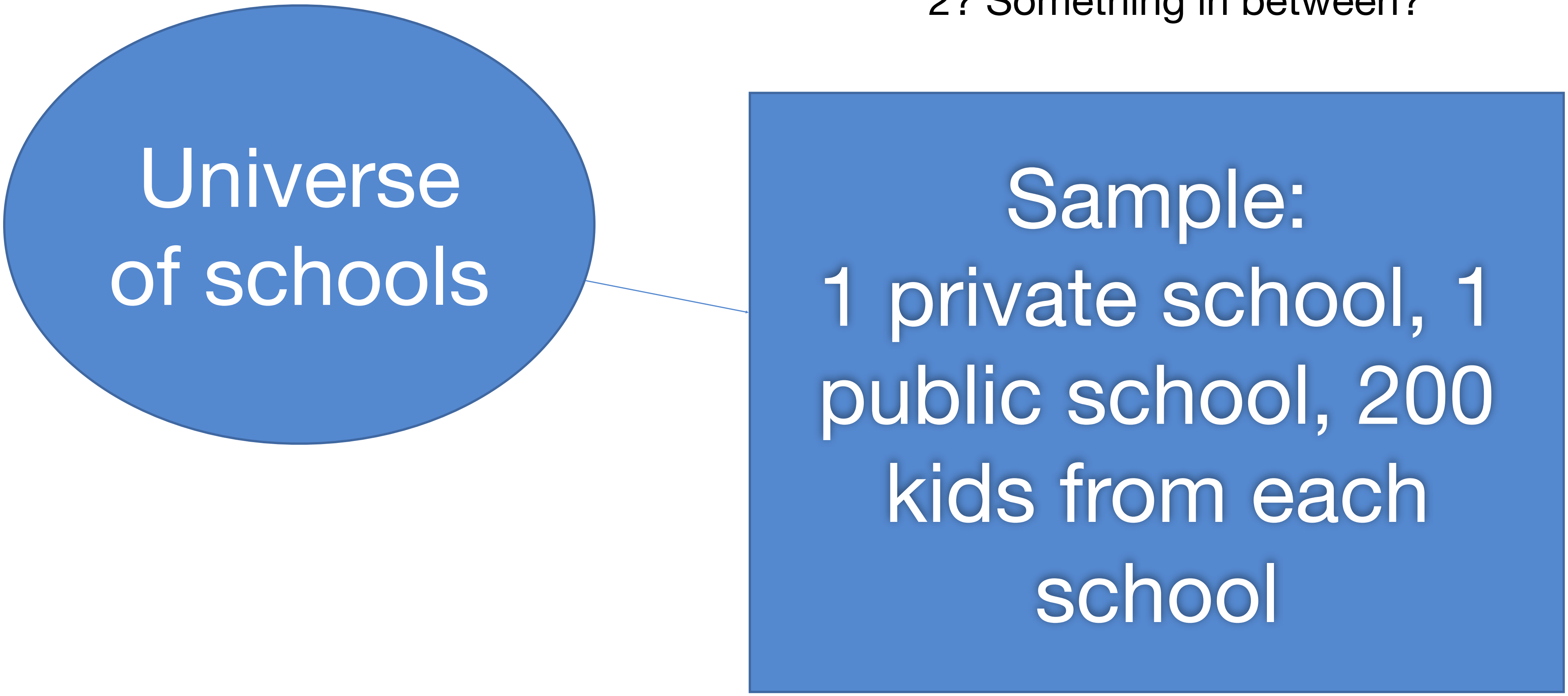
Some examples:

- Researchers sample schools from a district, then sample classes from those schools, then students from those classes (three levels)
- Researchers sample students, then measure each student at multiple different time points (two levels)

Multilevel data? Why stress?

Ignoring nested structure is perilous (okay, not really, but it doesn't work well)

What is our sample size? 400? 2? Something in between?



Universe of schools

Sample: 1 private school, 1 public school, 200 kids from each school

Multilevel data? Hooray!

Multilevel data let us ask interesting questions:

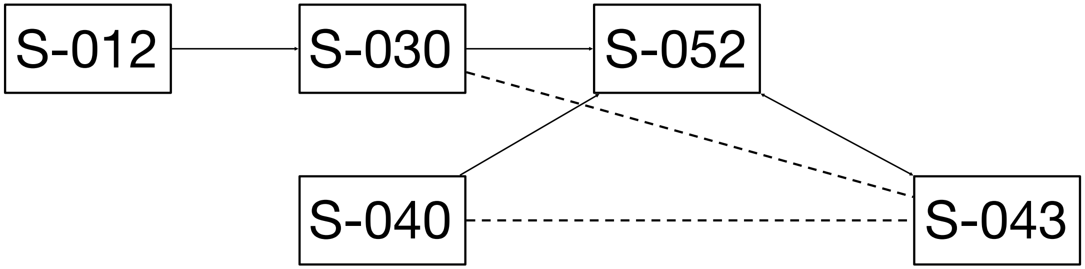
A. Do students from higher SES backgrounds tend to have higher levels of math achievement than students from lower SES backgrounds?

B. Controlling for student SES, do *schools* with higher *mean* student SES tend to have higher *mean* levels of math achievement?

C. How much did treatment impacts vary in my multi-site randomized controlled trial?

D. Are racial achievement gaps bigger or smaller in higher SES schools than lower SES schools?

Where does this course fit in the methods sequence?



This course…

- shows how to apply S-030/S-040 type analyses (specifically regression) to multilevel data
- Overlaps very slightly with S-052
- Uses much more mathematical formalism (greek letters) than S-052

To take this course, you should *definitely* know...

How to interpret estimates of linear models.

How to think critically about whether they are appropriate.

How to think at least a bit about testing and estimation.

How to use some statistical software (R ideal)

I.e., you should know:

S-040 or Stat-139

If these courses were hard for you, you should also have another more advanced quantitative course under your belt so you can have a stronger foundation.

Furthering your research

A major component of this course is the final project

This final project is ideally a vehicle to advance your research.

Ideally, you will come to this course with your own dataset to analyze.

(If not, you should start looking for one.)

You might also want…

- ★ to be able to use R for simple tasks, including data management and fitting models
- ★ some knowledge of generalized linear regression (e.g., logistic regression)
- ★ some mathematical comfort (e.g., summation notation and matrices)
- ★ an intuitive comfort with linear regression (especially interpretation)

These are not “must haves,” but having one or more of these will make the course easier.

To take this course

You don’t need to…

- ★be an expert using R
- ★be experienced with mathematical statistics (especially proofs)
- ★have any experience analyzing multi-level data

General Content of Course

1. Grapple with clustered data (and see some common tools to handle it).

Core of Course

2. Introduce a formal model (the multilevel model) to handle and describe such data.
3. Examine how to assess uncertainty (and do inference) using such models.
4. Examine how to learn about variation using such models.
5. Extend the above to longitudinal data.
6. Learn how to assess uncertainty using simulation and bootstrap.
7. Generalize and extend to nonlinear data (e.g., counts).

Sure, but what will we *actually* be doing?

Two lectures per week (T/Th, 2-3:30).

Some readings

Weekly section

About three short problem sets (individual)

Two longer analyses (pairs)

Final project (ideally pairs and *very* student-driven)

Primary Materials of the Course

- ★ Canvas - the central spot for announcements and resources. This is where you will submit data.
- ★ Raudenbush and Bryk Book - a concrete and clear textbook that focuses on analysis of specific forms of data
- ★ R Studio - a real tool for statisticians.

Why take this course?

So you can do these things:

- ★ Think critically about data-based claims and quantitative arguments.
- ★ Explore and analyze real data with hierarchical or longitudinal aspects.
- ★ Present such data numerically and visually.
- ★ Know which statistical methods to use when, and how to use them.
- ★ Use R, a statistical software package.
- ★ Learn new statistical analysis techniques on your own.

⚠ Finding Course Bureaucracy Details ⚠

Please see posted slides on the website.

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

Also,

YOU MUST

READ THE SYLLABUS

Further Details

How homework is made

Problem sets: Individually completed work

- ★ Collaboration encouraged.
- ★ Individual write-ups.

Projects: Mandatory groups of 2

- ★ Single write-up submitted for the group.

Final Project: Groups of 1-3 students, with an ideal of 2.

- ★ Individual projects only if due to using dissertation/ research work as one's final project (although we encourage collaboration nonetheless).

Further details, and further information on **collaboration vs. plagiarism**, in syllabus

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) because it negatively impacts others.

The work you will do

- ★ 2 Problem Sets
  - Smaller questions targeting specific issues of understanding.
- ★ 2 Projects
  - Analyze real-world data sets and write coherently about what you find.
- ★ Final Project
  - Analyze your own data and discover something new.
- ★ A few mini-assignments, such as a warm-up learning R assignment.

A Note on Papers and Examples

The original sources for the studies used in the course will be provided and linked when possible - if interested in the details of the study, please check out the original article!

An incomplete take on Longitudinal Data

One approach to longitudinal data is to view a sequence of observations from people across time as nested inside people.

Other approaches not covered:

- ★ Survival Analysis
- ★ Hazard Models
- ★ Cox Regression Models

See Singer and Willet's
"Applied Longitudinal Data Analysis"

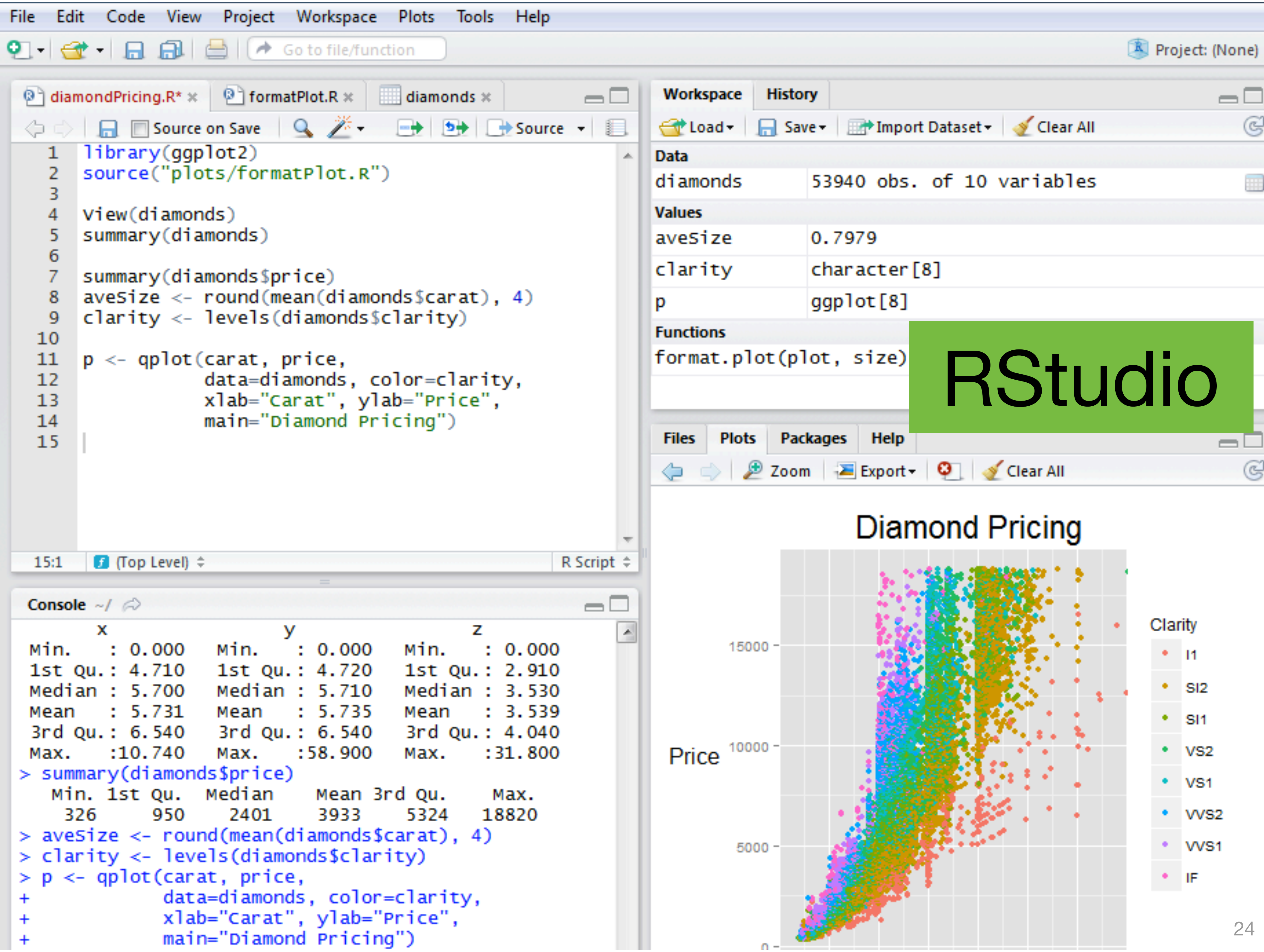
Where is the Math?

Goal: teach good intuition as to how these models work, how they are fitted, and what to worry about.

Goal: teach literacy so students can read papers in the field, communicate their models and results, and engage with the research.

Math and mathematical notation help these goals, therefore:

- ★ Mathematical notation will be used extensively.
- ★ Informal mathematical argument will be used in class sometimes.
- ★ Optional readings will be given out for those who want to dive deeper.



RStudio

```r
library(ggplot2)
source("plots/formatPlot.R")

View(diamonds)
summary(diamonds)

summary(diamonds$price)
aveSize <- round(mean(diamonds$carat), 4)
clarity <- levels(diamonds$clarity)

p <- qplot(carat, price,
           data=diamonds, color=clarity,
           xlab="Carat", ylab="Price",
           main="Diamond Pricing")
```

Data	
diamonds	53940 obs. of 10 variables
Values	
aveSize	0.7979
clarity	character[8]
p	ggplot[8]
Functions	
format.plot(plot, size)	

```
       x                 y                z
 Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
 1st Qu.: 4.710   1st Qu.: 4.720   1st Qu.: 2.910
 Median : 5.700   Median : 5.710   Median : 3.530
 Mean   : 5.731   Mean   : 5.735   Mean   : 3.539
 3rd Qu.: 6.540   3rd Qu.: 6.540   3rd Qu.: 4.040
 Max.   :10.740   Max.   :58.900   Max.   :31.800
> summary(diamonds$price)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    326     950    2401    3933    5324   18820
> aveSize <- round(mean(diamonds$carat), 4)
> clarity <- levels(diamonds$clarity)
> p <- qplot(carat, price,
+            data=diamonds, color=clarity,
+            xlab="Carat", ylab="Price",
+            main="Diamond Pricing")
```

Getting Started with R and RStudio on Your Computer

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

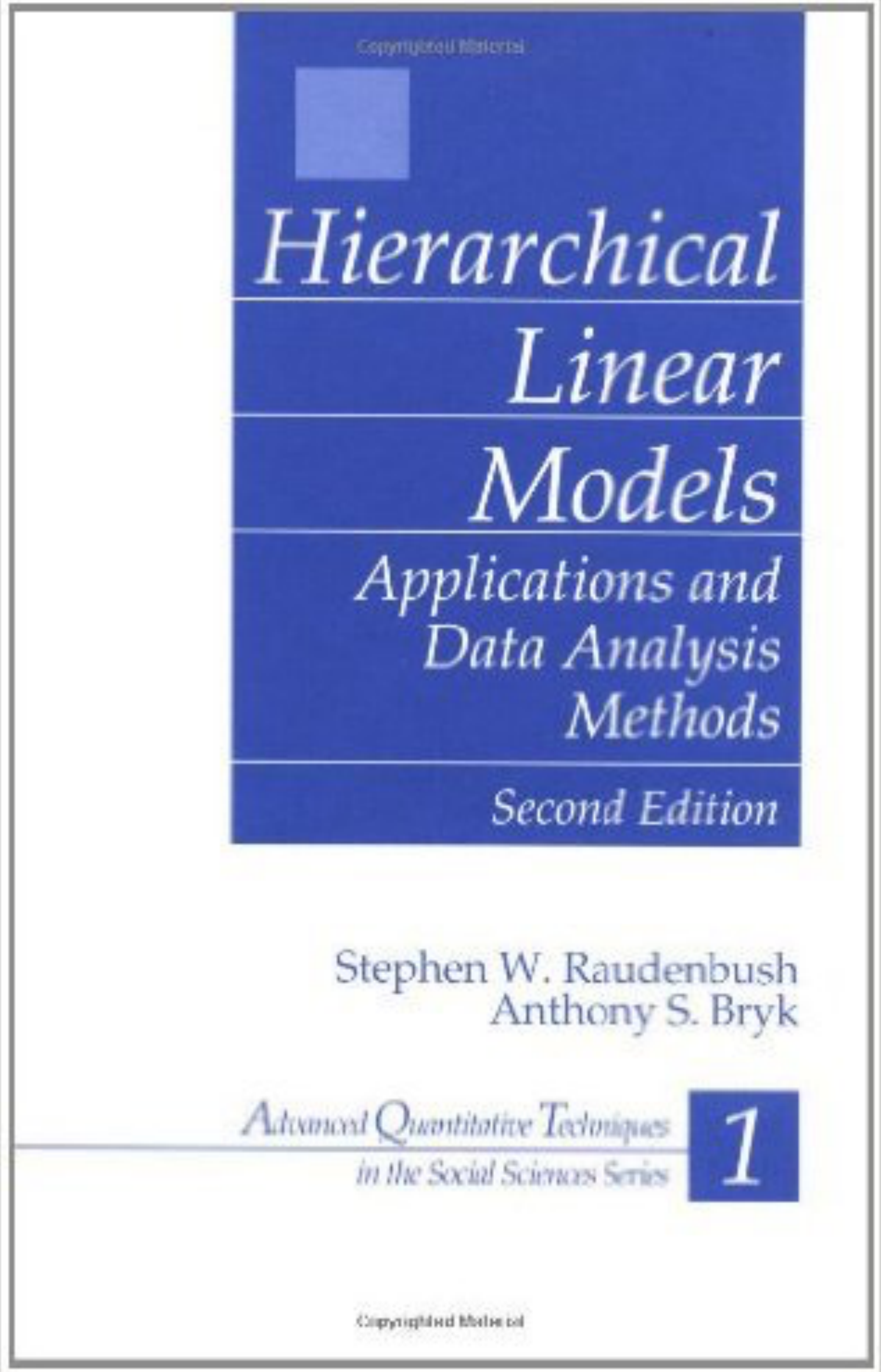
  http://lib.stat.cmu.edu/R/CRAN/

- ★ Once you install R, download and install RStudio.

  http://www.rstudio.com

We will help you with this during office hours or section if you get stuck.

Primary Textbook



Hierarchical Linear Models

Applications and Data Analysis Methods

Second Edition

Stephen W. Raudenbush
Anthony S. Bryk

Advanced Quantitative Techniques in the Social Sciences Series 1

Math, explanation, no code

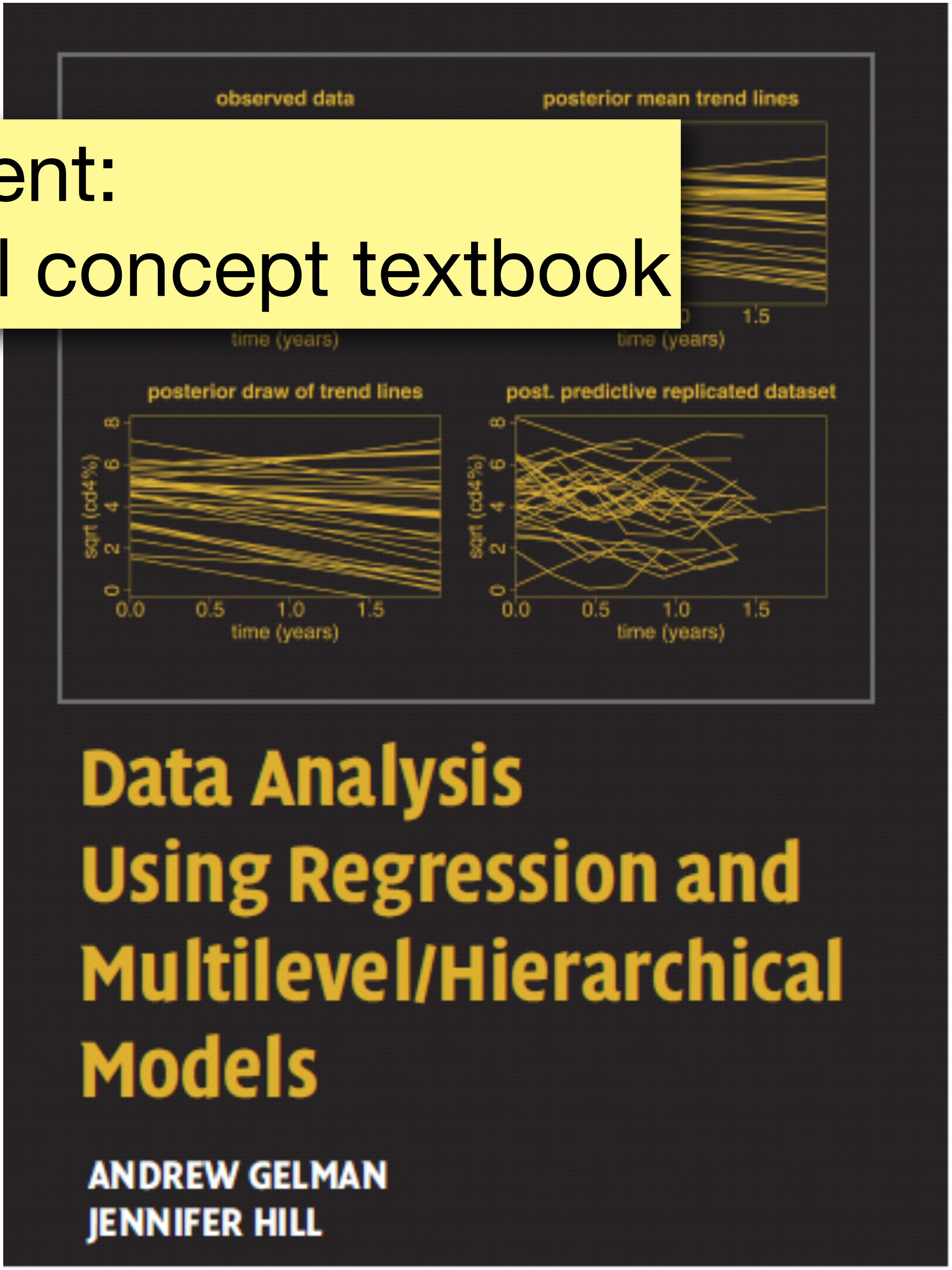
★ Some mathematics showing where MLMs come from.

★ Excellent worked examples, connecting the fitted models to conceptual questions.

★ Case studies that bring the models to life.

★ Discussion of many classic concerns that arise with modeling.

Supplement:
High level concept textbook



Data Analysis Using Regression and Multilevel/Hierarchical Models

ANDREW GELMAN
JENNIFER HILL

It is actually *two* textbooks

Part 1: Single Level Regression

- ★ Part A - Basic fitting, logistic, generalized linear models
- ★ Part B - Inference, simulation, causal inference

Part 2: Multilevel Regression

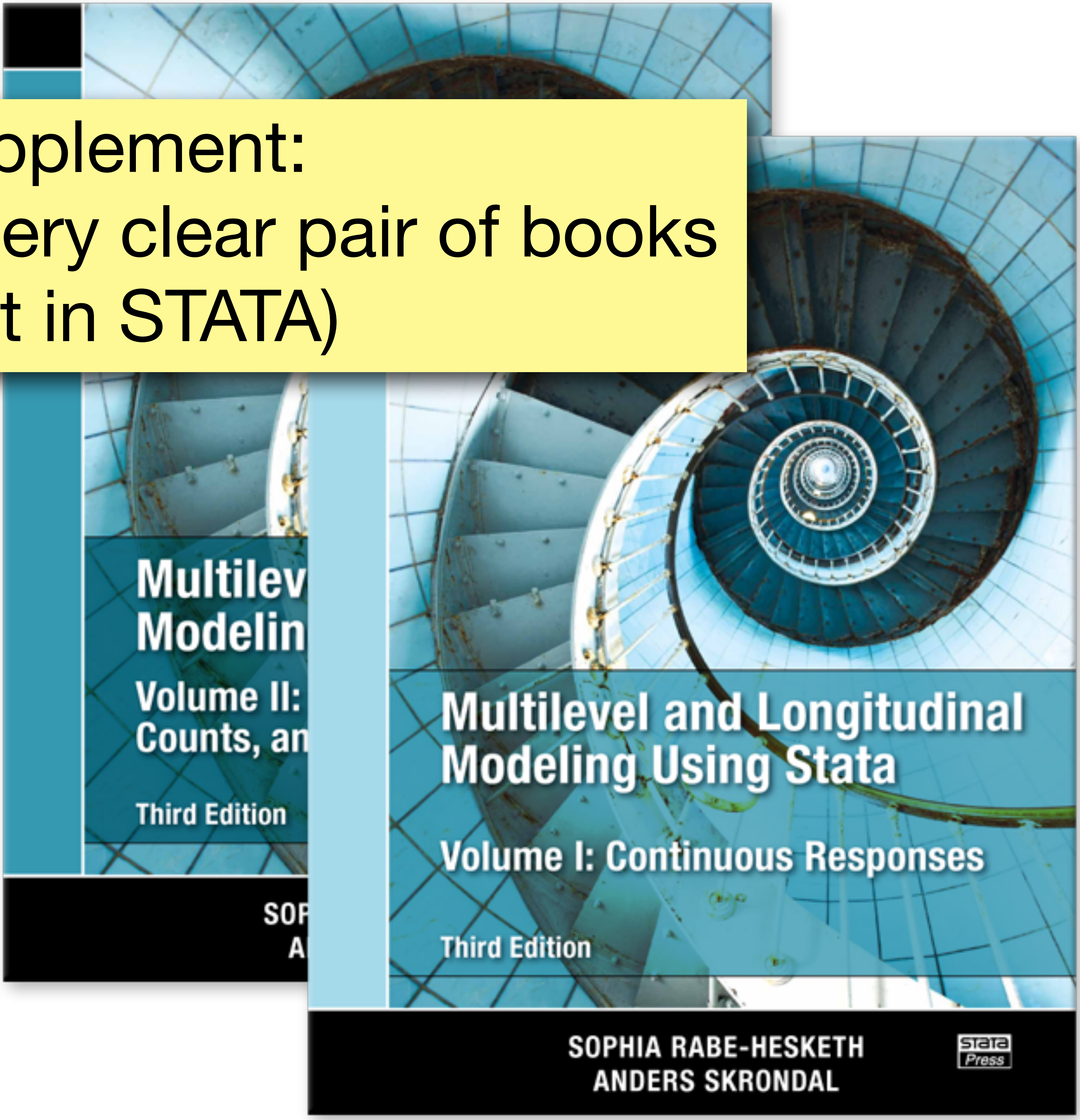
- ★ Part A - What these things are and how to think about them
- ★ Part B - Fitting them and getting Fancy

Part 3: Getting fancy

- ★ Power and sample size
- ★ Further conceptual issues of multilevel models
- ★ More on inference

The first part has some ideas of simulation for uncertainty that are important

Supplement:
A very clear pair of books
(but in STATA)



It is actually *two (three?)* textbooks

Volume 1: Continuous outcomes

- ★ Part I: An intro on linear regression
- ★ Part II: Multilevel models
- ★ Part III: Longitudinal and panel data

Part I should be review, basically.

Volume 2: Other outcomes

- ★ Each section starts with the “single level” (not multilevel) model and extends to multilevel models