Syllabus for S-043/Stat-151 (F2023)
Multilevel and Longitudinal Models

Last updated: 8/21/23

This syllabus is a game plan, not a contract. Changes may be made, with appropriate notice, at the discretion of the instructors to better facilitate student learning.

Table of Contents

Course Overview

S043/Stat 151 provides statistical tools for students engaged in ongoing research projects related to clustered or longitudinal data. The course focuses on the multilevel model, a statistical model that describes the relationships between outcome and covariates within and across clusters of data, where clustering is defined by location (e.g. school, district, hospital) and/or time (e.g. multiple observations for each student or patient). It also covers classic econometric approaches (cluster robust standard errors and fixed effects) for similar purposes.

This is a hands-on, applied course. The use of new statistical techniques will be first modeled in class, and then you will be asked to apply these techniques to real problems using real data. You will therefore learn the statistical software R as part of the course. You will be asked to interpret the outcomes of your data-analyses in words, and to communicate these interpretations clearly and concisely in writing. Time will therefore be spent on how to appropriately communicate, visualize, and summarize results and data. By the course end, you will be able to analyze data in the field and defend your analysis.

Some salient, very important aspects of the course:

- The course is targeted towards PhD students in education and other social sciences, but historically the course has been a mix of PhD students, master students, and undergraduates who are interested in these tools. The heterogeneity in the classroom provides many opportunities for everyone to learn from one another.
- The course uses R, a statistical programming environment, and resources for learning R are provided throughout. Sections generally focus on the technical aspects of using R.
- We assume no prior knowledge of R and will teach it both in lecture and section. That being said, starting with no R knowledge will make the course more intensive.
- The course culminates in a final project, which, for many graduate students, is a continuation or extension of their doctoral work. Those without such a project will have to locate data and generate research questions as part of this final project.
- Before the final project, there is a mix of individual and group assignments.
- The assignments are a mix of “projects,” which are essentially guided data analyses, and “problem sets/individual work” which are collections of smaller problems that target specific concepts or comfort with mathematical notation. This category also includes some assignments such as reviewing and critiquing a published paper.

High Level Structure of Course

The backbone of S-043 are the two 1.5 hour live sessions each week. The content of the course will be structured in a series of units, and the units further subdivided by these live sessions. We will call the material and whatnot associated with preparing for and doing one of these sessions a “packet.” Both unit and packet are numbered, so 3.2 would be in reference to the 2nd packet (including corresponding live session) for the 3rd unit.

Each packet will have a mix of tasks and so forth; please read the text surrounding these tasks carefully to orient yourselves so you do not get lost. Things you should do to prepare for the live session, including relevant readings and so forth, will be clearly marked. Many packets will have a video to watch before the live sessions; we usually begin live sessions with some activity that builds on said video.

Some tasks and resources for a packet are optional; it is frequently the case that you will return to a packet when working on your final project. This is when, sometimes, these resources will suddenly be much more relevant to you. Or even years from now they may save you in an analysis. We recommend downloading everything and organizing it on your own computer for future reference.

The live sessions will be a mixture of applied examples, mathematical discussion, and R coding. There will be weekly section meetings to cover topics in further detail and to teach R programming. Projects will be completed in groups of two or three students; problem sets and other work will be done individually. The course will culminate in final projects, completed by groups of students, on topics selected by students. These final projects will ideally be dissertation work or come from other research agendas.

In addition to all of this, there will be asynchronous and synchronous methods of getting help including

- an active discussion forum (Slack)
- sections where students actively work on exercises with TF support
- office hours where students can chat and ask questions.

Content Overview

Data often have structure that needs to be modeled explicitly. For example, when investigating student outcomes we need to take into account that students are nested inside classes that are in turn nested inside schools. If we are watching students develop over time, we need to account for the dependence of measurements across time. If we do, we can describe where variation is, and get improved understanding of our data's structure. If we do not, we will tend to believe that our inferences are more precise than they really are, and our estimates may be biased (systematically wrong).

This course will primarily focus on how to use multilevel (hierarchical) models for dealing with such problems. We will focus on specific versions of these tools for the most common forms of longitudinal and clustered data, but also keep an eye on the general themes behind the methodology, which will enable students to analyze more complex scenarios if needed.

We will primarily focus on the linear model with continuous outcomes (the classic regression framework), but also cover binary, count, and ordinal outcomes. We will discuss the applicability of these methods, how they might fail, and what one might do to protect oneself in such circumstances. We will also survey other methods for addressing these types of data, in particular the econometric approach of using cluster-robust standard errors.

See the course calendar for a more detailed list of content, and the Canvas pages for each unit for more detail beyond that.

Goals of Course

- Learn how to recognize, explore, and analyze multilevel data, primarily by fitting multilevel models.
- Learn how to use R to conduct an analysis in real life.
- Develop statistical thinking and deepen conceptual foundations of statistics.

Meeting time and contact information

Instructor: Luke Miratrix, Associate Professor in HGSE, affiliate faculty in Statistics

Email: luke_miratrix@gse.harvard.edu

Meeting: Tues/Thur 12 :00-1 :15pm Eastern

See Canvas site for further details of meeting times, etc.

Intended Audience

S-043 is primarily designed as a second-year quantitative methods course for PhD students in education but is also open to other interested students throughout the university. It could also be appropriate for statistics undergraduate concentrators who want to engage with applied projects and an open-ended course structure, or to education master's students with strong quantitative backgrounds. Willingness to learn R is a must. Familiarity with linear regression, hypothesis testing, and basic forms of inference is a must. You will need to be comfortable with the regression of continuous outcomes onto any type of covariates, including categorical predictors. Understanding interactions in such models is important. Familiarity with logistic regression is desirable, but not required.

Prerequisites and Assumed background

People from many, many different backgrounds take S043/Stat151. In general, the more background experience you have, the easier the course. Even if you don't have as solid a background, you can still get a lot out of this course if you are willing to sign on for an intense experience.

The stated prerequisite to this course is S052, Stat 139, or equivalent, i.e., you have gone a bit beyond an applied course on linear regression (S040). If you have taken S040 or equivalent, you could take this course, but this is not recommended unless you have a strong mathematical or computer programming background. More specifically, we expect you to have the following skills as a prerequisite for this course:

- You have some experience working with and thinking about real data.
- You can calculate summary statistics and visualizations for data (the mean, standard deviation, scatterplots, histograms) using some sort of statistical software.
- You can load data into some sort of statistical software, use that software to fit a regression model, and then interpret the results of the model.
- You know how to include categorical covariates in a linear regression.

- You know what an interaction term in a linear regression model is, and can add one to a model you are fitting.
- You can interpret confidence intervals and p-values.
- You can write down a regression equation and identify what the covariates and parameters are in a presented regression equation.

You will have an easier time if you have some of the following skills and experiences as well:

- You have experience doing quantitative research in a fairly independent manner.
- You have some experience with doing things in R (even a little helps here).
- You know what logistic regression is, and know how to fit a logistic regression model using some sort of statistical software.
- You have seen random intercept models before, as perhaps in S-052.
- You have seen regression with fixed effects for groups, as perhaps in an econometrics course.

If you have a strong mathematical background, or strong comfort with math, or if you have a strong computer programming background, or strong comfort with that, then you can likely get a lot out of this course even if you do not have some of the above skills and knowledge. If you do not, you may find the course very challenging and overwhelming.

Textbook & Reading

The primary text is Raudenbush and Bryk's *Hierarchical Linear Models: Applications and Data Analysis Methods (Advanced Quantitative Techniques in the Social Sciences)*. This book gives a lot of excellent advice about the core models used in Education. It has many examples, but no accompanying code, instead focusing on mathematical representation. It also has some mathematical derivations for multilevel models that are a good overview and reference for those interested. In particular, the central portion of this book is packed with great insights into different modeling contexts.

There is also an online "handouts textbook" that has a bunch of short pieces that target different aspects of the course such as coding tips, some mathematical derivations, or worked case studies.

Some additional books which cover the material we'll be exploring are:

- Sophia Rabe-Hesketh and Anders Skrondal's *Multilevel and Longitudinal Modeling Using Stata*, Second Edition (both volumes)

  This is a very clear text that describes how to use multilevel modeling to answer all sorts of questions. There are many examples. It is all in Stata, and thus can be used to translate this course to that statistical environment.

- Gelman & Hill's *Data Analysis Using Regression and Multilevel/Hierarchical Models*

This is a very good book on how to think in general terms about multilevel modeling. It has a strong focus on visualization and on larger themes rather than specifics. It is technically in R, but the style of R used is often hard to follow, so ignore that aspect.

- Singer and Willett's *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*

  This book deals with specific conceptual issues of longitudinal data, and is designed to be very non-technical and approachable (i.e., it is clear). It has several case studies of interest. It also covers some forms of longitudinal data we will not cover in this course, such as survival analysis.

These books will be excerpted and constitute part of the reading; scans of these parts are posted on Canvas. There will be other supplementary reading, such as sample academic papers illustrating how these methods are used in practice, assigned as the course progresses.

For reference the following abbreviations:

R&B – Raudenbush and Bryk, the primary textbook

G&H – Gelman & Hill

RH&S – Rabe-Hesketh and Skrondal

S&W – Singer and Willett

Handout – The handout textbook at https://lmiratrix.github.io/MLM/

How to Find Reading Resources

Often you will want to find a reading to help with a specific task you are working in. In general this is how I would find help in this case:

1) Scan the handouts textbook to see if there is a handout on that specific thing. That may solve your problem.
2) If not, find a lecture by looking at the packet list to identify where that specific thing was discussed. Then look at the R code, lecture slides, and/or the reading for that lecture.
3) You can also check the section materials for that week---there is often useful code and tips in the section handouts.
4) Email a TF or instructor to get a pointer as to where to look.

Also, I strongly recommend downloading the library reserves scans onto your own computer. Name those files to make finding things easier. Also grab the section handouts. Generally speaking, I think the case study documents posted in the handout textbook are possibly the most useful handouts.

Expectations for Reading

In general, you can choose to read before or after the lecture, although please note that some of the readings are marked as to be read at specific times. *But it is very important to actually read the readings*. We cannot cover all of the details of this material during class and lecture, and the reading both reinforces what we discussed as well as extends it in important directions.

The readings are marked for each course under different headings. The headings are:

> Reading – This means I expect you to read it, either before or after lecture (unless a given time is specified).
>
> Supplementary Reading – These are truly optional. Some of these are for strengthening foundational knowledge that we are building on in the lecture (generally read these before lecture). Some are more technical or detailed treatments of the content if you want a deeper dive. Some should merely be noted and remembered as potential resources depending on where your final project brings you.
>
> Coding/Visualization Reading – These readings are designed to help you with coding. They should be read, just like the main readings, but should often be read close to the time when embarking on doing some actual work so the ideas are fresh. These readings also serve as excellent references.

Office Hours

There are four general reasons you might want to go to office hours:

- **Get help with the assigned work** such as when you get stuck on a problem or want clarification to some question.
- **Talk about the course material**. You might want to do this because you are thinking about some research topic (for the final project or for other work), you are curious about some direction hinted at in lecture, you are trying to figure out why the material is actually relevant, or for a host of other reasons.
- **Career advice or help**. You might also want to come to office hours to talk about your career, or to ask for a letter of recommendation, or to request some form of accommodation.
- **You want to talk more broadly.** Maybe you want to talk about some other kind of statistics, or a good book, or about what it is like to be a professor. These are all fine reasons.

There are two forms of office hours: open office hours and private appointment. The posted office hours on the website are drop-in. For discussing the homework, I prefer the conversations be open with people listening in on others' questions. This provides learning opportunities for all, especially as people tend to be stuck on the same things. For other matters you can still drop in---frequently there are points where people are gone or not yet arrived and so a one-on-one conversation is natural (I recommend showing up at the very start or towards the end of the posted hours). You are also welcome and encourage to email

to set up an individual appointment time. Also, if you cannot make the posted office hours due to a standing conflict, you are welcome to email me to set up alternate times as well.

Having direct conversations and relationships with different professors can be an important and transformative component of your education, and office hours provides one way to form such connections. More strategically, this is a door towards things such as future letters of recommendation that actually speak to who you are (and are thus more effective), or hearing about poorly advertised opportunities, such as research assistantships, which can help advance your career.

Going to my office hours is part of your participation grade.

The Canvas course website

Bookmark the course website and check it often (especially in advance of every class and sometimes more frequently). The website is my primary means of taking care of "housekeeping" matters (eliminating the need to discuss deadlines, etc., in class). All course handouts, including lecture notes, assignments, etc., will be available online.

Student Behavior

In the following I outline expectations regarding student behavior and engagement.

Course Participation

Class participation is an important part of learning. If you have a question, it is likely that others do as well. I encourage active participation, both in class or on the discussion board. However, if time is tight or a comment takes us too far astray, do not be offended if I defer your contribution to another time or place. Students who feel uncomfortable speaking in class will *not* be penalized, as long as they participate in other ways.

As part of this class, you will be asked to make a *participation plan* where you define what participation means for you, and then commit to working to participate as you state.

Collaboration

Appropriate and judicious collaboration is encouraged. Discussion and the exchange of ideas are essential to academic work. Therefore, you are encouraged to consult with your classmates as you work. However, after discussions with peers, make sure that you can do the work yourself and ensure that any answers you submit for evaluation are the result of your own efforts (or possibly your group's efforts). Please list the names of students with whom you have collaborated when you submit work. Furthermore, for any writing assignments, if you received any help with your writing (feedback on drafts, etc.), you must acknowledge this assistance.

Student Accommodation

HGSE is committed to making course materials accessible to all students. Please contact the teaching team right away if you find any accessibility issues that prevent your full access and

participation in the course. For any questions you have regarding academic success, accessibility, and accommodations, we encourage you to contact KellyAnn Robinson, Sr. Associate Director of Student Support Services, at kellyann_robinson@gse.harvard.edu.

- GSE: Access and Disability Services (ADS)
- FAS: Accessible Education Office (AEO)

Course Grading

If possible, take the course SAT/UNSAT as this allows you to focus on learning. That being said, the course letter grade will be based on the following:

Problem Sets/Individual Work	20%
Projects	20%
Final Project	20%
"Best in Show"	15%
Participation & Attendance	15% (appx)
"Ready for Class" Surveys	10% (appx)

"Best in Show" takes your best performance across your three grades of Problem Sets, Projects, and Final Project and counts it extra towards your final grade (so, e.g., if you final project is the best, with an A, your "Best in Show" grade is also an A). This is to offset unforeseen circumstances such as a missed assignment, and allows students to be evaluated in the way that best showcases their abilities.

Assignments

The projects, problem sets, and other assignments are the most important opportunity for learning, as the course content become clear only when applied. We encourage students to consult with classmates while doing coursework, but please read the section on Collaboration, above.

You will usually have about a week and change to complete a given assignment. Work will be turned in online via Canvas as a pdf document and should be neatly presented. See the assignment style guide posted on Canvas for particulars. Graded work will be returned as soon as feasible.

Problem sets/individual work will contain a mixture of short answer questions, questions that require some calculation, simulation studies, open-ended questions, reviews of academic papers, and data sets to be analyzed. They are turned in **individually.**

Each **project** has you analyze a dataset using the methods covered in the course. The resulting document will be a well-formatted and edited report with salient and useful tables and figures. They will be turned in along with code to allow for full reproducibility. This code will not be directly graded but can aid graders in deciphering the main work so they can provide useful feedback. They are turned in as a **group.**

There will be about three problem sets of smaller questions and two projects. The two projects will be

1. An analysis of an observational multilevel data set
2. An analysis of a longitudinal data set

There will also be a few other smaller assignments targeting specific concepts and skills.

Late Work

We are generally reluctant to accept late work as that prevents us from releasing solutions in a timely manner, which undermines the learning experience of your classmates. We also find it is better for students to turn in what they have rather than getting increasingly behind. Our late policy also makes it easier for the teaching assistants to manage grading. All of this being said, we also recognize that life happens. If you need an extension of a day or two, please just write us to ask. If you need something beyond that, or if something happens, **turn in what you have by 48 hours after the due date and email us explaining the difficulty**. Then, especially if the delay is due to something larger, such as an unscheduled trip home, we can arrange alternative work or accommodations. A single assignment will not matter much in the long run; to help with this, we have the "Best in Show" grading policy which means that your other work will be upweighted to offset the potential harm due to a partially completed assignment, if it comes to that. Unforeseeable emergencies and circumstances can further override these policies. **Ask**.

Final Course Project

Final projects are ideally carried out in teams of two or three students. The final product of the project will consist of a short semi-formal paper and a presentation given to the class during reading week. Applying the course's methods to your own research is a fine project, as is doing methodological work. We will help you throughout the process of forming an idea through the finished work.

Statistical Programming

Statistical programming entails using a computer to analyze data in a way where you design and control the analysis and sometimes aspects of data generation (simulation). This is a step (or possibly several steps) beyond simply running built-in commands and interpreting output. It includes cleaning and transforming data to suit your needs, designing visualizations and plots to illustrate points you want to make, and conducting analyses that are tailor-made to your questions of interest.

The course will make extensive use of the statistical software package R, which runs on both PCs and Macs. The software is free and available online.[1] R is straightforward to learn, but is extremely powerful. It is used widely in many other statistics courses and also in research in such fields as education, psychology, economics, medical research, epidemiology, public health, and political science.

[1] See www.r-project.org

We highly recommend using RStudio, which makes using R easier. RStudio is an Integrated Development Environment (IDE) that structures your experience, helps keep things organized, and offers multiple time-saving features to make your programming experience better. You might also consider R Markdown. R Markdown allows for generating documents with embedded R code and R output in a clean format, which can greatly help report generation.

We will teach R and the associated tools as part of this course. We assume no background knowledge with statistical programming or any other programming.

An Agenda-Driven FAQ

Q: How do I prepare for lecture?

Read the Canvas Packet page for the lecture to get a sense of what the lecture is about. This will list what you need to do to prepare. Many of the readings can be read before or after lecture, as works best for you. Things that I will assume were read or done before lecture will be clearly marked as such.

Q: How do I review lectures?

Read the readings if you have not already. Also use the recorded video if that is helpful to review the lecture.

Q: Where are assignments posted?

We post all assignments on Canvas.

Q: Where do I submit assignments?

You submit all assignments on Canvas.

Q: Where do I getting help and ask questions?

There are three primary ways we recommend, other than your fellow students:

> **Quiet Questions/Chat**: Quiet questions and chat provide an alternative way for students to ask questions during lecture (you are also encouraged to ask questions directly). The TF will type back replies or ask me to delve into things as appropriate. The record of these are a resource to augment your notes.
>
> **Slack Discussion Forum:** This provides another way to engage in dialogue about the course material (and, in particular, problem sets and projects).
>
> **Office Hours**: Come to these! The small group setting proves invaluable for getting direct, targeted help.

Q: Where are the readings?

There are three types of readings: textbook, course handouts, and other third-party materials. The textbook is your physical textbook. Everything else is on the resources page, organized by unit. Most handouts are in the "handout textbook" which is posted online.

Q: Where are lecture slides, code from lecture, and demo data files?

On Canvas. See a packet for links to all material for that packet.

Q: Are the lectures recorded?

Yes, and the link to the recordings is on Canvas.

Q: Why R and not Stata?

A: There are several reasons to use R instead of Stata. These range from the practical to the pedagogical. An incomplete list of reasons are:

R is free. This might not matter much now, but it could matter depending on what future career you pick.

R is actually easier to use than Stata, especially if you want to do anything at all original (such as simulations).

R works a lot better with automatic report generation. Report generation is a nice way for coupling a data analysis with the write up, which makes reproducibility a snap. This is an increasingly important issue in the sciences.

Working in R develops better programming habits and relies less on the hacky solutions that Stata encourages.

R has a lot more packages than Stata, and the packages it has usually work better/do more than Stata's do. This means that you will have access to the cutting edge methods being developed, which gives you greater control over your data.

R is a very marketable skill. Stata is used in certain organizations, but R is quickly becoming a language of choice for research organizations around the world.

R allows you to be creative with your data. For example, it is better at making and customizing visual displays.

You can analyze multiple data sets of different formats and types at once. This allows for rapid exploration.

Using a flexible statistical environment is mind expanding.

Q: But I already know Stata. Do I need to learn R also?

A: I believe you should, but no. The RH&S book will give you the Stata code (we do not support Stata directly, however.) That being said, R allows for things Stata cannot easily do. Plus, all the course material is in R, and translating it will impede learning considerably. Also, knowing multiple languages for programming is itself important. Knowing multiple methods and approaches for answering a question encourages the separation of the content and the methods can deepen understanding.

Q: Will there be a lot of math in this course?

A: This is an applied course. We will endeavor to teach how to think about data and analyze it appropriately. To a certain extent this means talking about how to think about uncertainty in an appropriate way, pitfalls that one might encounter due to things such as sample size, and so forth. To make these conversations clear, we will use mathematical notation and formalism, but we will not be focused on mathematical derivations or asymptotic theory.

That being said, we will also occasionally discuss the mathematical underpinnings of the methods. For example, we will touch on maximum likelihood estimation, the machinery behind empirical Bayes, and so forth.

Q: Why is it important to learn quantitative methods in education (or anywhere)?

A: In these times it seems as if people who are comfortable with quantitative skills can hijack and railroad conversations about policy, science, or pretty much anything even when they are horribly, catastrophically wrong. Learning to not be intimidated and see what these arguments rest on is very important. On the flip side, because of the currently elevated status of quantitative argument, learning how to do it is important for defending one's ideas. Learning how to engage on this level is a means to empowerment. Also, on a good day data analysis can actually help answer questions about how the world works. That is always a very satisfying feeling.

Other Important Resources

Please see Canvas for the following:

- Assignment Style Guide: this reference gives specific requests for formatting work that makes grading much easier.
- Handout textbook: This "textbook" has almost all of the handouts of the course that we have written. It is an on-line textbook with a clickable table of contents.

Acknowledgements

This syllabus has traveled from Luke Miratrix to Joe McIntyre and back to Luke Miratrix. This syllabus has taken material from Andrew Ho's syllabus, which in turn borrows from an entire family tree of syllabi reaching back to courses of Judith Singer. It has been developed with generous input from many in the statistics department and school of education. Many arguments for R, below, originated with Joe McIntyre. Other philosophizing found in this syllabus is borrowed from earlier syllabi, interviews, and writings of Luke Miratrix.