S-040: Introduction to Applied Data Analysis

Harvard Graduate School of Education
Fall 2018

Syllabus

Class meets Tuesdays and Thursdays from 10AM to 11:30AM and from 11:30PM to 1PM in Larsen G08 (probably), the basement lecture hall at 14 Appian Way.

Instructor: Joe McIntyre (joseph_mcintyre@gse.harvard.edu) (please call me Joe)

Course overview

Often when quantitative evidence is being used to answer questions, scholars and decision-makers must either analyze empirical data themselves or thoughtfully manage and appraise the analyses of others. This course will cover the basic principles of analyzing quantitative data. Students will examine real data gathered to address questions in education, psychology, and other social science research settings, becoming acquainted with basic descriptive statistics, tabular and graphical methods for displaying data, the notion of statistical inference, and analytic methods for exploring relationships between variables. These topics together will provide students with a solid foundation for addressing research questions through statistical modeling using simple and multiple linear regression. There will be an emphasis on applying statistical concepts; in particular, how to: (1) select the appropriate statistical techniques for answering specific questions; (2) properly execute those techniques; (3) examine the assumptions necessary for the techniques to work appropriately; (4) interpret analytic results; (5) summarize the findings in a cogent manner; and (6) produce publication-style visual displays of results. Because quantitative skills are best learned through practice, computer-based statistical analyses will be an integral part of the course.

The assignments for this course will include: (1) several short problem sets involving the core concepts covered in class, (2) several longer assignments involving guided data analysis and the interpretation and reporting of research results, and (3) a final project involving a semi-structured data analysis and the interpretation and reporting of research results. The problems sets will be completed individually, while the longer assignments and final will be completed in pairs.

Our strategy will be to learn statistical analysis by *doing* statistical analysis. During the semester, we'll address a variety of substantive research questions by analyzing multiple data sets and fitting increasingly sophisticated regression models. As we learn how to use regression models in practice, we'll discuss their:

- **Purpose**. What types of substantive research questions are we answering?

- **Mathematical representation.** How does the model algebraically capture the relationship(s) we're trying to examine?

- **Assumptions**. What assumptions do we need to make to conduct a given analysis and trust its results? How do we determine whether these assumptions hold? What should we do when they don't?

- **Implementation**. How do we get the computer to do the calculations?

- **Interpretation**. How do we interpret the results? What inferences may we make? What inferences shouldn't we make?

- **Presentation**. How should we present results to a technical audience? To a non-technical audience?

- **Relationship to other statistical methods.** How is regression similar to and different from other methods you've learned or read about?

- **Implications for research design**. How should the next study be designed so that we'd be in better shape to address our research questions?

- **Limitations.** What are the limitations to our results, and how should we convey these to technical and non-technical audiences?

By the end of the semester, your statistical skills should be sufficiently developed that you can critically examine other people's analyses and carefully perform some of your own.

Prerequisites

No prior data analytic experience is required, but a working knowledge of basic algebra (GRE-level mathematics) is assumed, and some previous exposure to introductory statistics may be advantageous. This course is required for first-year Ph.D. students (those who feel this requirement should be waived based on their prior coursework in statistics should be in touch with Clara Lau at clara_lau@gse.harvard.edu) and recommended for any Ed.M. students wishing to enroll in a spring semester course that requires S-030 or S-040 as a prerequisite, such as S-052 or A-164. Permission of the Instructor is required. **Please consult with the Instructor if you have any questions about whether S-040 is right for you.**

Student survey

To give the teaching team a chance to get to know you, please begin the sign-up process by filling out this survey:

http://bit.ly/s040_2018_survey

NOTE that filling out this survey is optional but strongly recommended!

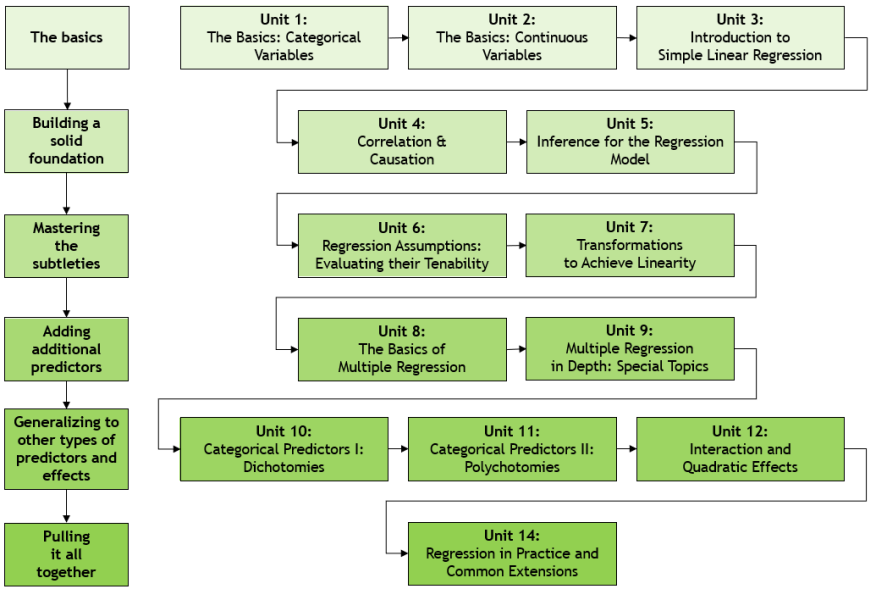
Course philosophy and content

Most of our time—both inside and outside of class—will be spent learning how to actually do data analysis. When we believe that your understanding will be enhanced by knowing something about the mathematical underpinnings of the techniques we use, we'll discuss them. Our goal will be to offer straightforward conceptual explanations that do not sacrifice intellectual rigor.

We will devote time to illustrating how to present results in words, tables, and figures. Good data analysis is craft knowledge; it involves more than using software to generate reams of output. Thoughtful analysis can be difficult and messy, raising delicate problems of model specification and parameter interpretation. We will confront such issues directly, offering concrete advice for sound decision making.

S-040 is structured around 13 learning units, shown below. Some units will take just one class session; others will take two or three. Each unit is supported by lectures and an in-class PowerPoint presentation, as well as a supplementary e-book chapter, which we will provide for you as part of your materials for each unit.[1] We also intend to offer a series of interactive visualizations for you to play with to deepen your understanding of the course material. We'll place each PowerPoint presentation on the course website at least 24 hours prior to the class session in which it will be used. **We strongly suggest that you download and/or print the presentations in advance of the relevant classes, bring them with you to class meetings, and plan on annotating them with detailed notes as you partake in lecture.**

Class participation is an important part of learning, even in a relatively large lecture course like S-040. If you have a question, it's likely that others do as well. We encourage active participation, however, if time is tight or a comment takes us too far astray, do not be offended if we defer your contribution to another time or place.



[1] E-book chapters are in development.

Course websites: http://bit.ly/s040_2018_section_1 (10-11:30) and http://bit.ly/s040_2018_section_2 (11:30-1)

Bookmark the course website and check it often. The website is our primary means of taking care of "housekeeping" matters (reducing the need to discuss deadlines, etc., in class). It also has resources designed to enhance your learning, including handouts, homework assignments, datasets, course notes, and web-based materials that help further explain statistical concepts.

Meeting times and the attendance policy

Consistent with HGSE policy, class begins at 10 minutes after the scheduled meeting and ends at the scheduled time. Please be seated and ready at the appointed time. **Attendance is mandatory and will be tracked throughout the semester; if you know you will be missing a class session, please be in touch with us to let us know ahead of time.**

Online class videos

Each class meeting will be taped, digitally encoded, and streamable online; videos are typically ready by the end of the day of each lecture. We provide the videos so that you can review the class material at your own pace. **Videos should supplement, not supplant, lectures. As noted before, attendance is mandatory. Sometimes the video capture technology doesn't work, and no video is recorded; don't rely exclusively on the videos!**

Professional behavior in a digital age

Many students bring laptops to class to take notes. **Non-course-related computer use (e.g., email, web surfing, Facebook, etc.) can be distracting to your classmates, so we request that you keep this behavior to an absolute minimum. Cell phones should be completely silenced, including loud vibrations, and they should not be used for texting in class.**

Statistical computing

Statistical computing is an integral part of S-040. To support your learning, the quantitative methods sequence at HGSE primarily uses Stata as its data analysis software. Although Stata will be the primary language of the course S-040 also offers support for R, the free statistical software. Most other statistics courses at HGSE are taught using Stata, so if you intend to continue to another class, you need to ensure that you're able to code using Stata, *even if you use R for S-040*. If you want to learn more about the different software options you have for this class, please speak with Joe.

We do not teach programming during class time, although code is threaded through the lecture slides. However, we provide resources to help you learn how to program on your own at your own pace. The Stata and R modules of the course website will provide a number of resources. Teaching Fellows will also cover coding issues in their sections. We also recommend this text for interested students: Kohler, U., & Kreuter, F. (2009). Data analysis using Stata (2nd ed.). College Station, TX: Stata Press. R has a number of free online textbooks available, but for a detailed introduction to the language, consider An Introduction to R by Venables, Smith, and the Core R Team. Another frequently overlooked resource is Google (or another search engine). **The truth is out there.**

There are two ways you can access Stata. The least expensive option is to use one of the networked workstations available on the HGSE campus (e.g., on the 2nd, 3rd, and 4th floors of Gutman Library). For students who would like to use Stata on their own PCs, you may purchase Stata following the links in the Stata Help section, or follow the link provided here: http://bit.ly/s040_stata. **Stata I/C will be sufficient for this course, and will be sufficient for most other classes at HGSE.**

R is free to download. If you choose to use R, we strongly encourage you to also download RStudio, an integrated development environment for R. RStudio is also free to download and use.

The course website has Stata and R code available for you to look at. This code produces all of the analyses we present in class. It's your responsibility to read and understand the code so that you can complete your assignments (though you're always welcome to ask questions about it).

Email correspondence

Students should check e-mail and the course website for updates, announcements, etc., from the Instructor and the TFs. The Instructor and TFs will make every effort to respond to email within ***2 working days***. If you have questions that require longer responses, you may be asked to attend office hours or schedule an in-person appointment. You will be assigned to a weekly section, and the Teaching Fellow who leads your section will be your initial "point person" for email correspondence. However, you are free to email and speak with any of the course TFs about anything related to your learning in the course. If you have an urgent question for the Instructor or TFs, please make the time-sensitive nature of your question apparent in the subject line of your email so we will know how to prioritize student requests. We cannot guarantee that we'll be able to respond quickly, but will try to do so. **If you have a question about an assignment or a question that may be of interest to other students, please use the Canvas discussion boards.**

Sections, office hours, and consultations with Instructor

Students will be required to attend weekly 1.5-hour sections led by the course Teaching Fellows. Timing and location of these sections will be determined during the first full week of class. There will be structured discussions and computing activities to enable a more interactive experience than is possible during lectures. Attendance is mandatory and the topics covered in section will be directly related to lecture, intended to supplement the teaching and learning that happens in class. There will be time in section to ask questions.

***Note: These are the times the TFs are available to help you. Please be respectful of their time and do not expect TFs to help with S-040 related matters when they are not on duty. You can ask, but they may not have time.**

In addition to section, there will be optional office hours available to meet with the Instructor. Students can sign up for office hour slots in the Scheduler section of the Calendar tab on the course website.

Finally, toward the end of the semester we will schedule one in-person consultation per pair of students to discuss your plans for the final project. These meetings will last approximately half

an hour, and we have found them to be extremely useful for students as they plan their final analyses.

Library guarantee

Joe will occasionally be found sitting in Gutman library. When he's in Gutman, he intends to be available to students. You may, therefore, feel completely free to approach him and ask him questions, even if he looks like he's really busy (honestly, he's probably just making memes). The only exception is if he's already meeting with someone, in which case you can just hover nearby until he's done.

Homework assignments

We believe that the best way to learn a statistical analytic technique is to ***do*** statistical analysis using the technique. To help you develop your skills, we will use a set of **homework assignments.** During the semester, we will assign up to six problem sets consisting of questions that generally require short, structured answers. We will also assign six data analytic assignments consisting of a research question, a dataset, and a sequenced set of questions that guide you through a complete statistical analysis. As part of those latter assignments, you will need to write and run a Stata program. You will also interpret the output and summarize your results in prose, tables and figures. Your writing should be clear and concise, integrating substance and statistics. To help focus your energies, we will indicate page limits and font-size guidelines for your assignments. **Further information on how we grade assignments will be posted on the course website.**

Problem sets will be posted no less than 1 week prior to the due date, and data analytic assignments will always be posted at least two weeks prior to the due date. Because we endeavor to read and return your assignments quickly, **all assignments must be turned in on time**. Late assignments will not be graded and will contribute zero points to your course grade. Exceptions to this rule may be made at the discretion of the Instructor only. If you find yourself in this position, please contact the Instructor *before* the deadline to discuss alternative arrangements. **Only the Instructor (and not the TFs) can grant extensions.**

Students often have questions about how their grades were assigned on problem sets and assignments. If you have a question, we ask that you do the following:

1) First, reach out to the TF who graded your assignment. Ask her or him to clarify any comments that you found confusing. Please keep in mind that TFs are almost never able to change grades after they have been assigned.
2) If, *after speaking to the TF*, you believe that your grade was unfair, you may contact the instructor for a regrade. The instructor will regrade the entire assignment, which may result in your grade on the assignment increasing, decreasing, or staying the same.

Tentative Assignment Schedule

Students will typically have assignments due each a week throughout the semester (we'll never have more than one assignment due in a week). Assignments will generally be due on Thursdays, with a digital submission deadline of 11:59 PM (though this may change depending on when the

instructional team finds time to meet). For the most part, we will alternate between problem sets and data analytic assignments. For each assignment, the instructor will make clear whether working in pairs is required or whether it is mandatory for students to work alone.

Because we want to be as responsive as possible to the pace of the course and to the unique collective learning and teaching experience that each semester brings, we do not want to present a list of specific due dates here. However, to help you plan your work load for the semester, we provide you with the estimated timing of the units and their corresponding assignments:

Unit	Unit 1:	Unit 2:	Unit 3:	Unit 4:	Unit 5:	Unit 6:	Unit 7:
Description	The Basics: Categorical Variables	The Basics: Continuous Variables	Introduction to Simple Linear Regression	Causation and Correlation	Inference for the Regression Model	Regression Assumptions: Evaluating their Tenability	Transfor-mations to Achieve Linearity
Length	~ 2 classes	~ 2 classes	~ 2 classes	~ 1 class	~ 2 classes	~ 2 classes	~ 2 classes
Problem set	**Problem Set A** (Individual)	**Problem Set B** (Individual)		**Problem Set C** (Individual)		**Problem Set D** (Individual)	
Assignment	**Homework 1** (Partners)	**Homework 2** (Partners)			**Homework 3** (Partners)		
	Final Project						

Unit	Unit 8:	Unit 9:	Unit 10:	Unit 11:	Unit 12:	Unit 14:
Description	The Basics of Multiple Regression	Multiple Regression in Depth: Special Topics	Categorical Predictors I: Dichotomies	Categorical Predictors II: Polychotomies	Interaction and Quadratic Effects	Regression in Practice and Common Extensions
Length	~ 2 classes	~ 1.5 classes	~ 1.5 classes	~ 1.5 classes	~ 2 classes	~ 2 classes
Problem set	**Problem Set E** (Individual)					
Assignment	**Homework 4** (Partners)		**Homework 5** (Partners)		**Homework 6** (Partners)	
	Final Project					

The course website contains a document with *tentative* assignment due dates.

Collaboration and study groups

Many people learn best when working in a group, and we encourage collaborative learning. Our primary goal in teaching S-040 is to help students improve their understanding of applied statistics and data analysis, and collaborative learning is a great way of achieving this goal. To mimic statistical work in the real world and to provide a chance for you to use statistical language actively, we mandate completion of data-analytic assignments in pairs throughout the course. **We will provide optional resources early in the semester to help students find data analysis partners. We will also communicate with pairs, at their request, throughout the semester to ensure that partnerships are working well for both individuals.**

We mandate collaboration for at least three reasons. First, learning statistics is like learning a language. To learn it, one must "speak" it actively and in a genuine context with other individuals. Second, collaborative statistical analysis is the norm and individual work is the exception in the world of statistical practice. Third, our experience has been that, on average, students who work in pairs both perform better and enjoy themselves more than students who work individually. Statistical collaboration is a case where the whole is greater than the sum of its parts.

Beyond pairs, study groups can be helpful to you as you prepare to do the assignments, both in terms of how to approach the work (including how to use the computer effectively) and in terms of how to think about important concepts. **However, students must turn in work as pairs or individuals, as specified explicitly on each assignment, not group work. Papers should be written in your own words—your text should reflect your own understanding of the material. If you discuss the assignment with other groups, please indicate whom you spoke with on your assignment.**

Each group will undoubtedly develop its own structure; nevertheless, here are a few suggestions:

- Groups with more than six members become less useful and may be harder to organize because finding common meeting times becomes increasingly problematic.

- Plan at least one session of 1½ to 2 hours for each homework assignment (early enough so that there is sufficient time if an additional session is necessary).

- Schedule the meetings so that you have sufficient time afterwards to write in pairs or individually. When we read your assignments, we focus on what you say and how you say it. The assignments have been devised to require not only computation and programming skills, but skills in analyzing and reporting the material.

- Use the groups to ask questions, try out interpretations, and so on. Often one person can explain something that makes you see something in a new way—or the other way around. Different people have different insights and strengths – some are good programmers, some ask good questions, others value contextual analysis—and you can learn from listening to what others in a group have to offer.

- **Be careful about sitting in groups at laptops or computers and simultaneously composing text.** You and your partner must write your own paper, on your own, using your own language. **Your papers should be written in your own words, not those of your study group.**

- Be sensitive to the distinction between collaboration to plan for and interpret the assignment and collaboration to write up the assignment. The former is encouraged; the latter is forbidden (besides, when applicable, your partner). If the distinction begins to feel murky, refocus your group's work on lecture content and course materials.

The problem of plagiarism

Please read the School's policy on plagiarism in the HGSE Student Handbook, which includes the statement, "Students who submit work either not their own or without clear attribution to the original source, for whatever reason, face sanctions up to and including dismissal and expulsion." Attention to this policy is particularly important in a course like S-040, in which collaboration with other students is encouraged. If you work closely with other students during the planning of your analyses—a process that we encourage and fully support—be sure to recognize the other students' contributions explicitly in your written account (a footnote is fine for this purpose). This helps avoid the natural questions that arise when similarities are detected at grading. **If you have any questions about what constitutes appropriate collaboration, or how to define what constitutes your own work, please see a Teaching Fellow. We are not trying to discourage you from working collaboratively or from getting help from the teaching team.**

Final project

The final project and accompanying data sets will be posted near the beginning of November, by which point you will have enough skills to at least formulate a research question. The final project will be due (uploaded to the S-040 website) **by midnight on the final day of the exam period.** Unlike the homework assignments, we do not structure this project with specific questions. Instead, we will give you some data sets and broad guidance for selecting a research question. You'll have the opportunity to pose your own research question, consult with the Instructor, develop an analytic plan, conduct the analyses, and report the results. As with assignments, final projects must be submitted on time. Extensions will not be granted, except in the case of personal emergency.

If none of the datasets are interesting to you, you may petition to use your own dataset.

Working with real data

S-040 provides students with the datasets they need to do the assignments. However, in real life analyses, you will often be responsible for finding, cleaning, and reshaping your own datasets. To help you learn to do this, if you're interested, S-040 offers detailed instructions showing you how to download and process the datasets we use for our assignments, as well as offering you self-checks to ensure that you've everything correctly. If you are considering a job which requires you to work with data, you may find it helpful to work through these resources.

Grades

You will be evaluated on the basis of your performance on the homework assignments (approximately one-half of your grade), problem sets (approximately one-quarter of your grade), the term project (approximately one-quarter of your grade), as well as attendance at lectures and sections (a small proportion of your grade). While we use arithmetic computations to arrive at a first approximation of your course grade, in the end, no individual assignment takes on undue weight, and the slope of individual trajectories is a factor we consider. We look at your whole portfolio of work when assigning course grades. Students may choose to take the course on a credit/no-credit basis. Satisfactory performance requires an average of B or better and completion of all assignments.

Accommodations

We encourage students needing accommodations in instruction or evaluation to notify us early in the semester. If you have a disability or health concern that may have some impact on your work in this class and for which you may require adjustments or accommodations, please contact Eileen Berger eileen_berger@gse.harvard.edu, Access and Disability Services (ADS) administrator in Gutman 124. No accommodations can be given without authorization from ADS, or without advance notice. If you already have a Faculty Contact Form for this course from ADS, please provide us with that information privately in our offices so that we can make those adjustments in a timely manner. All inquiries and discussions about accommodations will remain confidential.

Supplementary resources and texts

No books are required. The course website will contain supplemental resources. Students should be able to master the material by attending classes and sections, studying the accompanying slides, working (collaboratively) on the assignments, and using the other online resources, such as the e-book chapters that will be posted for each unit.

That said, many students report that they'd like to have a textbook for use both during the semester, to provide a different perspective on a topic being covered in class and/or for future reference. If you are interested in finding an (optional) text book, please see the instructor for recommendations.

As you'll soon see, S-040 is very writing intensive. To help support students who want to work on their writing skills, the following references and resources may be useful:

- HGSE Academic Writing Services in Gutman Library
- APA Online Tutorial: http://isites.harvard.edu/icb/icb.do?keyword=apa_exposed
- Writing Resources (including *Writing Like an Educator* Course and Reference Materials): http://isites.harvard.edu/icb/icb.do?keyword=awrs&pageid=icb.page48297
- Sign-up for Individual Sessions at the Writing Center: http://isites.harvard.edu/awrs

Personal request

Congrats on reading this far! This is a wicked long syllabus! But now I want to impose on you a little more. This course gets better when students help me to understand what works and what's missing. If you think there's a resource missing, please let me know. If you think an explanation was unclear, please tell me. Every year I learn a lot from students who are willing to share their opinions and perspectives with me.