# S-030: Intermediate Statistics:
# Applied Regression and Data Analysis

Harvard Graduate School of Education
Spring 2014

Class meets Monday and Wednesday from 8:30AM to 10:00AM, Larsen G08

Instructor: James Kim
Larsen Hall Room 505
james_kim@gse.harvard.edu
617-496-1517

Faculty Assistant: Melanie Reeves
Larsen Hall Room 509
melanie_reeves@gse.harvard.edu
617-384-5047

## Course overview

Are scores on high-stakes tests primarily a function of socioeconomic status? Do mandatory seat belt laws save lives? In this course, you will learn how to use a set of quantitative methods referred to as the general linear model--regression, correlation, analysis of variance, and analysis of covariance--to address these and other questions that arise in educational, psychological, and social research.

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 dozens of 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 research questions is the model most useful for addressing?
- **Mathematical representation.** How does the model algebraically capture the relationship(s) we're trying to examine?
- **Assumptions**. What assumptions need we make so that we can fit the model to data? 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 computer 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 cautions and caveats should we be aware of, and how should we convey these concerns 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

Everyone should have completed an introductory statistics course such as S-012. If your knowledge of basic statistics is rusty, please review the principles of estimation and inference in an introductory statistics textbook. There is also a Math Self-Test on the course website that represents the mathematical background that we expect of you (at the level of your quantitative GREs). **If you have not successfully completed an introductory statistics course, S-030 is not the right course for you. Please see me if you have any questions about whether your statistics background is sufficient.**

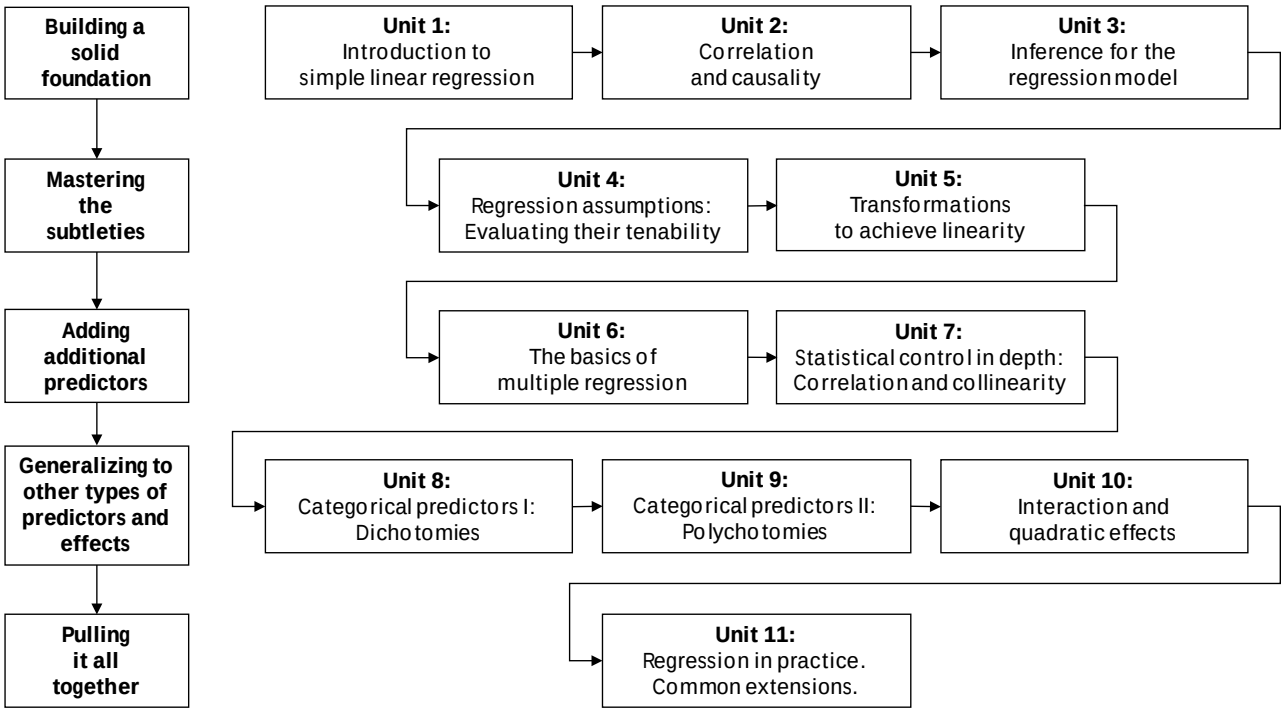
## Course philosophy and content

Most of our time—both inside and outside of class—will be spent learning how to do data analysis. When I believe that your understanding will be enhanced by knowing more about the mathematical underpinnings, I'll offer (what I hope are) 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'll confront such issues directly, offering concrete advice for sound decision making.

S-030 is structured around 11 learning units shown below. Some will take just one class session; others will take two or three. Instead of formal readings, each unit is supported by lectures and an in-class PowerPoint presentation. I'll place each PowerPoint presentations on the course website by **the morning of the class session in which it will be used**. **It is your responsibility to download and print the presentations in advance of the relevant classes and bring them with you to class meetings.**

Class participation is an important part of learning, even in a relatively large lecture course like S-030. If you have a question, it's likely that others do as well. I encourage active participation, and course grades will take into account students who make particularly strong contributions. 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.



**Course website: http://isites.harvard.edu/k101411**

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). It also has resources designed to enhance your learning, including handouts, homework assignments, datasets, 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 time (i.e., at 8:40 AM) and ends at the scheduled time (i.e., at 10:00 AM). Please be seated and ready at the appointed time. **I expect all students to attend every class meeting. Also, check your Harvard email for announcements.**

**Online class videos**

Each class meeting will be taped, digitally encoded, and streamable online; we endeavor to have the videos ready by the end of the day or midday the next day. I provide the videos so that you can review the class material at your own pace. **Please don't abuse the system: videos should supplement, not supplant, lectures.**

**Professional behavior in a digital age**

S-030 is technologically intensive and many students bring laptops to class to take notes. Personally, I'd find it difficult to take notes online because the notes I'd be writing would likely be less text-based and more graphical (equations, graphs and other sketches), but I'll leave that up to you. **What I do expect is professional behavior—that means no email, web surfing, instant messaging, or any other electronic activity during class.** It's not only rude, it's distracting to your classmates. If you will use a laptop, please sit to the sides of the lecture hall. The center of the lecture hall will be reserved as a laptop-free zone. This should go without saying, but 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-030. To support your learning, the quantitative methods sequence at HGSE uses Stata for Windows. I assume that everyone is comfortable using a computer to perform basic statistical analysis, although I don't assume that you've used Stata.

I do not teach programming during class time, although code is threaded through the lecture slides. We provide resources to help you learn how to program on your own at your own pace. The **Stata Help** section of the course website will provide a number of resources. Teaching fellows may also cover coding issues in their sections. If a reference is desired, I recommend this text: Kohler, U., & Kreuter, F. (2009). Data analysis using Stata (2nd ed.). College Station, TX: Stata Press. You can search for prices here: http://www.addall.com/New/compare.cgi?dispCurr=USD&isbn=1597180467

There are two ways you can access Stata. The least expensive option is to use one of the networked workstations available in the Learning Technology Center (LTC) on the 3rd floor of Gutman Library and elsewhere on the HGSE campus (e.g., on the 2nd 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. Stat/IC, which will be sufficient for this course, is available for $98 for a year-long license and $179 for a perpetual license. The so-called "Small Stata" is also sufficient and costs only $49 a year, although we recommend investing in a more substantial version.

## Homework assignments

I believe that the only way to learn how to conduct statistical analysis is to conduct statistical analysis. To help you develop your skills, we will administer and grade **six homework assignments.** Each assignment is a carefully constructed exercise consisting of a research question, a dataset, and a sequenced set of questions that guide you through a complete statistical analysis. As part of each assignment, you'll need to write and run a Stata program. You'll also interpret the output and summarize your results in prose, tables and figures. We grade assignments using a holistic approach—we are interested in not only whether you get the "right answer" but also in your reasoning and presentation. Your writing should be clear and concise, integrating substance and statistics. To help focus your energies, we indicate page limits and expect that you will adhere to them. **Further information on how we grade assignments will be found on the Assignments page of the course website.**

Here is a tentative schedule for homework assignments; these dates may well change. All assignments must electronically submitted by the date and time specified. Late assignments will not be graded and will contribute 0 to your course grade, so submit with time to spare in anticipation of unforeseen technical issues. To avoid last-minute panicking, I strongly encourage you to have the assignment complete or near-complete by the class period prior to the due date.

**Tentative Assignment Schedule**

| Assignment | Available on or about | Due by 1PM on | Collaboration Format |
|---|---|---|---|
| Assignment #1 | Monday, February 3 | Friday, February 14 | Pairs Mandatory |
| Assignment #2 | Monday, February 17 | Friday, February 28 | Pairs Mandatory |
| Assignment #3 | Monday, March 3 | Friday, March 14 | Individual Mandatory |
| Assignment #4 | Monday, March 17 | Friday, March 28 | Pairs Mandatory |
| Assignment #5 | Monday, March 31 | Friday, April 11 | Pairs Mandatory |
| Assignment #6 | Monday, April 14 | Friday, April 25 | Pairs Mandatory |
| Final Project | Monday, March 31 | Friday, May 9 | Individual Mandatory |

## Collaboration and study groups

Many people learn best when working in a group, and I encourage collaborative learning. My primary goal in teaching S-030 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, I mandate completion of assignments in pairs throughout the majority of the course, excepting only the third assignment and the final project.

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, my experience has been that, on average, students who work in pairs and groups 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 where specified above, not group work. Papers should be written in your own words—your text should reflect your own understanding of the material.**

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

- Groups with six or more 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 (early enough so that there is sufficient time if an additional session is necessary). After 2 hours of statistics, everyone's eyes will be glazing over.

- 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—you each represent each others' resources. 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 beyond, 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, ordinarily will be dismissed from the Harvard Graduate School of Education." Attention to this policy is particularly important in a course like S-030, in which collaboration with other students is encouraged. If you work closely with other students during the planning of your analyses—a process that I encourage and fully support—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.**

**I cannot overemphasize the need for all students to monitor their own behavior. Assignments are structured such that you can receive feedback on *your understanding of the material.* The consequences for plagiarism are appropriately severe.**

## Final project

The final project will be posted on **Monday, March 31** and is due **Friday, May 9, 2014.** Unlike the homework assignments, I do not structure this project with specific questions. Instead, I will give you some general background and a broadly stated research question. You'll have the opportunity to refine this research question further, develop an analytic plan, conduct the analyses, and report the results. As for assignments, final projects must be submitted on time. Extensions will not be granted, except in the case of personal emergency.

**Grades**

You will be evaluated on the basis of your performance on the homework assignments (approximately two-thirds of your grade) and the term project (approximately one-third 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 satisfactory/unsatisfactory basis. Satisfactory performance requires an average of B or better and completion of all assignments.

**Accommodations**

Students needing accommodations in instruction or evaluation must notify me early in the semester, and HGSE's policies must be followed. Late requests for accommodations will not be honored unless there is a pressing reason, such as a recent injury.

**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, studying the accompanying slides, watching the online Stata tutorials, working (collaboratively) on the assignments, attending TF sections, and using the other online resources.

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. To provide this perspective, I've ordered: Kleinbaum, D.G., Kupper, L.L., Nizam, A., & Muller, K.E. (2008). *Applied Regression Analysis and Other Multivariable Methods*, 4th ed. (Pacific Grove, CA: Duxbury Press). This is one of several "standard" applied regression textbooks for graduate students in the social sciences. Statistics books tend to be very expensive so be sure that you'll use this book before buying it. If you have another of the "standard" regression textbooks (e.g, Mendenhall & Sincich; Kutner, Nachtsheim, & Neter; a previous edition of this book, or any other book that has regression in the title), you may find another one unnecessary.

If you're looking for an alternative explanation of something covered in class, you might want to try one of the following two texts.

- Agresti, A., & Finlay, B. (2009). *Statistical methods for the social sciences* (4th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

- Allison, P.D. (1999). *Multiple Regression: A Primer*. (Thousand Oaks, CA: Pine Forge Press). Very well written with the same data analytic emphasis as S-030, but at a much more **elementary** level. A good place to look for relatively simple explanations.

- DeMaris, A. (2004). *Regression with Social Data: Modeling Continuous and Limited Response Variables.* (New York: Wiley). Very well written with the same data analytic emphasis as S-030, but at a much more **advanced** level. A good place to look for extensions of what we cover.

As you'll soon see, S-030 is very writing intensive. To help support students who want to work on their writing skills, the following references and resources may be useful:

- Williams, J.M. (1995). *Style: Toward Clarity and Grace*. (Chicago: University of Chicago Press). Who wouldn't want to write with clarity and grace? More a philosophical treatise than a list of dos and don'ts, but a book well worth reading.

• Miller, J.E. (2004). *The Chicago Guide to Writing about Numbers.* (Chicago: University of Chicago Press). Although I disagree with some of Miller's advice (especially about graphics), I still think this book can be helpful for students first learning how to write about quantitative research.

- 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://www.appointmentquest.com/provider/2030159020