S-052: Intermediate Statistics:
Applied Regression & Data Analysis

Harvard Graduate School of Education
Spring 2014

Class meets Tuesdays and Thursdays from 10:10AM to 11:30AM.

Instructor team

Courtney Pollack | Samuel Ronfard | Darrick Yee

Email: s052@gse.harvard.edu
Office: Gutman 416
Office hours: By appointment

Course overview

Welcome to *S-052: Applied Data Analysis*. This course is an integrated continuation of the fall course, *S-040*, and is part of the *HGSE* school-wide network of courses in quantitative methods. The *S-040* and *S-052* courses form the cornerstone of a sequence of courses in applied statistical methods for consumers and producers of rigorous educational, social, and psychological research.

The course is designed to develop and extend the data-analytic skills that you began to acquire in earlier courses and to help you learn to communicate your findings clearly to audiences of other empirical researchers, scholars, policy-makers, practitioners, students, and parents. We have designed *S-052* to contribute to the diverse data-analytic toolkit that you will need in order to perform sensible and believable analyses of complex educational, psychological, and social data.

Core topics such as multiple regression analysis, introduced earlier in our sequence, continue to be the foundation of *S-052*. However, we extend your use of these techniques to cover a wider variety of conditions encountered in the world of real data-analysis, including selected multivariate methods. A listing of major course topics is provided later in this document.

True to its name, *S-052* is an *applied* (not a *technical*) course in which you will learn by observing and engaging in the authentic activities of real applied data analysis. The use of new statistical techniques will be "modeled" in class, and then you will be asked to apply these new techniques to real problems using real data in "data-analytic memos" and a take-home examination.

You will also be asked to interpret the outcomes of your data-analyses in words, and to communicate these interpretations clearly and concisely in writing. Demonstrating that you have the computer skills necessary for good data analysis is an integral part of the course

Presentational structure

As a rough guide, presentation of each new statistical technique (or each extension of a statistical technique that you already know) will contain *seven components*. For each technique, we will:

- ***Provide A Relevant Research Question***. All analytic methods are secondary to, and exist because of, questions of substantive importance. So, stating the specific research question that is to be addressed is a critical precursor to any effective data-analysis.

- ***Obtain a Suitable Dataset***. For each new method described, we will provide one or more real datasets and use them to address the research question. We will introduce each dataset by describing its origins and by providing connections to related Internet resource materials where they are available. Our presentations may also include comments on the utility and reasonableness of the research design that led to the data-collection.

- ***Specify an Appropriate Statistical Model***. The basis of any effective data analysis is the specification of a statistical model that represents credibly the substantive process under investigation and embodies the researcher's hypotheses. We will describe the specification of such statistical models and the meaning and role of critical parameters in the model in terms of the stated research question.

- ***Fit the Statistical Model to Data***. Fitting the specified statistical model to data will provide the vehicle by which we will discuss the application and functioning of appropriate statistical methods. We will also comment on elements of computer programming for data analysis, in this case using the statistical program, Stata.

- ***Describe and Assess the Assumptions Underlying the Statistical Method***. All analytic methods are underpinned by assumptions. In fact, the particular assumptions underlying a statistical technique effectively constitute an additional source of information that is "input" implicitly by the technique into the data-analysis. This makes the adequacy of the assumptions essential to check. If they are violated, then the additional "information" that they have passed into the analysis will be incorrect and the findings dubious. We will describe the assumptions that underpin each new technique and illustrate how the credibility of the assumptions can be assessed using relevant diagnostics.

- ***Estimate, Test and Interpret Central Parameters of the Statistical Model***. At the end of any data analysis, estimates of model parameters and their associated tests provide the formal answers to the research questions. We will describe the estimation, testing, and interpretation of critical model parameters, along with the use of appropriate statistics for summarizing model goodness-of-fit.

- ***Answer the Research Question***. At its end, statistical analysis is not worth anything if its findings cannot be translated into common sense conclusions for an audience of intelligent consumers, whether they are other scholars, practitioners and policy makers, or parents and students. We will emphasize the authoring of cogent substantive interpretations of findings, based on the careful identification and translation of relevant parts of the data-analytic output.

Finally, and on a "need-to-know" basis, we will outline selected technical details of the data-analyses, connect new techniques with existing methods, and provide ways of dealing with the limitations of each new technique in practice.

In the rest of this document, we list the course content in greater detail. Important details of course are described on the course website.

Overview of course content (Subject to minor adjustment)

1. Fitting Sensible Taxonomies of Multiple Regression Models:

(a) *Deciding Which Regression Models To Fit*. Addressing research questions by fitting taxonomies of multiple regression models, and determining a sensible "final" model. Reviewing the specification of regression models in which a single substantive construct is represented by a system of "dummy" predictors. Reviewing the notion of a statistical interaction between predictors. This introductory section provides an opportunity to review your prior learning about multiple regression analysis.

(b) *Testing Complex Hypotheses About Regression Parameters*. Comparing nested multiple regression models. Using the *General Linear Hypothesis Test* (ΔR^2 *test*) to conduct formal tests of the joint impact of several predictors simultaneously on an outcome.

(c) *Detecting Influential Data-Points, and Assessing Their Impact on Model Fit*. Introducing the notion of influence statistics. Conducting sensitivity analyses.

(d) *Checking the Assumptions On The Residuals*. Understanding the importance of the assumptions on the residuals in a multiple regression analysis. Graphical methods for assessing distributional assumptions on the residuals, including the *Normal Quantile Plot* (*NQP*) and its companion *Wilks-Shapiro* test.

(e) *Interpreting Findings*. Using fitted plots to display and interpret the size and direction of detected effects for prototypical individuals in the population, especially in the presence of statistical interactions.

2. Regression Analysis When the Outcome/Predictor Relationship Is Non-Linear:

(a) *Dealing With Non-Linear Outcome/Predictor Relationships*. Using power transformations to linearize the outcome/predictor relationship. Introducing Tukey's *Ladder of Transformations* and the *Rule of the Bulge*.

3. Basic Logistic ("Binomial Logit") Regression Analysis:

(a) *Modeling the Relationship Between a Dichotomous Outcome and Predictors using a Linear Probability Model*. The problematic impact of specifying a linear regression model when the outcome is dichotomous – appropriate interpretation of fitted values and problems in the residual distribution.

(b) *Modeling the Relationship Between a Dichotomous Outcome and Predictors with Logistic Regression ("Logit") Analysis*. Using a non-linear logistic (or "logit") function to represent the hypothesized relationship between a dichotomous outcome and predictors. Goodness-of-fit statistics for logistic regression analysis.

(c) *Fitting Taxonomies Of Nested Logistic Regression Models*. Addressing research questions about the prediction of dichotomous outcomes by fitting and comparing nested logistic regression models using a *General Linear Hypothesis* (Dc^2) *Test*.

(d) *Interpreting Fitted Logistic Regression Models*. Using fitted *odds* and *odds-ratios*, and fitted trend lines plotted for prototypical individuals in the population, to demonstrate the size and direction of an effect detected via logistic regression analysis.

4. Extensions of the Basic Logistic Regression Approach:

(a) *Discrete-Time Survival Analysis*. Using logit analysis to examine the occurrence and timing of events in a person's life. Introducing the concepts of hazard and survivor probability, and the discrete-time hazard model. Using prototypical fitted hazard and survivor functions, and predicted median lifetimes, to interpret findings.

5. Regression Analysis When the Residuals Are Not Independent:

(a) *Introducing the Multilevel Regression Model*. Using a multilevel "random intercepts" regression model to account for the grouping of individuals within higher-level "units." Fitting the multilevel model using random-effects regression analysis. Partitioning residual variance into its within-group and between-group components, estimating and interpreting the intra-class correlation.

(b) *Using the Multilevel Regression Model to Analyze Longitudinal Data*. Using the multilevel model to analyze individual change over time. The issue of residual auto-correlation over time within-person. Handling more complex assumptions about residual autocorrelations over time.

(c) *Using the Multilevel Regression Model To Obtain Internal Consistency Estimates of Test Reliability*. Using replicate measurements on a construct, and the multilevel model, to obtain internal consistency, test-retest, and parallel-forms estimates of reliability.

(d) *The Notion of a "Fixed Effects" Multilevel Model*. A useful and robust alternative to the random-intercepts multilevel model. Implications and interpretation of the new specification.

6. Forming Composites from Multiple Indicators of a Construct:

(a) *Classical Methods For Compositing Multiple Indicators Of A Construct*. Traditional strategies for forming data-composites -- standardization of indicators, creating a weighted linear composite. Measurement error and internal-consistency reliability (Cronbach's alpha).

(b) *Using Principal Components Analysis (PCA) To Form An "Ideal" Data-Composite*. Introducing principal components analysis as an alternative to classical item-analysis in data-compositing. Creating a data-composite with maximum variance.

(c) *Using PCA To Peek Inside the Multivariate Structure Of Data*. Using the eigenvalues and a scree plot to estimate how many "dimensions of information" underlie a given set of indicators. Interpreting the underlying dimensions numerically, graphically and substantively.

7. Causal Inference:

(a) *Randomized Experiments*. Defining "cause" and "effect" using Rubin's potential outcomes framework. Using OLS regression to identify causal effects of treatment. Illustrating bias in observational data as a violation of residual assumptions.

(b) *Natural Experiments*. Defining a "natural experiment." Using arguably exogenous assignment to identify treatment effects via OLS. Brief overview of difference-in-differences, regression-discontinuity, and instrumental variables designs.

(c) *Selection on Observables*. Describing how to satisfy identifying assumptions using observed data. Using basic subclassification to identify treatment effects. Brief introduction to using propensity scores for subclassification and to correct for selection bias.

Prerequisites

Everyone should have completed an intermediate statistics course such as S-030 or S-040. If your knowledge of multiple regression is rusty, please review the principles of estimation and inference in an introductory regression textbook. **If you have not successfully completed a course that covers multiple regression, S-052 is not the right course for you.** You should be comfortable, for example, with the inclusion and interpretation of an interaction term in a regression model. **Please see us if you have any questions about whether your statistics background is sufficient.**

Course participation

Most of our time—both inside and outside of class—will be spent learning how to do data analysis. When we believe that knowing more about the mathematical underpinnings will enhance your understanding, we'll offer (what we 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.

Class participation is an important part of learning, even in a relatively large lecture course like S-052. If you have a question, it's likely that others do as well. We 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 we defer your contribution to another time or place.

Course website: http://isites.harvard.edu/icb/icb.do?keyword=k96469&login=yes

Bookmark the course website and check it often (especially in advance of every class and sometimes more frequently). The website is our 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 10:10 AM) and ends at the scheduled time (i.e., at 11:30 AM). Please be seated and ready at the appointed time. **We expect all students to attend every class meeting and arrive on time.**

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. We 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-052 is technologically intensive and many students bring laptops to class to take notes. Personally, we'd find it difficult to take notes online because the notes we'd be writing would likely be less text-based and more graphical (equations, graphs and other sketches), but we'll leave that up to you. **What we 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 try to sit to the sides of the lecture hall or toward the back of the classroom. The center-

front 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 cell phones should not be used for texting in class.

Statistical computing

Statistical computing is an integral part of S-052. To support your learning, the quantitative methods sequence at HGSE uses Stata for Windows. We assume that everyone is comfortable using a computer to perform basic statistical analysis, although we don't assume that you've used Stata.

We 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. Sections of the course website provide a number of resources. Teaching fellows may also cover coding issues in their sections. If a reference is desired, we recommend this text: Kohler, U., & Kreuter, F. (2012). Data analysis using Stata (3rd ed.). College Station, TX: Stata Press. You can search for prices here: http://www.addall.com/New/compare.cgi?dispCurr=USD&isbn=1597181102

There are two ways you can access Stata. The least expensive option is to use one of the networked workstations available on the 2nd, 3rd, or 4th floors of Gutman Library. For students who would like to use Stata on their own PCs, you may purchase Stata following this link: http://isites.harvard.edu/icb/icb.do?keyword=research_technologies&pageid=icb.page295634 . Stata/IC, which will be sufficient for this course, is available for $98 for a year-long license and $179 for a perpetual license. Note that "Small Stata" is not sufficient for this course.

Homework assignments

We believe that the only way to learn how to conduct statistical analysis is to actually conduct statistical analysis. To help you develop your skills, we will administer and grade **seven homework assignments.** A tentative schedule for homework assignments follows; 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, we 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 5PM on	Collaboration Format
DAM 1	Tuesday, January 28	Friday, February 7	Pairs
DAM 2	Friday, February 7	Friday, February 21	Pairs
DAM 3	Friday, February 21	Friday, February 28	Pairs
DAM 4	Friday, February 28	Friday, March 14	Pairs
DAM 5	Friday, March 14	Friday, April 4	Pairs
DAM 6	Friday, April 4	Friday, April 18	Pairs
DAM 7	Friday, April 18	Friday, May 2	Pairs
Final Exam	Monday, May 12	Wednesday, May 14	Individual

Collaboration and study groups

Many people learn best when working in a group, and we encourage collaborative learning. Our primary goal in teaching S-052 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 assignments in pairs throughout the course, excepting only the final exam.
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 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 other's 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-052, 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—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 one of the instructors or a Teaching Fellow.**

> **We 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 Exam

The final exam is a two-day affair that will be posted during exam week. It will run from Monday, May 12 through Wednesday, May 14. As with 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 one half of your grade) and the final exam (approximately one half 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 us 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, utilizing the additional resources on the iSite, working (collaboratively) on the assignments, attending TF office hours or special sections, and using the other online resources.