Political Science 531
Quantitative Political Analysis II
Linear Models and Statistical Inference

Jake Bowers
jwbowers@illinois.edu
Moodle: http://cho.pol.uiuc.edu/moodle/course/view.php?id=7

Spring 2011

General Information

Moodle enrollment key: ps531. I will be distributing readings and assignments on the Moodle.

Where/When We meet Mondays, 1:30–3:30pm in 126 Wohler's Hall.

Office Hours Thurs 1-3pm by appointment in 231 CAB or other times by appointment. If you know in advance that you want to come to office hours, please email me to reserve a 20 minute slot.

Overview

What does it mean to say "statistically significant" referring to the outcome of a linear regression? When is it reasonable to say this? When is it confusing?

Why can we report that a 95% confidence interval contains some set of plausible values for a quantity of interest? When would we mislead ourselves and others with such claims?

In your last course you practiced fitting linear models to data and gained the computational and conceptual foundation for answering such questions. In this course, you will deepen your understanding of statistical inference and estimation using linear models. This is a course in applied statistical theory. The approach here is to work toward understanding statistical theory by application. As such we will emphasize the hard work of writing computer programs rather than the hard work of proving theorems. After the hard work required by this class you will have developed strategies for answering the questions posed above and thus will be well-positioned to use linear models with confidence and creativity and good judgement.

Goals and Expectations

This class aims to help you learn to think about the linear model.

The point of the course is to position you to do the future learning that is at the core of your work as an academic analyzing data.

I also hope that this course will help you continue to develop the acumen as a reader, writer, programmer and social scientist essential for your future daily life as a social science researcher.

Expectations First and foremost, I assume you are eager to learn. Eagerness, curiosity and excitement will impel your energetic engagement with the class throughout the term. If you are bored, not curious, or unhappy about the class you should come and talk with me immediately. Graduate school is not the place to waste your time on courses that are not important to you.

Second, I assume you are ready to work. Learning requires work. As much as possible I will link practice directly to application rather than merely as a opportunity for me to rank you among your peers. Making work about learning rather than ranking, however, will make our work that much more difficult and time consuming. You will make errors. These errors are opportunities for you to learn — some of your learning will be about how to help yourself and some will be about statistics. If you have too much to do this term, then again, consider dropping the course. Graduate school is a place for you to develop and

begin to pursue your own intellectual agendas: this course may be important for you this term, or it may not. That is up for you to decide.

Third, I assume you are willing to go along with my decisions about the material and sequence. I will be open to constructive and concrete suggestions about how to teach the class as we go along, and I will value such evaluations at any point in the class. That said, if you do not think you need to take this course, then don't take it.

Fourth, I assume some previous engagement with high school mathematics, probability and statistical computing in R (see, for example, the syllabus for PS530 as taught last term).

Rules There aren't many rules for the course, but they're all important. First, read the assigned readings before you come to class. Second, turn everything in on time. Third, ask questions when you don't understand things; chances are you're not alone. Fourth, don't miss class or section.

All papers written in this class will assume familiarity with the principles of good writing in Becker (1986).

All final written work will be turned in as pdf files. I will not accept Microsoft, Apple, OpenOffice, or any other proprietary format. Work turned in using those formats will not be looked at and subsequent pdf files will be considered late work.

Late Work I do not like evaluation for the sake of evaluation. Evaluation should provide opportunities for learning. Thus, if you'd prefer to spend more time using the paper assignment in this class to learn more, I am happy for you to take that time. I will not, however, entertain late submissions for the subsidiary paper assignments that are due throughout the term. If you think that you and/or the rest of the class have a compelling reason to change the due date on one of those assignments, let me know in advance and I will probably just change the due date for the whole class.

Incompletes Incompletes are fine in theory but terrible at this university in practice. I urge you to avoid an incomplete in this class. If you must take an incomplete, you must give me *at least* 2 months from the time of turning in an incomplete before you can expect a grade from me. This means that if your fellowship, immigration status, or job depends on erasing an incomplete in this class, you should not leave this incomplete until the last minute.

Participation We will be doing hands-on work nearly every class meeting. I will lecture very little and instead will pose problems of statistical theory, research design, and data, which will require us to confront and apply the reading that prepared us for the day's work. I anticipate that you'll work in small groups and that I'll circulate and offer help when I can. I will break away to draw on the board or demonstrate on my own computer now and then if everyone is running into the same problem.

Papers Other than the reading, the main assignment for this term is for you to write a paper that assesses the usefulness/reasonableness of the standard linear model (and associated pre-packaged statistical inference measures) for your own work.

The ideal paper will take the pre-packaged regression table which I presume would be the centerpiece of a paper that you wrote last term, or which you are writing this term in some other class and would deeply scrutinize this table using simulations and theory: do you 95% confidence intervals reject true null hypotheses 5% of the time or less? how biased are your calculations of effects/slopes and/or standard errors? what is your target of inference? why is that target reasonable and/or useful?

After doing this task you will be prepared to assess any other paper you write in the future — you will know when, how, and why your linear model is or is not persuasive as an engine of statistical inference.[1]

[1] Causal inference from the linear model will only be touched on in this class. But, it will be a central concern of your class in research design as well as the third course in this series.

Grades I'll calculate your grade for the course this way: 50% class involvement and participation and 50% the paper.

Because moments of evaluation are also moments of learning in this class, I do not curve. If you all perform at 100%, then I will give you all As.

Involvement Quality class participation does not mean "talking a lot." It includes coming to class; turning in assignments on time; thinking and caring about the material and expressing your thoughts respectfully and succinctly in class.

As much as possible, we will be working in groups during the class meetings. This work will require that you have done the assigned reading in advance and that you are an active collaborator.

Computing We will be using R in class so those of you with laptops available should bring them. Of course, I will not tolerate the use of computers for anything other than class related work during active class time. Please install R (http://www.r-project.org) on your computers before the first class session.

Computing is an essential part of modern statistical data analysis — both for turning data into information and for conveying that information persuasively (and thus transparently and reliably) to the scholarly community. In this course we will pay attention to computing, with special emphasis on understanding what is going on behind the scenes. You will be writing your own routines for a few simple and common procedures.

Most applied researchers use two or three computing packages at any one time because no single language or environment for statistical computing can do it all. In this class, I will be using the R statistical language. You are free to use other languages, although I suspect you will find it easier to learn R unless you are already a code ninja in some other language that allows matrix manipulation, optimization, and looping.

As you work on your papers, you will also learn to write about data analysis in a way that sounds and looks professional by using either a WYSIWYG system like Word, OpenOffice, or Wordperfect, or a typesetting system like LaTeX, to produce documents that are suitable for correspondence, collaboration, publication, and reproduction. No paper will be accepted without a code appendix or reproduction archive attached (or available to me online). No paper will be accepted unless it is in Portable Document Format (pdf).[2] No paper will be accepted with cut and pasted computer output in the place of well presented and replicable figures and tables. Although good empirical work requires that the analyst understand her tools, she must also think about how to communicate effectively: ability to reproduce past analyses and clean and clear presentations of data summaries are almost as important as clear writing in this regard.

Books

I'm am requiring fewer books rather than more. The readings will be drawn from a variety of sources. I will try to make most of them available to you as we go if you can't find them easily online yourselves.

Required Fox, J. (2008a). *Applied regression analysis and generalized linear models*. Sage. [3]

Achen, C. H. (1982). *Interpreting and Using Regression*. Sage, Newbury Park, CA.

Fox, J. and Weisberg, S. (2011). *An R Companion to Applied Regression*. Sage. [4]

[2] Actually, I'm willing to consider HTML or Postscript although practice with pdf will help you most in submitting papers to journals and other forms of scholarly communication.

[3] For additional materials and appendices see http://socserv.socsci.mcmaster.ca/jfox/Books/Applied-Regression-2E/index.html

[4] http://socserv.socsci.mcmaster.ca/jfox/Books/Companion/index.html

Recommended No book is perfect for all students. I suggest you ask around, look at other syllabi online, and just browse the shelves at the library and used bookstores to find books that make things clear to you. Here are some recommendations:

Books much like Fox (2008a) with slightly different emphases and more R in the text:

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.[5] This book has a really nice few chapters on causal inference and on post-estimation model exploration and interpretation as well as many excellent chapters on multilevel models.

Lancaster, T. (2004). *An introduction to modern Bayesian econometrics*. Blackwell Pub. This book is a nice introduction to Bayesian inference (in addition to Gelman and Hill, which is also an introduction to Bayesian inference without being as explicit about it). Come and talk with me if you'd like pointers to more of the Bayesian literature.

Trosset, M. W. (2009). *An Introduction to Statistical Inference and Its Applications with R*. CRC Press. This book represents a nice modern take on what you'd learn in your first or second course in a statistics department. The linear model plays a relatively small role. However, the coverage of frequentist theory is very nicely done.

If you'd like books that more closely link the statistics with R:

Faraway, J. (2005). *Linear Models With R*. CRC Press

Faraway, J. (2006). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. CRC Press

Verzani, J. (2005). *Using R for Introductory Statistics*. Chapman & Hall/CRC

If you'd like different perspectives on the material and perhaps a bit less math I *highly* recommend the following books. I love them!

These books are particularly good to help you get clear on the fundamental concepts of statistical inference: what it means to test a hypothesis, construct a confidence interval, etc...

Berk, R. (2004). *Regression Analysis: A Constructive Critique*. Sage

Freedman, D., Pisani, R., and Purves, R. (2007). *Statistics*. W.W. Norton, New York, 4th edition

Gonick, L. and Smith, W. (1993). *The cartoon guide to statistics*. HarperPerennial New York, NY

Kaplan, D. (2009). *Statistical Modeling: A Fresh Approach*. http://www.macalester.edu/~kaplan/ism/[6]

If you'd like more math and theory try these:

Cox, D. R. (2006). *Principles of statistical inference*. Cambridge University Press, Cambridge. This is one of my favorite books on statistical theory at the moment.

Rice, J. (2007). *Mathematical Statistics and Data Analysis*. Duxbury Press, Belmont, CA, 3rd edition. This is commonly assigned for first year statistics ph.d. students.

Greene, W. H. (1997). *Econometric Analysis*. Prentice Hall, 3rd edition (Or any edition of Greene.). This is commonly assigned for first year economics ph.d. students.

Angrist, J. and Pischke, J. (2009). *Mostly harmless econometrics: an empiricist's companion*. Princeton Univ Pr

Kennedy, P. (2003). *A guide to econometrics*. The MIT Press Other editions of this surely exist.

[5] http://www.stat.columbia.edu/~gelman/arm/
[6] http://www.macalester.edu/~kaplan/ism/

Math books

You should also have at least one math book on your shelves. Some general recommendations for books that combine linear algebra and calculus among other topics:

Chiang, A. C. (1984). *Fundamental Methods of Mathematical Economics*. McGraw-Hill/Irwin; 3rd edition (February 1, 1984)

Fox, J. (2008b). *A mathematical primer for social statistics*. SAGE Publications Inc

Gill, J. (2006). *Essential mathematics for political and social research*. Cambridge Univ Pr

Simon, C. P. and Blume, L. (1994). *Mathematics for Economists*. W.W. Norton, New York, NY

Self-Help

If you discover any books that are particularly useful to you, please alert me and the rest of the class about them. Thanks!

Schedule

Note: This schedule is preliminary and subject to change. If you miss a class make sure you contact me or one of your colleagues to find out about changes in the lesson plans or assignments.

The idea behind the sequencing here is to start as simple as possible and complicate later. Many of you have already been "doing regression" and this class exists to help you understand more deeply what you are doing — to give you power over your tools, to enable creativity, flexibility, and, at minimum, to help you avoid errors.

This class emphasizes the linear model. There are mathematically simpler ways to introduce the concepts and techniques of statistical inference, but you are already using linear models and you'll continue to use them throughout your careers (where linear models include linear regression, logit, probit, poisson, multinomial logit, etc..). Since this class aims to help you do work and privileges such doing over deep theory, and since this is your second course, we'll thus focus on the mathematically and conceptually more complex but more commonly used linear model.

Data: I'll be bringing in data that I have on hand. This means our units of analysis will often be individual people or perhaps political or geographic units, mostly in the United States. I'd love to use other data, so feel free to suggest and provide it to me — come to office hours and we can talk about how to use your favorite datasets in the class.

Theory: This class is about statistical inference and thus statistical theory. Yet, statistics as a discipline exists to help us understand more than why the linear model works as it does. Thus, social science theory cannot be far from our minds as we think about what makes a given data analytic strategy meaningful. That is, while we spend a term thinking a lot about how to make meaningful statements about statistical inference, we must also keep substantive significance foremost in our minds.

I General Principles for Frequentist Statistical Inference: Randomize, Repeat, Reject

The first section of the class focuses as directly as possible on the foundations of statistical inference for the linear model. We need to know the target of our inference, and why we might be justified in inferring to such a target.[7] It turns out that computers make the job of doing such inference much easier, but in committing to computation we'll have to learn a bit more math so that we can communicate most effectively with our computers as they make our lives easier.

[7] Cobb (2007) provided the "randomize, repeat, reject" motto and otherwise articulates some of the inspiration for this course.

1 — January 24, 2011— What is the linear model for?

Reminder: Bring laptops if you have them. Those bringing laptops should have R installed or be able to access their Rstudio accounts and be able to access the net.

Topics: Review of the bivariate linear model; Linear model as description using smoothed conditional means; Uses of the the linear model (description, forecasting, statistical inference, causal inference); Dummy variables; Differences of means; Predicted values; Residuals; Linearity assessment.

Read: **Henceforth, "*" means "recommended" or "other useful" reading. The readings not marked with "*" are required.**

Fox, 2008a, Chapters 1,2,5.1

Berk, 2008, Pages 1–8[8]

*Berk, 2004, Chapter 1–3

Do: **"Do" means, "do in class." I am writing down sketches about what we might do in class so that you might be thinking about these tasks as you read.**

Practice the linear model and prove to selves that a linear model with dummy variables tells us something about a difference of means and that the proposed computational technique does minimize the sum of squared residuals. Consider and explore other ways to summarize the conditional distribution of an outcome on an explanatory variable (summaries of ranks? quantiles? something else?).

2—January 31—What is a hypothesis test?

How much evidence does some data summary provide against a substantively relevant hunch about the process under study? How can we formalize and communicate the plausibility of such hunches in light of our observation?

Topics: Two bases out of at least three bases for frequentist statistical inference (random assignment, random sampling); randomization distributions and sampling distributions of test statistics. Today focus on random assignment and randomization distributions of test statistics under the sharp null hypothesis of no relationship. Generating randomization distributions for hypotheses about aspects of the linear model using enumeration (aka permutation) and simulation (shuffling). Introduction to significance level of a test versus size of a test.

Read: Kaplan, 2009, Chap 15,16.1,16.6,16.7,17.5,17.7,17.8 discusses tests of hypotheses in the context of permutation distributions of linear model based test statistics. He wants to emphasize the F-statistic and R^2 and the ANOVA table, but his discussion of permutation based testing will apply to our concern with the effect of an experimental treatment on an outcome.

Gonick and Smith, 1993, Chap 8 explains the classical approach to hypothesis testing based on Normal and t-distributions.

Imbens and Rubin, 2009, Chap 5 explains Fisher's approach to the sharp or strict null hypothesis test in the context of the potential outcomes framework for causal inference.

*Berk, 2004, Chap 4 provides an excellent and readable overview of the targets of inference and associated justifications often used by social scientists.

*Fox, 2008a, Chap 21.4 explains about bootstrap hypothesis tests (i.e. sampling model justified hypothesis tests).

*Fisher, 1935, Chap 2 explains *the* invention of random-assignment based randomization inference in about 15 pages with almost no math.

*Rosenbaum, 2002b, Chap 2–2.4 explains and formalizes Fisher's randomization inference.

[8] http://www.library.uiuc.edu/proxy/go.php?url=http://dx.doi.org/10.1007/978-0-387-77501-2

*Rosenbaum, 2002a explains how one might use Fisher-style randomization inference with linear regression.

Do: TBA

3—February 7—What is a confidence interval?

Given a reasonable data summary, what other guesses about said quantity are plausible?

Topics Continuing on statistical inference; Inverting hypothesis tests; null hypotheses and alternatives; Introduce the weak null and the average treatment effect; The bootstrap; Today focus on sampling models for statistical inference but link back to assignment models via hypothesis test inversion. More on concepts of level of test versus size of test, Type I and Type II errors, power of tests.

Read: Kaplan, 2009, Chap 14

Gonick and Smith, 1993, Chap 7

Fox, 2008a, Chap 21

*Imbens and Rubin, 2009, Chap 6 discusses and compares Fisher's approach to Neyman's approach. We will defer discussion about the the parts of the discussion regarding Normality until later in the course. Review their chapter 5.8 for discussion about inversion of the hypothesis test to create confidence intervals.

*Neyman, 1990; Rubin, 1990 *the* invention of random-sampling based randomization inference.

*Lohr, 1999, Chap 2.7 a clear exposition of the random-sampling based approach.

Do: TBA. Notice some of the limitations of the each computational approach to generating confidence intervals: the sampling model as approximated by the bootstrap has problems with small samples (introduce ideas about collinearity and efficiency); the assignment model as approximated with shuffling (or enumeration) becomes computationally expensive. Both require models of effects.

Due: First round of regression tables (or sketches thereof) for the final paper. The substantive topic addressed by your tables should be pretty fascinating and important to you so as to buoy your interest and energy for the rest of the term.

4—February 14—"Controlling for" and "Holding Constant": Statistical and Stratification.

For the next three classes we step away from statistical inference and talk more about causal inference and explanation (and computation) as we engage with multiple regression.

Topics: Confounding/Omitted variable bias; Efficiency of linear model estimators; Extrapolation/Interpolation; "Post-treatment" or intermediate variables versus covariates; Post-stratification versus residualization; Specification issues including collinearity and continuing engagement with dummy/categorical covariates variables; introduce interaction terms as a way to do stratification.

Read: Fox, 2008a, Chap 5.2 (multiple regression scalar form).

Berk, 2004, Chap 6–7 (skipping stuff on standardized coefs)

Gelman and Hill, 2007, Chap 9.0–9.2 (on causal inference)

*Hanushek and Jackson, 1977, Chap TBA (more on multiple regression in scalar form)

*Kaplan, 2009, Chap 8,10,11 (on a geometric interpretation of residualization)

Do: TBA; write a program to fit a least squares line as if you only knew the scalar form; interpret the results of a regression "holding constant" via residualization versus stratification.

Due: Second round of regression tables. By now you should know what regression results you aim to investigate for the rest of the term.

5—February 21—How can we compute reasonable guesses without fuss? (try matrices).

Topics: Basic matrix algebra (also called "linear algebra") [matrices and vectors introduced; addition, subtraction, multiplication, transposition and inversion]; Matrix algebra of the linear model (the importance and meaning and source of $\mathbf{X}\hat{\boldsymbol{\beta}}$ and $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$); Matrix algebra for estimating and interpreting the linear model in R; More engagement with collinearity and dummy variables.

Read: Fox, 2008a, Appendix B.1.0–B.1.3 and Chap 9

*Fox, 2008a, Chap 10 (another geometric interpretation)

*Fox, 2008a, Appendix B (more on matrices)

Do: Explain, explore and unpack the function we've been using to produce slopes. What limitations on the $\mathbf{X}$ matrix are required by the least squares criterion? How might we manage them. Prove to ourselves that our functions work.

6—February 28—Challenges to reasonable guessing/description and inference: Overly Influential points and Overly (Multi)Collinear predictors.

Topics: Influence, Leverage, Hat Matrix; Methods for handling highly influential points; Methods for handling overly multicollinear predictors.

Read: Fox, 2008a, Chap 11 on Overly Influential Points

Fox, 2008a, Chap 13 on Overly (Multi)Collinear predictors.

*Achen, 2002 (on why kitchen sink regression are a problem)

*Fox, 2008a, Chap 7 on a common case of collinearity: categorical predictors/dummy variables

*Fox, 2008a, Chap 19 on making linear models resistant to overly influential points.

7—March 7—Inference for combinations of predictors.

Topics: Interaction terms; Constrained linear regression and linear contrasts; F-tests, Wald-type-tests and related confidence regions; Inference for predictions; Goodness of fit (part 1) (R^2 and standard error of the regression). Understanding more about the algebra of expectations/variances.

Read: Fox, 2008a, Chap 5.2.3,7, 8.5 (on R^2 and dummy variables and interactions)

Kaplan, 2009, Chap 16 (on F-tests and R^2)

TBA for more on R^2 and goodness of fit.

*Brambor et al., 2006 (on large-sample based approach for interaction terms)

II Connections to Large-Sample Statistical Theory

When we don't have the time for our computers to do the "repeat" phase of "randomize, repeat, reject", what can we do? Luckily for us, the mathematical underpinning of "repeat" after "randomize" has been well developed. It is this foundational mathematics that enables the standard regression table to exist (you know, the one that you get when you type `summary(myregression)` in R). Much of the time this table is an excellent approximation to what we did with repetitive computing in the previous section of the course. Sometimes it is a terrible approximation. This part of the course aims to connect the computationally intensive but conceptually clear and mathematically simple theory that we learned and

applied above to the computationally simple but mathematically complex theory that provides most of the information social scientists currently use from linear models.

Since we have little time, we will not do proofs; instead we will convince ourselves that the mathematicians and statisticians working between roughly 1690 and 1940 invented reasonable approximations using simulations. More importantly, we'll learn how to evaluate when those analytic results help us and when they do not.

8—March 14—The Mathemagic of Sums.

Topics: Foundations of the large-sample theory: laws of large numbers and central limit theorems; more explanation for the pervasiveness of the mean as a data summary. Approximation.

Reading: Freedman et al., 2007, Chap 16–18 (especially Chapter 18)

TBA from Rice?

*Trosset, 2009, Chap 8

*Lohr, 1999, Chap 2.7–2.8 Recall the use of a sampling theory approach requiring the Central Limit Theory and compare to an approach positing a distribution for outcomes.

Do: Prove to ourselves that one Central Limit Theorem and one Law of Large Numbers actually works; Show how we can approximate our simulation based results very well (or very badly) using this mathemagic.

March 21—Spring Break

9—March 28—Sampling based Large sample/Asymptotic theory for the linear model.

Topics: Gauss-Markov theorem and associated classic linear model assumptions (introducing notions of non-constant variance, dependence); The different roles of Normality in the theory of the linear model; The t-distribution and t-test; the F-distribution and F-test; The usefulness of the large sample theory in flexible interpretation and assessment of the linear model (i.e. the ease of simulation from the implied sampling distribution of the coefficients).

Read: Fox, 2008a, Chap 6,9

Achen, 1982

Berk, 2004, Chapter 4,6

Gelman and Hill, 2007, Chap 7 (using the large sample theory to interpret and assess the linear model)

*Trosset, 2009, Chap 9 (not about the linear model, but nice on large sample hypothesis testing in general)

*Fox, 2008a, Chap 12 (on approaches to adjusting for violations of the large-sample theory assumptions. (WLS, GLS)

Do: Design simulations to assess how well the large-sample theory approximates the simulation based results in some common datasets and designs. Begin to develop some intuitions for when the standard regression table is fine and when it is worrisome. Notice how useful these results are in research design (before we can collect data we cannot shuffle or re-sample). Discuss how we might design studies to enhance statistical inference. Notice the role of assumptions — especially the additional assumptions.

10—April 4— A general, large sample, based approach to making reasonable guesses: Maximum Likelihood for the linear model.

Frequentists make inferences to control groups based on experimental design (following Fisher), to a population based on sampling design (following Neyman). They also make inferences to *a model of the population* often called a *data generating process*. Such models are at the core of the likelihood approach to statistical inference (also credited to Fisher).

Topic: A third frequentist mode of inference; Role of the central limit theorem and Normality in this approach; OLS is MLE; MLE as useful for binary outcomes in addition to continuous outcomes.

Read: Fox, 2008a, Chap 9.3.3

Gelman and Hill, 2007, Chap 18.1–18.3 (skimming over the stuff about Bayes, Gibbs sampling, etc...)

*TBA from Rice (2007) or other more canonical and mathematical treatments

*King, 1989, Chap 4

*Cox, 2006, Chap 1,2

Do: Re-estimate our linear models using our own likelihood maximizing function (first by examining the profile likelihood function graphically and second by asking the computer to find the maximum). Plug in a Bernoulli likelihood and explore use with binary outcomes. Assess the statistical inferences from MLE compared to those arising from shuffling and/or bootstrapping (or enumerating, or even Normal approximations to the shuffles).

Due: A plan for your final paper specifying why the regression table you chose is substantively meaningful; your simulation and assessment plans; plans for tables and/or figures; and your plans for producing reproducible analyses.

III Dependent Observations and Causal Inference: Challenges to Statistical Inference

The previous sections have prepared us to engage with some of the more realistic designs used by political scientists: designs in which each row in the dataset is related to other rows by design. Recall that we had to make assumptions about independence whether we used simulations or large-sample theory. Here we engage with some of the ways that one may handle dependent research designs, and perhaps, at the end, introduce the topic of dealing with missing data. Thus, this course ends with some very real-world situations in which we desire statistical inferences (let alone causal inference) but that challenge the simple methods of justifying said inferences.

11—April 11—Difference-in-Differences/Before-After Designs

One of the simplest ways we might violate independence assumptions is when we want to strengthen causal inference by observing the same unit twice (perhaps both before and after some event/intervention/change).

Topics: Back to causal inference — Consider benefits of observing each unit twice; Consider difficulties for statistical inference posed by observing each unit twice; Consider solutions; Generalized Least Squares (GLS) and "cluster-robust" or "heteroskedasticity consistent" methods.

Read: Bertrand et al., 2004

Allison, 1990

Gelman and Hill, 2007, Chap 10.7 (on longitudinal designs.)

*Gelman and Hill, 2007, Chap 9–10 (on causal inference including some discussion of instrumental variables.)

*TBA something from Campbell and Stanley on the strengths and weaknesses of such designs.

Do: Assess the usefulness of the strategies of (a) ignoring dependence [?how badly mislead might we be? what simple strategies might we use? (remember the advice in Achen (1982))] and (b) the corrections proposed above using our simulation based methods that we developed earlier in the course.

Due: A proposal for assessing the properties of your regression table.

12—April 18—Corrections for unstructured dependence (multilevel/longitudinal designs part 1)

The first set of approaches to multilevel designs (which include some panel designs, measurement models, etc..) focuses on correcting the statistical inferences from models estimated ignoring the dependence.

Topics: Huber/White/ "cluster-robust" standard errors; Issues arising with causal and statistical inference from multilevel designs.

Read: Fox, 2008a, Chap 12 (especially 12.2)

Long and Ervin, 2000 (on the general idea of estimating a model while ignoring dependence and then correcting some piece of that model.)

Wooldridge, 2003 (on the block-diagonal variance-covariance matrix methods ("cluster robust") types of corrections.)

Freedman, 2006 (for some criticism of these kinds of methods.)

*Cribari-Neto, 2004 (on other methods for making the simple correction and problems arising from the initial ideas of Huber and White.)

*Green and Vavreck, 2007 (for an exemplar of using simulation to check the performance of such methods in some particular designs.)

*Faes et al. (2009) (on different conceptions of what "degrees of freedom" and "sample size" might mean in a complex design)

Do: Explore and discover how many cluster/groups is enough for us to trust that confidence intervals based on this type of method have good coverage.

13 —April 25—Instrumental Variables

Topics: The potential outcomes approach to causal inference; causal inference in a randomized experiment; causal inference with instruments.

Read: Angrist et al., 1996

Dunning, 2008

Sovey and Green, 2011

Gelman and Hill, 2007, Chap 10.4–10.5 (you might also want to revist their Chap 9 and previous part of Chap 10).

*Imbens and Rosenbaum, 2005 On weak instruments

*Angrist and Pischke, 2009[Chapter 4]

Do: Work on understanding the requirements of an instrumental variable and their use in statistical inference. Perhaps discuss the dilemma of weak instruments.

14 —-May 2—Matching and Poststratification

Topics: Matching is a nonparametric way to compare like with like. Although there are many kinds, since we only have one day we will focus on post-stratification — a kind of matching that does not repeat units and that thus enables more or less conventional statistical inference for causal effects after matching is complete.

Read: Rosenbaum, 2010[Chap 3, 7–13]

Hansen, 2004

Lu et al., 2011

Do: Practice producing matched sets, assessing balance, assessing causal effects.

May 9—Final Papers Due

IV References

Achen, C. H. (1982). *Interpreting and Using Regression*. Sage, Newbury Park, CA.

Achen, C. H. (2002). Toward A New Political Methodology: Microfoundations and Art. *Annual Review of Political Science*, 5:423–450.

Allison, P. (1990). Change scores as dependent variables in regression analysis. *Sociological methodology*, 20:93–114.

Angrist, J. and Pischke, J. (2009). *Mostly harmless econometrics: an empiricist's companion*. Princeton Univ Pr.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables (Disc: p456-472). *Journal of the American Statistical Association*, 91:444–455.

Becker, H. S. (1986). *Writing for Social Scientists: How to Start and Finish Your Thesis, Book, or Article*. University of Chicago Press.

Berk, R. (2004). *Regression Analysis: A Constructive Critique*. Sage.

Berk, R. (2008). *Statistical learning from a regression perspective*. Springer.

Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How Much Should We Trust Differences-in-Differences Estimates? *The Quarterly Journal of Economics*, 119(1):249–275.

Brambor, T., Clark, W. R., and Golder, M. (2006). Understanding interaction models: Improving empirical analyses. *Political Analysis*, 14:63–82.

Chiang, A. C. (1984). *Fundamental Methods of Mathematical Economics*. McGraw-Hill/Irwin; 3rd edition (February 1, 1984).

Cobb, G. (2007). The introductory statistics course: A ptolemaic curriculum? *Technology Innovations in Statistics Education*, 1(1).

Cox, D. R. (2006). *Principles of statistical inference*. Cambridge University Press, Cambridge.

Cribari-Neto, F. (2004). Asymptotic inference under heteroskedasticity of unknown form. *Computational Statistics and Data Analysis*, 45(2):215–233.

Dunning, T. (2008). Model specification in instrumental-variables regression. *Political Analysis*, 16(3):290.

Faes, C., Molenberghs, G., Aerts, M., Verbeke, G., and Kenward, M. (2009). The effective sample size and an alternative small-sample degrees-of-freedom method. *The American Statistician*, 63(4):389–399.

Faraway, J. (2005). *Linear Models With R*. CRC Press.

Faraway, J. (2006). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. CRC Press.

Fisher, R. (1935). *The design of experiments. 1935*. Oliver and Boyd, Edinburgh.

Fox, J. (2008a). *Applied regression analysis and generalized linear models*. Sage.

Fox, J. (2008b). *A mathematical primer for social statistics*. SAGE Publications Inc.

Fox, J. and Weisberg, S. (2011). *An R Companion to Applied Regression*. Sage.

Freedman, D., Pisani, R., and Purves, R. (2007). *Statistics*. W.W. Norton, New York, 4th edition.

Freedman, D. A. (2006). On the So-called "Huber Sandwich Estimator" and "Robust Standard Errors". *The American Statistician*, 60(4):299–302.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gill, J. (2006). *Essential mathematics for political and social research*. Cambridge Univ Pr.

Gonick, L. and Smith, W. (1993). *The cartoon guide to statistics*. HarperPerennial New York, NY.

Green, D. and Vavreck, L. (2007). Analysis of Cluster-Randomized Experiments: A Comparison of Alternative Estimation Approaches. *Political Analysis*.

Greene, W. H. (1997). *Econometric Analysis*. Prentice Hall, 3rd edition.

Hansen, B. B. (2004). Full matchining in an observational study of coaching for the sat. *Journal of the American Statistical Association*, 99:609.

Hanushek, E. and Jackson, J. (1977). *Statistical Methods for Social Scientists*. Academic Press, Inc., San Diego, CA.

Imbens, G. and Rosenbaum, P. (2005). Robust, accurate confidence intervals with a weak instrument: quarter of birth and education. *Journal of the Royal Statistical Society Series A-Statistics In Society*, 168:109–126.

Imbens, G. and Rubin, D. (2009). Causal inference in statistics. Unpublished book manuscript. Forthcoming at Cambridge University Press.

Kaplan, D. (2009). *Statistical Modeling: A Fresh Approach*. http://www.macalester.edu/~kaplan/ism/.

Kennedy, P. (2003). *A guide to econometrics*. The MIT Press.

King, G. (1989). *Unifying Political Methodology: The Likelihood Theory of Statistical Inference*. Cambridge University Press, New York.

Lancaster, T. (2004). *An introduction to modern Bayesian econometrics*. Blackwell Pub.

Lohr, S. (1999). *Sampling: Design and Analysis*. Brooks/Cole.

Long, J. S. and Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, 54(3):217–224.

Lu, B., Greevy, R., Xu, X., and Beck, C. (2011). Optimal nonbipartite matching and its statistical applications. *The American Statistician*, 65(1):21–30.

Neyman, J. (1923 [1990]). On the application of probability theory to agricultural experiments. essay on principles. section 9 (1923). *Statistical Science*, 5:463–480. reprint. Transl. by Dabrowska and Speed.

Rice, J. (2007). *Mathematical Statistics and Data Analysis*. Duxbury Press, Belmont, CA, 3rd edition.

Rosenbaum, P. R. (2002a). Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, 17(3):286–327.

Rosenbaum, P. R. (2002b). *Observational Studies*. Springer-Verlag, second edition.

Rosenbaum, P. R. (2010). *Design of Observational Studies*. Springer.

Rubin, D. B. (1990). [on the application of probability theory to agricultural experiments. essay on principles. section 9.] comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, 5(4):472–480.

Simon, C. P. and Blume, L. (1994). *Mathematics for Economists*. W.W. Norton, New York, NY.

Sovey, A. J. and Green, D. P. (2011). Instrumental variables estimation in political science: A readers guide. *American Journal of Political Science*, 55(1):188—200.

Trosset, M. W. (2009). *An Introduction to Statistical Inference and Its Applications with R*. CRC Press.

Verzani, J. (2005). *Using R for Introductory Statistics*. Chapman & Hall/CRC.

Wooldridge, J. (2003). Cluster-Sample Methods in Applied Econometrics. *The American Economic Review*, 93(2):133–138.