Statistical Inference:
Linear Models, Descriptive, Causal and Statistical Inference

Jake Bowers
jwbowers@illinois.edu
Online: http://jakebowers.org/

Spring 2018

Overview

Where/When We will meet Tuesdays, 3:30pm – 5:30pm in 404 David Kinley Hall.

Office Hours I will have office hours 3:30–4:30pm on Mondays. Please make an appointment if you want to come to office hours to ensure that we can meet and talk.

Introduction What does it mean to say "statistically significant"? When is it reasonable to say this? When is it confusing? Why can we report that a 95% confidence interval excludes implausible hypotheses 95% of the time? 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 foundations for thinking about statistical inference and for thinking about specifying, fitting, and interpreting linear models. In this course, you will deepen your understanding of statistical inference and estimation. This is a course in applied statistical theory focusing on linear models. You will work toward understanding the basic theory of estimation and testing by application. We will emphasize the hard work of writing computer programs rather than the hard work of proving theorems. By the end of the term you will have developed strategies for answering the questions posed above and thus will be well-positioned to use linear models to learn about politics with confidence and creativity and good judgement.

Goals and Expectations

This class aims to help you learn to think about what it means to do statistical inference for both descriptive and causal claims.

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 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. Energetic engagement manifests itself in meeting with your classmates outside of the class, in asking questions during the class, and in taking the term paper seriously.

Second, I assume you are ready to work. Learning requires work. As much as possible I will encourage you to 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 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. I have made changes to this course in the middle of the term upon hearing great and useful ideas from students. I am happy to do so. 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 the R statistical programming language. If you love learning computing languages then you can still get a lot out of this course — you will learn a lot about R as kind of laboratory for learning about statistical theory.

Rules There aren't many rules for the course, but they're all important. First, ask questions when you don't understand things; chances are you're not alone. Second, don't miss class or section. Third, do the work. This does not mean divide the work up among your classmates so that you only do part of the work. Each person should engage with all of the work even if the people who writes it up changes from week to week.

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 unless we have another specific arrangement.[1] I will not accept Microsoft, Apple, or any other proprietary format.

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 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 While I choose the topics for this class, the discussion in the class is mostly led by you. I will lecture very little. Instead we will use the assignments given out the previous week to raise questions about 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 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 or is asking a question that raised by the work.

Explorations Every week or so, I will ask you to complete a short assignment that encourages you to engage creatively with the topics of interest. I anticipate that you will work on these assignments in groups and that each of you will come to class prepared to discuss them. I don't think that the groups should have more than 3 people in them. However, I'm willing to have larger groups if you talk with me about it. All of the work should be collaborative. Do not divide up the work into parts.

I'll ask that you turn these in by Sundays at 5pm.

Weekly Answers After class each week I'll ask each of you to answer one question on your own. The idea is for you to learn whether you got the main concept for each week by writing one paragraph as an answer to a question that I will pose.

I'll ask that you turn these in by Thursdays at 5pm.

Term Papers Other than the explorations, and in-class participation, the main assignment for this term is for you to write a short paper that teaches us about and/or advances the ball regarding some problem of statistical methodology and/or research design that is important to your own work or subfield. For example, you might choose an existing published paper and ask, "Are the statistical inferences and estimators used in this paper appropriate for the question asked? Do they perform as they should? What other approaches might this person have used? Do those other approaches produce different answers?" So, this is a methods paper, not a substantive paper. It can be very short.

[1] For example, if you have some reason why pdf files make your life especially difficult, then of course I will work with you find another format.

Grades are Feedback Humans need feedback to close the intention to action gap. They also need feedback to feel good about their progress and to motivate them. In this class I will use grades as feedback. All grades except for the final grade will be satisfactory, unsatisfactory (with the possibility of "outstanding"), and fail. These map roughly onto A=satisfactory, C=unsatisfactory, and F=fail (i.e. you didn't try).

I'll calculate your grade for the course this way: 20% explorations (everyone in the group receives the same grade, A if you are creative and thoughtful and diligent, C if you are not or if you don't seem to be getting the concepts, no late work accepted); 20% in class participation (satisfactory if you ask good questions that show that you have thought about the material, a good question can be a simple question; unsatisfactory if your questions show that you are not doing the reading and/or are not actively involved in the explorations; fail if you are not in class); 10% attendance (A if you show up, 0 if not); 20% Weekly answers; 30% final paper.

You can miss two classes without grade penalty although I suspect you'll be sad to miss the discussion.

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. This work will require that you come prepared to meetings with your classmates and also to the class itself and that you are an active collaborator. Involvement also means meeting with your classmates at least once per week outside of class to complete the work begun in the class.

This class is an opportunity to practice courage: I expect you to make a guess when I ask a question (in writing or in person), I expect that you will ask a question when you have a problem.

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: your own likelihood functions, your own least squares solvers, your own bootstrapping and permutation statistical inference routines, and your own posterior distribution for Bayesian statistical inference.

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 by using either R markdown or Sweave (R+LaTeX). 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.

[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.

Books

I'm am not requiring any particular books this term. 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.

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:

Fox, J. (2008a). *Applied regression analysis and generalized linear models*. Sage. [3] This book does a great job of combining mathematical clarity with readability for social scientists.

Achen, C. H. (1982). *Interpreting and Using Regression*. Sage, Newbury Park, CA. This book is a crucial resource. Highly highly recommended.

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

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] A nice supplement to Fox and Achen especially with the 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

Imai, K. (2017). *Quantitative Social Science: An Introduction*. Princeton University Press

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

Angrist, J. D. and Pischke, J.-S. (2014). *Mastering'metrics: The path from cause to effect*. Princeton University Press

Kaplan (2012)[6]

[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`

[5] `http://www.stat.columbia.edu/~gelman/arm/`

[6] Second edition:`http://mosaic-web.org/go/StatisticalModeling/`, First edition still has interesting resources: `http://www.macalester.edu/~kaplan/ism/`

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 or Casella, G. and Berger, R. L. (2011). Statistical inference. *Duxbury Press*, pages 99–100. These are 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 Now canonical in applied economics. Very accessible introduction to an econ perspective on applied statistics.

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

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 as you confront the ever changing landscape of statistics and data analysis and social science in the future, 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.[7]

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

[7]Linear models here include linear regression, logit, probit, poisson, multinomial logit, etc ....

thinking a lot about how to make meaningful statements about statistical inference, we must also keep substantive significance foremost in our minds.

I Review, fit criteria, descriptive inference, causal inference, and adjustment

Last term you engaged with the idea of “controlling for” and fitting lines to scatterplots (or planes to clouds of points). In this section of the course we revisit the idea of adjustment.

1 Tuesday, January 16—Adjustment:The problem(s) of adjustment of observational studies using the linear model.

Topics: The problem of knowing when one has "controlled for" enough. The curse of dimensionality. The problem of overly influential points — especially if the data description model uses least squares. The problem of sensitivity to functional form.

Read: **Henceforth, “*” means “recommended” or “other useful” reading. The readings not marked with “*” are most important.**

Rosenbaum, 2010, Chap 1,3,7,8,9,13[8]

Achen, 2002 (on the problem of kitchen sink regressions)

*Berk, 2010

*Gerber and Green, 2012, Chap 1 and 2

*Holland, 1986

Fox, 2008a, Chap 11 on Overly Influential Points

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

*Hansen, 2004 on full matching for adjustment

*Hansen and Bowers, 2008 on assessing balance.

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

2 Tuesday, January 23—Adjustment: How can we overcome some of the problems in adjusting using the linear model while still adjusting and clarifying our inferences?

Read: Rosenbaum, 2010, Chap 1,3,7,8,9,13[9]

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

Hansen, 2004 on full matching for adjustment

Hansen and Bowers, 2008 on assessing balance.

3 Tuesday, January 30—Estimation: What is the linear model for? What does “unbiased" mean?

Topics: Introduction of the property of unbiasedness since you already know about estimators. How might we assess claims about unbiasedness? (Also, the idea of potential outcomes and the fundamental problem of causal inference.)

[8] http://www.springerlink.com/content/978-1-4419-1212-1/contents/

[9] http://www.springerlink.com/content/978-1-4419-1212-1/contents/

Read: Berk, 2008, Pages 1–8[10]

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

Shalizi, 2017, Chap 1–2

Achen, 1982 (The focus here is on consistency rather than on unbiasedness)

*Gerber and Green, 2012, Chap 1 and 2

James et al., 2013, skim Chap 3 (on linear models) and Chap 6.2.2 and 6.2.3 on the lasso [11]

*Fox, 2008a, Chapters 1,2,5.1 *Fox, 2008a, Chap 5.2 (multiple regression scalar form).

4 Tuesday, February 6—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.

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

This 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.[12] 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.

5 Tuesday, February 13—Testing: 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: Fisher, 1935, Chap 2 explains *the* invention of random-assignment based randomization inference in about 15 pages.

Rosenbaum, 2010, Chap 2 (`http://www.springerlink.com/content/978-1-4419-1212-1/contents/`)

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.

[10] `http://www.library.uiuc.edu/proxy/go.php?url=http://dx.doi.org/10.1007/978-0-387-77501-2`
[11] `http://www-bcf.usc.edu/~gareth/ISL/getbook.html`
[12] Cobb (2007) provided the "randomize, repeat, reject" motto and otherwise articulates some of the inspiration for this course.

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).

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

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

6 Tuesday, February 20—Testing: 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

7 Tuesday, February 27—No Class

Rescheduled.

8 Tuesday, March 6—Testing: Design based statistical inference from resampling.

Topics 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

*Neyman, 1937 invents the confidence interval.

*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: 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.

9 Tuesday, March 13—No Class—Review operating characteristics of statistical inference methods.

Read: Rosenbaum, 2010, Glossary (`http://www.springerlink.com/content/978-1-4419-1212-1/contents/`)

Bertrand et al., 2004 (for an example of why paying attention to dependence matters)

*Trosset, 2009, Chap 9 and especially 9.3

10 Thursday, March 15—Information: Challenges to reasonable guessing/description and inference: More on high dimensional models

Topics: Penalized linear models, overfitting, cross-validation.

Read: James et al., 2013, Chap 3.6 and Chap 6.4–6.6

On 'micronumerosity': `http://davegiles.blogspot.com/2011/09/micronumerosity.html`

*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

III 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.

11 Tuesday, March 20—Spring Break

No Class.

12 Thursday, March 29—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.

IV From Populations and Sampling Processes to Models

13 Tuesday, April 3—No Class, Linowes Lecture

14 Tuesday, April 10—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.

Read: *Fox, 2008a, Chap 9.3.3

Green, 1991, Use the 2009 Version of from https://sites.google.com/site/donaldpgreen/plsc504

*Fox and Weisberg, 2011, Chap 5 and see also http://socserv.socsci.mcmaster.ca/jfox/Courses/SPIDA/index.html

*King, 1989, Chap 4

*Cox, 2006, Chap 1,2

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

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). Assess the statistical inferences from MLE compared to those arising from shuffling and/or bootstrapping (or enumerating, or even Normal approximations to the shuffles).

15 Tuesday, April 17—Logit, Probit, Poisson, Oh My!

Topics: We continue to work with maximum likelihood as a method for generating closed-form estimators and for providing closed-form estimators of standard errors, and thus as a very useful approach to statistical inference. We will focus on understanding some common parameterizations of binomial and Poisson outcome generating functions.

Read: *Fox, 2008a, Chap 14

Gelman and Hill, 2007, Chap 5

*Gelman and Hill, 2007, Chap 6

*Fox, 2008a, Chap 15

Do: Write our own logit fitting routine. Assess the circumstances under which we would prefer to find mle estimates versus rely on consistency results and large sample theory of simple linear regression models versus use some form of resampling for statistical inference with binary outcomes.

16 Wednesday, April 18–Full Drafts of Reports Due

By 5pm.

17 Tuesday, April 24—Introduction to Bayesian Statistical Inference

Now that we have traveled quite far from the idea of repeating a physical operation, we can take one more step: the idea that our model of the outcome might not actually reflect some population, but rather that our model represents our beliefs about some population. Or, perhaps, that our model of the population ought to be seen as only one of many possible but related models.

Topics: Bayes rule and MLE; posteriors, priors, hyper-priors. The idea of sampling from a posterior and of the posterior distribution of a parameter as a basis for statistical inference.

Read: Gelman and Hill, 2007, Chap 18.1–18.3

Lynch, 2007, Chap 3,7 (see it all via a campus machine/VPN connection at `http://www.springerlink.com/content/978-0-387-71264-2/contents/`)

*Imbens and Rubin, 2009, Chap 8 (for a perspective on Bayesian estimation of counterfactual causal parameters)

*Lancaster, 2004, Chap 1, 3, 5.2, Appendix 1

*Albert, 2009, Chap 9

*Kruschke, 2010

If this were a semester long course in Bayes, we'd use these next textbooks in addition to the readings above.

*Gill, 2002

*Jackman, 2009

*Gelman et al., 2004

Do: Link the ideas about likelihood to Bayesian conceptions of statistical inference; fit a Normal outcomes, linear parameters, Bayesian model (i.e. a Normal mle model, i.e. a least squares model, i.e. a difference of means model).

18 Tuesday, May 1—Bayesian Statistical Inference for Linear Models: Normal, Binomial, Poisson Models

Topics:

Read:

Do:

19 Tuesday, May 8—TBA

20 Thursday, May 10—Extra Final Paper Consulting and/or Presentations

21 Monday, May 15—Final Papers Due

If you would like to turn in your final paper after this date, please let me know at least a week in advance. Final grades are due to the college by May 18.

V 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.

Albert, J. (2009). *Bayesian computation with R*. Springer Verlag.

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

Angrist, J. D. and Pischke, J.-S. (2014). *Mastering'metrics: The path from cause to effect*. Princeton University Press.

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.

Berk, R. (2010). What you can and can't properly do with regression. *Journal of Quantitative Criminology*, pages 1–7.

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.

Casella, G. and Berger, R. L. (2011). Statistical inference. *Duxbury Press*, pages 99–100.

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.

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.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition.

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

Gerber, A. and Green, D. (2012). *Field experiments: Design, analysis, and interpretation*. WW Norton.

Gill, J. (2002). *Bayesian Methods: A Social and Behavioral Sciences Approach*. Chapman & Hall/CRC, first edition.

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. P. (1991). Maximum likelihood for the masses. *The Political Methodologist*.

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

Hansen, B. (2004). Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association*, 99(467):609–618.

Hansen, B. and Bowers, J. (2008). Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science*, 23:219.

Holland, P. W. (1986). Statistics and causal inference (with discussion). *Journal of the American Statistical Association*, 81:945–970.

Imai, K. (2017). *Quantitative Social Science: An Introduction*. Princeton University Press.

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

Jackman, S. (2009). *Bayesian analysis for the social sciences*, volume 846. Wiley.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer.

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

Kaplan, D. (2012). *Statistical Modeling A Fresh Approach*. Daniel Kaplan, Macalester College, St. Paul, MN, second edition.

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.

Kruschke, J. (2010). Doing bayesian data analysis. *Manuscript, Indiana University*.

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

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

Lynch, S. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists*. Springer Verlag.

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.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 236(767):333–380.

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.

Shalizi, C. (2017). Advanced data analysis from an elementary point of view. Online textbook.

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

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.