Statistical Inference and Adjustment for the Linear Model

Instructor: Jake Bowers
jwbowers@illinois.edu
http://jakebowers.org/
Methods Preceptor: Gustavo Diaz
diazdia2@illinois.edu
http://gustavodiaz.org/

Spring 2021

Overview

Where/When We will meet Thursdays, 1:00pm – 3:20pm on Zoom (link in the Moodle).

Office Hours Please make an appointment on http://calendly.com/jakebowers if you want to come to office hours to ensure that we can meet and talk. I'm happy to schedule other times if those don't work for you.

Methods Preceptor Office Hours: Tuesdays 10–11am via Zoom (link on the Moodle).

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 and adjustment by application. In order to solidify and internalize the key concepts surrounding testing and estimation, 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.

Another way to think about this class: This class does not aim to help you learn to use logistic or OLS (or poisson etc.) models. Rather this class aims to help you (1) **decide** when you should or should not use a given approach to statistical adjustment, estimation or testing and (2) **evaluate** your decision: the course should help you learn how to find out whether the approach to adjustment, estimation, or testing that you chose operates well or whether it is likely to mislead you. Thus the focus of the course is on **operating characteristics** and **diagnostics** and not on techniques.

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.

The **specific goals** of the course are that students:

- Explain in their own words key concepts in statistics like ”statistical inference”, ”hypothesis testing”, ”point estimation”, ”p-value”, ”confidence interval”, ”statistical adjustment (‘controlling for’)” and describe how such concepts fit together in applied research.
- Understand the differences between testing and estimation and be able to articulate the major criteria for good testing and good estimation.

- Evaluate testing and estimation strategies in their own work and the work of others.
- Understand statistical adjustment using the linear model and statistical adjustment by stratification and be able to articulate the major criteria for good statistical adjustment.
- Evaluate statistical adjustment tactics in their own work and the work of others.
- Explain in their own words the differences between the three main frequentist modes of statistical inference: Fisher's randomization based approach, Neyman's randomization and sampling based approach, and Maximum Likelihood.
- Develop judgement about the different choices and arguments required in each of those approaches (test statistics, asymptotic approximations, details of permutation and bootstrap algorithms, outcome data generating probability functions, link functions, functional forms).
- Apply all three approaches to statistical inference to applied problems and evaluate their operating characteristics.
- Practice scientific computing using R and writing in R+markdown.
- Produce a pre-analysis plan for a research project they are or might pursue.

Expectations I assume you are eager to learn. Eagerness, curiosity and excitement will impel your energetic engagement with the class throughout the term. If you feel bored, not curious, or unhappy about the class you should come and talk with me as soon as you can. 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.

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.

Ask questions when you don't understand things; chances are you're not alone.

Don't miss class or section.

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.

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.

I assume some previous engagement with high school mathematics, probability and statistical computing in the R statistical programming language. If you haven't had experience with R but 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.

All papers written in this class will use reproducible and/or literate programming practices (Jake Bowers and Voors 2016) and will include a code appendix.

All final written work will be turned in as pdf files unless we have another specific arrangement.[1]

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

[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.

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 any subsidiary paper assignments or other homeworks 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.

In-Class Exercises Because Zoom fatigue is real, we will hope to have some exercises for you to complete in class each week using R or perhaps just writing a sentence or in response to some question.

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. The point of the explorations is for you to (1) practice learning on your own and in a group (this is how you will learn about statistics for the rest of your career, so I'm happy for you to practice it now), (2) engage with the topic of the week so that you are prepared to come to class with questions and ideas, (3) practice coding and confronting coding errors.

Pre-analysis plans Other than the explorations, and in-class participation, the main assignment for this term is for you to write the pre-analysis plan for your first year paper or some other research project.

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 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: 30% explorations (everyone in the group receives the same grade, no late work accepted); 30% in-class participation ("A" if you ask good questions that show that you have thought about the material, a good question can be a simple question; "C" if your questions show that you are not doing the reading and/or are not actively involved in the explorations or if you are silent; "F" if you are not in class); 10% attendance ("A" if you show up, "F" if not); 30% final pre-analysis plan.

You can miss two classes without grade penalty. I drop the lowest exploration grade, too.

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.

You can redo any evaluation or the final paper in order to increase your grade on that assignment. If you want to resubmit something already graded, you need to let me know in advance so that I can make time to grade it again.

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.

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 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.

Academic Integrity According to the Student Code, 'It is the responsibility of each student to refrain from infractions of academic integrity, from conduct that may lead to suspicion of such infractions, and from conduct that aids others in such infractions.' Please know that it is my responsibility as an instructor to uphold the academic integrity policy of the University, which can be found here: http://studentcode.illinois.edu/article1_part4_1-401.html.

Disability Accomodations To ensure that disability-related concerns are properly addressed from the beginning, students with disabilities who require assistance to participate in this class should see me as soon as possible. To obtain disability-related academic adjustments and/or auxiliary aids, students with disabilities must contact the course instructor and the Disability Resources and Educational Services (DRES) as soon as possible. To contact DRES you may visit 1207 S. Oak St., Champaign, call 333-4603 (V/TTY), or e-mail a message to disability@illinois.edu

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:

John Fox (2008). *Applied regression analysis and generalized linear models*. Sage. [2] This book does a great job of combining mathematical clarity with readability for social scientists.

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

John Fox and Sanford Weisberg (2011). *An R Companion to Applied Regression*. Sage.[3]

Kosuke Imai (2017). *Quantitative Social Science: An Introduction*. Princeton, NJ: Princeton University Press, p. 408. ISBN: 0691175462

Books much like John Fox 2008 with slightly different emphases and more R in the text: A. Gelman and J. Hill (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.[4] 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.

T. Lancaster (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.

Michael W. Trosset (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.

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

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

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

If you'd like books that more closely link the statistics with R : J.J. Faraway (2005). *Linear Models With R*. CRC Press

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

J. Verzani (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 …

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

David Freedman, Robert Pisani, and Roger Purves (2007). *Statistics*. 4th. New York: W.W. Norton

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

Daniel Kaplan (2012). *Statistical Modeling A Fresh Approach*. Second. Macalester College, St. Paul, MN: Daniel Kaplan[5]

If you'd like more math and theory try these: D. R. Cox (2006). *Principles of statistical inference*. Cambridge: Cambridge University Press. This is one of my favorite books on statistical theory at the moment.

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

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

J.D. Angrist and J.S. Pischke (2009). *Mostly harmless econometrics: an empiricist's companion*. Princeton Univ Pr. ISBN: 0691120358 Now canonical in applied economics. Very accessible introduction to an econ perspective on applied statistics.

P. Kennedy (2003). *A guide to econometrics*. The MIT Press. ISBN: 026261183X 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:

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

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

Jeff Gill (2006). *Essential mathematics for political and social research*. Cambridge University Press Cambridge

Carl P. Simon and Lawrence Blume (1994). *Mathematics for Economists*. New York, NY: W.W. Norton

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

[5] Second edition:http://mosaic-web.org/go/StatisticalModeling/

throughout your careers (where linear models include linear regression, logit, probit, poisson, multinomial logit, etc …).

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 Review and Statistical adjustment

1— January 28— Review and Overview

2— February 4— Parametric Statistical Adjustment via the Linear Model

What kinds of questions should one ask about a linear model? Why are those questions relevant to our substantive work? What does it mean to use the linear model ”to control for”? Why would anyone use the linear model for this purpose if it does such a bad job at adjustment?

Useful Reading: **Henceforth, “*” means “recommended” or “other useful” reading. The readings not marked with “*” are especially useful in my experience.**

* Richard Berk (2010)

More review on the problems of adjustment in observational studies

Achen 2002 (on the problem of kitchen sink regressions)

John Fox 2008, Chap 11 on Overly Influential Points

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

3— February 11—Nonparametric Statistical Adjustment via Stratification

Topics: We discovered that covariance adjustment or ”controlling for” using the linear model removes only linear and additive relationships, may create new problems in regards overly influential points, and has clear diagnostic for when we have ”controlled for enough”. Stratification and balance assessment combine to create one simple solution to some of those problems.

Useful Reading: P R Rosenbaum 2010, Chap 1, 3, 7, 8, 9, 13

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

Hansen 2004 on full matching for adjustment

Hansen and J. Bowers 2008 on assessing balance.

Pashley and Miratrix 2020 on issues in estimating standard errors and overall average effects after matching

* For matching with more than one level see J. Zubizarreta 2012 or J. R. Zubizarreta and Keele 2017 plus Pimentel et al. 2018.

4— February 18—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.

Useful Reading: *John Fox 2008, Appendix B.1.0–B.1.3 and Chap 9

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

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

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

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?).

Be able to explain how OLS ‘controls for‘ (a) with a single binary indicator (”dummy”) control variable (i.e. a weighted average of stratum specific mean differences or coefficients) and (b) with a single continuous control variable (i.e. removing linear and additive relationships)

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.[6] 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— February 25—No Class

6— March 4—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 in a randomized experiment. 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.

Useful Reading: Fisher 1935, Chap 2 explains *the* invention of random-assignment based randomization inference in about 15 pages.

P R Rosenbaum 2010, Chap 2

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.

[6] Cobb 2007 provided the “randomize, repeat, reject” motto and otherwise articulates some of the inspiration for this course.

Imbens and D. 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.

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

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

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

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

7— March 11—What is an estimator? What does "unbiased" mean? What does "consistency" mean? What is MSE?

Topics: Introduction of the property of unbiasedness, consistency, mean squared error. How might we assess claims about unbiasedness? (Also, the idea of potential outcomes and the fundamental problem of causal inference.)

Useful Reading: R.A. Berk 2008, Pages 1–8[7]

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

Gerber and Green 2012, Chap 1–3

James et al. 2013, skim Chap 3 (on linear models) and Chap 6.2.2 and 6.2.3 on the lasso [8]

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

8— March 18—What is a confidence interval? How do they relate to hypothesis tests? What is the bootstrap? What is a standard error?

Topics Given a reasonable data summary, what other guesses about said quantity are plausible? Continuing on statistical inference; Inverting hypothesis tests; null hypotheses and alternatives; a Central Limit Theorem.

Useful Reading: Kaplan 2009, Chap 14

Gonick and Smith 1993, Chap 7

*Jerzy Neyman 1937 invents the confidence interval.

*John Fox 2008, Chap 21

*Imbens and D. 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.

*J. Neyman 1990; D. B. 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.

III Connections between Finite Sample, Design-Based Approaches and 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

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

[8] http://www-bcf.usc.edu/~gareth/ISL/getbook.html

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.

9— March 25—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); recap of Central Limit Theorems.

Useful Reading: *John Fox 2008, Chap 6, 9

Achen 1982

Richard 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)

*John Fox 2008, 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 1—Review

This week allows you to go back to the previous parts of the course to review the properties of estimators and tests. It also introduces the DeclareDesign approach to assessing those properties. So, no new readings recommended.

IV From Inference to Unobserved Counterfactuals and Populations to Inference to Unknown Model Parameters

11— April 8—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.

Useful Reading: *John Fox 2008, Chap 9.3.3

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

*John 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).

12— April 15—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; how to choose a link function.

Useful Reading: *John Fox 2008, Chap 14

Gelman and Hill 2007, Chap 5

*Gelman and Hill 2007, Chap 6

*John Fox 2008, 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.

13— April 22—Pre-Analysis Plans

Topics: Today we focus on what goes into a pre-analysis plan and why we use them. The final project for this class is a pre-analysis plan that departs from the norm by involving investigations into the justifications for and operating characteristics of your statistical adjustment, estimation, and testing plans.

Assignment before class: (1) Do some of the reading below so that you can generate some questions about pre-analysis plans. (2) Find and read a pre-analysis plan for a study of interest to you. You can see some here at the the EGAP Registry as well as at The Registry for International Development Impact Evaluations, The American Economic Association RCT registryand the Center for Open Science Registries. We might ask you to report on it during the class including any questions that came up for you about it.

Useful Reading:

- The Research Design Process This module and associated slides discuss pre-registration of designs and pre-analysis plans.
- 10 Things to know about Pre-Analysis Plans A guide to making pre-analysis plans.
- The template for the OES analysis plans
- Resources to learn more about pre-analysis plans from J-PAL
- A perspective from the WorldBank
- Presentations about pre-analysis planning from BITSS

14— April 29—Review of Fisher, Neyman and Likelihood

We review (1) what statistical inference means when we are inferring to or targeting (a) a counterfactual, (b) a population, or (c) a probability model of outcomes; and (2) how we assess the operating characteristics of the techniques we use for such inference (a) estimators (bias, MSE, consistency, efficiency) and (b) tests (false positive rates, power).

15— May 14—Pre-Analysis Plans Due

If you would like to turn in your final paper after this date, please let me know at least a week in advance.

V References

References

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

– (2002). "Toward A New Political Methodology: Microfoundations and ART". In: *Annual Review of Political Science* 5 (1), pp. 423–450.

Angrist, J.D. and J.S. Pischke (2009). *Mostly harmless econometrics: an empiricist's companion*. Princeton Univ Pr. ISBN: 0691120358.

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

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

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

– (2010). "What you can and can't properly do with regression". In: *Journal of Quantitative Criminology* 26.4, pp. 481–487.

Bowers, Jake and Maarten Voors (2016). "Six Steps to a Better Relationship with Your Future Self, V 2.0". In: *Revista de Ciencia Política* 36.3, pp. 829–848.

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

Cobb, G.W. (2007). "The Introductory Statistics Course: A Ptolemaic Curriculum?" In: *Technology Innovations in Statistics Education* 1.1.

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

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

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

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

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

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

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

Freedman, David, Robert Pisani, and Roger Purves (2007). *Statistics*. 4th. New York: W.W. Norton.

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

Gerber, Alan S and Donald P Green (2012). *Field experiments: Design, analysis, and interpretation*. WW Norton.

Gill, Jeff (2006). *Essential mathematics for political and social research*. Cambridge University Press Cambridge.

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

Green, Donald P (1991). *Maximum Likelihood for the Masses*. Tech. rep. Yale University ISPS. URL: https://sites.google.com/site/donaldpgreen/plsc504.

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

Hansen, B.B. (Sept. 2004). "Full matching in an observational study of coaching for the SAT". In: *Journal of the American Statistical Association* 99.467, pp. 609–618.

Hansen, B.B. and J. Bowers (2008). "Covariate Balance in Simple, Stratified and Clustered Comparative Studies". In: *Statistical Science* 23, p. 219.

Imai, Kosuke (2017). *Quantitative Social Science: An Introduction*. Princeton, NJ: Princeton University Press, p. 408. ISBN: 0691175462.

Imbens, G. and D. Rubin (2009). “Causal Inference in Statistics”. Unpublished book manuscript. Forthcoming at Cambridge University Press.

James, Gareth et al. (2013). *An introduction to statistical learning*. Springer.

Kaplan, Daniel (2009). *Statistical Modeling: A Fresh Approach*. http://www.macalester.edu/~kaplan/ism/. ISBN: 978-1448642397.

– (2012). *Statistical Modeling A Fresh Approach*. Second. Macalester College, St. Paul, MN: Daniel Kaplan.

Kennedy, P. (2003). *A guide to econometrics*. The MIT Press. ISBN: 026261183X.

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

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

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

Neyman, J. (1990). “On the application of probability theory to agricultural experiments. Essay on principles. Section 9 (1923)”. In: *Statistical Science* 5. reprint. Transl. by Dabrowska and Speed, pp. 463–480.

Neyman, Jerzy (1937). “Outline of a theory of statistical estimation based on the classical theory of probability”. In: *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences* 236.767, pp. 333–380.

Pashley, Nicole E and Luke W Miratrix (2020). “Insights on Variance Estimation for Blocked and Matched Pairs Designs”. In: *Journal of Educational and Behavioral Statistics* XX.X, pp. 1–26. DOI: 10.3102/1076998620946272.

Pimentel, Samuel D et al. (2018). “Optimal multilevel matching using network flows: An application to a summer reading intervention”. In: *The Annals of Applied Statistics* 12.3, pp. 1479–1505.

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

Rosenbaum, P R (2010). “Design of observational studies”. In: *Springer series in statistics*.

Rosenbaum, Paul R. (2002a). “Covariance adjustment in randomized experiments and observational studies”. In: *Statistical Science* 17.3, pp. 286–327.

– (2002b). *Observational Studies*. Second. Springer-Verlag.

Rubin, Donald 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”. In: *Statistical Science* 5.4, pp. 472–480.

Simon, Carl P. and Lawrence Blume (1994). *Mathematics for Economists*. New York, NY: W.W. Norton.

Trosset, Michael 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.

Zubizarreta, J.R. (2012). “Using Mixed Integer Programming for Matching in an Observational Study of Kidney Failure After Surgery”. In: *Journal of the American Statistical Association, Forthcoming*.

Zubizarreta, José R and Luke Keele (2017). “Optimal multilevel matching in clustered observational studies: A case study of the effectiveness of private schools under a large-scale voucher system”. In: *Journal of the American Statistical Association* 112.518, pp. 547–560.