Reading guidelines for Econ 671, Fall 2014

Gray Calhoun

August 26, 2014, version 1.0.0

This document gives short introductions to the required reading for each module of Econ 671. My hope is that you'll be able to understand the material better and finish the reading faster if you know why I've assigned it and what the most important parts are. The specific reading assignments and RAT dates are listed on the course syllabus.

In several places in this guide, I'll discuss the level of understanding that I want you to take away from the reading (i.e., I might say, "become familiar with this material, but don't get bogged down in the details.") When I do this, I am referring to the level of understanding that I expect you to have before taking the RAT for this material; i.e. it is based on the readings alone. I will expect you to understand the material better once we have covered it in class.

[1] This explanation is taken from the human-readable summary of the license provided by Creative Commons at «http://creativecommons.org/licenses/by/4.0/». The complete text of the license is available at «http://creativecommons.org/licenses/by/4.0/legalcode».

Contents

1 *Probability*

The expectation is that, after completing this part of the course, you'll be able to set up reasonable probability models for random events, and to derive exact and approximate sampling properties for statistical estimators.

[Gre12] Appendix B covers the basics of finite-sample probability theory and should be viewed as the absolute minimum a PhD economist should know about that subject. [CB02] covers this material in more detail in Chapters 1 – 4 and is available if you want better explanations.

[CB02] Chapter 5 is crucial and is the most important part of this reading assignment. [Gre12] Appendix D is the least important: much of the material is presented in [CB02] but there is some important new material. The CLT and LLN for sequences of heterogeneous random variables are important conceptually, for example, as are the inequalities.

Recommended but optional material

- [CB02] Chapters 1 – 4 as well as the rest of Chapter 5
- [Gal97] Chapters 1 – 4

2 *Statistical estimation*

The goal here is to learn how to construct an estimator if you are given a probability model, and to be able to understand how to derive its properties.

The first three parts of 7.2 cover the three basic methods of constructing estimators that we use in this course. The EM Algorithm (7.2.4) is optional. Method of Moments (in 7.2.1) is usually the easiest way to propose an estimator, but Maximum Likelihood (7.2.2) is usually more efficient.

Section 7.3 discusses how to evaluate estimators; 7.3.1 and 7.3.4 discuss how to evaluate an estimator given an explicit formula for penalizing estimation errors (a loss function). Usually the best estimator with respect to a given loss function is going to be different for different parameter values — which makes finding a uniformly best estimator impossible. Section 7.3.2 shows that we can (sometimes) find a uniformly best estimator after restricting the class of estimators under study. Here we look at unbiased estimators, and find conditions under which an unbiased estimator can be the "best" (minimum variance). There is a strong connection to MLE, which is part of the reason why unbiasedness and MLE are widely used — this is discussed in several results and examples related to the Cramér-Rao Lower Bound. Section 7.3.3 is optional.

Chapter 10 looks at asymptotic properties of point estimation. You should think of these asymptotic properties as approximations to an

estimator's finite sample properties. Under many conditions (e.g. those similar to the assumptions of the CLT), the finite sample distribution of a statistical estimator is not affected very much by the distribution of its underlying random variables. Most of the estimators used in economics can only be evaluated asymptotically, so this material is important.

Recommended but optional material

- [Gre12] Chapters 12 – 16 (13.4 is especially relevant)
- [Gal97] Chapter 5

3 Statistical inference

Chapter 8 introduces hypothesis testing. The presentation is a little strange — issues like test size and power are not discussed until Section 8.3, after the test statistics have been covered. It might help to read 8.3.1 first and possibly skim all of 8.3 before reading 8.2.

Chapter 9 covers confidence intervals. Again, it might help to skim Section 9.3 (how to evaluate confidence intervals) before reading 9.2 (how to construct them). There is a deep conceptual link between hypothesis tests and confidence intervals that's laid out in Theorem 9.2.2, which is why we're covering both of them together.

Chapter 10 covers the asymptotic properties of tests and intervals. The Likelihood Ratio Test, Wald test, and score test (usually called the LM test in econometrics) are especially important and widely used in economics, but you won't be responsible for the math behind them until next semester.

Recommended but optional material

- The optional reading listed for *Statistical Estimation* also covers inference.

4 Introduction to linear regression

A key issue in applied economics is trying to understand causality, and we're going to address it in Parts 5 and 6 of the class. Before that, we will focus on algebraic and mathematical properties of the linear regression model, so that we understand how the estimator works and what it returns when we use it, but we'll be focusing on relatively small and specialized problems. The material in this part should be a review of material that you covered as an undergraduate, but the notation may be slightly different since we will be using matrix notation.

This part of the course covers the math behind the OLS estimator — algebraic properties and statistical properties. "Algebraic" properties refers

to properties of the OLS estimator that are always true, simply because of algebraic identities. What happens when another variable is added to the model, etc. "Statistical" properties are properties of the estimator that depend on the underlying data generating process. Right now we're only concerned with the math behind the OLS estimator, we're not worried about interpretation yet.

- [Gre12] Chapter 2 covers the OLS assumptions; you've almost certainly seen them before.
- [Gre12] Chapter 3 covers algebraic properties of OLS. You should know what they are and should be familiar with the results, but you certainly do not need to memorize them.
- Chapters 4 and 5 of [Gre12] cover the important statistical properties of OLS — finite sample properties and asymptotic properties. This is the main part of the reading. Note that the asymptotic properties are almost exactly the same as the finite-sample properties under normality, but the justification is via the CLT rather than finite-sample arguments.
- You're only required to read a small part of Chapter 9 of [Gre12] — how to derive consistent estimators of the OLS standard errors if the ϵ's are heteroskedastic.

This is a fair bit of reading, but you should have seen all of this material in undergraduate econometrics already. The only difference is the notation. Most undergraduate textbooks don't use matrix notation, but we will now.

Recommended but optional material

- [CB02] Chapter 11 (along with ANOVA)
- [Gre12] Appendix A reviews matrix algebra
- [Fre09] Chapters 1 – 5

5 Regression modeling

In this part of the class, you're going to learn how to build a regression model for data analysis — what variables should be used as regressors, how variables should be transformed, etc. We're still going to focus mostly on OLS. There are many other modeling strategies that you'll study later in the program, but the intuition from the OLS model will be still be useful then. Much of the math will be useful as well.

To think of a specific example, imagine that we're interested in the labor-market decisions of college-age Americans. There are many factors

that could/should influence their decisions, so deciding which variables should be used as regressors is difficult. Deciding on the dependent variables may be difficult as well, and there may be several outcomes that you'd like to understand. The material you reviewed in Part 4 might be appropriate if all of those decisions were handed to you (although not always), but they usually are not. So, in this part of the class, we'll look at different strategies for making those decisions, and how they can influence or bias the empirical results.

We'll also start to look at the difference between *forecasting* and *policy analysis*. Forecasting is usually the easier problem — once we have built a model of labor-market decisions (for example), we can use it to predict decisions that will be made in the future. But we may not be able to easily build a model that predicts hypothetical outcomes of hypothetical decisions that will be made in the future; there are issues related to "omitted variable bias" and "selection bias" (and many other forms of bias) that make predicting hypothetical counterfactuals more difficult than predicting the outcome of a decision that's already been made. So we will look at those issues as well, and discuss when a model is appropriate for policy analysis (e.g. hypotheticals) and when it is only appropriate for forecasting.

- The readings in [Gre12] Chapter 5 cover the basics of model selection very briefly.
- [Gre12] Chapter 6 discusses transformations of the regressors and should be review from undergrad.
- The additional reading in [Gre12] Chapter 9 covers weighting strategies that can be more efficient than OLS under heteroskedasticity. You have probably already seen them in undergrad. Regardless, you just need to be familiar with these approaches.

  You can think of this chapter as modeling the conditional variance of a system, while most of the other readings discuss modeling the conditional expectation.
- [Gre12] Chapter 10.1 – 10.3 shows that there can be efficiency gains when estimating several different regression models simultaneously. Again, be familiar with this material but do not get bogged down in too many details.
- [GH07] Chapter 6 covers a different approach than OLS. If the dependent variable (y_i, usually) is not continuous and unbounded, OLS can be inappropriate. There is an extension of linear regression called the *Generalized Linear Model* that can be appropriate for applications where y_i is, for example, a positive integer that represents the number of times a particular event happened under different conditions. Since

y_i would be discrete in this case, OLS is likely to be inappropriate, and interpreting the estimated coefficients could be difficult.

Read over the chapter to get a general impression of the technique. Many economists aren't taught GLMs explicitly which can lead to ad hoc analysis when the dependent variable needs to be transformed. This reading is meant to help you avoid that mistake.

Recommended but optional material

- [Fre09] Chapters 6 and 7
- [LP09] is a review article that covers model selection in finance.

6 Program evaluation

In this part of the class, we're concerned with settings where there is a *specific* effect that we want to understand — the effect of a particular training program on future employment, for example — but we are much less interested on other effects unless they change the effect of the treatment — we don't necessarily care whether or not job-seekers face age discrimination, for example, but we may care whether or not the training program is only effective for younger workers. This as a fundamentally different approach than we covered in Part 5.

Continuing the example, we know that if individuals choose whether or not they participate in the training program, it can bias our estimates of its effect. Even if program graduates find jobs easily, we're not going to be able to determine whether that's because of the program itself, or because of other factors that both caused them to participate in the program and also cause them to be more employable (which is pretty likely here — one would expect that motivated and educated people are good employees and also tend to seek out additional training).

The approach we studied in Part 5 tells us that this is only a problem if we can't observe these additional factors, and that we should try to proxy for them by including demographic information (race and age, etc.), measures of education and experience, etc. Once we have all of the variables in the model, we can run OLS with a "training program" dummy variable and interpret it as the causal effect of the training program on future employment. But there are a few obvious problems with this approach:

1. We often can't observe all of the relevant variables.
2. To determine whether or not a training program works, now we're expected to have a complete model of discrimination and employability.

The best way to address these issues would be running an experiment: if people are assigned to the training program randomly, it is unrelated to

other factors that affect employability, and we can get consistent (often unbiased too) estimates of its effectiveness. The two catches are:

1. Even experiments will typically violate the exogeneity assumptions we've been making so far, which implies that our assumptions have been stronger than necessary. And they have: we've made assumptions that let us estimate the entire parameter vector without bias/consistently. Since we only want to estimate a single parameter here, we can weaken those assumptions.
2. Experiments in economics are often impossible.

So the reading is going to help us understand exactly why experiments work and to give us a mathematical framework for thinking about experimental intervention: Rubin's *Potential Outcomes* framework. (It is discussed in several of the papers, not just Rubin's.) You need to become familiar with it and understand how it relates to exogeneity in the linear regression model.

- [Fre91] is a skeptical assessment of regression in the social sciences. Take it to heart. I'll only ask very basic, high level questions about it on the RAT, though — just to verify that you have actually read it. You can think of the tools we're studying in this part of the class as attempts to address these shortcomings.
- [Fis26] is a short and nontechnical paper that discusses some of the key aspects of experimental design. Since the tools we study in this section are meant to approximate an experiment, you should know something about experiments as well. Again, I'll ask only basic, high level questions.
- [Ros09] makes the bridge between well-run observational studies and experiments, and lays out clear steps that you can apply in your own research. Again, I'll ask only basic, high level questions. Although this book focuses on matching estimators — which we won't cover — the main points apply equally well to regression.
- [Rub05] is a short overview of the *potential outcomes framework*. This is a notation for thinking about counterfactual outcomes that simplifies a lot of conceptual problems.
- [IW09] shows how to use the potential outcomes framework to study linear regression and "natural experiments." As before, you can ignore material on matching estimators or the propensity score and focus on the material related to regression and instrumental variables.
- [Gre12] is a very conventional treatment of Instrumental Variables and is useful because it spells out the math and formulas behind the estimator. The assumptions here are fairly strong and are analogous to the

assumptions [Gre12] presents for linear regression. We'll think about relaxing these assumptions in class, but first you have to familiarize yourself with the estimator.

Recommended but optional material

- [Fre09] Chapters 9 and 10 (and reread 1)

References

[CB02] George Casella and Roger L Berger. *Statistical Inference*. Duxbury Press, 2nd edition, 2002.

[Fis26] R.A. Fisher. The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain*, 33:503–513, 1926. Available at «http://digital.library.adelaide.edu.au/dspace/handle/2440/15191».

[Fre91] David A. Freedman. Statistical models and shoe leather. *Sociological Methodology*, 21:291–313, 1991. Available at «http://www.jstor.org/stable/270939».

[Fre09] David A. Freedman. *Statistical Models: Theory and Practice*. Cambridge University Press, revised edition, 2009.

[Gal97] A. Ronald Gallant. *An Introduction to Econometric Theory*. Princeton University Press, 1997.

[GH07] Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2007.

[Gre12] William H. Greene. *Econometric Analysis*. Prentice Hall, 7th edition, 2012.

[IW09] Guido W. Imbens and Jeffrey M. Wooldridge. Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1):5–86, March 2009. Available at «http://ideas.repec.org/a/aea/jeclit/v47y2009i1p5-86.html».

[LP09] Hannes Leeb and Benedikt M. Pötscher. Model selection. In *Handbook of Financial Time Series*, pages 889–925. Springer, 2009.

[Ros09] Paul R. Rosenbaum. *Design of Observational Studies*. Springer, 2009.

[Rub05] Donald B. Rubin. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469), 2005. Available at «http://www.jstor.org/stable/27590541».