

Topics in Advanced Statistical Methods
Political Science 597B
Fall 2004

101 Pond Laboratory Tuesday and Thursday 1:00-2:15

Professor De Boef

sdeboef@psu.edu

Office Hours Fridays, 9:30-11:50

320 Pond Laboratory

863-9402

Course Description.

This course is designed as part of the graduate methods sequence for students in political science. As such it builds on the analytical and theoretical background developed in 502 and 503. This is not a required course, but most students doing empirical work should enroll and all students wishing to take other methods courses offered in the department are likely to find it a requisite for specific classes. The course content is broken down into 3 parts. We begin with a review of the basic linear model beyond the problem of estimation. Our focus will be on the Gauss-Markov assumptions. Our treatment will be theoretical as well as practical. We will talk about remedial measures, but we will also talk about what features of political processes can give rise to assumption violations and how these features may lead to us to select alternative models that more accurately represent the underlying data generating process as we understand it. So armed, we move to the next sections of the course on models for time series data and for limited dependent variables. While these models only begin to touch the surface, they represent important classes of models that seem well matched to the kinds of processes and data social scientists regularly encounter.

Content in More Detail.

The Gauss-Markov Assumptions. This section of the course will begin with a quick review of the Gauss-Markov regression assumptions, the estimators of the regression parameters, their variances, and other quantities of interest. We will then work from Wooldridge's text and move quickly through the regression assumptions. Class sessions will focus on when, in practice, the assumptions are likely to be violated and on deriving the properties of estimators when we violate the regression assumptions. We'll then present solutions to the problems. Much of the class session will be presented in theoretical terms. Homework will be a mix of theoretical and applied problems.

Time series analysis. The time series analysis section of the course will introduce students to the fundamental importance of time dependence in regression analysis; we will talk about the important concepts of (non)stationarity and unit roots, how to test for stationarity (and unit roots), and then will maintain the regression analysis tradition a la Hendry (or pretty much any standard econometric textbook treatment) for model building. We will focus on a typology of dynamic regression models for stationary time series and how to test for and interpret multiplier effects. We will then move to models for nonstationary and cointegrated time series. We will also talk about Granger causality and a variety of important concepts of endogeneity. We will rely on Wooldridge for this section of the course, but will supplement heavily with a variety of other texts for this

part of the course. The bookstore will have the Enders text, but you probably do NOT want to purchase this text unless you plan on taking time series analysis down the road. I will use it when I teach time series next year, but will not use it as much as I originally planned for this course.

Limited dependent variables. The section of the course devoted to limited dependent variables will follow Scott Long's book fairly closely. I will assume that students have already had a basic introduction to probit and/or logit and will as such go somewhat quickly through a review of this material so that we can spend more time covering models for multinomial choice, ordered choice, censored/truncated data, and count models.

Required Texts.

These have been ordered. The texts are expensive and you may want to share them. Wooldridge, Jeffrey. 2002. *Introductory Econometrics: A Modern Approach*, 2nd edition. Thomson, South-Western: Mason, Ohio.

Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Sage Publications, Inc.: Thousand Oaks, California.

Recommended Texts. I'll make copies of these available for you to copy sections from as recommended.

Enders, Walter. 2004. *Applied Econometric Time Series*, 2nd edition. John Wiley and Sons, Inc.: Hoboken, New Jersey.

Gujarati, Damodar. *Basic Econometrics*.

Long, J. Scott and Jeremy Freese. 2003. *Regression Models for Categorical Dependent Variables Using Stata*, Revised Edition. Stata Press: College Station, Texas.

I have extra copies of Scott, Enders, Wooldridge, and Gujarati that you may borrow for short periods of time as well.

Course Requirements.

There will be 4 types of regular assignments for this course---reading assignments, equation-based pencil/paper assignments, data assignments, and free-writing assignments---as well as a final poster. Weekly reading assignments should be read before (and again after!) the class period in which they are covered. Typically the readings are quite short relative to readings in other classes, but you will find that a solid understanding of the material will require you to read and reread the material. With the exception of the poster, the remainder of the assignments must be kept and turned in to me in a homework notebook or labbook. I recommend a 3-ring binder (preferably a flexible and skinny one that is easy for both you and me to lug back and forth!).

Rather than a final exam, you will present a poster at the end of the semester as if you were presenting a poster at the APSA, for example. We will hold a special class session and students and faculty from the department will be invited to attend. We'll talk more about the posters in class in the first few weeks.

Grading:

<i>Weekly homework assignments:</i>	65%
<i>Poster:</i>	25%
<i>Participation:</i>	10%

The Lab Sessions.

Lab sessions are scheduled each week. There is, however, no one scheduled to teach them. The plan for these lab sessions is thus a mixture of things. First, we will schedule the labs as working sessions that we will plan to supplement the class and that you will run with my help. Second, I am planning some sessions to coincide with activities run by the Statistics Partnership. These will cover issues of data presentation/interpretation (making tables, graphs, etc...). Third, I've asked Kyle Joyce to plan some presentations on using Clarify to present estimation results. This latter may coincide with the second or may simply be something Kyle does for the class. We may also use these sessions to play catch up if we fall behind or to discuss applied papers on techniques covered in class. We'll see what works. In any case, however, we will be rescheduling the time for the lab as it conflicts with TA assignments of multiple students enrolled in the course.

A word on attendance. It is important that you do not miss class unless absolutely necessary. When possible, please let me know in advance if you are ill or cannot make class.

Tentative Schedule

The schedule below is listed without dates. Roughly we'll go through each topic in 5 weeks. Since this is a new course, we'll set the pace together. At the end of each class period we'll confirm that where we are, what we've accomplished so far, and what needs to be read for the next class.

Part I. The Linear Model—Beyond Estimation: The Cross-Sectional Case

1. Introduction
2. The estimators and the Gauss-Markov assumptions in the cross-sectional case

Required Readings:

- Wooldridge Chapters 2 and 3, skip pages 89-95

Recommended Readings:

- Gujarati Chapters 2-9

3. Violating the assumptions in the cross-sectional case
4. Issues of sample selection

Required Readings:

- Wooldridge sections 9.4 and 17.5

Recommended Readings:

- Dougherty pages 297-298
- Greene pages 778 and surrounding

5. Violating the zero conditional mean assumption

Required Readings:

- Wooldridge Chapter 3 pages 89-103, and Chapter 9, beginning through-9.3
- 6. Issues pertaining to collinearity
Required Readings:
 - Gujarati, pages TBA
- 7. Violating the homoscedasticity assumption
Required Readings:
 - Wooldridge Chapter 8, beginning through-8.4Recommended Readings:
 - Gujarati, pages TBA
- 8. Issues relating to the normality of the errors
Required Readings:
 - Wooldridge Chapter 4, introduction

Part 2: Time Series Analysis

- 9. Introduction to time series
- 10. A typology of time series models.
Required Readings:
 - Wooldridge Chapter 10, beginning through-10.2Recommended Readings:
 - TBA
- 11. The Gauss-Markov Assumptions for Time Series Regression
Required Readings:
 - Wooldridge Chapter 10, 10.3-10.5
- 12. Understanding Time Dependence.
Required Readings:
 - Wooldridge Chapter 11, all.
 - Wooldridge, Chapter 18, sections 18.2 and 18.3
- 13. Violating the no serial correlation assumption
Required Readings:
 - Wooldridge Chapter 12, all.
- 14. Spurious Regression and Cointegration
Required Readings:
 - Wooldridge, Chapter 18, section 18.4, Enders pages 320-350.
- 15. Endogeneity/Exogeneity and Granger Causality.
Required Readings:
 - TBA

Part 3: Limited Dependent Variables

- 16. Introduction to limited dependent variables.
- 17. Maximum Likelihood Review(Introduction?)
Required Readings:
 - Long Chapters 1-2.Recommended Readings:
 - King, pages TBA
- 18. Binary Choice, estimation.
Required Readings:

- Long Chapter 3, beginning – 3.6.
19. Binary Choice, interpretation.
Required Readings:
- Long Chapter 3-7 to end.
20. Binary Choice, hypothesis testing.
Required Readings:
- Long Chapter 4, skip 4.2.
21. Ordered Choice.
Required Readings:
- Long Chapter 5, skip 5.6.
22. Multinomial Choice.
Required Readings:
- Long Chapter 6, skip 5.6.
23. Limited Outcomes.
Required Readings:
- Long Chapter 7.
24. Count Models.
Required Readings:
- Long Chapter 8.

Academic Dishonesty¹

The Department of Political Science, along with the College of the Liberal Arts and the University, takes violations of academic dishonesty seriously. Observing basic honesty in one's work, words, ideas, and actions is a principle to which all members of the community are required to subscribe.

All course work by students is to be done on an individual basis unless an instructor clearly states that an alternative is acceptable. Any reference materials used in the preparation of any assignment must be explicitly cited. In an examination setting, unless the instructor gives explicit prior instructions to the contrary, whether the examination is in-class or take-home, violations of academic integrity shall consist of any attempt to receive assistance from written or printed aids, or from any person or papers or electronic devices, or of any attempt to give assistance, whether the one so doing has completed his or her own work or not.

Other violations include, but are not limited to, any attempt to gain an unfair advantage in regard to an examination, such as tampering with a graded exam or claiming another's work to be one's own. Violations shall also consist of obtaining or attempting to obtain, previous to any examinations, copies of the examination papers or the questions to appear thereon, or to obtain any illegal knowledge of these questions. Lying to the instructor or purposely misleading any Penn State administrator shall also constitute a violation of academic integrity.

In cases of a violation of academic integrity it is the policy of the Department of Political Science to impose appropriate penalties that are consistent with University guidelines.

¹Much of the text above has been directly obtained from the sections of the Princeton University website <http://www.princeton.edu/pr/pub/rrr/99/pages/OI.htm> concerning academic integrity (Rights, Rules, Responsibilities introductory text as well as pages 55-69) as well as from the website of the Department of Economics at The Pennsylvania State University.

Disabilities

The Pennsylvania State University encourages qualified people with disabilities to participate in its programs and activities and is committed to the policy that all people shall have equal access to programs, facilities, and admissions without regard to personal characteristics not related to ability, performance, or qualifications as determined by University policy or by state or federal authorities. If you anticipate needing any type of accommodation in this course or have questions about physical access, please tell the instructor as soon as possible. Reasonable accommodations will be made for all students with disabilities, but it is the student's responsibility to inform the instructor early in the term. Do not wait until just before an exam to decide you want to inform the instructor of a learning disability; any accommodations for disabilities must be arranged well in advance.

Visit our web site

The Political Science Department is in the process of upgrading its web site and will continue to do so during this and future academic years. In the undergraduate section you will find a wealth of information including course schedules, faculty office hours, faculty home pages describing their areas of teaching and research activities, answers to questions about advising, internship opportunities, announcements, and much, much, more. Check back often: we will continuously update our information about internships and career opportunities: <http://polisci.la.psu.edu/>

Homework Assignment Overview and Guidelines.

De Boef
Political Science 597B
Fall 2004

As I noted, there will be 4 types of regular assignments for this course---reading assignments, equation-based pencil/paper assignments, data assignments, and free-writing assignments. I can't stress enough how important the reading assignments are for the course. I recommend reading the assignments at least 3 times. If you find that you don't like Wooldridge or that you like Gujarati or some other text (Greene, Maddala, or someone else), then read it several times, you'll be way ahead of the game as we progress if you do! There will likely be a small number of equation-based pencil/paper assignments, but these will be smaller in number than the remaining types of assignments. I'll say more about the data assignments below. The free writing assignments are the most unusual for a methods class. These assignments are designed to complement the pencil/paper assignments and get you to think about the big picture. Free-writing assignments will take a variety of forms, but early on you can expect me to ask you to write about key terms or the process of analysis using your own words, without using books or notes.

More on the data assignments

Throughout the semester we will be working with a variety of different types of data structures. Early on we will be working with cross sectional data---subscripts will be of the little-i variety---and dependent variables will be continuous. Our discussions will feature countries or voters and the kinds of problems we'll worry most about will tend to be spatial in nature. Some of you will find that the kinds of problems you are most interested in will lead you to work most often with this type of data. I'll ask you to share your data with the class for homework assignments in this part of the course. As we move to time series analysis the subscripts will shift from "i" to "t" and the unit of analysis will less often be individuals and more often be countries. The kinds of statistical problems we focus on will deal with time dependence. Most likely a different set of you will find that the research questions that inspire you most are best answered with time series data and for those of you, I'll ask that you share your data for homework assignments in this portion of the course. The third section of the course will shift back to i subscripts but this time we'll have interval or nominal data. Those of you interested in (voter) choice are likely to find this section of the course best matched to the kinds of questions you are most interested in. Here I'll ask you to make your data available to your classmates for homework assignments. Throughout the semester you will work with a variety of datasets. I'll expect you to work together to familiarize yourselves with the data and independently as you select models to run and as you write up interpretations of estimation results.

Making your data available to your classmates/assignment number one.

Part of your first homework assignment will be to prepare a codebook/dictionary of a dataset you are interested in/willing to share with your classmates this semester. The codebook must include variable names and basic descriptive information (substantive as well as statistical) of enough variables to estimate a reasonably well specified (and interesting) model. The dependent variable must be suited to one of the types of data we will be modeling this semester: continuous dependent variables across space, time series (with at least 30 time points and preferably 50 or more), limited dependent variables (ideally ordered choice that we could also collapse to a dichotomous choice).

Homework Notebooks

Each homework (including free-writing and pencil/paper, as well as the data) assignment must be turned in with a running/cumulative notebook of homework assignments and lab notes. The contents of the notebook will include a variety of types of information including the following: dictionary or codebook information on all variables used in any models estimated in any homework assignments (I should always be able to check back easily and find basic summary statistics for any variable used), results from previous homework estimating simpler models, results from previous homework assignments if I've asked you to redo homework, any comments I've written up and handed back, and other things I'm not currently thinking of! This might also be a good place to keep handouts on computing hints I give you, too.

Hints

Homework grades will reflect interpretation and presentation as well as accuracy and detail. When homework assignments require the use of statistical software, you will need to turn in the program file, which should include a header detailing the purpose of the program file, the date, your name, the name of the program file, the data file, and any other pertinent information, the body of the program file should also contain comments detailing the purpose of sections of the code (this is just good computing practice). I recommend that you use Stata (or R). I will expect estimation results to be presented in tables, with stars denoting levels of significance and the like (Stata will produce results in tabular format for you so this isn't a bit deal). Tables should stand alone with real variable names (no contg for contiguity, for example). When graphical presentation of results would serve the reader the best, that's what I'll want to see.