

**Applied Statistics 450**  
**Topics in Applied Statistics: Panel Data**  
**Washington University in Saint Louis, Spring Semester 2007**  
Lecture: TTh 1300–1430, Eliot 115

Instructor: Robert W. Walker, Ph. D.  
Office: 317 Eliot Hall  
Phone: (314) 935-5881  
Email: rww@wustl.edu  
WWW: <http://rww.wustl.edu>  
Office Hours: M 1130-1330

**Course Description:** This topics course in applied statistics develops models for the analysis of data collected from distinct units at multiple points in time. Panel data can be generically described as containing multiple units observed at multiple points in time. Because panel data require attention to both heterogeneity (arising from units) and dynamics (arising from time), we will cover both topics individually, in summary form, before considering their interaction and developing intuitions for situations that require greater attention to one than the other. Though a host of other topics will receive attention, we will focus on the following issues: (1) Can individual time series be pooled and under what conditions? (2) Deterministic vs. random sources of variation arising from units or time points; (3) If the aforementioned effects are present, do they manifest in intercepts or coefficients; and (4) What issues arise in translating techniques for panel data to censoring, truncation, and other pathologies that result in limited dependent variables? While the Classical Linear Regression [CLR] Model has wonderful small sample properties and is often a useful model for a whole host of data classes, there are also occasions when generalizations of this model are to be preferred.

This course has two general foci: (1) to prepare students with an understanding of the unique challenges posed by longitudinal/panel data and (2) to provide students with tools to implement extant models from statistics and econometrics or develop their own when extant models prove inappropriate. In general, this course is to be an undergraduate survey of longitudinal and panel data.

The general design of the course can best be described as a lecture-based seminar. Though lectures will cover key material and derivations, we will work through examples and new problems in a collaborative fashion. With this in mind, the classroom is but a small fraction of the course; you will learn by doing problem sets, readings, replications, or programming in statistical computing languages.

In practical terms, we will begin with definitions and an overview of panel data. We will then examine nonparametric techniques for summarizing panel data and test hypotheses regarding the homogeneity of cross-sections. We will then set about the meat of the course, extending the linear model to the host of pathologies that arise from panel data with models of dynamics and heterogeneity. Our final discussion of standard panel data models will focus on causal interpretation of panel data models (difference-in-difference and the like). Most of this course will focus

on conventional estimators for panel data, we will only briefly extend the course topics to models of discrete Markov chains and state-space transition models, limited dependent variables, and other data types. If time permits, we can extend the discussion of random effects models to multilevel settings.

The course will rely on three small monographs and a series of published articles. The three monographs are:

Luke, Douglas A. *Multilevel Modeling* Thousand Oaks, CA: Sage Publications (ISBN: 0761928790)

Markus, Gregory. 1997. *Analyzing Panel Data* Thousand Oaks, CA: Sage Publications (ISBN: 0803913729)

Saysr, Lois 1993. *Pooled Time Series Analysis* Newbury Park, CA: Sage Publications (ISBN: 0803931603)

**Grade Determination:** Final course grades will be determined on the basis of problem sets (25%), a midterm examination (25%), and a research project (50%). Because this is a course in statistics and grades have discrete and ordered statistical distributions, we will periodically discuss the details of the grading rules. The mix of undergraduates and graduates implies considerable diversity in the range of assignments.

**Statistical Software and Computation:** All of the models covered in this class can be estimated using standard software packages. You can use whatever software you wish, but I will only support two packages: Stata and R. Both of these packages contain most of the models necessary for this course, and provide programming functionality that allows the implementation of other models. I encourage each student to choose a software package, and use the package throughout the semester to estimate each type of model. Highly motivated students should perform all work in both Stata and R [or simply write estimation algorithms to maximize likelihoods directly in one or the other, maybe both]. Both Stata and R are available on Windows machines in the Social Science Computing Facility (SSCF) in Eliot Hall. If you want to become a serious Stata user, I strongly recommend that you purchase copies of the manuals. Stata has an excellent set of manuals: the Users Guide details how the software is structured; the Reference Manuals contain a detailed description of how each command works and each model is estimated; and the Stata Programming Reference Manual which is geared toward those who want to program their own models. All of these can be purchased from Stata directly using the GradPlan.

R is the best choice for graphical and exploratory data analysis, and is a powerful statistical programming language. R can be downloaded for free for Windows, Macintosh, Linux, and Unix operating systems from <http://www.r-project.org>. The R manual is also available for free on the web. The texts by Venables and Ripley (2002) and Dalgaard (2002) are also quite useful.

As a final note, some relatively simple models can be estimated by hand. Though the computer can find answers to these problems, you should be very cautious of understanding the underlying math and mechanics as you might just be asked to solve a computationally tractable variant by hand at some point.

**Problem Sets:** Students may complete (more or less) problem sets collaboratively, but each student will turn in their own solutions. Working together brings benefits to all, but it is imperative that each student have a satisfactory understanding of all portions of collective work. Because of the (potentially) collective nature of homework, late assignments will not be accepted without **prior**<sup>1</sup> consent of the instructor. It is my belief that experiences with team production are a beneficial part of both academic and professional life. I contend that you will best learn the material by learning and studying, but also by teaching, debating and working through problems with a diverse array of colleagues.

**Midterm Exam:** TBA. A take home examination to be distributed in the first half of March.

**Other Research Activity:** Each student will produce a manuscript that applies or develops the appropriate statistical model to an important substantive problem. Students will choose their own topics, although it would make sense to choose something that might lead to a publishable manuscript and/or thesis.

We can work out a schema for this and precise dates in the next two class meetings. In general, the assignment will come in three parts. (1) A rough draft, (2) a simultaneously constructive and critical memorandum on a classmate's paper, and (3) a final draft with appropriate revisions in accordance with (2).

### Reminders:

1. I expect that everyone will maintain a classroom conducive to learning. Thus, everyone is expected to behave with basic politeness, civility, and respect for others. This includes respecting the beginning and ending times for the course.<sup>2</sup>
2. My goal is to take you through a complete logically structured course that combines the key issues into a cohesive and self-contained whole. This plan is a bit ambitious and I reserve the right to alter the course readings beyond the midterm if it becomes infeasible.
3. I promise not to accept late assignments. This is my solemn vow.

**Suggestions:** Suggestions for improvement are welcome at any time. Any concern about the course or your performance in this course should be brought to my attention.

---

<sup>1</sup>Prior here means prior to the start of class on the date that the assignment is due.

<sup>2</sup>While probably unnecessary, but as an honest expression of an infuriating feature of modern life, I demand that cellular phones be set to silent (if not turned off) during my class.

We will keep a calendar posted on the syllabus but cannot really plan more than one month at a time.

Date	Lecture Title
01.16.07	Syllabus, Structure, and Review Materials
01.18.07	Summarizing and Describing Panel Data
01.23.07	A Review of the Linear Model: Heterogeneity
01.25.07	A Review of the Linear Model: Dynamics
01.30.07	To Pool or not to Pool? Hicks (1994)
02.08.07	UCSD – Classes cancelled.
02.22.07	SPPC – Classes cancelled.
03.12 - 03.16: SPRING BREAK	
04.12.07	MPSA – Classes cancelled.

- Fixed Effects and Pooling  
The Dirty Pool controversy in *International Organization* involving Green, Kim and Yoon (2001), Oneal and Russett, Beck and Katz, and King (2001).
- Standard Errors and Nuisance  
Beck and Katz (1995)
- Overviews for Political Scientists  
Stimson (1985)
- Random Coefficients  
Beck and Katz (2006)
- Bayesian Variation  
Western (1998); Western and Jackman (1994).
- Dynamic Panel Data Models  
Wawro (2002)
- Rarely Changing Indicators and Fixed Effects: Plümper and Troeger (N.d.)

## References

- Beck, Nathaniel L. and Jonathan N. Katz. 1995. "What to Do (and Not to Do) with Time-Series-Cross-Section Data in Comparative Politics." *American Political Science Review* 89(3):634–647.
- Beck, Nathaniel L. and Jonathan N. Katz. 2006. "Random Coefficient Models for Time-Series-Cross-Section Data: Monte Carlo Experiments." *Political Analysis* Forthcoming.

- Green, Donald P., Soo Yeon Kim and David H. Yoon. 2001. "Dirty Pool." *International Organization* 55(2):441–68.
- Hicks, Alexander M. 1994. Introduction to Pooling. In *The Comparative Political Economy of the Welfare State*, ed. Thomas Janoski and Alexander M. Hicks. Cambridge University Press.
- King, Gary. 2001. "Proper Nouns and Methodological Propriety: Pooling Dyads in International Relations Data." *International Organization* 55(2):497–507.
- Plümper, Thomas and Vera E. Troeger. N.d. "Efficient Estimation of Rarely Changing Variables in Fixed Effects Models." *SSRN eLibrary*. Forthcoming.
- Stimson, James A. 1985. "Regression in Space and Time: A Statistical Essay." *American Journal of Political Science* 29(4):914–47.
- Wawro, Gregory. 2002. "Estimating Dynamic Panel Data Models in Political Science." *Political Analysis* 10(1):25–48.
- Western, Bruce. 1998. "Causal Heterogeneity in Comparative Research: A Bayesian Hierarchical Modelling Approach." *American Journal of Political Science* 42(4):1233–59.
- Western, Bruce and Simon Jackman. 1994. "Bayesian Inference for Comparative Research." *American Political Science Review* 88(3):412–23.