Analyzing Panel Data

GSERM - St. Gallen

10-14 June 2024

Professor Christopher Zorn
Pennsylvania State University
E-mail: zorn@psu.edu

Overview

This is the syllabus to the *Analyzing Panel Data* course offered in the Global School in Empirical Research Methods (GSERM) during summer 2024. This syllabus contains details about course content, materials, and logistics. Clickable links are in HSG green.

Course Content

Analysts often find themselves presented with data that vary both over cross-sectional units and across time. Such *panel* data provides unique and valuable opportunities to address substantive questions in the economic and social sciences. This course begins with a discussion of some of the challenges and opportunities that such data provide, along with an exposition of the relevant dimensions of variation in panel data, and of approaches and techniques for description and visualization of such data. It will then progress to models for one-way and two-way unit effects (fixed, between, and random), models for complex panel error structures, dynamic panel models, models that leverage panel data to make causal inferences in observational contexts, and nonlinear panel models for discrete dependent variables. Students will learn the statistical theory behind the various models, details about estimation and inference, and techniques for the visualization and substantive interpretation of their statistical results. Students will also develop statistical software skills for fitting and interpreting the models in question, and will use the models in both simulated and real data applications. Students will leave the course with a thorough understanding of both the theoretical and practical aspects of conducting analyses of panel data.

Logistics

For summer 2024, our class will meet daily, in person, from 09:15-12:00 and 13:00-15:15 CET. Class materials will be made available primarily through the course's dedicated Github repository, at https://github.com/PrisonRodeo/GSERM-Panel-2024. Course materials will also be available through CANVAS at the University of St. Gallen.

Prerequisites (Knowledge of Topic)

Students should have a comfortable familiarity with univariate differential and integral calculus, basic probability theory, and linear algebra is required. Students should have completed Ph.D.-

level courses in introductory statistics and linear regression models, up to the level of *Regression III*. Familiarity with discrete and continuous univariate probability distributions will be helpful.

Hardware

Course exercises and exams will be completed on the students' own laptop computers. For purposes of this class, a laptop running any widely-used operating system (Windows, OS-X, Linux) will be acceptable.

Software

All lecture materials, slides, and in-class examples will be conducted using the R statistical language. Students are encouraged to come to class with current versions of both R and RStudio on their laptops. The instructor will also provide support for students wishing to use Stata for their analyses. Students may also elect to use other statistical software (e.g., Python, SAS, Julia, PSPP, etc.).

Literature

Mandatory

The course has one required text:

- Hsiao, Cheng. 2014. *Analysis of Panel Data*, 3rd Ed. New York: Cambridge University Press.

Alternatively, students may substitute:

- Croissant, Yves, and Giovanni Millo. 2018. *Panel Data Econometrics with R*. New York: Wiley.

Additional readings will also be assigned as necessary, all of which will be available on the GSERM StudyNet / CANVAS site and at the course github repository. See below for a detailed schedule of course topics and readings.

Supplementary / Voluntary

None.

Mandatory Readings Before Course Start

None.

Examination

Students will be evaluated on one written homework assignment that will be completed during the course (30 percent) and a final examination (70 percent). The homework assignment will involve a combination of simulation-based exercises and "real data" analyses, and will be completed during the evenings while the class is in session. For the final examination, students will complete the final examination either on the afternoon of the class's final day, or during the week following the end of the course (due date: June 21, 2024). Additional details about the final examination will be discussed on the first day of the course.

Supplementary Aids

The exam will be a "practical examination" (see below for content). Students will be allowed access to (and encouraged to reference) all course materials, notes, help files, and other documentation in completing their exam.

Examination Content

The final examination will involve the application of the techniques taught in the class to one or more "live" data example(s). These will typically take the form of either (a) a replication and extension of an existing published work, or (b) an original analysis of observational data with a longitudinal / panel component. Students will be required to specify, estimate, and interpret various forms of panel data models, to conduct and present diagnostics and robustness checks, and to give detailed justifications for their choices.

Workload

At least 24 units, 45 minutes each, on 5 consecutive days.

Course Structure and Content

Note that each day is accompanied by "Useful Readings." They are meant to be exactly that: readings that are useful for learning about the topic discussed. It is **not** expected that you will have read all (or even most) of the "Useful Readings" prior to that class meeting. All readings are available in the "Readings" folder in the course Github repo.

Day One: Pooling Data, Variation, and Visualization

Useful Readings:

- Bartels, Larry M. 1996. "Pooling Disparate Observations." *American Journal of Political Science* 40(August):905-42.

- Gassen, Joachim. 2020. "Using `ExPanD` for Panel Data Exploration." R vignette: https://cran.r-project.org/web/packages/ExPanDaR/vignettes/use_ExPanD.html (accessed 31 May 2022).
- Hsaio, Cheng. 2014. *Analysis of Panel Data*. Chapter 1.
- Hsiao, Cheng. 2007. "Panel Data Analysis – Advantages and Challenges." *Test* 16:1-22.
- Neuhaus, J. M., and J. D. Kalbfleisch. 1998. "Between- and Within-Cluster Covariate Effects in the Analysis of Clustered Data." *Biometrics* 54:638-45.
- Nuamah, Nicholas N. N. N. 1986. "Pooling Cross Section and Time Series Data." *The Statistician* 35:345-51.

Day Two: Unit Effects Models: One- and Two-Way Fixed-, Between-, and Random-Effects Models

Useful Readings:

- Bell, Andrew, and Kelvyn Jones. 2015. "Explaining Fixed Effects: Random Effects Modeling of Time-Series Cross-Sectional and Panel Data." *Political Science Research and Methods* 3:133-153.
- Bell, Andrew, Malcolm Fairbrother, and Kevin Jones. 2019. "Fixed and Random Effects Models: Making an Informed Choice." *Quality and Quantity* 53:1051-1074.
- Bliese, P. D., D.J. Schepker, S.M. Essman, and R.E. Ployhart. 2020. "Bridging Methodological Divides Between Macro- and Microresearch: Endogeneity and Methods for Panel Data." *Journal of Management* 46:70-99.
- Clark, Tom S. and Drew A. Linzer. 2015. "Should I Use Fixed Or Random Effects?" *Political Science Research and Methods* 3:399-408.
- Collischon, Matthias, and Andreas Eberl. 2020. "Let's Talk About Fixed Effects: Let's Talk About All the Good Things and the Bad Things." *Kólner Zeitschrift für Soziologie und Sozialpsychologie* 72:289-299.
- Hazlett, Chad, and Leonard Wainstein. 2022. "Understanding, Choosing, and Unifying Multilevel and Fixed Effect Approaches." *Political Analysis* 30:46-65.
- Hsaio, Cheng. 2014. *Analysis of Panel Data*. Chapter 3 (*or* Croissant and Millo, *Panel Data Econometrics with R*, Chapter 2).
- Kropko, Jonathan, and Robert Kubinec. 2020. "Interpretation and Identification of Within-Unit and Cross-Sectional Variation in Panel Data Models." *PLoS ONE* 15(4): e0231349. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0231349

- Mummolo, Jonathan, and Erik Peterson. 2018. "Improving the Interpretation of Fixed Effects Regression Results." *Political Science Research and Methods* 6:829-835.
- Plümper, Thomas, and Vera E. Troeger. 2019. "Not So Harmless After All: The Fixed-Effects Model." *Political Analysis* 27:21-45.
- Zorn, Christopher. 2001. "Estimating Between- and Within-Cluster Covariate Effects, with an Application to Models of International Disputes." *International Interactions* 27(4):433-45.
- Good tweets, blog posts, etc.:
  - What Panel Data Is Really All About, by Robert Kubinec.
  - Fixed Effects Infatuation, at Greed, Green and Grains.
  - Why I don't use the terms "fixed" and "random" (again), at Statistical Modeling, Causal Inference, and Social Science.

Day Three: Dynamic Panel Data Models

Useful Readings:

- Beck, Nathaniel, and Jonathan N. Katz. 1995. "What To Do (And Not To Do) With Time-Series Cross-Section Data." *American Political Science Review* 89:634-647.
- Croissant and Millo, *Panel Data Econometrics with R*, Chapter 7.
- Keele, Luke, and Nathan J. Kelly. 2006. "Dynamic Models for Dynamic Theories: The Ins and Outs of Lagged Dependent Variables." *Political Analysis* 14:186-205.
- Mehic, Adrian. 2021. "FDML versus GMM for Dynamic Panel Models with Roots Near Unity." *Journal of Risk and Financial Management* 14:405.
- Natalie Mizik and Eugene Pavlov. 2018. "Panel Data Methods in Marketing Research." In Natalie Mizik and Dominique M. Hanssens, Eds. *Handbook of Marketing Analytics*. Northampton, MA: Edward Elgar.
- Nickell, Steven. 1981. "Biases In Dynamic Models With Fixed Effects." *Econometrica* 49:1417-1426.
- Pickup, Mark, Paul Gustafson, Davor Cubranic, and Geoffrey Evans. 2017. "`OrthoPanels`: An `R` Package for Estimating a Dynamic Panel Model with Fixed Effects Using the Orthogonal Reparameterization Approach." *The R Journal* 9:60-76.
- Pickup, Mark and Vincent Hopkins. 2022. "Transformed-Likelihood Estimators for Dynamic Panel Models with a Very Small T." *Political Science Research & Methods* 10:333-352.

- Wawro, Gregory. 2002. "Estimating Dynamic Panel Data Models in Political Science." *Political Analysis* 10:25-48.
- Good tweets, blog posts, etc.:
  - Dynamic Panel Models, at the Social Science Statistics Blog.
  - Ignore at your Peril: Dynamic Panel Data, at the Econometrics Tutorial for Stata.

Day Four: Causal Inference with Panel Data

Useful Readings:

- Abadie, Alberto. 2005. "Semiparametric Difference-in-Differences Estimators." *Review of Economic Studies* 72:1-19.
- Abadie, Alberto. 2021. "Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects." *Journal of Economic Literature* 59:391-425.
- Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2010. "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program." *Journal of the American Statistical Association* 105:493-505.
- Blackwell, Matthew, and Adam N. Glynn. 2018. "How to Make Causal Inferences with Time-Series Cross-Sectional Data under Selection on Observables." *American Political Science Review* 112:1067-1082.
- Callaway, Brantley, and Pedro H.C. Sant'Anna. 2021. Difference-in-Differences with Multiple Time Periods. *Journal of Econometrics* 225:200-230.
- Doudchenko, Nikolay, and Guido Imbens. 2017. "Balancing, Regression, Difference-In-Differences and Synthetic Control Methods: A Synthesis." Working paper: Graduate School of Business, Stanford University.
- Gibbons, Charles E., Juan Carlos Suarez Serrato, and Michael B. Urbancic. 2018. "Broken or Fixed Effects?" *Journal of Econometric Methods* 8:20170002.
- Imai, Kosuke, and In Song Kim. 2019. "When Should We Use Unit Fixed Effects Regression Models for Causal Inference with Longitudinal Data?" *American Journal of Political Science* 62:467-490.
- Imai, Kosuke, and In Song Kim. 2021. "On the Use of Two-Way Fixed Effects Regression Models for Causal Inference with Panel Data." *Political Analysis* 29:405-415.
- Li, Fan. STAT 640: Causal Inference (course slides: `https://www2.stat.duke.edu/~fl35/CausalInferenceClass.html`).

- Liu, Licheng, Ye Wang, and Yiqing Xu. 2024. "A Practical Guide to Counterfactual Estimators for Causal Inference with Time-Series Cross-Sectional Data." *American Journal of Political Science* 68:160-176.
- Xu, Yiqing. 2017. "Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models." *Political Analysis* 25:57-76.bi
- Good tweets, blog posts, etc.:
  - Fixed Effects vs Difference-in-Differences, by Doug Johnson.
  - Problems with two-way fixed-effects event-study regressions, by Pedro H. C. Sant'Anna and Brantly Callaway.
  - The Causal Representation of Panel Data: A Comment On Xu (2022), by Robert Kubinec.
  - DiD Paper Thread by Nikolaj Harmon.
  - Staggered Difference-in-Differences with Two-Way Fixed Effects and Interactions by Vincent Arel-Bundock.

Day Five: Panel Data Models for Discrete Dependent Variables

Useful Readings:

- Ballinger, Gary A. 2004. "Using Generalized Estimating Equations for Longitudinal Data Analysis." *Organizational Research Methods* 7:127-150.
- Cameron, A. Colin, and Pravin K. Trivedi. 1998. *Regression Analysis of Count Data*. New York: Cambridge University Press. Chapter 9.
- Cook, Scott J., Jude C. Hays, and Robert J. Franzese. 2020. "Fixed Effects In Rare Events Data: A Penalized Maximum Likelihood Solution." *Political Science Research and Methods* 8:92-105.
- Crisman-Cox, Casey. 2021. "Estimating Substantive Effects in Binary Outcome Panel Models: A Comparison." *The Journal of Politics* 83:532-546.
- Hsaio, Cheng. 2014. *Analysis of Panel Data*. Chapter 7 (*or* Croissant and Millo, *Panel Data Econometrics with R*, Chapter 9).
- Muff, Stefanie, Leonhard Held, and Lukas F. Keller. 2016. "Marginal or Conditional Regression Models for Correlated Non-Normal Data?" *Methods in Ecology and Evolution* 7:1514-1524.
- Williamson, Tyler, and Pietro Ravani. 2017. "Marginal Structural Models in Clinical Research: When and How to Use Them?" *Nephrology Dialysis Transplantation* 32:ii84–ii90,

- Zorn, Christopher. 2001. "Generalized Estimating Equation Models for Correlated Data: A Review with Applications." *American Journal of Political Science* 45:470-90.
- Good tweets, blog posts, etc.:
    - Generalized Estimating Equations (GEE), at Practical Statistics.
    - An introduction to Generalized Estimating Equations at Towards Data Science.
    - Generalized estimating equations (GEE) and multilevel models at Statistical Modeling, Causal Inference, and Social Science.
    - Using inverse probability of treatment weights & Marginal structural models to handle time-varying covariates, by Mark Bounthavong.