ELPS 998: Applied Data Analysis in Policy Studies

Department of Educational Leadership and Policy Studies
University of Kansas, School of Education

Mondays, 7:00-10:00pm, Joseph R. Pearson Hall 147

Argun Saatcioglu Office hours: By appointment
(argun@ku.edu) (JRP 407)

Objective

This seminar is designed to help graduate students develop applied econometric skills in quantitative research. It focuses largely on analytical techniques for examining large scale, complex datasets common in policy studies and sociology, with particular emphasis on education. A variety of datasets are used at the state and federal level, such as Education Longitudinal Study (ELS) and Integrated Public-Use Microdata Series (IMPUS). Several substantive topics are applied to the datasets, such as teacher and school effects on student performance, faculty productivity in higher education, race and class issues in the K-12 achievement gap, differentiation of school and nonschool problems, neighborhood effects on schooling, the dropout problem, regional and metropolitan effects on attainment, minority over-representation in special education, and the effects of inclusion and exclusion levels on the performance of special education students.

As part of an applied seminar, most sessions in the class will follow a sequence where (1) a substantive policy- and theory-related research issue is addressed, (2) the process of question development and argumentation regarding the issue are discussed, (3) the set of common analytical techniques applied to that issue in the existing research are described, (4) the properties of the data structures commonly applicable to the questions on the issue are explained, and most importantly, (5) the mechanics and the practical application of the relevant analytical techniques are addressed. The typical session will include hands-on application of techniques on the computer as well as other exercises conducted in class.

Text and materials

(4) Other readings such as published research articles, book excerpts, and chapters assigned for various sessions will be provided by the instructor in electronic format.
Software

(1) We will use STATA. Versions 10 IC or 11 IC, or even older version (all the way down to 8.0) will work for the class. Please do not purchase Small STATA, as it will not have the capacity to support large datasets. STATA IC (intercooled) allows over 2,000 variables. You can also work with version SE (special edition) or MP (multi-core), both of which have higher capacities than IC. STATA 11 IC costs $100 under KU’s grad plan.

(2) We may possibly use some SAS 9.1.3, particularly for data manipulation, and for specific categorical regression and simultaneous regression techniques. SAS is available in many labs on campus, and could also be purchased from KU for about $70. It may not be necessary for you to obtain a perpetual personal copy of SAS since we will not use this software intensively, and may use it even less depending on how the class goes.

Prerequisite

A solid foundation in graduate-level basic statistics is required. This involves decent familiarity with univariate and bivariate statistical concepts and tests (e.g., common distributions, central tendency measures, variance-covariance, correlation, t test, significance level and hypothesis testing, and linear regression). Experience with any popular statistical analysis package is also helpful though not essential.

Evaluation

Weekly assignments (10x4) 40%
Midterm exam 20%
Final exam 20%
Attendance and class participation 20% (don’t miss class!!)

* The details of each weekly assignment and the midterm and final exams will be further discussed in class.
Class Schedule

1/24. Session 1—Introduction

Methodological topics:
- Review of univariate and bivariate statistics, t test, oneway ANOVA, simple linear OLS and its assumptions.

Substantive topics:
- Faculty productivity and compensation in higher education.
- Match achievement in high school.

Readings:
- Wooldridge, chapters 2 and 3
- Acock, chapters 4, 5, and 8 (helpful in getting familiar with STATA)

Datasets:
- faculty.dta
- hsb.dta

1/31. Session 2—Fundamental threats to unbiased estimation in large datasets (1)

Methodological topics:
- Exogeneity/endogeneity in model specification, consistency and efficiency in coefficient estimation, omitted variable bias, selecting control variables, joint F and LM tests, outliers and influentials.

Substantive topics:
- Faculty productivity and compensation in higher education.
- Neighborhood effects on student achievement.

Readings:
- Wooldridge, chapters 4, 5, 6, and 9
- Chow, Julian and Claudia Coulton. 1998. “Was There a Social Transformation of Urban

Datasets:
- faculty.dta
- neighborhood.dta
- CMSD_set1.dta

Assignment 1:
- Fit OLS models using CMSD_set1.dta to predict neighborhood effects on GPA and truancy (unexcused absence). The particular neighborhood indicators will be specified in class. Propose and use various control variables as you see fit. Explain your reasoning. Discuss issues of model specification, potential mediation issues, omission bias, and influential observations. Implement joint F tests on your models as necessary. After fitting the models for all students, repeat the entire exercise by race (white/black). Be sure to construct informative, efficient tables and thoroughly discuss your findings at each step.

2/7. Session 3—Fundamental threats to unbiased estimation in large datasets (2)

Methodological topics:
- Collinearity, heteroskedasticity, sampling weights, and clustered observations.

Substantive topics:
- Effects of high school performance on college GPA.
- Racial achievement gap in the public schools.

Readings:
- Wooldridge, chapter 8

Datasets:
- ELS0204_set1.dta
- CMSD_set2.dta
- gpa1.dta

Assignment 2:
- a) Fit an OLS model using ELS0204_set1.dta to predict the effect of grade 10 math score (base year) on grade 12 test score (first follow up). Control for available racial, socioeconomic, and other important background factors. Test for heteroskedasticity and discuss the potential problems that may arise under heteroskedastic error distribution. Re-estimate the model if necessary with appropriate corrections.
b) Fit the same OLS model, this time addressing the issue of sampling weights concerning selection of students from base year into the first follow up.

c) Fit a model using CMSD_set2.dta that predicts the effect of neighborhood conditions on GPA in grade 10. Control for other relevant and available factors in model specification, and test for whether weighting by race is necessary (i.e., OLS or WOLS).

2/14. Session 4—Regression with categorical and count data

Methodological topics:
- Logit/probit regression, ordinal and multinomial models, poisson model.

Substantive topics:
- The dropout problem.
- Race and class biases in special education identification.
- Predictors of truancy in middle and high school.

Readings:
- Wooldridge, chapter 17 (553-65; 573-578).
- Acock, chapter 11.

Datasets:
- ELS0204_set2.dta
- CMSD_set3.dta
- CMSD_set4.dta

Assignment 3:
- a) Fit logit and probit models, using ELS0204_set2.dta, to predict the effects family and other background factors on dropping out of high school. Use the appropriate weight variable for the first follow-up (F1) data collection for ELS. Discuss fundamental differences in the assumptions and results, if any, between logit and probit estimates.
- b) Fit an ordinal logit model, using CMSD_set3.dta, to estimate the effects of race, class, gender, family, neighborhood, and school factors in the inclusion level (1, 2, 3) of SPED students in regular education classrooms.
- c) Fit a multinomial logit model to estimate the effects of the same predictors on the student’s identification as BH, DH, LD, other SPED, and regular education. (Bonus: Another multinomial model that accounts for inclusion levels in the outcome variable).
- d) Fit a poisson model using CMSD_set4.dta that predicts the effects the same set of predictors on excused and unexcused absences. Compare and discuss the results of the two analyses.
2/21. Session 5—Examining moderation and mediation effects

Methodological topics:
- Interaction terms, path models, testing time effects.

Substantive topics:
- Faculty productivity in higher education.
- Effects of residential mobility on academic performance.
- Changes in the effects of race and class on achievement over time.

Readings:

Datasets:
- faculty.dta
- achievement_set1.dta
- ELS0204_set3.dta
- CMSD_set5.dta

Assignment 4:
- a) Fit an OLS model, using ELS0204_set3.dta, to predict the interaction effects of race and socioeconomic status on base year math and reading test scores (use base year student weights and control for other relevant background factors).
- b) Fit an OLS model, using ELS0204_set3.dta to predict the mediated effect of family composition on math and reading test scores (use base year student weights and control for other relevant background factors). As mediators, test the role of various family processes and cultural capital variables.
- c) Fit logit and OLS models, using achievement_set1.dta, to estimate the effects of residential mobility’s interaction with grade level on grade retention and on math achievement.
- d) Fit an OLS model, using CMSD_set5.dta, predicting the effects of residential mobility on GPA over time (years). Use interaction terms to specify the time effect.
2/28. Session 6: Understanding longitudinal and other cluster structures in large datasets

Methodological topics:
- Introduction to repeated measures regression, standard error adjustment to clustering, autoregressive structures, general estimating equations, simple panel models.

Substantive topics:
- Racial achievement gap in the public schools.
- Neighborhood effects on performance.
- Student and school effects on motivation to pursue graduate degree.
- School finance and student performance.

Readings:
- Wooldridge, chapter 13.

Datasets:
- gpa3.dta
- neighborhood.dta
- postdoc.dta
- CMSD_set6.dta
- papke05.dta

Assignment 5:
- a) Fit an OLS model, using neighborhood.dta, to predict the effect of family background on achievement. Compare your results to the same model fitted with standard errors cluster-corrected for neighborhood nesting and school nesting (separately).
- b) Fit a logit and a GEE model, using postdoc.dta, predicting the odds of postdoctoral training given the available predictors in the dataset. Display and discuss the results and interpret the differences (students nested within universities).
- c) Fit an OLS and a simple fixed-effects panel model (difference-in-difference), using CMSD_set6.dta, predicting the effects of race, class, neighborhood and family factors on GPA.

3/7. Session 7—Review and midterm exam

3/14. Session 8—Advanced models for longitudinal panel data

Methodological topics:
- Dummy variable regression, fixed effects regression, random effects regression, the Hausman test.
Substantive topics:
- Racial achievement gap in the public schools.
- Value added teacher effects on learning.
- School effects on student performance.

Readings:
- Wooldridge, chapter 14.

Datasets:
- gray_set1a.dta; gray_set1b.dta
- gray_set2a.dta; gray_set2b.dta
- CMSD_set7a.dta; CMSD_set7b.dta
- CMSD_set8a.dta; CMSD_set8b.dta

Assignment 6:
- a) Fit OLS, fixed, and random effects models, using gray2.dta, to predict teacher value added effects on student performance. Compare and discuss the results. Which one is empirically more robust?
- b) Use the preferred results from above to estimate the effects of the available predictors of teacher value added stored in gray_set2.dta.
- c) Fit OLS, fixed, and random effects models, using CMSD_set8.dta to predict school value added effects on student GPA. Compare and discuss the results. Which one is empirically more robust?
- d) Use the preferred results from above to estimate the effects of the available predictors of school value added stored in CMSD_set8.dta.

3/21. No class, spring break
3/28. Session 8: Models with systems of equations, limited outcomes, and censored samples

Methodological topics:
- Instrumental variables regression (2SLS), truncated regression, Tobit and Heckman selection (Heckit) models.

Substantive topics:
- Family and neighborhood effects on the schools.
- Teacher and principal turnover and migration in the public schools.
- The urban dropout problem.

Readings:
- Wooldridge, chapter 15.

Datasets:
- CMSD_set9a.dta, CMSD_set9b.dta
- CMSD_set10.dta
- charlie.dta

Assignment 7:
- a) Fit OLS and 2SLS models, using CMSD_set9a.dta, to estimate the effects of family structure on school value-added. In 2SLS, discuss your choice of instruments, the first stage results, and test for exogeneity. Also, discuss the conceptual and technical rationale for your choice of controls in both stages. Finally, compare and discuss your results from OLS and 2SLS, discuss.
- b) Fit OLS and 2SLS models, using CMSD_set9b.dta, to replicate Michael and McLanahan’s (1996) findings regarding neighborhood effects on finishing high school. As a proxy for their city-level measures, use ZIP code-level predictors. Compare and discuss the results of the two approaches. Any comments/critiques on Michael and McLanahan’s study?
- c) Fit a Heckit model, using CMSD_set10.dta, to predict the effects of family, neighborhood, and income factors on yearly GPA gain in high school. Adjust the regression for the probability of zero gain due to dropping out. Include relevant controls and compare your results to findings from an OLS model for the same analysis.
4/4. Session 10—Hierarchical regression with large datasets (1)

Methodological topics:
- Decomposition of variance components in multi-level data structures, specification of fixed and random effects, cross-level interactions.

Substantive topics:
- School and neighborhood effects on achievement and attainment.
- School value-added.
- Geo-spatial effects on attainment.

Readings:

Datasets:
- achievement_set2.dta
- ELS0204_set4.dta
- CMSD_set11.dta
- IPUMS4080.dta

Assignment 8:
- a) Fit an OLS and a hierarchical model, using ELS0204_set4.dta, predicting the effects of student background factors on math achievement (students nested in schools). Compare and discuss your results from the two regressions.
- b) Fit a hierarchical model, using ELS0204_set4.dta, to predict individual- and school-level socioeconomic effects on GPA in high school. Discuss your “centering” choices, and interpret your results thoroughly.
c) Use CMSD_set11.dta to fit a hierarchical linear model to predict school value-added effects on GPA.

d) Use CMSD_set11.dta to fit a hierarchical logistic model to predict student and neighborhood effects on high school graduation. Compare your insights to those of Stewart, Stewart, and Simmons (2007).

4/11. Session 11—Hierarchical regression with large datasets (2)

Methodological topics:
- Growth models and specification of time effects in multilevel data structures, three-level models.

Substantive topics:
- Hierarchical effects of students and schools and achievement over time.
- The urban dropout problem.

Readings:

Datasets:
- achievement_set3.dta
- CMSD_set12.dta
- CMSD_set13.dta

Assignment 9:
a) Fit a hierarchical linear model, using `achievement_set3.dta`, predicting the effects of student background factors on math achievement over time (students nested in schools).


4/18. Session 12—Duration/hazard models, event history structures in large datasets

Methodological topics:
- Life tables, Kaplan-Meier estimation, proportional hazards model, Cox regression.

Substantive topics:
- Policy diffusion across the states.
- The urban dropout problem.
- Teacher turnover.
- Deviance in school.

Readings:

Datasets:
- `teacher.dta`
- `cigarette.dta`
- `CMSD_set14.dta`
- `CMSD_set15.dta`

Assignment 10:
- a) Using `teacher.dta`, produce a life table and KM estimates for teacher turnover; plot the survival function as well.
- b) Using `cigarette.dta`, fit discrete-time model assuming continuous time hazards to be proportional, to predict the effects of social resistance and TV on smoking in school.
c) Fit an event history model, using \textit{CMSD\_set15.dta}, to predict the effects of race, class and other student background factors on the hazard of dropping out.

4/25. Session 13—\textit{Review and final exam}

5/2. Session 14—Final exam review