Applied Regression using STATA

Instructors: Greg Duncan and Paul Yoo
Introductory videos from Greg and Paul.

emails: gduncan@uci.edu; pyyoo@uci.edu

This 10-week course was taught in the Winter quarter of 2021 as: Education 265: Applied Regression Analysis for Education and Social Scientific Research.

It was designed to serve two purposes: i) teach students the basics of STATA in the context of social and behavioral research; and ii) provide students with a practical understanding of how regression analysis is used in social and behavioral research.

The STATA portion of the course consists of STATA tutorials in Weeks 1 and 2, eight problem sets scattered over the 10 weeks of the course, and a weekly lab led by Paul Yoo. All but the last problem set are based on a class data set constructed from the Early Childhood Longitudinal Study. We provide STATA code which, when applied to the public use ECLS-K data, will give you the class data set. The regression portion of the course consists of weekly reading assignments plus class lectures and discussions.

We want to acknowledge that the STATA portion of the course has been developed and refined by many TA's prior to 2020 (Tran Keyes, Chin-Hsi Lin, Teya Rutherford, Elizabeth Miller, Tyler Watts, Tutrang Nguyen, Diane Hsieh). I have tried to honor that tradition of further developing and refining the materials; any and all errors in this version are mine (Paul Yoo).

Links to YouTube recordings of the labs and lectures are embedded in the syllabus, below. Here is a more formal explanation of the course:

Class goals. The purpose of this course is to provide students with an introduction to research uses of various regression models, with particular emphasis on prudent application and intuitive interpretation. Topics will include a review of and extensions to the OLS regression model, logistic regressions models, causation and natural experiments, descriptive techniques for longitudinal data, and various techniques for analyzing longitudinal data. The approach taken in the class will be a relatively intuitive one, with plenty of hands-on exercises between classes to promote familiarity with the material.

We hope that by the end of the course students will be able to understand the statistical underpinnings of most current social-science education and psychological research using cross-sectional and longitudinal data. This is also a stepping stone to more advanced quantitative courses.

Mastering the material will require a careful attention to the lecture discussions, reading of the course texts and articles as well as to the exercises that will be handed out each week.

Here is a list of the lecture videos:

Introduction from Greg Duncan (0 hours:7 minutes)
Introduction from Paul Yoo (0:05)
Lecture for Week 1 (Part 1): Regression coefficients (0:53)
Lecture for Week 1 (Part 2): Spline functions (1:05)
Lecture for Week 2 (Part 1): Every experiment is a regression (1:24)
Lecture for Week 2 (Part 2): Forward and reverse causal thinking (0:38)
Lecture for Week 3 (Part 1): Fun with standard errors (0:55)
Lecture for Week 3 (Part 2): Different ways of expressing regression coefficients (0:37)
Lecture for Week 3 (Part 3): OLS vs. ANOVA and dummy variables (0:41)
Lecture for Week 4 (Part 1): Parabolas and dummy variable interactions (1:23)
Lecture for Week 4 (Part 2): Moderation and Mediation (0:50)
Higher-quality video starts here:
Lecture for Week 5 (Part 1): Logistic regression (1:03)
Lecture for Week 5 (Part 2): Introduction to causal analysis (0:51)
Lecture for Week 6 (Part 1): TOT vs. ITT (0:58)
Lecture for Week 6 (Part 2): Regression discontinuity (1:08)
Lecture for Week 7 (Part 1): Difference in difference (0:45)
Lecture for Week 7 (Part 2): Fixed effects (:55)
Week 7: Midterm answer review (0:41)
Lecture for Week 8 (Part 1): Longitudinal models (2:01)
Lecture for Week 8 (Part 2): A dour look at growth models (0:52) (It is recommended that you watch Part I first)
Lecture for Week 9: Getting it wrong and right on skill growth models (2:05)
Lecture for Week 10 (Part 1): Power analysis (1:36)
Lecture for Week 10 (Part 2): Beyond OLS (0:41)

Here is a list of Lab Videos

Getting Started Part 1 (0:07): how to create the class dataset
Getting Started Part 2 (0:03): all the lab files
Lab for Week 1 (1:41): writing do files, exploring the data, basic cleaning
Lab for Week 2 (1:36): loops, bi-variate analysis
Lab for Week 3 (1:25): printing tables, testing group differences, egen
Lab for Week 4 (1:49): regressions, f-tests, and interactions
Lab for Week 5 (2:39): binary outcomes, lpm, and logits
Lab for Week 6 (NO LAB)
Lab for Week 7 (1:22): collapse, binned scatter, spline
Lab for Week 8 (1:24): extras on some graphing

Lab for Week 9 (1:44): change models, fixed effects
Lab for Week 10 (NO LAB)

Here is a zip file with all the lab files
please note, all of the do files have been provided as .txt files (a website limitation) so you will need to rename the script files with the .do extension

Papers, exams and grading. I will require two papers during the quarter, a very short one due at the beginning of week 8 and a more substantial one due by the end of Monday, March 15th of finals week. The first, one-page paper requires students to invent a "natural experiment" that could be used to test a hypothesis of interest. The second paper will require an analysis of data (either the data set we will be working with as part of the class or your own data). There will be one exam taken during the lab period of Week 6.

Your class grade will depend on the following criteria:
10% — quality of contributions to class discussion
30% — quality of lab homework and responses to weekly class assignments
10% — first paper
20% — exam
30% — second paper

My grading policy adheres to regulations in the UCI Academic Senate manual.

There are two recommended textbooks:

Rachel A. Gordon, *Regression Analysis for the Social Sciences*, New York: Routledge, 2010. (There is a 2nd edition available, but the 1st addition is fine for our purposes and less expensive)

Jeffrey M. Wooldridge, *Introductory Econometrics,* South-Western College Publishing. (There are many editions to this book. Any will do for our purposes.)

I strongly recommend that you have one or both of these books in your possession. Gordon is fairly new and written at a more elementary level. Wooldridge is a comprehensive and fairly accessible beginning econometrics text. I list readings from both below.

Recommended:

Paul D. Allison, *Multiple Regression: A Primer*, Pine Forge Press, 1999.

Week 1, January 4: Introduction to the course and the class data set, review of bivariate OLS

We know that the first class is on the first day of the quarter and that you will not have had a chance to read the assigned readings before class. At a minimum please concentrate on readings needed to answer the two assignment questions which are due one hour before class time.

Readings for class:

Chapters 1 and 2 in Gordon, Appendix A, B and C – review as needed. Note: these appendices are not included in the 2nd edition of Gordon's book.

Chapters 3 and 4 in Gordon

Chapter 1 in Allison's "Multiple Regression: A Primer"

Duncan, G., W. Yeung, and J. Brooks-Gunn. (1998). How Much Does Childhood Poverty Affect the Life Chances of Children? *American Sociological Review*, 63(3), pp. 406-423. Pay particular attention to interpreting the numbers in Table 3. Duncan Yeung Brooks-Gunn

Introduction to the course data set: ECLSK Manual, pages xxv through xxxiv; 1-1 through 1-7; 7-9 & 7-10 ECLSK_K8_Manual_part1

Lecture:

Lecture notes Lecture notes Week 1

Lecture video:
Lecture for Week 1 (Part 1): Regression coefficients (0:53)
Lecture for Week 1 (Part 2): Spline functions (1:05)

STATA instruction, problem sets and labs

To make things easier, all the relevant lab files have been placed into a single zip file, but they are also individually provided throughout the syllabus.

1) Take STATA tutorials at https://stats.idre.ucla.edu/stata/seminars/notes15/ (Links to an external site.) (3 modules under "Class Notes"—Entering, Exploring, Modifying). This will take approximately 2 hours.

2) Read/watch the "getting started" materials (below). They will walk you through how to create the class dataset and introduce you to all of the files you will need for the STATA lab sessions.

Getting started instructions (written instructions)
Getting Started Video Part 1 (0:07): how to create the class dataset
Getting Started Video Part 2 (0:03): all the lab files

Create_classdata script (to run in STATA)
Overview of the class dataset (power point slides)

3) Lab Video and do file

Week 1 Lab Video (1:41): writing do files, exploring the data, basic cleaning
Week 1 do file

4) Problem Set 1: review of basic STATA commands, browsing through the dataset, histogram & kdensity plots:

please note, all of the do files have been provided as .txt files (a website limitation) so you will need to rename the script files with the .do extension

Week 2, January 11, Review/extensions of bivariate and multivariate OLS models

Readings for class:

Chapters 5 and 6 in Gordon

Chapters 2 and 3 in Allison's "Multiple Regression: A Primer"

Chapters 1 and 2 (Sections 2.1-2.3), Sections 7.1-7.3 of Wooldridge, *Introductory Econometrics*

Duncan, G., Magnuson, K. and Ludwig, J. The Endogeneity Problem in Developmental Studies *Research in Human Development*, Vol. 1, Nos. 1&2, 2004, pp. 59-80. Duncan Magnuson Endogeneity Problem

Gelman, A. and Imbens, G. (2013) *Why ask Why?* Working paper. Gelman Imbens reversecausal_13oct05

Heresy 101 notes Heresy 101 062921

Lecture:

Lecture notes Week 2

Lecture video:
Lecture for Week 2 (Part 1): Every experiment is a regression (1:24)
Lecture for Week 2 (Part 2): Forward and reverse causal thinking (0:38)

post-lecture video (0:10) on the regression-based group differences lecture example with STATA (the excel file used in this video)

STATA instruction, problem sets and labs

1) Take STATA tutorial on https://stats.idre.ucla.edu/stata/seminars/notes15/ (Links to an external site.) (2 modules under "Class Notes"— Managing, Analyzing Data). This will take approximately 1.5 hours.

2) Lab Video and do file

Week 2 Lab Video (1:36): loops, bi-variate analysis
Week 2 do file

3) Problem Set 2: data cleaning, correlation and bivariate regressions

please note, all of the do files have been provided as .txt files (a website limitation) so you will need to rename the script files with the .do extension

Week 3, January 18: Standard errors, interpreting coefficients and regression vs. ANOVA

Readings for class:

Chapter 7 in Gordon

Sections 7.1 – 7.3 and Appendix A from Wooldridge, *Introductory Econometrics*

Fryer Jr., R & Levitt, S. (2006). The Black-White Test Score Gap Through Third Grade. *American Law and Economics Review,* 8(2), pp. 249-281. Fryer and Levitt Gap through Grade 3 full article

Handout on standard errors Standard error example output 011621

Interpreting regression coefficients Interpreting coefficients 010921

COVID prediction example COVID example 011721

Regression and ANOVA comparison Regression and ANOVA pptx 011721

Lecture:

Lecture notes Week 3

Lecture videos:
Lecture for Week 3 (Part 1): Fun with standard errors (0:55)
Lecture for Week 3 (Part 2): Different ways of expressing regression coefficients (0:37)
Lecture for Week 3 (Part 3): OLS vs. ANOVA and dummy variables (0:41)

STATA instruction, problem sets and labs

1) Lab Video and do file

Week 3 Lab Video (1:25): printing tables, testing group differences, egen
Week 3 do file

2) Problem Set 3: data cleaning, descriptive table, standardized outcome, dummy variable predictors, regression table

please note, all of the do files have been provided as .txt files (a website limitation) so you will need to rename the script files with the .do extension

Week 4, January 25: Interactions, dummy variables and mediated models

Readings for class:

Chapter 8 in Gordon

Chapter 8 in Allison's "Multiple Regression: A Primer"

Section 7.4 from Wooldridge, *Introductory Econometrics*

Sample exam from a past year that includes some content we may not get to before the midterm – sample exam

Lecture:

Lecture notes Week 4

Lecture video:
Lecture for Week 4 (Part 1): Parabolas and dummy variable interactions (1:23)
Lecture for Week 4 (Part 2): Moderation and Mediation (0:50)

STATA instruction, problem sets and labs

1) Lab Video and do file

Week 4 Lab Video (1:49): regressions, f-tests, and interactions
Week 4 do file

2) Problem Set 4: interactions, covariates, bar graphs

please note, all of the do files have been provided as .txt files (a website limitation) so you will need to rename the script files with the .do extension

Week 5, February 1: Logistic regression and introduction to causal analysis

Readings for class:

Chapter 9 in Gordon

Section 7.5 of Wooldridge, *Introductory Econometrics*

De Maris, Alfred. "A Tutorial in Logistic Regression" *Journal of Marriage and the Family*, 57, Nov. 1995, 956-968 DeMaris Logistic regression

Study for the midterm! Sample exam questions

Lecture:

Lecture notes Week 5

Lecture video:
Lecture for Week 5 (Part 1): Logistic regression (1:03)
Lecture for Week 5 (Part 2): Introduction to causal analysis (0:51)

STATA instruction, problem sets and labs

1) Lab Video and do file

Week 5 Lab Video (2:39): binary outcomes, lpm, and logits
Week 5 do file

2) Problem Set 5: binary outcomes, non-linear relationships

please note, all of the do files have been provided as .txt files (a website limitation) so you will need to rename the script files with the .do extension

Week 6, February 8: Causality analysis continued: TOT vs. ITT and instrumental variables

Readings for class:

Sections 3.3, 3.4 of Wooldridge, *Introductory Econometrics*

Chapters 10 & 11 in Gordon

Chapter 6 in Allison's "Multiple Regression: A Primer"

Cortes, K. E., Goodman, J. S., & Nomi, T. (2015). Intensive math instruction and educational attainment long-run impacts of double-dose algebra. *Journal of Human Resources*, 50(1), 108-158. Cortes Goodman Double Dose

Wong, V. C., Cook, T. D., Barnett, W. S., & Jung, K. (2008). An effectiveness-based evaluation of five state pre-kindergarten programs. *Journal of Policy Analysis and Management: The Journal of the Association for Public Policy Analysis and Management*, *27*(1), 122-154. Wong Cook Barnett preK

Exercise assigned: *Problem set 6: Logistic regression and Margins* *Problem Set 6*

Lecture:

Lecture notes Week 6

Lecture video:
Lecture for Week 6 (Part 1): TOT vs. ITT (0:58)
Lecture for Week 6 (Part 2): Regression discontinuity (1:08)

STATA instruction, problem sets and labs

1) NO Lab Video and do file, but there is a problem set.

2) Problem Set 6: Logistic regression and Margins

Midterm during the Tuesday lab session Midterm exam Midterm exam answer video

Week 7, February 15: Causality and quasi-experimental designs: difference-in-differences, fixed effects

Readings for class:

Currie, J. and Walker, R. (2011). Traffic Congestion and Infant Health: Evidence from E-Z Pass. *American Economic Journal: Applied Economics*, 3, pp. 65-90. Currie Walker EZPass

Colen, C. and Ramey, D. (2014) Is breast truly best? Estimating the effects of breastfeeding on long-term child health and wellbeing in the United States using sibling comparisons. *Social Science & Medicine,* 109, pp. 55-65. Colen Ramey Breast feeding

In class, we will also talk about this policy brief on New Orleans charter schools: Brief

Discussion of class paper — see these guidelines.

Cassie Hart tables: Hart tables

Lecture:

Lecture notes Week 7

Lecture video:
Lecture for Week 7 (Part 1): Difference in difference (0:45)
Lecture for Week 7 (Part 2): Fixed effects (:55)

STATA instruction, problem sets and labs

1) Lab Video and do file

Week 7 Lab Video (1:22): collapse, binned scatter, spline
Week 7 do file

2) Problem Set 7: Graphical approaches to functional form, regression discontinuity

please note, all of the do files have been provided as .txt files (a website limitation) so you will need to rename the script files with the .do extension

Week 8, February 22: Causality and quasi-experimental designs: conceptual and empirical models of longitudinal processes

Readings for class:

Sections 13.1, 13.2 & 14.4 from Wooldridge, *Introductory Econometrics*

NICHD Early Child Care Research Network and Greg Duncan, (2003). "Modeling the Impacts of Child Care Quality on Children's Preschool Cognitive Development". *Child Development,* 74(5) pp. 1454-1475. NICHD Duncan childcare

Exercise assigned: No problem set. Finish short paper on natural experiments! Make progress on end-of-term paper.

Lecture:

Lecture notes Week 8

Lecture video:
Lecture for Week 8 (Part 1): Longitudinal models (2:01)
Lecture for Week 8 (Part 2): A dour look at growth models (0:52) (It is recommended that you watch Part I first)

STATA instruction, problem sets and labs

1) Lab Video and do file

Week 8 Lab Video (1:24): extras on some graphing
Week 8 do file

NO problem set for week 8

Week 9, March 1: Longitudinal analysis and the joys of being really wrong

Short paper on natural experiments due Sunday the 28th at 3pm.

Readings for class:

Chapter 14 (sections 14.1-14.2) in Wooldridge, *Introductory Econometrics*

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., ... & Sexton, H. (2007). School readiness and later achievement. *Developmental Psychology*, 43(6), 1428. School readiness pdf

Class powerpoint on Duncan et al. 2007: School_readiness for class

Bailey, D. H., Duncan, G. J., Watts, T., Clements, D. H., & Sarama, J. (2018). Risky business: Correlation and causation in longitudinal studies of skill development. *American Psychologist*, 73(1), 81.Bailey Risky Business Online Version

Class powerpoint on Risky Business: Skill Building paper for class

Lecture:

Lecture notes Week 9

Lecture video:
Lecture for Week 9: Getting it wrong and right on skill growth models (2:05)

STATA instruction, problem sets and labs

1) Lab Video and do file

Week 9 Lab Video (1:44): change models, fixed effects
Week 9 do file

2) Problem Set 8: Longitudinal models

please note, all of the do files have been provided as .txt files (a website limitation) so you will need to rename the script files with the .do extension.

* This is the last STATA problem set, but there is one more problem set for week 10 below.

Week 10, March 8: Clustered data, standard errors, and power calculations

Readings for class:

Users' manual for the Optimal design software program The general link to the program and manual is: http://hlmsoft.net/od/ (Links to an external site.)

Download G*-Power here (Links to an external site.)

Bloom, H. S. (1995). Minimum detectable effects a simple way to report the statistical power of experimental designs. *Evaluation review*, *19*(5), 547-556. Bloom MDES Eval Rev-1995-Bloom

Power section from Baby's First Years proposal Power section from Baby's First Years proposal

Beyond OLS — an overview Regression with STATA Chapter 4 – Beyond OLS

Exercise assigned: *Problem Set 9: Power analysis*

Lecture:

Lecture notes Week 10

Lecture video:
Lecture for Week 10 (Part 1): Power analysis (1:36)
Lecture for Week 10 (Part 2): Beyond OLS (0:41)

Lab for Week 10 (NO LAB)

Monday, March 15, 11:59pm—*SECOND PAPER DUE*