The stated purpose of the course on the syllabus:

The purpose of this course is to provide students with an introduction to research uses of regression models and the statistical analysis of longitudinal data. Topics will include a review of the OLS regression model, logistic regressions models, causation and natural experiments, descriptive techniques for longitudinal data, event-history methods, and various other techniques for analyzing longitudinal data. The approach taken in the class will be a relatively intuitive one, with plenty of computer exercises between classes to become familiar with the material.

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

Mastering the material will require a careful reading of the course texts as well as persistent attention to the exercises that will be handed out each week. There is one formal class session each week—on Monday between 9:00 and 11:30. Each week’s problem set will be posted in advance and is due in electronic form in Duncan’s EEE dropbox by 9 am Friday of the week in which they are assigned.

I will require two papers during the quarter, one just before the middle and another due on the last day of class. The first paper is very short and requires students to invent a “natural experiment” to test a hypothesis of interest. The second paper will require an analysis of the data set we will be working with as part of the class. Students who wish to use their own data sets to do their papers are encouraged to do so. There will be one in-class exam around the middle of the quarter.

The class grade will depend on the following criteria:

10% — quality of contributions to class discussion
25% — quality of homework
10% — first paper
20% — exam
35% — second paper

My grading policy adheres to regulations in the UCI Academic Senate manual.
2016 Syllabus for:
Education 265: Applied Regression Analysis for Education and Social Research

There are two textbooks:
Jeffrey M. Wooldridge, Introductory Econometrics, 2nd or 3rd edition, South-Western College Publishing.

It is strongly recommended that you have one of these two 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:

All articles listed in the syllabus are available through EEE or the UC-eLinks in Google Scholar.

---

**Week 1**, March 28: Introduction to the course, the class data set and Stata

Readings for class:
Chapters 1 and 2 in Gordon, Appendix A, B and C – review as needed
Chapters 3 and 4 in Gordon
“Introduction to the … class data set,” in EEE
“Heresy 101: Worthy statistical concerns for educational and developmental studies,” in EEE

Exercises assigned: 1) Problem Set 1
2) Take Stata tutorial on [http://www.ats.ucla.edu/stat/stata/notes/](http://www.ats.ucla.edu/stat/stata/notes/) (3 movies under “Class Notes”—Entering, Exploring, Modifying). This will take approximately 2 hours.

---

**Week 2**, April 4: Review/extensions of OLS model

Readings for class:
Chapters 5 and 6 in Gordon – review as needed
Chapters 7, 8 and 9 in Gordon
Chapters 1 and 2 (Sections 2.1-2.3) of Wooldridge, Introductory Econometrics

Exercise assigned: Problem Set 2
2) Take Stata tutorial on [http://www.ats.ucla.edu/stat/stata/notes/](http://www.ats.ucla.edu/stat/stata/notes/) (2 movies under “Class Notes”—Managing, Analyzing Data). This will take approximately 1.5 hours.
**Week 3, April 11: Continued review of OLS**

Readings for class:
Chapters 10 and 11 in Gordon
Chapters 2 (sections 2.4-2.6), 3, 6 (especially sections 6.1 and 6.2), 7 (sections 7.1-7.4) and Appendix A from Wooldridge, *Introductory Econometrics*

*Exercise assigned: Problem Set 3*

---

**Week 4, April 18: Logistic regression**

Readings for class:
Chapter 7 (section 7.5) and 17 (section 17.1) of Wooldridge, *Introductory Econometrics*

*Exercise assigned: Problem Set 4*

---

**Week 5, April 25: Natural experiments and regression discontinuity models**

Readings for class:

*Exercise assigned: Problem Set 5*

---

**Week 6, May 2: Change and other simple longitudinal models**

Readings for class:
*Longitudinal Data Analysis Using Stata*, by Paul D. Allison, November 2007 (pdf, available on EEE)
Chapter 14 (sections 14.1-14.2) in Wooldridge, *Introductory Econometrics*
*Journal of Consulting and Clinical Psychology*, 65(1): 130-140.

Duncan, Boisjoly, Kremer, Levy and Jaque Eccles (2005) “Peer Effects in Drug Use and Sex 

**Exercise assigned:** NONE — study for the mid-term

---

**Week 7, May 9:** More change models

**Mid-Term Exam:** Second half of class (75 minutes)

**Readings for class:**
Allison, Paul D. “Change Scores as Dependent Variables in Regression Analysis” in C.C. Clogg 

Duncan, Greg, Engel Mimi, Claessens, Amy and Chantelle Dowsett, “The Value of Replication 
for Developmental Science,” *Developmental Psychology*, published online on October 2, 
2014, DOI - 10.1037/a0037996

“Paper writing guidelines”

**Exercise assigned:** Problem Set 6

---

**Week 8, May 16:** Event history methods

**Short paper on natural experiments due**

**Readings for class:**

**Recommended:** Paul D. Allison, *Event History Analysis*, Sage University Papers Series, Sage 
Publications, 1984

**Exercise assigned:** Problem Set 7

---

**Week 9-10, May 23 & May 30 (need to reschedule owing to Memorial Day):** Clustered data, 
standard errors and power calculations

**Readings for class:**
Users’ manual for the Optimal design software program (http://sitemaker.umich.edu/group-
based/optimal_design_software)

Stata Library: Analyzing Correlated (Clustered) Data 
(http://statistics.ats.ucla.edu/stat/stata/library/cpsu.htm)--concentrate on sections relevant 
to Stata.

Bloom, H. S. (1995). Minimum detectable effects a simple way to report the statistical power of 

**Exercise assigned:** Problem Set 8

---

**Monday, June 6—SECOND PAPER DUE**
Week 1, March 28: Introduction to the course, the class data set and Stata

A couple of weeks before the beginning of the class I send out the syllabus and strongly suggest that students view the Stata tutorials listed for the first week before classes begin.

I spend the bulk of the class going over the “ASR Table for first class” pdf page (everything is in the Dropbox that I believe that I gave you access to) in an interactive way.

So…what does the first entry in the table mean (income is scaled in $10,000)? Is a .14 year per $10,000 increase a large response? Might it be bigger for lower than higher-income households? – hence the spline form, log form and, most flexibly, the dummy variable form. What do those numbers mean? Are they telling the same story?

Continuing with the middle columns of the table, for regressions applied to dichotomies, we have the logit results for high school completion. What do these numbers mean? Why might we want to convert them to probabilities?

I don’t spend much time on the Cox model results in the last set of columns. Mostly I point out that the whole table is just the basic regression approach modified with tricks and procedures to handle the kind of data and problems that they are likely to encounter.

Miraculously, going over this one table takes at least an hour and a half and provides a preview of things to come, and of the course’s emphasis on coming up with a commonsense interpretation of coefficients in research articles. “Fear no regression table!”, I tell them.

Apart from administrative details and some comments from the TA, I also spend time handing out and going over the points in “Heresy 101.” I don’t explain any of them in real detail, of course, but just a few minutes on each one will be revelatory for students who have already acquired a feel for how educational or developmental research is done and now see that much of what they thought about how the game is played is not at all what they will be learning this quarter.

Tyler: The first lab and problem set are focused on basic introductory issues in Stata. Although the students have used Stata before, I found that they didn’t have a really solid grasp of the basic programming syntax of Stata. I go over file directories, if statements, descriptive commands, and management commands like “sort,” “order,” and “assert.” We also spend time talking about the ECLSK dataset, why it was collected and what types of information it contains.

Week 2, April 4: Review/extensions of OLS model

I start this class with an introduction to counterfactual causal thinking, for which the Winship and Morgan reading is quite useful. I expand on the points listed in the third bullet in the “Heresy 101” handout and emphasize that it is the height of dishonesty to say that only RCTs establish causality, that one’s correlational analysis do not provide causal estimates but then
reel off a long set of policy recommendations. Policy recommendations require something close to causal estimates and so, apart from demographic-type big picture description, we should aspire to coming as close to causal estimates as possible. This too is revelatory.

Then I begin trudge through what should be a review of basic OLS stuff but, from their inability to answer simple questions, isn’t so much of a review after all. The simple linear model, quadratics (this is tricky – I have a nice example of a quadratic fit to two subgroups where both the linear and squared term coefficients are of the opposite signs within and across groups and yet both are essentially telling the same story over the relevant range of data), interactions (even simple ones are tricky to track main and interactive effects), dummy variables.

**Tyler:** In the second lab, we focus on making descriptive tables, regression tables (both with estout commands) and interpreting dummy variable coefficients. This problem set is a big push forward, so I remember grades being a little worse than usual here. For the rest of the class, they will continue to make well-formatted tables, so they keep practicing this for the next 8 weeks.

---

**Week 3, April 11: Continued review of OLS**

This week continues the review of the OLS model. I get them to read and interpret the tables in the Fryer and Levitt article, which requires understanding a lot of the regression ideas. As a bonus, it provides fixed effect estimates, so I try to explain those. (I try again later in the quarter as well). As a double bonus, the article models how much information should go into a table in order to make it self-contained and how an introduction to a paper should look – motivating the problem in a compelling way and describing the results, all done clearly and concisely.

**Tyler:** On this problem set, the students replicate the main results table from Fryer and Levitt (2006). We don’t weight the models, so the sample sizes slightly differ. I gave them the final tables on the problem set with the correct n’s, and their job was to use what we learned in Weeks 1 and 2 to produce exactly the same tables. On average, students spent the most time on this problem set (I believe), but I think it will be a great teaching tool for Tutrang because it’s easy to diagnose where people are going wrong with this type of assignment. By the end of week 3, students should now be able to run analyses and produce tables on their own.

---

**Week 4, April 18: Logistic regression**

This week stresses the uninterpretability of the log odds coefficients and even odd ratios, and the value of converting logit, etc. results into a probability change interpretation.

**Tyler:** This lab and problem set closely follows the lecture. We start with interpreting interactions, then we move to estimating outcomes on dichotomous variables (i.e., logits, LPM’s). If the problem set is done correctly, students see that the LPM should produce the same results as the marginal effects interpretation of a logistic regression.

---

**Week 5, April 25: Natural experiments and regression discontinuity models**
Now we start getting into the fun stuff, which will be quite foreign to almost all students. The Duncan, Magnuson and Ludwig article sets up the omitted-variables problem in the context of developmental (and much of the educational) literature. I like Hanushek and Raymond because it lays out a model of education production and then identifies it with a triple difference approach. As next week argues, starting with a model is NOT what most are used to. Wong et al provide a nice example of regression discontinuity, which is really just a simple example of dummy variable interactions. This is a lot to cover in a single class session and it sometimes spills over into the next week.

**Tyler:** This lab and problem set focuses on graphing in Stata, which nicely matches the lecture’s focus on RD. In the lab, I walk students through the importance of visualizing your data, and we talk about the collapse commands and binning the data. I then give them a bunch of graphing code, and on the problem set, they have to use it and alter it for a new set of variables. I think it is probably one of the more difficult coding assignments.

---

**Week 6, May 2: Change and other simple longitudinal models**

My main goal in this class is to highlight the distinction between *conceptual* models of change and *empirical* models of change. This is fairly comfortable for economists and TOTALLY ALIEN to everyone else. They think that a HLM latent growth model is how you should model change, even if it bears no connection to any conceptual model of interest.

For the conceptual discussion I introduce “instantaneous change” and “accumulation of inputs” (and, less importantly, “lagged change”) models of the processes by which dependent and independent variables might be related. These are on slide 41-43 of the “Causation Session” powerpoint slides (those slides were used for a workshop and I don’t use them for class – in fact I almost never use slides but instead work through the ideas on the board).

In each case you start with the simplest way to express them and then emphasize that estimating them probably produces a lot of omitted-variables bias. So then first difference each of them to get rid of some of that bias. Simple algebra gives you a correspondence between the parameters of the level and change models – truly a profound revelation for most!

The NICHD Early Child Care Research Network and Greg Duncan article walks through the level and change version of the accumulation of inputs model and shows that the most common way of estimating these models (with a lagged dependent variable) is not interpretable in light of the accumulation of inputs model. Not covered in the article is that results from a latent growth model can’t be readily interpreted with any sort of conceptual model either. Of course, there are more models than just instantaneous change and accumulation of inputs, so the goal here is to get student to distinguish between conceptual and empirical models and not to go off blindly in the direction that empirical models take them. The Allison reading for this week is helpful as well.

The Curran and Duncan et al readings are examples of SEM vs. natural experiments that I try to get to as well.
**Tyler:** No problem set this week because there was a midterm. I just reviewed for the midterm during the lab.

### Week 7, May 9: More change models

I usually lecture/discuss for an hour and then give the midterm exam for the rest of the class. The exam comes kind of late in the quarter, but since there is no final exam, it functions as something in between a mid-term and a final and it really coaxes the students into consolidating a lot of loosely-understood material.

Sometimes I finish up with the discussion of change models, helped by the Allison reading this week.

There is no good place for talking about the (to them) strange culture of economists valuing robustness checking, post data, 80-page supplemental materials, so I try to stick it here, using the Duncan et al. article on replication.

This is the point when students need to start to get serious about their papers, so I distribute the paper writing guidelines.

Again this seems a little awkward, but due next week is a short (one page single spaced) paper on natural experiments. The idea is to get them to think up some kind of natural experiment for a research question of interest to them. It doesn’t have to be feasible, only embody the logic of a natural experiment, which will also be TOTALLY ALIEN to all of them.

**Tyler:** This problem set focuses on change models. I tried to find an example in the ECLSK data that closely matches Greg’s approach in the NICHD childcare article. We basically try to replicate that analysis with different variables focused on kindergarten instruction. Students struggle with these models, so I ended up working through a lot of the material from Greg’s lecture again. We also introduce school and classroom fixed effects in this problem set.

### Week 8, May 16: Event history methods

Event history methods expand or contract to fit the space available. I find it astonishing that developmentalists, who study transitions, don’t have them in their methods tool kits. So something on the language and logic of hazard models is useful. The class readings provide an overview and a good empirical example of breaking up the data into a person-year format and approaching the event history problem as a conventional logistic regression. That is particularly helpful when you have time-varying covariates.

**Tyler:** This problem set diverges from what’s happening in lecture in order to address approaches to missing data. So far, all of the problem sets have asked students to use listwise deletion. Here, we redo the analyses from the previous problem set and try out various approaches to missing data (FIML, missing dummy variables). By now, we are talking a lot in during the lab about focusing on the key result from your analysis. The change model stuff has introduced students to the idea that you are setting up an analysis with the hopes of estimating
one causal effect. Here, I argue that when you estimate models with missing data adjustments, you should still be focused on your one key result. You simply want to make sure that missing data doesn’t affect your key estimate.

**Week 9-10, May 23 & May 30 (need to reschedule owing to Memorial Day):** Clumped data, standard errors and power calculations

This is a fairly recent addition and seems very much appreciated. The Bloom reading is very useful, as is the problem set, which walks them through doing power analyses for unclustered and clustered data. The problem set also put two good freeware programs into their hands and heads.

If time allows, I go over the Heresay 101 notes as a review, and then have a discussion of future courses they can consider, a pep talk about how everyone can do it, and that, as with all things, mastery comes with practice, practice and more practice.

**Tyler:** The last problem set is focused on power analysis, and this is the only problem set that I didn’t change. I believe Greg made it a few years ago. It doesn’t use the ECLSK data, but it asks students to generate power analyses for various scenarios using G*Power and Optimal Design. In the final week, I just show students a bunch of analytic stuff in Stata (e.g., weighting, measurement models, etc.) that we hadn’t covered, but there is no problem set – everyone is focused on writing course papers.