Introduction to Multivariate Regression & Econometrics

HED 612

Lecture 12

1. Linear Probability Model
2. Creating (nearly!) Publication Quality Regression Tables
3. Final Projects
4. 10 Min Break
5. Learning to Read Quantitative Empirical Work

Prepare for class

We'll be using the ELS Data today; no need to re-download!

1. Download the Lecture 12 PDF and R files for this week
   - Place all files in HED612_SP21 »> lectures »> lecture12
2. Open the RProject (should be in your main HED612_SP21 folder)
3. Once the RStudio window opens, open the Lecture 12 R script by clicking on:
   - file »> open file... »> [navigate to lecture 12 folder] »> lecture12_v2.R
4. **Install the "stargazer" package in lecture12_v2.R**
   - If R prompts you to install `stargazer` and "its dependencies", go ahead and click install
   - If R doesn't prompt you, install via line 8 `install.packages("stargazer")`

Finishing the Semester...

- Today, 4/14
  - Linear Probability Models for 0/1 DV
    - Exporting regression "nearly-publication" ready tables!
  - Mini lesson on what each section of final project should accomplish!
  - Finish Reviewing Powers(2004)

- 4/21 [No Class/Reading Day]
  - Reading for 4/28
    - Klasik, D., Blagg, K., & Pekor, Z. (2018). Out of the Education Desert: How Limited Local College Options are Associated with Inequity in Postsecondary Opportunities. Social Sciences, 7(9), 2018.

- 4/28
  - Reviewing Empirical Research:
    - Klasik, D., Blagg, K., & Pekor, Z. (2018). Out of the Education Desert: How Limited Local College Options are Associated with Inequity in Postsecondary Opportunities. Social Sciences, 7(9), 2018.
  - Introduction to non-linear functions [we'll get as far as we can...]

- Last class session, 5/5
  - Student presentations!
    - Don't stress these; don't need to be perfect!
    - Just an opportunity to learn from each other and get some practice "presenting"
  - **Final paper is due on our "final exam day/time": 5/7/21 by 5:30pm**

Linear Probability Model

Linear Probability Model

- Binary Variables (i.e., dummies, indicators) as dependent variables are very common in education research!
  - Y = Retention (0=dropped out, 1= persisted)
  - Y = Graduation (0= did not graduate, 1= graduated)
  - Y = Pass/Fail (0=Failed, Passed=1)

- Regression models with a binary dependent variable attempt to interpret the effect of X on the *probability* of "success" (Y=1)
  - Or in some cases the probability of "failure"

- Most social science disciplines model binary dependent variables via non-linear regression models
  - logistic regression [will cover in HED 649 Spring 2022]
  - but interpretation can be difficult because its measured via odds ratios

- Econometrics models binary dependent variables via **linear probability model**
  - Population parameters can be estimated via OLS!
  - Simple to estimate and interpret!
  - Only "tool" that doesn't carry over? R^2; but program evaluation is less concerned with model fit than hypothesis testing about the population parameter β_1
  - It can ONLY be used for binary, categorical dependent variables
    - If more than two categories, then we have to use different models covered in HED 649

Linear Probability Model, with Continuous X

- RQ: What is the effect of hours spent on homework on the probability of attending college?
- **Pop Reg Model:** $Yi = \beta_0 + \beta_1 X_{1i} + ui$
  - Y = `college` (1= attended college, 0= did not attend college)
  - X = average hours spent on homework per week
- Run in R
- **OLS Prediction Line**
  - w/o estimates: $Yi = \hat{\beta}_0 + \hat{\beta}_1 X_{1i}$
  - w estimates: $Yi = -0.083759 + 0.019575 * X_{1i}$
- **Interpretation of** $\hat{\beta}_1$
  - General: On average, a one-unit increase in X is associated with a $\hat{\beta}_1 * 100$ percentage point change in the probability of Y=1
  - On average, a one-hour increase in average homework hours spent per week is associated with a ~2 ($0.02 * 100$) percentage-point increase in the probability of going to college.
- **Interpretation of** $\hat{\beta}_1$ when you SCALE X
  - General: On average, a N-unit increase in X is associated with a ((N-unit$*\hat{\beta}_1$)*100) percentage point change in the probability of Y=1
  - On average, a five-hour increase in average homework hours spent per week is associated with a 10 percentage-point increase $((5 * 0.02) * 100)$ in the probability of going to college.

Linear Probability Model, with Categorical X

- RQ: What is the effect of high school extracurricular participation on the probability of attending college?
- **Pop Reg Model:** $Yi = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ui$
  - Y = `college` (1= attended college, 0= did not attend college)
  - X1 = `excurr` (1= participated in extracurriculars, 0= did not participate [reference group])
  - X2 = average hours spent on homework per week
- Run in R
- **OLS Prediction Line**
  - w/o estimates: $Yi = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i}$
  - w estimates: $Yi = -0.227656 + 0.350495 * X_{1i} + 0.016144 * X_{2i}$
- **Interpretation of** $\hat{\beta}_1$
  - General: On average, being in the "non-reference group" as opposed to the "reference group" is associated with a 100*$\hat{\beta}_1$ percentage point change in the probability of Y=1
  - On average, participating in extracurriculars as opposed to not participating in extracurriculars is associated with a 35 (0.350495*100) percentage point increase in the probability of attending college, holding hours spent on homework constant

Creating (nearly!) Publication Quality Regression Tables

Stargazer Library

- Need to install first via `install.packages("stargazer")`
- Package used to create publication ready tables
- Some resourses
  - CRAN Package Documentation
  - CRAN Package Vignettes
  - Helpful Presentation & Summary
  - Google! Google! Google!
- Show in R!

Final Projects

Approach to final projects...

- I'm super chill about final project! :)
  - No pressure opportunity to learn and "de-mystify" the quant research process!
  - I know we're all running on fumes...
  - I just want you to get what you can out of the assignment given space/time/energy you have!
- "Presentations" on 5/5
  - Do not need to be polished presentations
  - Just a way to share with the class what you did
  - We will learn from the process of putting together our own project and from eachother's work too!
  - No requirements in terms of "structure": power point; share screen of your paper draft; run regression in RStudio for us, etc.

No More Problem Sets, Working Towards Final Projects

- By 4/21:
  - Finalize RQ
  - Identify dataset, clean DV/IV
  - Run bivariate regression/Interpret $\hat{\beta}_1$
  - Identify 3-5 control variables that satisfy OVB
    - Clean control variable(s) if available in your dataset (& if you have the time)
    - Run multivariate regression/Interpret $\hat{\beta}_1$
- By 4/28
  - Write up methods section and results
- By 5/5
  - Read and summarize 5 empirical articles that belong in your literature review
  - Presentation does not need to be "polished"; just sharing your RQ and results!

Final Project: Introduction

- An introduction is (in my humble opinion) the hardest section to write! Why?
- It sets the tone for the entire study!
- An introduction should:
  - Convince the reader of the significance and contributions of the study via a "hook" (i.e., Who should care about this study? and Why?)
  - Present the research question
  - Summarize what the study does

- **Your final paper's introduction should (at least) present the research question and summarize the study**

- If you want try to "hook" the reader, here are some strategies:
  - Actionable research: You can take results and *directly* make recommendations for changes to policy/practice
  - Interesting research: New analyses on an important topic, resolving scholarly/public debates, introduces new ways to looking at old problems, has implications for theory

Final Project: Sometimes you may have a "background" section

- Policy/Background sections are very common in econometrics research.
  - Because econometrics research usually focuses on understanding the effect of some program/ policy /intervention etc.
  - The background section introduces the reader to the history and specific details of the program, policy, intervention being studied
    - Powers 2004 summarized all the legal background and precedence of the Williams v. State of CA case in this background section (p.766)

- For example, if you are studying the effect of participating in Upward Bound, Gear Up, Head Start, College Prep Programs etc on some outcome….
  - Assume the reader has NO KNOWLEDGE of the program
  - What is the program?
  - What are the goals of the program?
  - When was it created and for who?
  - How has the program changed over time (i.e., eligibility, curriculum, financial support from government, etc.)

If your independent variable of interest is a program, policy, intervention, your final paper should include a background section describing the program, policy, or intervention

Final Project: Literature Review

- Literature reviews should accomplish the following main tasks:
  - Reviewing previous literature in your study's area
  - Identifying the "gap" in previous literature
  - Addressing how your study fills that "gap"
- Literature reviews take a lot of time...
  - A lot of reading; but also really fun!
- You of course won't have the time to write out a developed literature review and I don't expect you to!

For your final paper, I expect you to identify and briefly summarize 5 empirical, peer-reviewed articles that belong in a fully developed literature review. Explain WHY these belong in your literature review (i.e., what are their contributions)

Final Projects: Methods and Data

Methods sections are straightforward but technical...

For your final paper, the Methods Section should have the following sections:

- **Empirical Strategy**
  - Your aim is to analyze the effect of X on Y...
  - The gold standard of econometrics research is a randomized control experiment; explain how YOUR independent variable of interest would be hypothetically randomized/why it's not possible (e.g., financial aid is awarded based on need, Upward Bound was created to serve low-income students, etc)
  - Thus, you are working with observational data and trying to get as close as possible to a causal effect (all you can do is get as close as possible to isolating the relationship of X on Y; but you can claim a causal relationship) by using control variables that would otherwise result in omitted variable bias (make sure you define OVB!)
  - Cite Stock and Watson textbook; Cite Cellini (2007)
- **Data and Variables**
  - What data are you using? If NCES Survey, provide some details about the survey, the sample, years conducted, etc.
  - Provide detailed information about your independent variable of interest (what was the survey question asked, how did you deal with missing values, did you transform this variable in any other way)
  - Provide detailed information about your dependent variable (what was the survey question asked, how did you deal with missing values, did you transform this variable in any other way)
  - Identify 3-5 control variables. Explain WHY they should be included in the model (i.e., how do they satisfy both conditions of OVB).

Final Projects: Methods and Data

Analytical Strategy

- Because your DV is continuous, you will run an Ordinary Least Squares Linear regression model
  - [or] Because your DV is binary, you will run a Linear Probability regression model
- Write out the population regression model and label what every piece of the equation represents.
- Specify that given your RQ, your focus is on analyzing β_1 because it represents the relationship between your X and Y
  - The magnitude of the coefficient shows the magnitude of the relationship:
    - the average effect on Y for a one-unit increase in X (continuous X, continuous Y)
    - the average change in Y for the non-reference group in comparison to reference group (categorical X, continuous Y)
    - the change in probability for going from Y=0 to Y=1 for a one unit increase in X (continuous X, binary Y)
    - the change in probability for going from Y=0 to Y=1 for the non-reference group in comparison to reference group (categorical X, binary Y)
  - The p-value of the hypothesis test on the coefficient shows the statistical significance of the relationship.
    - State the Null and Alternative Hypotheses!

A great example of how to write this section that you can emulate [posted on D2L]:

- Nolan L. Cabrera, Jeffrey F. Milem, Ozan Jaquette, & Roland W. Marx. (2014). Missing the (Student Achievement) Forest for All the (Political) Trees: Empiricism and the Mexican American Studies Controversy in Tucson. American Educational Research Journal, 51(6), 1084–1118. https://doi.org/10.3102/0002831214553705

Final Projects: Findings

This section will actually be very short...

- Start by describing the descriptive statistics of your indepedent and dependent variable (mean, standard deviation, min, and max)
- Interpret $\hat{\beta}_1$
  - In words and state the statistical significance

Create and include two tables to show the results:

- Table 1 should include the mean, standard deviation, min, max (counts and % for categorical variables) for your independent variable of interest and dependent variable
- Table 2 should include results of the bivariate regression and multivariate regression models

10 Min Break

Learning to Read Quantitative Empirical Work

Powers (2004): Overall RQ, Analytical and Writting Approach!

- **Williams v. State of California**
  - Class action lawsuit against the State of California in 2000; public school students represented by a coalition of law firms, civil rights organizations
  - Argued that if schools and students are judged on the basis of their test scores (high stakes accountability policies), then students and schools should be provided with equal access to school-related resources needed for academic success
  - Specifically addressed need for qualified teachers, sufficient/up-to-date textbooks, and adequate/safe facilities
- **Old legal approach to school financing**: focused on levels and formulas of spending
  - Didn't work because the state has minimal oversight of local educational control/"property taxes"; caps were eliminated via voter overrides; and the small proportion of state funding was given "equally"
- **New legal approach to school financing**: *Williams v. State of California* reframed the issue of school financing around the *conditions* of education rather than funding formulas!
  - Worked because the state is responsible for ensuring resources are used to produce desirable outcomes
- State of California argued that there was little empirical support that increased spending on resources identified in the case would increase student achievement
  - **RQ**: What is the relationship(s) between school and district characteristics on school's academic performance?

Powers (2004): Overall RQ, Analytical and Writting Approach!

- **Empirical Strategy**
  - Powers is not trying to establish a causal effect of school/resource characteristics on API score explicitly (impossible given the observational data)
  - But she is trying to isolate the Williams Case variables "relationship" on schools' academic performance by controlling for other factors that may be driving variation in schools' academic performance!
    - See on See pg.766 for rationale in including control variables
    - "To ensure the findings related to Williams variables are robust, that is, they are not systematically related to student, school, and district characteristics"
  - But still analyzing "descriptive relationships", not "causal relationships"
- **Literature Review Strategy**
  - HOW you format your literature review depends on your RQ and what is substantively important

(1) Focus on reviewing scholarship about your X (most econometrics studies do this)
  - Ex: What is the effect of receiving a Pell Grant on achievement?; review scholarship on the effect of Pell grants on all student outcomes (GPA, graduation, labor market, etc.)

(2) Focus on reviewing scholarship about your Y (many prediction/descriptive studies do this)
  - EX: What is the (predicted) probability of a student persisting into their 2nd year of college?; review scholarship on what we know about persistence (more likely to persist if you're involved on campus, live on campus, etc.)

(3) Focus on reviewing scholarship that specifically speaks to the *relationship* between X and Y
  - Ex: POWERS (2004, p.772): What do we know about the relationship between school resources and academic achievement

Powers (2004)

- **Data**: California DOE Data, school-level; Census Data; Federal School Data
  - Dependent variable(s): ?
  - Independent variable(s) of interest: ?
    - Hint there are three groupings and they are related to the Williams Case
  - Control Variables: ?
    - Hint there are two groupings!
    - Some control variables left out due to "collinearity"
    - **Collinearity**: When two or more of your independent variables are highly correlated; so correlated that adding just one of them explains both of their "effect" on the dependent variable. In other words including BOTH independent variables that are highly correlated is redundant and takes up statistical power
    - R will automatically drop variables if there is collinearity!
- **Model**: OLS Linear Regression
  - Powers doesn't write out the equation for the study!
  - Let's write it ourselves! [at least for the three groups of variables for Williams Case]

Powers (2004)

- Table 2: Nested Regression Models Using 1999 API Index as DV, pg. 781
  - The most "standard way" to format regression results
  - You show the "progression" towards your final regression model via multiple models
    - Intercept-only model (no independent variables) [Powers didnt show this]
    - Model 1 can have your independent variable of interest (standard for program eval) or some variables (Powers did student characteristics)
    - Models 2+... shows the addition of control variables
  - Shows beta coefficients, standard errors, and significant levels!
- Does show model fit statistics!
  - Powers is not writing from a program evaluation standpoint
  - Rather she's trying to show how variables from Williams Case explain (predict!) API scores
  - In this case, measures of model fit are important
  - R^2 increases from Model 1 to Model 2
    - Change in R^2 doesn't mean that Williams case IVs add "little" explanation; see pg 782!
  - In text, she mentions overall F-test is significant from Model 1 to Model 2
  - F-test is related to R^2
    - H_0: model with no independent variables (or limited IVs) fits the data just as well as model with full variables
    - H_a: model with more variables fits the data better than model with no IVs

Powers (2004)

- Interpretation for Table 2 for Williams Case Variables
- In text, she hardly interprets beta coefficients beyond "significance and direction"
  - Only sometimes mentions the magnitude of the coefficient
  - This approach is common in "descriptive research"
- **Teacher training** (reference group is fully credentialed teachers)
  - **Emergency Credentialed $\hat{\beta}$ Coefficient: -1.12*** SE= 0.09**
    - Interpretation: "On average, one-percentage-point increase in a school teaching staff's proportion of emergency credentialed teachers (as opposed to fully credentialed teachers) is associated with a 1.12 point decrease in API score, holding all covariates constant"
- **Teacher Experience**
  - **Years teaching $\hat{\beta}$ Coefficient: .80** SE= 0.26**
    - Interpretation: "A one year increase in a school teaching staff's average years of experience, on average, is associated with a 0.80 point increase in API score, holding all covariates constant"
- **Teacher Education** (reference group is less than M.A)
  - **Greater than MA $\hat{\beta}$ Coefficient: .42*** SE= 0.06**
    - Interpretation: "On average, a one-percentage-point increase in a school teaching staff's proportion of teachers with greater than MA degree (as opposed to lower than MA teachers) is associated with a 0.42 point increase in API score, holding all covariates constant"
- **School Calendar** (reference group is traditional year)
  - **Concept 6 $\hat{\beta}$ Coefficient: -34.46*** SE= 4.52**
    - Interpretation: "On average, being a school with a Concept 6 Calendar as opposed to a Traditional Calendar is associated with a 34.46 point decrease in API score, holding all covariates constant"
- **Textbooks**
  - **Per-pupil textbook expenditures $\hat{\beta}$ Coefficient: .11*** SE=0.003**
    - Interpretation: "A $1 increase in per-pupil expenditures for textbooks is associated with a 0.1 point increase in API score, holding all covariates constant"
    - Interpretation: "A $10 increase in per-pupil expenditures for textbooks, on average, is associated with a 1 point increase in API score, holding all covariates constant"
    - Interpretation: "A $100 increase in per-pupil expenditures for textbooks, on average, is associated with a 10 point increase in API score, holding all covariates constant"

Powers (2004)

Why do I like this piece of scholarship?

- The methods are simple!
  - Impactful research does not need to have complex methods!
  - Study does not use any analyses/methods beyond what we have learned so far in an introductory regression class; yet it is published in the AERA flagship journal
- Example of how to deal with "metrics" that I fundamentally disagree with but are part of our educational reality
  - Metrics like SAT/ACT scores, standardized K-12 test, are "firmly entrenched part of the political landscape" that schools and students must navigate
  - In many cases, they are institutionalized within federal, state, and university/school policies
  - "Aim is to make a strong argument for equity by marshaling the very data that dominate the political discourse" (p. 765)
- Direct response to policies
- Cool understanding of legal logic when it comes to education finance!