Week 1: Overview of class

Marcelo Coca Perraillon

University of Colorado
Anschutz Medical Campus

Health Services Research Methods I
HSMP 7607
2020

Outline

- Overview of this class
- Causal inference
- Regression models
- Topics we will cover
- Things you should know
- Logistics

Overview of this class

- This class is a combination of two applied topics: **regression analysis** and **methods for observational data**
- Or maybe better: regression models applied to the analysis of observational data
- By observational data I mean situations in which the information was collected without manipulation from an investigator. Alternatively, think of observational data as the **opposite of experimental data**
- We'll see much more about this but the main issue with observational data is that sometimes we want to answer research questions that are about **cause and effect**
- And answering **cause and effects** questions with observational data makes life complicated, but much more exciting (classic life trade-off?)

Overview of this class

- This class is also about regression models because regression modeling is the most common tool to we use to answer **causal** questions
- You need to become a regression modeling ninja as a doctoral student
- Ninja skills are acquired with a lot of study, time, practice, and a **mastering of the basics**
- The basics includes understanding what **regression model estimates mean**. Sounds simple, but it is not. And not knowing the basics is going to come back to **haunt you** – that I can promise with certainty
- In this class we will take advantage of the vanilla linear model, often inaccurately referred as OLS or confusingly as the "general" linear model (inaccurate because OLS is an estimation method; confusing because it's not the same as the Generalized Linear Model or GLM)

Overview of this class

- This class in an **applied** methods class
- I don't mean by applied that we will ignore statistical theory. We just won't be proving theorems
- Understanding theory beyond the proofs is the goal of this class. To be a good applied empirical researcher you need to understand different methods and know what theory tells you about them
- You don't need to know how to proof the asymptotic properties of an estimator, but you need to understand the concept
- The statistics (econometrics eventually follows) revolution of the last 40 years has been the result of raw **computing power**. We will use it too

Regression models

- The basic linear model or OLS or general linear model is $y_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_j x_{ji} + \epsilon_i$, where we assume $\epsilon \sim N(0, \sigma^2)$, which is the same as assuming $y \sim (\beta_0 + \beta_1 x_1 + ... + \beta_j x_j, \sigma^2)$
- We use this model a lot because it's incredibly flexible and helpful even in situations in which clearly assuming that $y \sim (\boldsymbol{\beta}' \boldsymbol{X}, \sigma^2)$ is wrong
- When this assumption doesn't hold, we can use other models like logistic, probit, Poisson. Basically, Generalized Linear Models (GLM)
- GLM is a framework or method that encompasses many different types of models. You need to define the **link** and the **family** of exponential distributions
- For example, a GLM with identity link and family Normal or Gaussian is the vanilla linear model. One with log link and Binomial family is the logistic model

Digression: Why do we use regression models?

- There are several reasons to run a regression (see Schmueli, 2010):
  - **Describe** a relationship
  - Make **causal** statements and test hypothesis
  - Make **predictions**
- The purpose of running a model **changes the way you think about assumptions and the way you conduct the analysis**
- We will often come back to issues of **causal inference**. It's central to health services research
- Anybody can analyze randomized data; working with **observational data** is a lot harder
- **Big picture**: The mechanics of regression are the same regardless of whether the model makes causal sense or not
- And by the way, don't get easily seduced by **machine learning** or "**data science**" or "**big data**": regression models are still the basics of these "new" fields. You can use regression models to make predictions

Causal inference

- You will become ninja like at using regression models because we use them in **causal inference** for observational data
- Causal inference has become a field in itself in the last 15 years or so in statistics, epidemiology, and other fields
- It has **always** been central in econometrics: think of data on prices and quantities and the notion of demand and supply curves
- But the new causal inference field in statistics has also changed econometrics
- It's an exciting time to be learning these methods now as they are being developed. But it's also **confusing**
- Assumptions come in many names, shape, and form. Sometimes there is no agreement on basic issues. I'll do a lot of translating among econometrics, biostatistics, and some, but not much, epi

The big picture

- We will start by **defining** causal effects and introducing some new language
- We will review regression results focusing on interpretation and practical issues at the beginning and during the semester
- Then we will talk about **research designs** and their assumptions without using statistical models
- I want you to be thinking about datasets and ways in which you could apply these methods. You will continue in Methods II
- During the second part of the class we will go deeper into **estimation**

Research designs

- By **research design** I mean the overall strategy used to answer a research question
- This is different than the **estimation** of effects or the particular regression model or statistical approach we use to answer a research question. I'm a big fan of separating the **design** from the **estimation**. Read Rubin (2007)
- Traditional econometrics (as in Wooldridge) tends to to combine both, but I find that more confusing than illuminating
- For example, whether causality is established or not is discussed in terms of assumptions about the error term $\epsilon \sim N(0, \sigma^2)$ (i.e. **conditional independence assumption**)
- But then in other models like logistic, probit, or Poisson there is **no additive error term** and so the intuition of conditional independence gets you tied in knots
- I'll cover both ways of understanding causal inference. Angrist and Pischke (2009) is a decent bridge. See also Imbens and Wooldridge (2009)

Research designs

- **A caveat**: It's not a clean separation
- In some circumstances our understanding of a research design needs to be adapted to a particular method of estimation
- Difference-in-difference is an example. Some assumption hold in the linear world, others can get complicated with nonlinear models
- See for example Lechmer (2010) page 196 (we will talk about this when we cover difference-in-difference models)
- **Jargon warning**: **Linear** means a lot of things. Sometimes literally a straight line; sometimes it's figuratively. But it also means **constant**. Sometimes people say linear when they really mean **additive** (as opposed to multiplicative like interactions)

Research designs

- We will cover the most important and widely used research designs in health services research classified on one main assumption:
- Selection on observables (ignorability):
  1. Regression adjustment
  2. Propensity scores and matching
- No selection on observables (ignorability):
  1. Difference-in-differences
  2. "Fixed" effects
  3. Instrumental variables
  4. Regression discontinuity
- I'll say more about what "selection on observables" means, but it's essentially whether we observe all **confounders** or not. If we do, regression adjustment and propensity scores are the right tools. If not, then there is an arsenal of tools with pros and cons. One simple rule of thumb is that **with observational data, almost for sure there is no selection on observables** (but could get close enough)

Selection on observables

- When the assumption that we observe, and therefore can control for, all confounders hold (selection on observables, ignorability), we can use regression models to obtain causal effects. With observational data we are seldom in this world, but sometimes we can get close enough
- BUT, regression adjustment relies on correct functional form, additive effects that may not hold when there is no **common support or lack of overlap**
- Think about this example: Most treated observations are older but most controls are younger. Age is a counfounder. What could be the problem with running $Y = \beta_0 + \beta_1 T_i + \beta_2 age_i + \boldsymbol{\beta' X} + \epsilon_i$?
- We are using older people to **extrapolate** information to the young and vice versa. We rely on the functional form to get the extrapolation right
- We are going to use the propensity score and matching to solve this problem. These are **new (or refined) methods that have given regression adjustment a facelift**

The Golden Rule of Causal Inference

- When we cover the methods, I want you to pay attention to the **Golden Rule of Causal Inference**:

  If you know how treatment was assigned, then you might be able to estimate (some type) of causal effect(s)
- (Digression: In my head I invented the term, which probably means that I read it or heard it somewhere, but I can't remember where – email me if you do)
- For each methods there will be **two types of assumptions**. Assumptions that can be verified with data and assumptions that you can't verify with with data (so called "**exclusion restrictions**")
- **No free lunches in statistics/econometrics**: for each method, you have to pay attention to the meaning of the estimates. They result not on **average** treatment effects but **local** average treatment effects (**LATE**)

Stata

- This class is also about Stata: Stata is like a knife to a cook, hammer to a carpenter, racket to a tennis player, essential oils to a new agey person, spiralizer for the paleo diet, misinformation to a politician
- In other words, it is a **tool** to achieve an end. The end is to manipulate and analyze data
- Stata is the tool of choice in health services research, policy analysis, econometrics, and health economics
- What makes Stata great is that the **syntax is consistent** (as opposed to SAS), you get the **output you need**, not everything under the sun (SAS), and the **documentation and help are fantastic** (as opposed to R and SAS)
- I have used SAS and R a lot. I do have strong opinions: SAS is great for manipulating data, but it makes you a bad analyst
- R is great for new methods, Bayesian statistics, and it's free, but the documentation is poor, difficult to manipulate large datasets. **And no, using R doesn't make you cooler, smarter, or a statistician**

Stata

- There are two features of Stata, among others, that are worth its price:
  1. Marginal effects (**margins** command)
  2. Tables with model output for comparison of models (**estimates** command)
- We will talk a lot about marginal effects because they are a great way of interpreting model estimates in any model other than the simple vanilla linear/OLS model with linear terms
- So maybe think of this class also as a class on Stata

Things to know

- We will go over the **basics** several times
- We will cover **a lot** of material; I do **expect you to work hard** – you are doctoral students. Eight or nine hours a week studying would be reasonable
- ...but there will be **no surprises**. It will always be clear what I expect from you. If not, just ask
- I don't expect you to find the resources you need to learn the material by yourself. That's why I'm here and that's why we have lecture notes, textbooks, and supplemental materials
- But I do expect you to figure some things out by yourself (it's not the opposite of previous point)
- **Ask questions**. Always ask questions even if you feel that it's too basic
- A note on my notes: use them as a **guide** to read the textbook and supplemental material but**you must read the assigned readings**

How to prepare for this class

- Attend class, although attendance is not mandatory
- Read the lecture notes, then the readings
- Work hard on problem sets (start early); **learn from your classmates**
- Statistics/econometrics is about math. Make every effort to know the math but...
- ...**math is a language**, not an end in itself. **Understanding what the math is telling you is much more important** for this class
- Look at/re-work a formula until it makes sense:

$$\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n}[(x_1 - \bar{x})^2 + ... + (x_n - \bar{x})^2]$$

Logistics

- Syllabus: work in progress but mostly done. I'll let you know if there is a significant change
- Homework: your chance to learn and practice
- Don't underestimate how upset I get about cheating on homeworks. **It's getting a lot worse with age**
- **No late homeworks, period**. If you have a valid excuse, let me know before the deadline

Grading and research proposal

- We will have weekly homework (30%). One midterm (25%) and a final (30%)
- Also a short study proposal (15%). The idea is for you to start thinking now about a research topic and a dataset. Obviously using some of the methods we will cover
- You can continue working on this project in the Methods II class
- Several students have written papers and/or dissertations based on these projects
- I like to separate design from estimation, but it also fits the class well because I want you to start thinking about a research proposal early in the class
- Schedule meetings with me to discuss your project over the semester

Readings and next steps

- I do expect you will do the readings. Some are optional (marked with an asterisk * in syllabus)
- Please review Wooldridge Appendices A to C
- You may find this lecture helpful: `https://perraillon.com/s/week-2-probability.pdf`
- You need to understand properties of an estimator: **bias**, **efficiency**, and **consistency**
- Review the basics of linear models

The End

- Next class: brief intro to Stata and review of linear/OLS
- Then causal inference
- Questions?