Week 1: Overview

Marcelo Coca Perraillon

University of Colorado
Anschutz Medical Campus

Health Services Research Methods I
HSMP 7607
2019

Outline

- Overview of this class
- What is regression analysis?
- Type of models we will cover
- Syllabus and goals
- Logistics

Overview of this class

- This class is about **regression analysis**
- Mostly about linear regression, aka ordinary least squares (OLS)
- We will also cover logistic and probit regression and a bit of Generalized Linear Models (GLM)
- Mostly about **cross-sectional** data (no clustered or longitudinal data) but you should make an effort to learn more about longitudinal data
- The different types of models –OLS, logit, Poisson– are driven by assumptions about the **distribution of the outcome or dependent variable**
- 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 [insert eye roll]

What is regression analysis?

- Textbook: "A conceptually simple method for investigating functional relationships among variables"
- A general regression model can be written as:

$$Y = f(X_1, ..., X_p) + \epsilon$$

- $f()$ can be any function relating the independent variables X to the outcome variable Y
- ϵ is a random error representing an unexplained source of discrepancy in the relationship between X and Y (could be **omitted** variables, too)
- Note that the equation is not a statistical model yet; just a general relationship among the variables. We need to be more specific about the details and the **data generating process**
- A **goal of this class is to be a stickler about writing statistical models**

Notation matters a lot

- Learning to write down models **matters a lot** (I'll repeat this like a parrot during the semester)
- You may think that it matters because it helps you communicate your research to **others**
- I actually think that it matters most because it helps **YOU** interpret your own models
- If you carefully write down your models, then you can **interpret coefficients** in the relevant scales
- If you interpret coefficients correctly, then you can figure out **which hypothesis test you need and what it means**
- All of the above will become clearer during the semester. For now, mark this slide with an asterisk, a heart, or many !!!!!!!

Jargon

- Another goal of this class is to help you **read the health services research literature**, which is based on econometrics and statistics/biostatistics (with a dash of epi for confusion)
- The variables $X_1, ..., X_p$ are called independent variables, *covariates*, explanatory variables, *predictors*, right-hand-side variables, factors, control variables, regressors...
- The variable Y is called the response variable, *outcome*, dependent variable, left-hand-side variable, regressand, explained variable...
- Sometimes ideas get lost in translation because econometricians and statisticians use terms in different ways
- The common offenders: **fixed effects**, **endogeneity**, **marginal effects**

Example

- Age and sex are related to a health index (HI):

$$HI = f(Age, Sex)$$

- A specific functional relationship:

$$HI = \beta_0 + \beta_1 Age + \beta_2 Male + \epsilon$$

- Another functional relationship:

$$HI = \beta_0 + \beta_1 Age + \beta_2 Age^2 + \beta_3 Male + \epsilon$$

- Yet another:

$$HI = \beta_0 + \beta_1 Age + \beta_2 Male + \beta_3 Age \times Male + \epsilon$$

- The $\beta_1, ..., \beta_p$ are the **regression parameters** or **coefficients** of the model
- We will spend most of the class working with a model in which we assume that $\epsilon \sim N(0, \sigma^2)$, which is the same as assuming that $Y \sim (\beta_1 X_1 + \cdots + \beta_p X_p, \sigma^2)$

Example

- I could have been more precise with the data structure and write:
- $HI_i = \beta_0 + \beta_1 Age_i + \beta_2 Male_i + \beta_3 Age_i \times Male_i + \epsilon_i$
- I could have added a sentence like "where i indexes an individual"
- With that, we know that we have a dataset with one observation per person. If "i" instead indexes a state, the model is interpreted in different way. In that case, *Age* would probably be average age by state and *Male* would probably be the proportion of males in the state
- This model is different, too:
  $HI_{it} = \beta_0 + \beta_1 Age_{it} + \beta_2 Male_i + \beta_3 Age_{it} \times Male_i + \epsilon_{it}$, where t denotes time in years
- This implies longitudinal data (we won't cover these models this semester)
- Why no subscript for *Male*?

Why do we use regression?

- There are many reasons to run a regression:
  - **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**
- **Big picture**: The mechanics of regression are the same regardless of whether the model makes causal sense or not
- 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
- You may as well try models like $Age = f(HI, Male)$ or $Male = f(HI, Age)$

Steps in regression analysis

- From Chatterjee:
  1. Statement of the problem, selection of variables
  2. Data collection
  3. Model specification, choice of fitting method
  4. Model fitting
  5. Model validation and criticism
- From Wooldridge:
  1. Develop a formal economic model (not always) or
  2. Rely on intuition or simple model to choose variables
  3. Model specification
  4. Testing of hypothesis
- Some philosophical differences are apparent here: no **model validation**?
- **Causality** is implicit in steps 1 and 2

Causality

- Causal inference has become a field in itself in the last 15 years or so in statistics and epidemiology
- It has **always** been central in econometrics: think of data on prices and quantities and the notion of demand and supply curves
- **Big picture**: estimating a statistical model and attaching a causal interpretation to it are **separate steps**
- The Chatterjee textbook focus on fitting and modeling but has little in terms of causality
- This has to do with the different origins of statistics and econometrics (experimental versus observational data)
- Wooldridge pays attention to causality. Economists have models of how people or firms interact, but what we see in the (**observational**) data are outcomes of the interactions, not the individual actions or components

Stata

- Mastering a statistical package will help you learn statistics
- Stata is a great statistical package
  1. Simple to use
  2. Clean interface
  3. Extensive and superb documentation
  4. Fairly cheap
- All the code for the examples I use in class will be posted on Canvas
- I'll introduce commands in lecture notes
- Make an effort to play with Stata. **Ask questions** when in doubt
- And no, using *R* doesn't make you cooler, smarter, or a statistician
- On the other hand, SAS is likely to make you a bad analyst

Topics we will cover

1. Exploratory data analysis
2. Causal inference
3. Simple linear regression (ordinary least squares, OLS)
4. Multiple linear regression
5. Maximum likelihood
6. Regression diagnostics
7. Variable transformation and weighted least squares
8. Qualitative variables
9. Linear probability models; logistic and probit regression
10. Variable selection
11. Dealing with autocorrelation
12. Dealing with collinearity

Methods sequence

- Methods I:
  - Introduction to causal inference
  - Linear regression, logit/probit, and a tiny bit of Generalized linear models (GLMs)
  - Interpretation of coefficients, modeling, and hypothesis testing
  - Proficiency with Stata
- Methods II:
  - Research design using observational data (and more causal inference)
  - Propensity scores (matching, weighting)
  - Panel/longitudinal data: difference-in-difference, fixed/random effects
  - Instrumental variables
  - Regression discontinuity
  - GLMs

Goals

1. Understand how to explore a dataset, including graphical techniques
2. Understand the statistical theory and assumptions underlying regression methods
3. Understand how to interpret regression parameters
4. Understand how to perform hypothesis testing
5. Understand under which circumstances a regression model can have a causal interpretation
6. Understand when and how various methods should be used, including the different uses of regression analysis (causal, descriptive, predictive)
7. Understand how to correctly write a statistical model
8. Become proficient in the use of Stata (aka Stata ninja)
9. Prepare you for Methods II

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**...
- ...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 textbooks**

How to prepare for this class

- Attend class, although attendance is not mandatory
- Read the textbook and supplemental material
- Work hard on problem sets (start early); learn from your classmates
- **Ask questions**!
- 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: always a work in progress but mostly done. I'll let you know if there is a significant change
- Homeworks: your chance to learn and practice
- You are violating university rules if you use previous year's homeworks or get answers online. Not to mention that I'll know about it because I make small changes in the data... So you are **warned**
- Also, 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
- Questions?