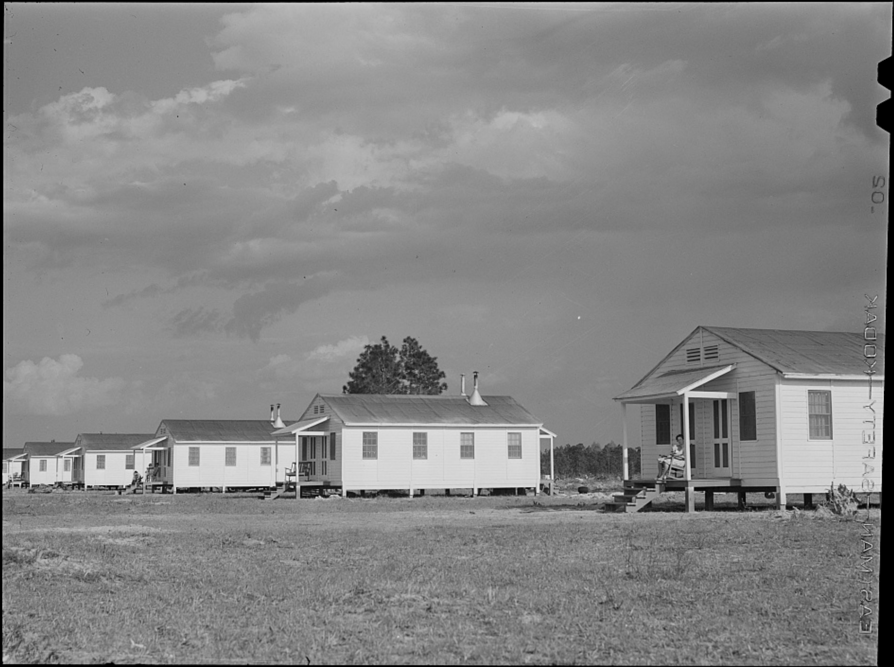
Gov 2002: Introduction

Fall 2023

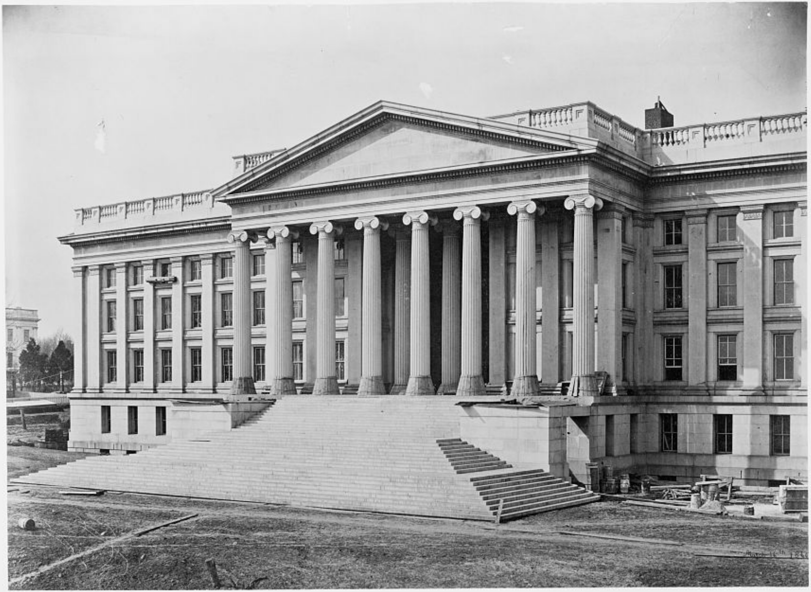
Matthew Blackwell

Gov 2002 (Harvard)

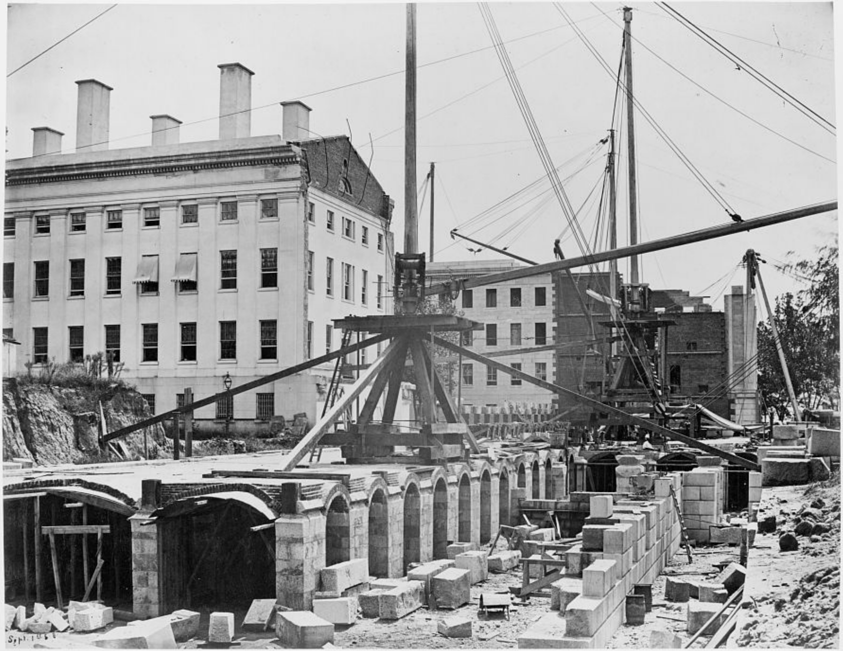
- Methods popular since I started grad school:
    - Machine learning, deep learning, text-as-data, audio-as-data, video-as-data, regression-discontinuity designs, Bayesian nonparametrics, design-based inference, spatial econometrics, network analysis, and so many more.



43698-D



123675


Sept. 1858

Why build a foundation?

- Extremely difficult to use or understand new methods without a **strong foundation in rigorous statistics**.
- You will be using methods for the rest of your career ⇝ you best invest!
    - Understanding your tools will make you better at your craft.
- You should never have to abandon a project because “you don’t know how to do it.”

Being asked a question about a method you don’t understand in a job talk.

Goals of this course

Have solid, core understanding of three topics:

1. Probability
2. Statistical Inference
3. Linear Regression

Overall goal: be empowered to learn any new method with relative ease.

Where we're going

Today:

- Understand the goals and logistics of the course
- Understand the basic definition of probability

1/ Course Details

Staff

- Instructor: Matthew Blackwell
- Teaching Fellow: Mitsuru Mukaigawara
- Course Assistant: Noah Dasanaike

Prerequisites

- Math we'll use in the course:
    - Knowledge of basic algebra and some exposure to basic statistics.
    - Calculus (limits, derivatives, integrals)
    - Linear algebra (vectors, matrices, etc)
    - Basically what's covered in Gov Math Prefresher (see syllabus for link)
- Computing:
    - We'll assume knowledge of R Math Prefresher.

How much time?

- The first year of grad school is a **marathon**:
  - Past students spent 5–20 hours per week on the HWs alone.
  - This can be painful, but it is **completely normal**
- Success in academia is a mix of: luck, creativity, knowledge, and **consistent hard work**
  - Becoming “fluent” in methods will pay off in the long (and short) run

Teaching resources

- Lecture: theoretical topics, example, etc.
- Sections: more specific targeted examples with an eye toward assignments
- Course Site: contains most of the course materials
    - Syllabus, schedule, lecture materials, etc.
- Ed Discussion Board: discussions about course material
- Slack: logistical and social discussions, DMs for help/study groups
- Office hours: ask even more questions.

Textbooks

- Responsibility = material covered in lectures.
- For those that want longer form writing,
    - Probability: Blitzstein and Hwang. Stat 110 textbook.
    - *A User's Guide to Statistical Inference and Regression* by me, basically a longer form version of my lecture notes. **In progress!**
- Other good book referenced on syllabus.

Grading

- Weekly homework assignments (55%)
- Take-home midterm exam (15%)
- Cumulative take-home final (20%)
- Participation (10%)
- PhD students: grades don't matter.

Things to do today

- Log into Ed and poke around.
- Join the course Slack.
- Make sure R, RStudio, and rmarkdown are all updated and work.

Outline of topics

- The basic outline of our semester, in backwards order:
    - **Regression**: core tool to estimate the relationship between variables.
    - **Inference**: how to learn about things we don't know from the things we do know.
    - **Probability**: what data we would expect if we did know the truth.
- Probability → Inference → Regression

2/ Overview of Probability and Statistics

Deterministic versus stochastic

- Key idea about statistics: **quantifying uncertainty**
- Imagine someone comes to us and says, “what is the relationship between voter turnout and campaign spending?”
- **Deterministic** account of voter turnout in a district:

$$\text{turnout}_i = f(\text{spending}_i).$$

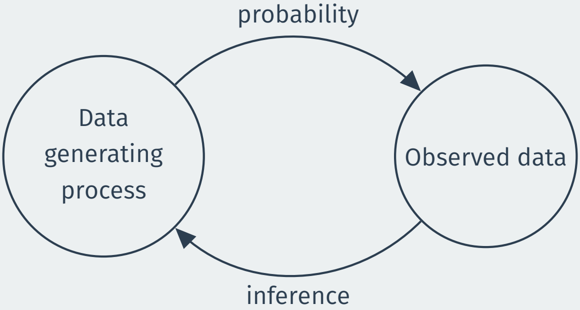
- What’s the problem with this? Omits all other determinants:
    - open seat, challenger quality, weather on election day, having the local college football team win the previous weekend, whether or not Jimmy had to stay home sick from school

Stochastic models

- Measure everything and then add it to our model:

$$\text{turnout}_i = f(\text{spending}_i) + g(\text{stuff}_i)$$

- Treat other factors as direct interest as **stochastic**:
  - They affect the outcome, but are not of direct interest.
  - We think of them as part of the chance variation in turnout.
- How do we quantify chance variation: **probability**



Why probability?

- Next few weeks: **probability**
  - Not a punishment.
  - Probability helps us study stochastic events.
  - Important for all of statistics.
- Statistical inference is a **thought experiment**.
- Probability is the logic of these thought experiments.
- Thought experiments: assume men and women were paid the same on average, but there was chance variation from person to person.
  - If true, how likely is the observed wage gap in this hypothetical world?
  - What kinds of wage gaps would we expect to observe in this hypothetical world?
- Probability to the rescue!