# Statistical Methods III: Spring 2013

Jonathan Wand

Stanford University

Introduction

# Outline

1. What is this course about?
2. Topics and questions addressed in this course
    - Likelihood-based inference
    - Estimation of unknown functions
    - Inference at the level of a model
3. Requirements
    - Replication Paper
    - Replication (and critique) of Replication Paper
    - Problem sets and quizzes
    - Participation
4. Philosophy
5. Questions?
6. Talking about your research interests / questions

# What is this course about?

- mathematical and statistical methods
  - formalizing theory, identification of parameters
  - estimation of unknown quantities and inference
- practical tools
- judgement about methodological approach
  - improving how you translate theory into statistical model
  - choosing the appropriate machinery for evaluating a theory
- how to better critique work of others
  - how well do the statistical models capture competing theories?
  - what is the power of test(s) to discriminate among theories?
  - what are threats to inference?
- how to enhance collaborative research and write scholarly work
- taking ownership of research and learning

# Likelihood-based inference

- A likelihood is a model of a data generating process.
- Standard linear model,

$$y_i = x_i^\top \beta + \epsilon_i$$
$$\epsilon_i \sim N(0, \sigma^2)$$

- Alternative, equivalent

$$Y_i \sim f(y_i \mid \mu_i, \sigma^2) \qquad \text{Q: what is } f?$$
$$\mu_i = x_i^\top \beta$$

- Consider $Y_i \in \{0, 1\}$, what have we got here:

$$Y_i \sim f(y_i \mid \pi_i) \qquad \text{Q: what is } f?$$
$$\pi_i = g(x_i^\top \beta) = 1/(1 + e^{-x_i \beta})$$

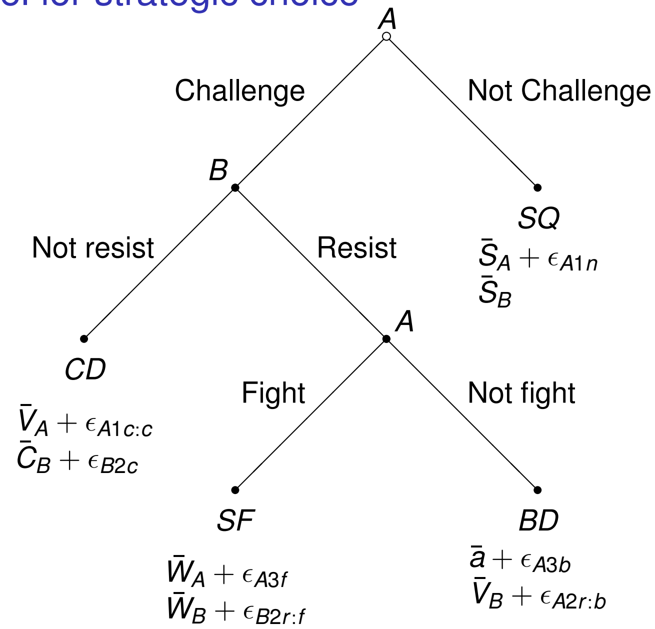
- Where do we get a likelihood? A theory.

# Likelihood-based inference

We will use theories of choice to motivate statistical models.

- dichotomous choice sets $Y_i \in \{0, 1\}$
- multiple unordered choice sets $Y_i \in \{0, 1, ..., K\}$
- ordered choice sets
- models of indifference and alienation in voting
- nested choices
- strategic choice

# Model for strategic choice



*A*

Challenge

Not Challenge

*B*

*SQ*

$\bar{S}_A + \epsilon_{A1n}$
$\bar{S}_B$

Not resist

Resist

*A*

*CD*

$\bar{V}_A + \epsilon_{A1c:c}$
$\bar{C}_B + \epsilon_{B2c}$

Fight

Not fight

*SF*

$\bar{W}_A + \epsilon_{A3f}$
$\bar{W}_B + \epsilon_{B2r:f}$

*BD*

$\bar{a} + \epsilon_{A3b}$
$\bar{V}_B + \epsilon_{A2r:b}$

# Likelihood-based inference

- Questions for each model,
  - what is known, what is unknown?
  - what can we identify?
- Questions that we solve in generality,
  - how do we estimate unknown quantities?
  - what are the properties of estimates?
- We will focus on Maximum Likelihood Estimators (MLE)
- ...details differ for other estimators, but many lessons/tools generalize

# Estimation of unknown functions

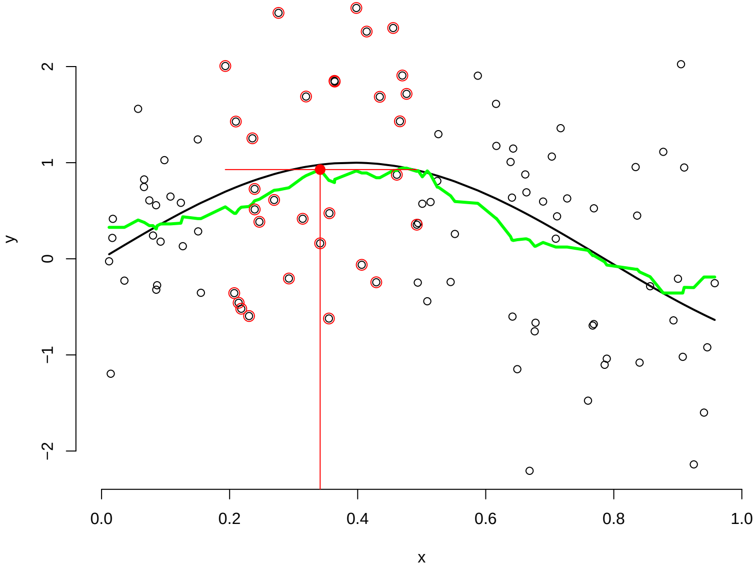
In regression classes, what do we do?

- the workhorse specification of the conditional expectation

$$E(Y_i|x_i) = x_i^\top \beta$$

- if $x_i$ are at least ordinal (polity, sort of), then treat as real numbers, $\rightarrow x_{ij}\beta_j$ describes a line
- if a variable is categorical, then perhaps create indicator values $\rightarrow x_{ij}\beta_j$ produces a bunch of mean shifts, one for each value of $x_i$
- we can do better than assuming everything is either a bunch of mean shift sor linear function, structure of mean shifts
  $\rightarrow$ and we can also test fitness of linearity

# Kernel-NN

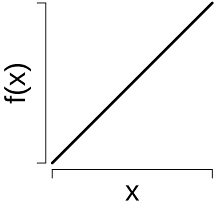


# Fitting shapes: a plethora of methods

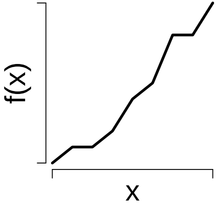
- Polynomials (e.g., Ostrom and Aldrich, 1978)
  - + few parameters, ease of implementation
  - – weakness: hard to impose shapes; global fit
- Smoothers/local regression (e.g., Fan 1990; Wand 1995)
  - + flexible
  - – overfitting; high dimensional; hard to test shapes; choice of polynomial order (local regression), bandwidth
- Isotonic regression (e.g., Barlow et al, 1972)
  - + discrete data; non-smoothness
  - – ill-defined on continuous data
- splines (e.g., Dierckx 1993)
  - + flexibility within limits; finite parameters; classical testing
  - – choice of knots and order

Cf. Keele (2008) and Hastie et al (2001) for unified overviews.

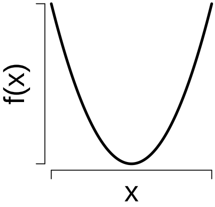
# Functional relationships



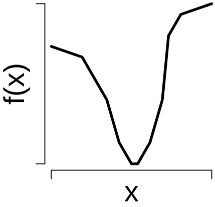
Linear, $f'(x) = \beta$



Monotonic, $f'(x) \geq 0$

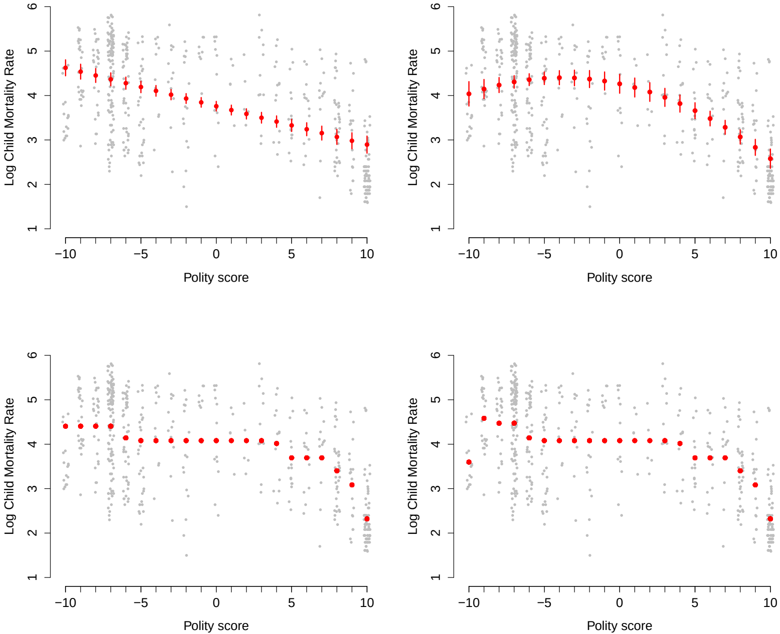


Quadratic $f' \propto \beta x$
Convex, $f''(x) > 0$

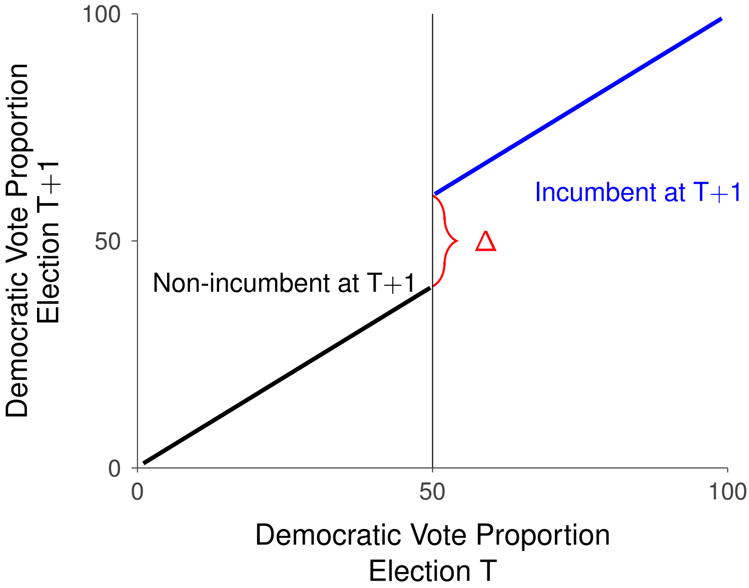


Single minima,
$f'(x) < 0$ then $f'(x) > 0$

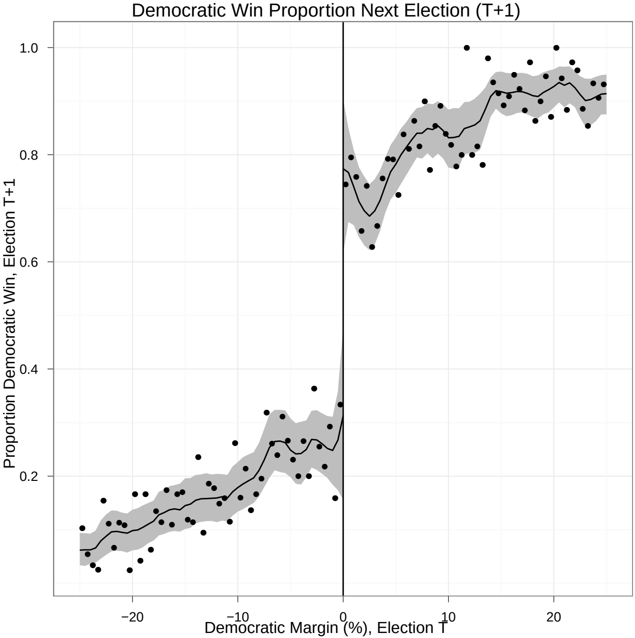
# Polity scores and child mortality



# RDD


100
50
0
0
50
100
Democratic Vote Proportion
Election T+1
Incumbent at T+1
Δ
Non-incumbent at T+1
Democratic Vote Proportion
Election T

# RDD



Democratic Win Proportion Next Election (T+1)
Proportion Democratic Win, Election T+1
1.0
0.8
0.6
0.4
0.2
−20
−10
0
10
20
Democratic Margin (%), Election T

# Inference at the level of a model

- What are we testing?
  - Already you know how test a hypothesis about a single parameter. Even a joint hypothesis about a set of parameters.
  - Here, we will think in terms of comparing models/theories.
  - at this point, we will already have tools for comparing a pairs of nested models (e.g., LRT).
- We will now generalize.
  - What happens if we have more than two theories?
  - (and you should always have at least two theories...)
  - and then add in a model that is purely data driven, as a specification test. What do you do now?
  - what if models are non-nested?
- We will also touch on the issue of DGP being composed of multiple models
  - which is itself a model
  - we will think in terms of mixtures of likelihoods

# Replication Paper

For a paper you find interest

- critique (what would do differently? what is at stake?)
- collect (get data from archive, author, or rebuild)
- replicate (rerun exactly what they said they found)
- implement correction implied by the

Logistics

- this is collaborative project, producing:
- a paper
- a replication archive
- (find a partner, papers ideally will be done in pairs)

# Replication (and critique) of Replication Paper

After replication papers are submitted,

- you will be assigned a paper to review
- this is a time-bounded exercise
- one day to produce 1-2 page review of paper, and replication archive
- akin to a journal review process

# Problem sets

For both problem sets and quizzing

- work together to figure out principles and concepts involved
- you must execute the answering of the problem set by yourself
- the work submitted must be your own

Problem sets

- weekly,
- submit replicable solutions via dropbox

Quizzes

- will accompany lecture notes
- you are expected to do these before the lecture
- not submitted, a guide and diagnostic

# Participation

- a key part of the course
- in-class and on-line
- use piazza to
    - ask questions of each other...
    - ... and answer each others questions
    - discuss lectures, readings...
    - ... and shape where we spend time in lectures

# Philosophy

Our goals in this course are for you

- to learn fundamentals, principles
- to gain practice generalizing to specific cases
- such that you gain the ability to produce new knowledge

This course is just the beginning.

# Your research

Let's talk about your interests.

- briefly, what is your (main) research question that you are thinking about
  NOTE: this obviously can be tentative—the point of this course is to help you to improve how you ask questions and the tools with you can test them
  - (if you have more than one, pick one)
  - if you do not currently have a research question in mind:
    what is a puzzle about the world you would like to answer?
- what is a key quantity of interest in this theory?
- what is a key hypothesis?
- what is the greatest obstacle to testing this hypothesis (e.g., confoundedness, data collection, ...)?

To clarify—we are looking for a question rather than a topic.