Soc504: Introduction

Brandon Stewart[1]

Princeton

February 6, 2017

[1] These slides are heavily influenced by Gary King, Matt Salganik and Teppei Yamamoto.

Where We've Been and Where We're Going...

- Last Week
  - a week long respite
- This Week
  - Monday
    - introduction
  - Wednesday
    - probabilistic infrastructure
- Next Week
  - likelihood inference
- Long Run
  - likelihood $\rightarrow$ GLMs $\rightarrow$ advanced methods

Questions?

Welcome and Introductions

Welcome and Introductions

- Soc504: Advanced Social Statistics

Welcome and Introductions

- Soc504: Advanced Social Statistics
- Your Preceptors

Welcome and Introductions

- Soc504: Advanced Social Statistics
- Your Preceptors
  - Rebecca Johnson
  - Ian Lundberg

Welcome and Introductions

- Soc504: Advanced Social Statistics
- Your Preceptors
  - Rebecca Johnson
  - Ian Lundberg
- Any newcomers?

Interesting Comments on the Final

Interesting Comments on the Final

I got the sense that the true test of the final was a philosophical point:

Interesting Comments on the Final

I got the sense that the true test of the final was a philosophical point: we were asked not just to recall and apply the many concepts we learned, but to dwell on the complexities of applying quantitative methods to study the real world, and the persnickety (and probably omnipresent) situation of when the numbers look right but the substantive takeaway is wrong.

Interesting Comments on the Final

I got the sense that the true test of the final was a philosophical point: we were asked not just to recall and apply the many concepts we learned, but to dwell on the complexities of applying quantitative methods to study the real world, and the persnickety (and probably omnipresent) situation of when the numbers look right but the substantive takeaway is wrong. I wondered what you were trying to show us by constantly having us discover that respected and published studies were, if not flawed, then at least imperfect. And what does it really mean for a study to be imperfect?

Interesting Comments on the Final

I got the sense that the true test of the final was a philosophical point: we were asked not just to recall and apply the many concepts we learned, but to dwell on the complexities of applying quantitative methods to study the real world, and the persnickety (and probably omnipresent) situation of when the numbers look right but the substantive takeaway is wrong. I wondered what you were trying to show us by constantly having us discover that respected and published studies were, if not flawed, then at least imperfect. And what does it really mean for a study to be imperfect? I spent a lot of time on some low point value questions, like fixing the nonlinearity in problem 1. And because of the more philosophical thing I mentioned above, I was never really sure whether I had the right answer to something, or whether we were meant to be sure...

Interesting Comments on the Final

I got the sense that the true test of the final was a philosophical point: we were asked not just to recall and apply the many concepts we learned, but to dwell on the complexities of applying quantitative methods to study the real world, and the persnickety (and probably omnipresent) situation of when the numbers look right but the substantive takeaway is wrong. I wondered what you were trying to show us by constantly having us discover that respected and published studies were, if not flawed, then at least imperfect. And what does it really mean for a study to be imperfect? I spent a lot of time on some low point value questions, like fixing the nonlinearity in problem 1. And because of the more philosophical thing I mentioned above, I was never really sure whether I had the right answer to something, or whether we were meant to be sure...

I will add though that after reviewing for and taking the final I have new doubts that I did not have during the semester. I am sure my classmates feel the same. It'd be great if we could review them in the Spring at some point.

Advice from you to you

- Be ready to spend a lot of time

Advice from you to you

- Be ready to spend a lot of time
- Definitely take it! And be prepared to set aside a lot of time for it.

Advice from you to you

- Be ready to spend a lot of time
- Definitely take it! And be prepared to set aside a lot of time for it.
- Ask questions if you don't know what's going on!

Advice from you to you

- Be ready to spend a lot of time
- Definitely take it! And be prepared to set aside a lot of time for it.
- Ask questions if you don't know what's going on!
- Study hard, work hard, review the slides.

Advice from you to you

- Be ready to spend a lot of time
- Definitely take it! And be prepared to set aside a lot of time for it.
- Ask questions if you don't know what's going on!
- Study hard, work hard, review the slides.
- Investing a considerable amount of time in getting familiar with R and its various tools will pay off in the long run!

Advice from you to you

- Be ready to spend a lot of time
- Definitely take it! And be prepared to set aside a lot of time for it.
- Ask questions if you don't know what's going on!
- Study hard, work hard, review the slides.
- Investing a considerable amount of time in getting familiar with R and its various tools will pay off in the long run!
- Go over the lecture slides each week. This can be hard when you feel like you're threaded water and just staying afloat, but I wish I had done this regularly.

Advice from you to you

- Be ready to spend a lot of time
- Definitely take it! And be prepared to set aside a lot of time for it.
- Ask questions if you don't know what's going on!
- Study hard, work hard, review the slides.
- Investing a considerable amount of time in getting familiar with R and its various tools will pay off in the long run!
- Go over the lecture slides each week. This can be hard when you feel like you're threaded water and just staying afloat, but I wish I had done this regularly.
- It's challenging but very doable and rewarding if you put the time in. There are plenty of resources to take advantage of for help.

Advice from you to you

- Be ready to spend a lot of time
- Definitely take it! And be prepared to set aside a lot of time for it.
- Ask questions if you don't know what's going on!
- Study hard, work hard, review the slides.
- Investing a considerable amount of time in getting familiar with R and its various tools will pay off in the long run!
- Go over the lecture slides each week. This can be hard when you feel like you're threaded water and just staying afloat, but I wish I had done this regularly.
- It's challenging but very doable and rewarding if you put the time in. There are plenty of resources to take advantage of for help.
- Take it, you won't regret it!

What's this course about?

What's this course about?

- Specific statistical methods for many research problems

What's this course about?

- Specific statistical methods for many research problems
  - How to learn (or create) new methods

What's this course about?

- Specific statistical methods for many research problems
  - How to learn (or create) new methods
  - Inference: Using facts you know to learn about facts you don't know

What's this course about?

- Specific statistical methods for many research problems
  - How to learn (or create) new methods
  - Inference: Using facts you know to learn about facts you don't know
- How to write a publishable scholarly paper

What's this course about?

- Specific statistical methods for many research problems
  - How to learn (or create) new methods
  - Inference: Using facts you know to learn about facts you don't know
- How to write a publishable scholarly paper
- All the practical tools of research — theory, applications, simulation, programming, word processing, plumbing, whatever is useful

What's this course about?

- Specific statistical methods for many research problems
  - How to learn (or create) new methods
  - Inference: Using facts you know to learn about facts you don't know
- How to write a publishable scholarly paper
- All the practical tools of research — theory, applications, simulation, programming, word processing, plumbing, whatever is useful
- $\rightsquigarrow$ Outline and class materials:

What's this course about?

- Specific statistical methods for many research problems
  - How to learn (or create) new methods
  - Inference: <u>Using facts you know to learn about facts you don't know</u>
- How to write a publishable scholarly paper
- All the practical tools of research — theory, applications, simulation, programming, word processing, plumbing, whatever is useful
- $\rightsquigarrow$ Outline and class materials:
  - The syllabus gives topics, not a strict weekly plan.

What's this course about?

- Specific statistical methods for many research problems
  - How to learn (or create) new methods
  - Inference: Using facts you know to learn about facts you don't know
- How to write a publishable scholarly paper
- All the practical tools of research — theory, applications, simulation, programming, word processing, plumbing, whatever is useful
- $\rightsquigarrow$ Outline and class materials:
  - The syllabus gives topics, not a strict weekly plan.
  - We will go as fast as possible subject to everyone following along

What's this course about?

- Specific statistical methods for many research problems
  - How to learn (or create) new methods
  - Inference: <u>Using facts you know to learn about facts you don't know</u>
- How to write a publishable scholarly paper
- All the practical tools of research — theory, applications, simulation, programming, word processing, plumbing, whatever is useful
- $\rightsquigarrow$ Outline and class materials:
  - The syllabus gives topics, not a strict weekly plan.
  - We will go as fast as possible subject to everyone following along
  - We cover different amounts of material each week

Modes of Learning

Modes of Learning

1. Weekly assignments

Modes of Learning

1. Weekly assignments
   - Readings (via Perusall which we will return to)

Modes of Learning

1. Weekly assignments
   - Readings (via Perusall which we will return to)
   - Problem Sets (likely 8 in total)

Modes of Learning

1. Weekly assignments
   - Readings (via Perusall which we will return to)
   - Problem Sets (likely 8 in total)
2. One “publishable” coauthored paper.

Modes of Learning

1. Weekly assignments
   - Readings (via Perusall which we will return to)
   - Problem Sets (likely 8 in total)
2. One “publishable” coauthored paper.
   - Long tradition of these being published from similar classes

Modes of Learning

1. Weekly assignments
   - Readings (via Perusall which we will return to)
   - Problem Sets (likely 8 in total)
2. One "publishable" coauthored paper.
   - Long tradition of these being published from similar classes
   - Some key mid-point assignments along the way

Modes of Learning

1. Weekly assignments
   - Readings (via Perusall which we will return to)
   - Problem Sets (likely 8 in total)
2. One "publishable" coauthored paper.
   - Long tradition of these being published from similar classes
   - Some key mid-point assignments along the way
   - First reading: "Publication, Publication" will help

Modes of Learning

1. Weekly assignments
   - Readings (via Perusall which we will return to)
   - Problem Sets (likely 8 in total)
2. One "publishable" coauthored paper.
   - Long tradition of these being published from similar classes
   - Some key mid-point assignments along the way
   - First reading: "Publication, Publication" will help
   - You won't be alone: you'll work with each other and us

Modes of Learning

1. Weekly assignments
   - Readings (via Perusall which we will return to)
   - Problem Sets (likely 8 in total)
2. One "publishable" coauthored paper.
   - Long tradition of these being published from similar classes
   - Some key mid-point assignments along the way
   - First reading: "Publication, Publication" will help
   - You won't be alone: you'll work with each other and us
3. Participation and collaboration:

Modes of Learning

1. Weekly assignments
   - Readings (via Perusall which we will return to)
   - Problem Sets (likely 8 in total)
2. One "publishable" coauthored paper.
   - Long tradition of these being published from similar classes
   - Some key mid-point assignments along the way
   - First reading: "Publication, Publication" will help
   - You won't be alone: you'll work with each other and us
3. Participation and collaboration:
   - Come to precept and office hours

Modes of Learning

1. Weekly assignments
   - Readings (via Perusall which we will return to)
   - Problem Sets (likely 8 in total)
2. One "publishable" coauthored paper.
   - Long tradition of these being published from similar classes
   - Some key mid-point assignments along the way
   - First reading: "Publication, Publication" will help
   - You won't be alone: you'll work with each other and us
3. Participation and collaboration:
   - Come to precept and office hours
   - Collaborate for problem sets

Modes of Learning

1. Weekly assignments
   - Readings (via Perusall which we will return to)
   - Problem Sets (likely 8 in total)
2. One "publishable" coauthored paper.
   - Long tradition of these being published from similar classes
   - Some key mid-point assignments along the way
   - First reading: "Publication, Publication" will help
   - You won't be alone: you'll work with each other and us
3. Participation and collaboration:
   - Come to precept and office hours
   - Collaborate for problem sets
   - Ask questions

Modes of Learning

1. Weekly assignments
   - Readings (via Perusall which we will return to)
   - Problem Sets (likely 8 in total)
2. One "publishable" coauthored paper.
   - Long tradition of these being published from similar classes
   - Some key mid-point assignments along the way
   - First reading: "Publication, Publication" will help
   - You won't be alone: you'll work with each other and us
3. Participation and collaboration:
   - Come to precept and office hours
   - Collaborate for problem sets
   - Ask questions
   - Build class camaraderie: prepare, participate, help others

Modes of Learning

1. Weekly assignments
   - Readings (via Perusall which we will return to)
   - Problem Sets (likely 8 in total)
2. One "publishable" coauthored paper.
   - Long tradition of these being published from similar classes
   - Some key mid-point assignments along the way
   - First reading: "Publication, Publication" will help
   - You won't be alone: you'll work with each other and us
3. Participation and collaboration:
   - Come to precept and office hours
   - Collaborate for problem sets
   - Ask questions
   - Build class camaraderie: prepare, participate, help others
4. Help us help you.

Course strategy

Course strategy

- We could teach you the latest and greatest methods,

Course strategy

- We could teach you the latest and greatest methods, but when you graduate they will be old

Course strategy

- We could teach you the latest and greatest methods, but when you graduate they will be old
- We could teach you all the methods that might prove useful during your career,

Course strategy

- We could teach you the latest and greatest methods, but when you graduate they will be old
- We could teach you all the methods that might prove useful during your career, but when you graduate you will be old

Course strategy

- We could teach you the latest and greatest methods, but when you graduate they will be old
- We could teach you all the methods that might prove useful during your career, but when you graduate you will be old
- Instead, we teach you the fundamentals, the underlying theory of inference, from which statistical models are developed:

Course strategy

- We could teach you the latest and greatest methods, but when you graduate they will be old
- We could teach you all the methods that might prove useful during your career, but when you graduate you will be old
- Instead, we teach you the fundamentals, the underlying theory of inference, from which statistical models are developed:
  - We will reinvent existing methods by creating them from scratch.

Course strategy

- We could teach you the latest and greatest methods, but when you graduate they will be old
- We could teach you all the methods that might prove useful during your career, but when you graduate you will be old
- Instead, we teach you the fundamentals, the underlying theory of inference, from which statistical models are developed:
  - We will reinvent existing methods by creating them from scratch.
  - We will learn: its easy to invent new methods too, when needed.

Course strategy

- We could teach you the latest and greatest methods, but when you graduate they will be old
- We could teach you all the methods that might prove useful during your career, but when you graduate you will be old
- Instead, we teach you the fundamentals, the underlying theory of inference, from which statistical models are developed:
  - We will reinvent existing methods by creating them from scratch.
  - We will learn: its easy to invent new methods too, when needed.
  - The fundamentals help us pick up new methods created by others.

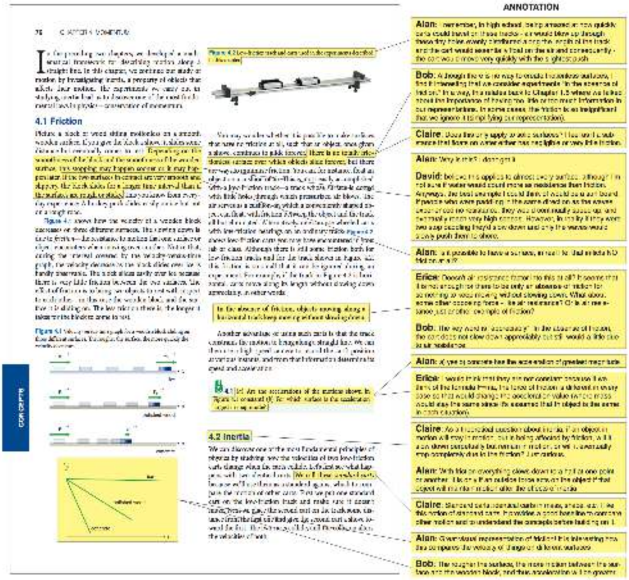
Course strategy

- We could teach you the latest and greatest methods, but when you graduate they will be old
- We could teach you all the methods that might prove useful during your career, but when you graduate you will be old
- Instead, we teach you the fundamentals, the underlying theory of inference, from which statistical models are developed:
  - We will reinvent existing methods by creating them from scratch.
  - We will learn: its easy to invent new methods too, when needed.
  - The fundamentals help us pick up new methods created by others.
- This helps us separate the conventions from underlying statistical theory.

Syllabus Talk Through

Perusall Example

Perusall Example



The Evolving Role of Reading

- Reading is much more essential this semester.

The Evolving Role of Reading

- Reading is much more essential this semester.
- Understanding statistics reading is a bit of a challenge at times. Biggest advice: don't skip the equations.

The Evolving Role of Reading

- Reading is much more essential this semester.
- Understanding statistics reading is a bit of a challenge at times. Biggest advice: don't skip the equations.
- Ask for help on Perusall if you don't understand pieces.

The Evolving Role of Reading

- Reading is much more essential this semester.
- Understanding statistics reading is a bit of a challenge at times. Biggest advice: don't skip the equations.
- Ask for help on Perusall if you don't understand pieces.
- We will still cover everything in class, but the reading will be important for complementing your understanding.

Help

Help

- Participate in Collaborative Annotation, Piazza, Precept, Office Hours

Help

- Participate in Collaborative Annotation, Piazza, Precept, Office Hours

Help

- Participate in Collaborative Annotation, Piazza, Precept, Office Hours
- In-

Help

- Participate in Collaborative Annotation, Piazza, Precept, Office Hours
- In-ter-

Help

- Participate in Collaborative Annotation, Piazza, Precept, Office Hours
- In-ter-rupt

Help

- Participate in Collaborative Annotation, Piazza, Precept, Office Hours
- In-ter-rupt me as often as necessary

Help

- Participate in Collaborative Annotation, Piazza, Precept, Office Hours
- In-ter-rupt me as often as necessary
- (Got a dumb question? Assume you are the smartest person in class and you eventually will be!)

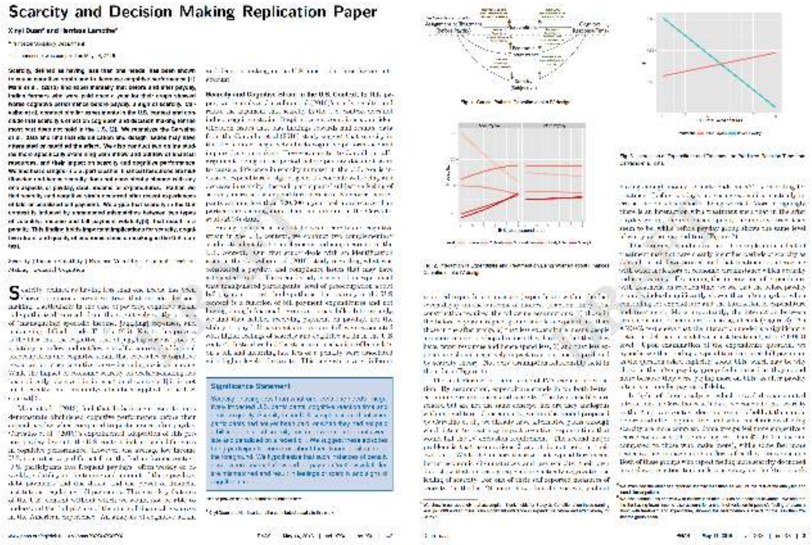
Help

- Participate in Collaborative Annotation, Piazza, Precept, Office Hours
- In-ter-rupt me as often as necessary
- (Got a dumb question? Assume you are the smartest person in class and you eventually will be!)
- When are Brandon's office hours?

Help

- Participate in Collaborative Annotation, Piazza, Precept, Office Hours
- In-ter-rupt me as often as necessary
- (Got a dumb question? Assume you are the smartest person in class and you eventually will be!)
- When are Brandon's office hours?
  - Come whenever you like; if you can't find me or I'm in a meeting, come back or email any time

PNAS Format



Scarcity and Decision Making Replication Paper

Significance Statement

Topics

- Week 1: Introduction and Theories of Inference
- Week 2: Maximum Likelihood Inference
- Week 3: Qois and Binary Outcome Models
- Week 4: Generalized Linear Models, Probit/Logit
- Week 5: Ordered DVs, Zero Inflation
- Week 6: Event Counts and Duration Modeling

Topics

- Week 1: Introduction and Theories of Inference
- Week 2: Maximum Likelihood Inference
- Week 3: Qois and Binary Outcome Models
- Week 4: Generalized Linear Models, Probit/Logit
- Week 5: Ordered DVs, Zero Inflation
- Week 6: Event Counts and Duration Modeling
- Week 7: Mixture Models and Expectation Maximization
- Week 8: Missing Data

Topics

- Week 1: Introduction and Theories of Inference
- Week 2: Maximum Likelihood Inference
- Week 3: Qois and Binary Outcome Models
- Week 4: Generalized Linear Models, Probit/Logit
- Week 5: Ordered DVs, Zero Inflation
- Week 6: Event Counts and Duration Modeling
- Week 7: Mixture Models and Expectation Maximization
- Week 8: Missing Data
- Week 9: Model Dependence and Matching
- Week 10: Mediation Analysis

Topics

- Week 1: Introduction and Theories of Inference
- Week 2: Maximum Likelihood Inference
- Week 3: Qois and Binary Outcome Models
- Week 4: Generalized Linear Models, Probit/Logit
- Week 5: Ordered DVs, Zero Inflation
- Week 6: Event Counts and Duration Modeling
- Week 7: Mixture Models and Expectation Maximization
- Week 8: Missing Data
- Week 9: Model Dependence and Matching
- Week 10: Mediation Analysis
- Week 11: Regularization and Hierarchical Models
- Week 12: Multilevel and Hierarchical Modeling

Why a Replication Project?

Why a Replication Project?

- A great way to get into writing, publishing research.

Why a Replication Project?

- A great way to get into writing, publishing research.
- Helps you see the literature in a new way

Why a Replication Project?

- A great way to get into writing, publishing research.
- Helps you see the literature in a new way
- Prepares you for the 2nd year empirical paper (For Soc grad students)

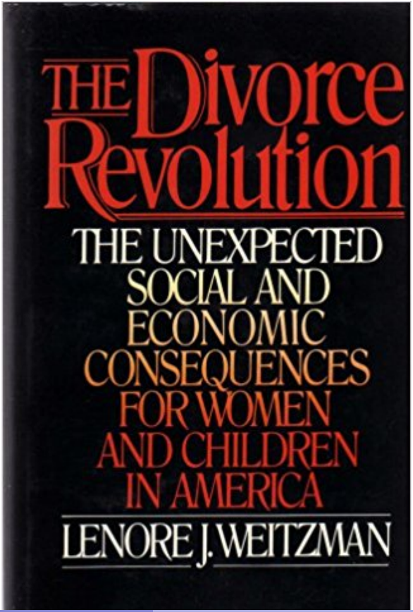
Why a Replication Project?

- A great way to get into writing, publishing research.
- Helps you see the literature in a new way
- Prepares you for the 2nd year empirical paper (For Soc grad students)
- Enforces better replicability practices

Why Replicability?

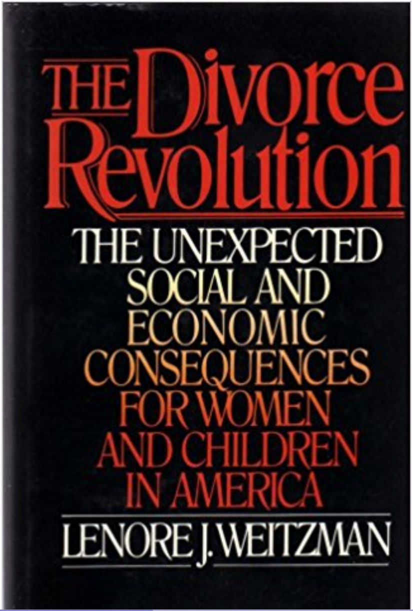
Why Replicability?

- Changes in living standard after divorce



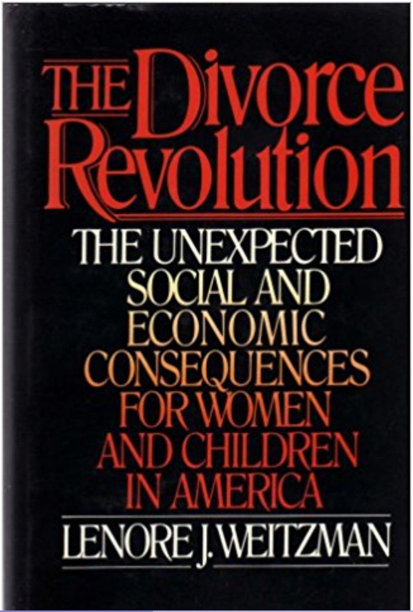
Why Replicability?

- Changes in living standard after divorce
  - for women declines 73%



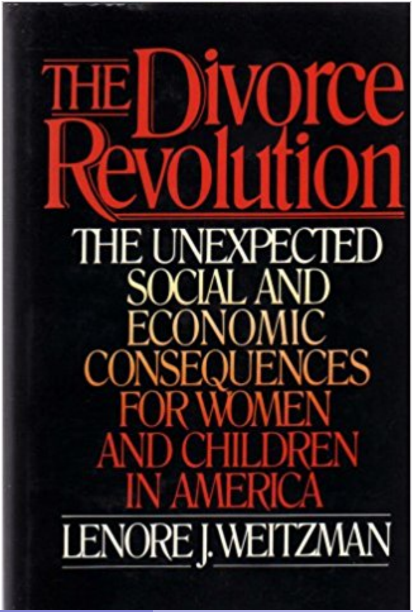
Why Replicability?

- Changes in living standard after divorce
  - for women declines 73%
  - for men increases 42%



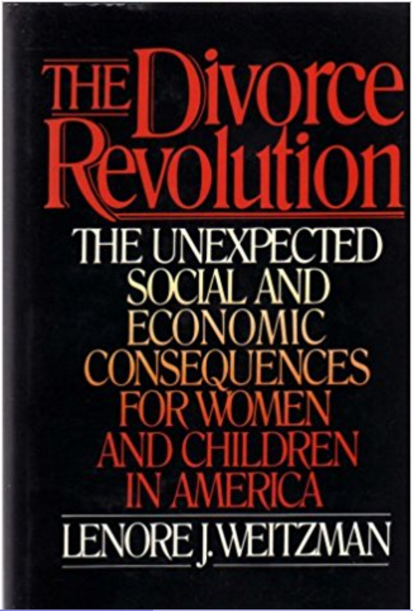
Why Replicability?

- Changes in living standard after divorce
  - for women declines 73%
  - for men increases 42%
- ASA Book Award in 1986



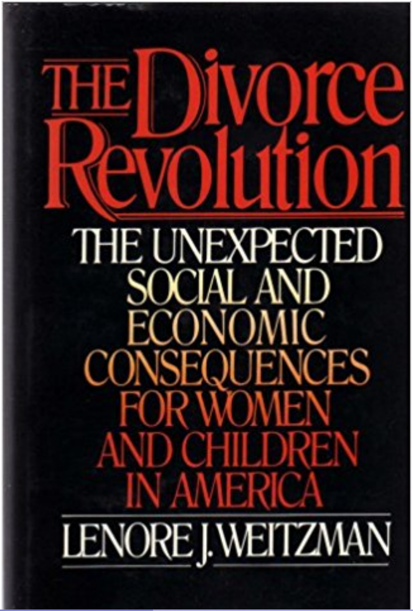
Why Replicability?

- Changes in living standard after divorce
  - for women declines 73%
  - for men increases 42%
- ASA Book Award in 1986
- Between 1986 and 1993, cited in 348 social science articles and 250 law review articles



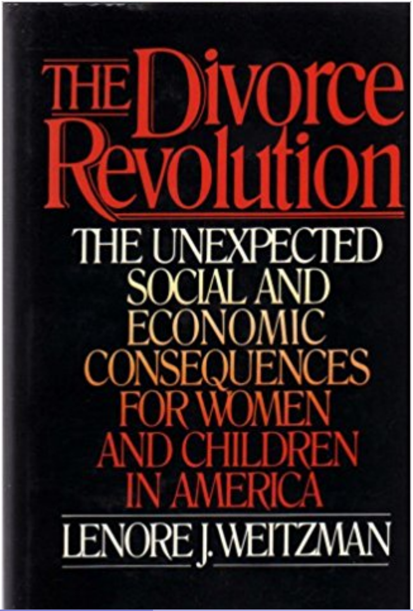
Why Replicability?

- Changes in living standard after divorce
  - for women declines 73%
  - for men increases 42%
- ASA Book Award in 1986
- Between 1986 and 1993, cited in 348 social science articles and 250 law review articles
- Between 1986 and 1993, cited in 24 legal cases and by the Supreme Court



Why Replicability?

- Changes in living standard after divorce
  - for women declines 73%
  - for men increases 42%
- ASA Book Award in 1986
- Between 1986 and 1993, cited in 348 social science articles and 250 law review articles
- Between 1986 and 1993, cited in 24 legal cases and by the Supreme Court
- Led to changes in divorce law in California



Revisited

A RE-EVALUATION OF THE ECONOMIC CONSEQUENCES OF DIVORCE*

Richard R. Peterson
Social Science Research Council

Over the last 20 years, researchers have focused considerable attention on the economic consequences of divorce. One book, Weitzman's The Divorce Revolution *(1985), reports a 73 percent decline in women's standard of living after divorce and a 42 percent increase in men's standard of living. These percentages, based on data from a 1977–1978 Los Angeles sample, are substantially larger than those from other studies. I replicate* The Divorce Revolution*'s analysis and demonstrate that the estimates reported in the book are inaccurate. This reanalysis, which uses the same sample and measures of economic well-being as* The Divorce Revolution, *produces estimates of a 27 percent decline in women's standard of living and a 10 percent increase in men's standard of living after divorce. I discuss the implications of these results for debates about divorce law reform.*

Revisited

"First, let me begin with Peterson's implied question: Was this responsible research and did I meet professional standards in analyzing these data?" (Weitzman, 1996)

Revisited

"…Changes to the original raw data file resulting from this data cleaning process were made by a series of programming statements on a master SPSS system file. *The raw data file that is stored at the Murray Center is the original 'dirty data' file and does not include these cleaning changes…*" (Weitzman, 1996)

Revisited

"Unfortunately, the original cleaned master SPSS system file no longer exists. I assumed it was being copied and reformatted as I moved for job changes and fellowships from the project's original offices in Berkeley to Stanford (in 1979), then to Princeton (in 1983), back to Stanford (in 1984) and then to Harvard (in 1986). With each move, new programmers worked on the files to accommodate different computer systems." (Weitzman,1996)

Revisited

"Before I left Stanford I instructed my programmers to prepare all my data files for archiving. I know now (but did not know then) that the original master SPSS system file that I used for my book had been lost or damaged at some point and was not included among these files. The SPSS system file that I thought was the master SPSS system file was the result of the merging of many smaller subfiles that had been created for specific analyses. It later became apparent that a programming error had been made, and the subfiles were not "keyed" correctly: Not all of the data from each individual respondent were matched on the appropriate case ID number, and data from different respondents were merged under the same case ID. At present it is not possible to disentangle exactly what mismatch occurred for any specific respondent." (Weitzman, 1996)

Revisited

"When I could not replicate the analyses in my book with what I had mistakenly assumed was the archived master SPSS system file, I hired an independent consultant, Professor Angela Aidala from Columbia University, to help me untangle what had happened. She reviewed all of the project files, documentation, and codebooks, as well as the available data and programming files to determine a possible computational error in the standard of living statistic. But she could not do this without an accurate data file to work with. We then went back to the original questionnaires and recoded a random sample of about 25 percent of the cases. There were so many discrepancies between the questionnaires and the 'dirty data' raw data file, and between the questionnaires and the mismatched SPSS system file, that we finally abandoned the effort and left a warning to all future researchers *that both files at the Murray Center were so seriously flawed that they could not be used*. It was a very sad, time consuming, and frustrating experience."

Revisited

Here's a good rule of thumb: If you are trying to solve a problem, and there are multi-billion dollar firms whose entire business model depends on the solving the same problem, you might want to figure out what the experts do and see if you can't learn something from it. (Gentkow and Shapiro 2014)

How to get started?

- Start by reading "Publication, Publication"
  `gking.harvard.edu/files/gking/files/paperspub.pdf`

How to get started?

- Start by reading "Publication, Publication"
  `gking.harvard.edu/files/gking/files/paperspub.pdf`
- Peruse the additional notes at `http://gking.harvard.edu/papers`

How to get started?

- Start by reading "Publication, Publication"
  `gking.harvard.edu/files/gking/files/paperspub.pdf`
- Peruse the additional notes at `http://gking.harvard.edu/papers`
- Find a partner

How to get started?

- Start by reading "Publication, Publication"
  `gking.harvard.edu/files/gking/files/paperspub.pdf`
- Peruse the additional notes at `http://gking.harvard.edu/papers`
- Find a partner
- Start looking for data!

Additional Things to Do

- Signup for Perusall
- Readings for Next Monday: pg 6-58 of *UPM*

Statistical Models: Variable Definitions

Statistical Models: Variable Definitions

- Dependent (or "outcome") variable

Statistical Models: Variable Definitions

- Dependent (or "outcome") variable
  - Y is $n \times 1$.

Statistical Models: Variable Definitions

- Dependent (or "outcome") variable
  - Y is $n \times 1$.
  - y_i, a number (after we know it)

Statistical Models: Variable Definitions

- Dependent (or "outcome") variable
  - Y is $n \times 1$.
  - y_i, a number (after we know it)
  - Y_i, a random variable (before we know it)

Statistical Models: Variable Definitions

- Dependent (or "outcome") variable
  - Y is $n \times 1$.
  - y_i, a number (after we know it)
  - Y_i, a random variable (before we know it)
  - Commonly misunderstood: a "dependent variable" can be

Statistical Models: Variable Definitions

- Dependent (or "outcome") variable
  - Y is $n \times 1$.
  - y_i, a number (after we know it)
  - Y_i, a random variable (before we know it)
  - Commonly misunderstood: a "dependent variable" can be
    - a column of numbers in your data set

Statistical Models: Variable Definitions

- Dependent (or "outcome") variable
  - Y is $n \times 1$.
  - y_i, a number (after we know it)
  - Y_i, a random variable (before we know it)
  - Commonly misunderstood: a "dependent variable" can be
    - a column of numbers in your data set
    - the random variable for each unit i.

Statistical Models: Variable Definitions

- Dependent (or "outcome") variable
  - Y is $n \times 1$.
  - y_i, a number (after we know it)
  - Y_i, a random variable (before we know it)
  - Commonly misunderstood: a "dependent variable" can be
    - a column of numbers in your data set
    - the random variable for each unit i.
- Explanatory variables

Statistical Models: Variable Definitions

- Dependent (or "outcome") variable
  - Y is $n \times 1$.
  - y_i, a number (after we know it)
  - Y_i, a random variable (before we know it)
  - Commonly misunderstood: a "dependent variable" can be
    - ★ a column of numbers in your data set
    - ★ the random variable for each unit i.
- Explanatory variables
  - aka "covariates," "independent," or "exogenous" variables

Statistical Models: Variable Definitions

- Dependent (or "outcome") variable
  - Y is $n \times 1$.
  - y_i, a number (after we know it)
  - Y_i, a random variable (before we know it)
  - Commonly misunderstood: a "dependent variable" can be
    - a column of numbers in your data set
    - the random variable for each unit i.
- Explanatory variables
  - aka "covariates," "independent," or "exogenous" variables
  - $X = \{x_{ij}\}$ is $n \times k$ (observations by variables)

Statistical Models: Variable Definitions

- Dependent (or "outcome") variable
  - Y is $n \times 1$.
  - y_i, a number (after we know it)
  - Y_i, a random variable (before we know it)
  - Commonly misunderstood: a "dependent variable" can be
    - ★ a column of numbers in your data set
    - ★ the random variable for each unit i.
- Explanatory variables
  - aka "covariates," "independent," or "exogenous" variables
  - $X = \{x_{ij}\}$ is $n \times k$ (observations by variables)
  - A set of columns (variables): $X = \{x_1 \ldots, x_k\}$

Statistical Models: Variable Definitions

- Dependent (or "outcome") variable
  - Y is $n \times 1$.
  - y_i, a number (after we know it)
  - Y_i, a random variable (before we know it)
  - Commonly misunderstood: a "dependent variable" can be
    - ★ a column of numbers in your data set
    - ★ the random variable for each unit i.
- Explanatory variables
  - aka "covariates," "independent," or "exogenous" variables
  - $X = \{x_{ij}\}$ is $n \times k$ (observations by variables)
  - A set of columns (variables): $X = \{x_1 \ldots, x_k\}$
  - Row (observation) i: $x_i = \{x_{i1}, \ldots, x_{ik}\}$

Statistical Models: Variable Definitions

- Dependent (or "outcome") variable
  - Y is $n \times 1$.
  - y_i, a number (after we know it)
  - Y_i, a random variable (before we know it)
  - Commonly misunderstood: a "dependent variable" can be
    - ★ a column of numbers in your data set
    - ★ the random variable for each unit i.
- Explanatory variables
  - aka "covariates," "independent," or "exogenous" variables
  - $X = \{x_{ij}\}$ is $n \times k$ (observations by variables)
  - A set of columns (variables): $X = \{x_1 \ldots, x_k\}$
  - Row (observation) i: $x_i = \{x_{i1}, \ldots, x_{ik}\}$
  - X is fixed (not random).

Equivalent Linear Regression Notation

Equivalent Linear Regression Notation

- Standard version (last semester)

Equivalent Linear Regression Notation

- Standard version (last semester)

$$
\begin{aligned}
Y_i &= x_i\beta + \epsilon_i \qquad \text{systematic} + \text{stochastic} \\
\epsilon_i &\sim f_N(0, \sigma^2)
\end{aligned}
$$

Equivalent Linear Regression Notation

- Standard version (last semester)

$$
\begin{aligned}
Y_i &= x_i\beta + \epsilon_i \qquad \text{systematic} + \text{stochastic} \\
\epsilon_i &\sim f_N(0, \sigma^2)
\end{aligned}
$$

Equivalent Linear Regression Notation

- Standard version (last semester)

$$
\begin{aligned}
Y_i &= x_i\beta + \epsilon_i \qquad \text{systematic} + \text{stochastic} \\
\epsilon_i &\sim f_N(0, \sigma^2)
\end{aligned}
$$

Equivalent Linear Regression Notation

- Standard version (last semester)

$$
\begin{aligned}
Y_i &= x_i\beta + \epsilon_i && \text{systematic} + \text{stochastic} \\
\epsilon_i &\sim f_N(0, \sigma^2)
\end{aligned}
$$

- Alternative version

$$
\begin{aligned}
Y_i &\sim f_N(\mu_i, \sigma^2) && \text{stochastic} \\
\mu_i &= x_i\beta && \text{systematic}
\end{aligned}
$$

Equivalent Linear Regression Notation

- Standard version (last semester)

$$\begin{aligned} Y_i &= x_i\beta + \epsilon_i && \text{systematic} + \text{stochastic} \\ \epsilon_i &\sim f_N(0, \sigma^2) \end{aligned}$$

- Alternative version

$$\begin{aligned} Y_i &\sim f_N(\mu_i, \sigma^2) && \text{stochastic} \\ \mu_i &= x_i\beta && \text{systematic} \end{aligned}$$

Equivalent Linear Regression Notation

- Standard version (last semester)

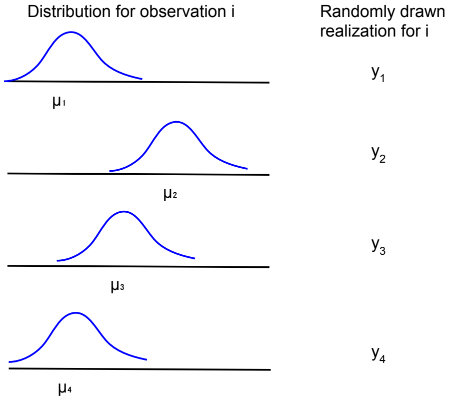
$$
\begin{aligned}
Y_i &= x_i\beta + \epsilon_i && \text{systematic} + \text{stochastic}\\
\epsilon_i &\sim f_N(0, \sigma^2)
\end{aligned}
$$

- Alternative version

$$
\begin{aligned}
Y_i &\sim f_N(\mu_i, \sigma^2) && \text{stochastic}\\
\mu_i &= x_i\beta && \text{systematic}
\end{aligned}
$$

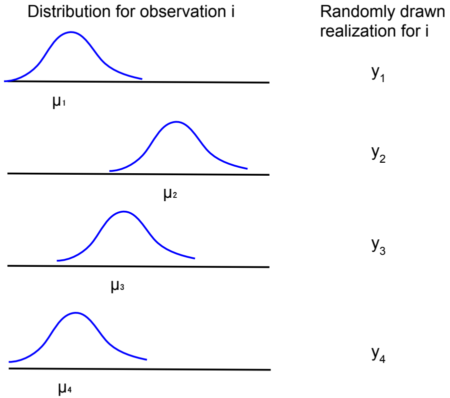
Understanding the Alternative Regression Notation

Understanding the Alternative Regression Notation



where $\mu_i = X_i\beta$.

Understanding the Alternative Regression Notation



where $\mu_i = X_i\beta$.

A Test: Is a histogram of y a test of normality?

Generalized Alternative Notation

Generalized Alternative Notation

$Y_i \sim f(\theta_i, \alpha)$ stochastic

Generalized Alternative Notation

$$Y_i \sim f(\theta_i, \alpha) \qquad \text{stochastic}$$
$$\theta_i = g(X_i, \beta) \qquad \text{systematic}$$

Generalized Alternative Notation

$$\begin{aligned} Y_i &\sim f(\theta_i, \alpha) && \text{stochastic} \\ \theta_i &= g(X_i, \beta) && \text{systematic} \end{aligned}$$

where

Generalized Alternative Notation

$$Y_i \sim f(\theta_i, \alpha) \qquad \text{stochastic}$$
$$\theta_i = g(X_i, \beta) \qquad \text{systematic}$$

where

Y_i random outcome variable

Generalized Alternative Notation

$$
\begin{aligned}
Y_i &\sim f(\theta_i, \alpha) && \text{stochastic} \\
\theta_i &= g(X_i, \beta) && \text{systematic}
\end{aligned}
$$

where

Y_i random outcome variable

$f(\cdot)$ probability density

Generalized Alternative Notation

$$\begin{aligned} Y_i &\sim f(\theta_i, \alpha) && \text{stochastic} \\ \theta_i &= g(X_i, \beta) && \text{systematic} \end{aligned}$$

where

Y_i random outcome variable

$f(\cdot)$ probability density

θ_i a systematic feature of the density that varies over i

Generalized Alternative Notation

$$
\begin{aligned}
Y_i &\sim f(\theta_i, \alpha) & \text{stochastic} \\
\theta_i &= g(X_i, \beta) & \text{systematic}
\end{aligned}
$$

where

Y_i random outcome variable

$f(\cdot)$ probability density

θ_i a systematic feature of the density that varies over i

α ancillary parameter (feature of the density constant over i)

Generalized Alternative Notation

$$\begin{aligned} Y_i &\sim f(\theta_i, \alpha) && \text{stochastic} \\ \theta_i &= g(X_i, \beta) && \text{systematic} \end{aligned}$$

where

Y_i random outcome variable

$f(\cdot)$ probability density

θ_i a systematic feature of the density that varies over i

α ancillary parameter (feature of the density constant over i)

$g(\cdot)$ functional form

Generalized Alternative Notation

$$Y_i \sim f(\theta_i, \alpha) \qquad \text{stochastic}$$
$$\theta_i = g(X_i, \beta) \qquad \text{systematic}$$

where

Y_i random outcome variable

$f(\cdot)$ probability density

θ_i a systematic feature of the density that varies over i

α ancillary parameter (feature of the density constant over i)

$g(\cdot)$ functional form

X_i explanatory variables

Generalized Alternative Notation

$$
\begin{aligned}
Y_i &\sim f(\theta_i, \alpha) && \text{stochastic} \\
\theta_i &= g(X_i, \beta) && \text{systematic}
\end{aligned}
$$

where

Y_i random outcome variable

$f(\cdot)$ probability density

θ_i a systematic feature of the density that varies over i

α ancillary parameter (feature of the density constant over i)

$g(\cdot)$ functional form

X_i explanatory variables

β effect parameters

Forms of Uncertainty

Forms of Uncertainty

$Y_i \sim f(\theta_i, \alpha)$ stochastic

Forms of Uncertainty

$$Y_i \sim f(\theta_i, \alpha) \qquad \text{stochastic}$$

$$\theta_i = g(X_i, \beta) \qquad \text{systematic}$$

Forms of Uncertainty

$$Y_i \sim f(\theta_i, \alpha) \qquad \text{stochastic}$$
$$\theta_i = g(X_i, \beta) \qquad \text{systematic}$$

- Estimation uncertainty: Lack of knowledge of β and α. Vanishes as n gets larger.

Forms of Uncertainty

$$Y_i \sim f(\theta_i, \alpha) \qquad \text{stochastic}$$
$$\theta_i = g(X_i, \beta) \qquad \text{systematic}$$

- Estimation uncertainty: Lack of knowledge of β and α. Vanishes as n gets larger.
- Fundamental uncertainty: Represented by the stochastic component. Exists no matter what the researcher does; no matter how large n is.

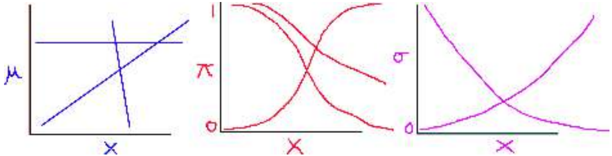
Forms of Uncertainty

$$
\begin{aligned}
Y_i &\sim f(\theta_i, \alpha) && \text{stochastic} \\
\theta_i &= g(X_i, \beta) && \text{systematic}
\end{aligned}
$$

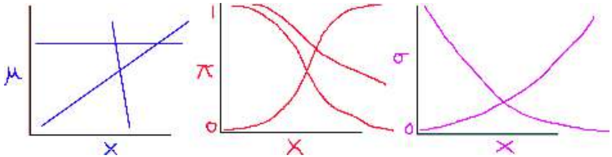
- Estimation uncertainty: Lack of knowledge of β and α. Vanishes as n gets larger.
- Fundamental uncertainty: Represented by the stochastic component. Exists no matter what the researcher does; no matter how large n is.
- (A Test: If you know the model, is $R^2 = 1$?)

Systematic Components: Examples

Systematic Components: Examples

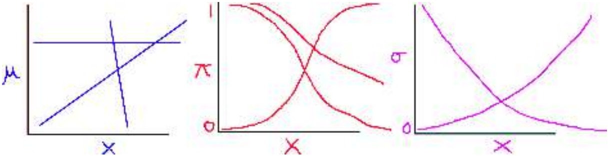


Systematic Components: Examples



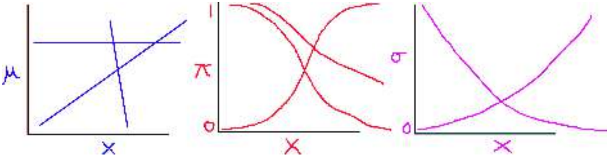
- $E(Y_i) \equiv \mu_i = X_i\beta = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}$

Systematic Components: Examples



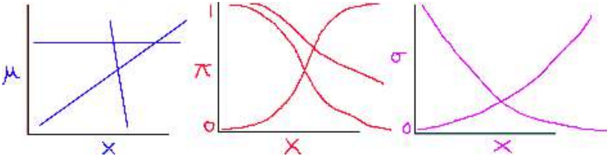
- $E(Y_i) \equiv \mu_i = X_i\beta = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}$
- $\Pr(Y_i = 1) \equiv \pi_i = \frac{1}{1+e^{-x_i\beta}}$

Systematic Components: Examples



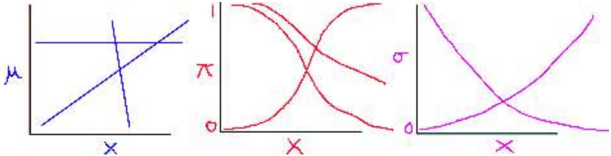
- $E(Y_i) \equiv \mu_i = X_i\beta = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}$
- $\Pr(Y_i = 1) \equiv \pi_i = \frac{1}{1+e^{-x_i\beta}}$
- $V(Y_i) \equiv \sigma_i^2 = e^{x_i\beta}$

Systematic Components: Examples



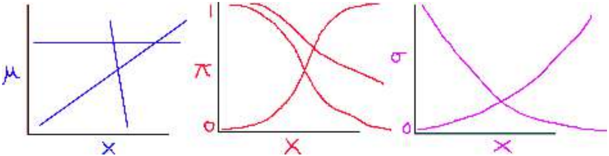
- $E(Y_i) \equiv \mu_i = X_i\beta = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}$
- $\Pr(Y_i = 1) \equiv \pi_i = \frac{1}{1+e^{-x_i\beta}}$
- $V(Y_i) \equiv \sigma_i^2 = e^{x_i\beta}$
- Interpretation:

Systematic Components: Examples



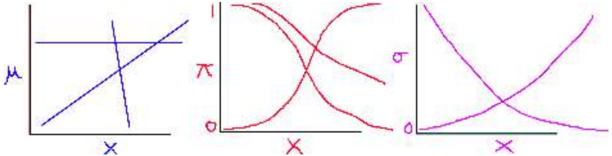
- $E(Y_i) \equiv \mu_i = X_i\beta = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}$
- $\Pr(Y_i = 1) \equiv \pi_i = \frac{1}{1+e^{-x_i\beta}}$
- $V(Y_i) \equiv \sigma_i^2 = e^{x_i\beta}$
- Interpretation:
  - Each is a class of functional forms

Systematic Components: Examples



- $E(Y_i) \equiv \mu_i = X_i\beta = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}$
- $\Pr(Y_i = 1) \equiv \pi_i = \frac{1}{1+e^{-x_i\beta}}$
- $V(Y_i) \equiv \sigma_i^2 = e^{x_i\beta}$
- Interpretation:
  - Each is a class of functional forms
  - Set β and it picks out one member of the class

Systematic Components: Examples



- $E(Y_i) \equiv \mu_i = X_i\beta = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}$
- $\Pr(Y_i = 1) \equiv \pi_i = \frac{1}{1+e^{-x_i\beta}}$
- $V(Y_i) \equiv \sigma_i^2 = e^{x_i\beta}$
- Interpretation:
  - Each is a class of functional forms
  - Set β and it picks out one member of the class
  - β in each is an "effect parameter" vector, with different meaning

Systematic Components: Examples

Systematic Components: Examples

- We will:

Systematic Components: Examples

- We will:
  - Assume a class of functional forms (called link functions) from theory (each form is flexible and maps out many potential relationships)

Systematic Components: Examples

- We will:
  - Assume a class of functional forms (called link functions) from theory (each form is flexible and maps out many potential relationships)
  - Estimate β from data to choose a member of the class

Systematic Components: Examples

- We will:
  - Assume a class of functional forms (called link functions) from theory (each form is flexible and maps out many potential relationships)
  - Estimate β from data to choose a member of the class
  - Be uncertain (Because of sampling, measurement, & fundamental uncertainties) about:

Systematic Components: Examples

- We will:
  - Assume a class of functional forms (called link functions) from theory (each form is flexible and maps out many potential relationships)
  - Estimate β from data to choose a member of the class
  - Be uncertain (Because of sampling, measurement, & fundamental uncertainties) about:
    - the family $\rightsquigarrow$ model dependence

Systematic Components: Examples

- We will:
  - Assume a class of functional forms (called link functions) from theory (each form is flexible and maps out many potential relationships)
  - Estimate β from data to choose a member of the class
  - Be uncertain (Because of sampling, measurement, & fundamental uncertainties) about:
    - the family $\rightsquigarrow$ model dependence
    - the member of the chosen family $\rightsquigarrow$ sampling error

Systematic Components: Examples

- We will:
  - Assume a class of functional forms (called link functions) from theory (each form is flexible and maps out many potential relationships)
  - Estimate β from data to choose a member of the class
  - Be uncertain (Because of sampling, measurement, & fundamental uncertainties) about:
    - the family $\rightsquigarrow$ model dependence
    - the member of the chosen family $\rightsquigarrow$ sampling error
- If we choose the family of functional forms wrong, we:

Systematic Components: Examples

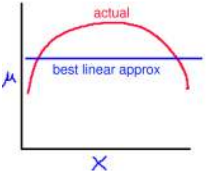
- We will:
  - Assume a class of functional forms (called link functions) from theory (each form is flexible and maps out many potential relationships)
  - Estimate β from data to choose a member of the class
  - Be uncertain (Because of sampling, measurement, & fundamental uncertainties) about:
    - the family $\rightsquigarrow$ model dependence
    - the member of the chosen family $\rightsquigarrow$ sampling error
- If we choose the family of functional forms wrong, we:
  - Have specification error, and potentially bias

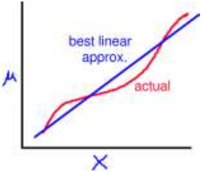
Systematic Components: Examples

- We will:
  - Assume a class of functional forms (called link functions) from theory (each form is flexible and maps out many potential relationships)
  - Estimate β from data to choose a member of the class
  - Be uncertain (Because of sampling, measurement, & fundamental uncertainties) about:
    - the family $\rightsquigarrow$ model dependence
    - the member of the chosen family $\rightsquigarrow$ sampling error
- If we choose the family of functional forms wrong, we:
  - Have specification error, and potentially bias
  - Still get the best [linear,logit,etc] approximation to the correct functional form.

Systematic Components: Examples

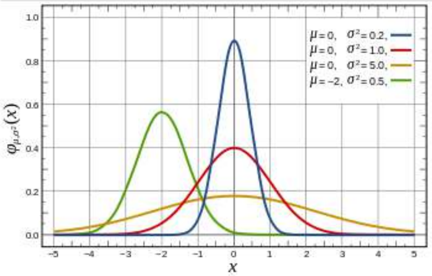
- We will:
  - Assume a class of functional forms (called link functions) from theory (each form is flexible and maps out many potential relationships)
  - Estimate β from data to choose a member of the class
  - Be uncertain (Because of sampling, measurement, & fundamental uncertainties) about:
    - ⋆ the family $\rightsquigarrow$ model dependence
    - ⋆ the member of the chosen family $\rightsquigarrow$ sampling error
- If we choose the family of functional forms wrong, we:
  - Have specification error, and potentially bias
  - Still get the best [linear,logit,etc] approximation to the correct functional form.
  - May be close or far from the truth:





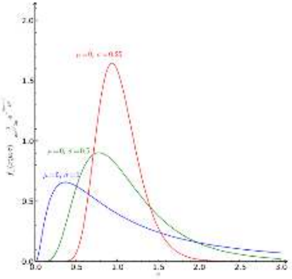
Overview of Stochastic Components

Overview of Stochastic Components



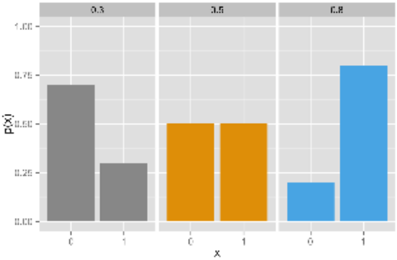
- Normal — continuous, unimodal, symmetric, unbounded

Overview of Stochastic Components



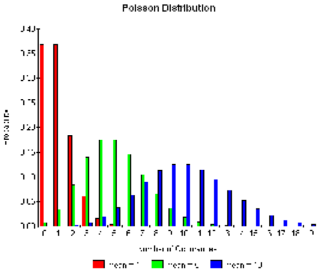
- Normal — continuous, unimodal, symmetric, unbounded
- Log-normal — continuous, unimodal, skewed, bounded from below by zero

Overview of Stochastic Components



- Normal — continuous, unimodal, symmetric, unbounded
- Log-normal — continuous, unimodal, skewed, bounded from below by zero
- Bernoulli — discrete, binary outcomes

Overview of Stochastic Components



- Normal — continuous, unimodal, symmetric, unbounded
- Log-normal — continuous, unimodal, skewed, bounded from below by zero
- Bernoulli — discrete, binary outcomes
- Poisson — discrete, countably infinite on the nonnegative integers

Choosing systematic and stochastic components

Choosing systematic and stochastic components

- If one is bounded, so is the other

Choosing systematic and stochastic components

- If one is bounded, so is the other
- If the stochastic component is bounded, the systematic component must be globally nonlinear (though possibly locally linear)

Choosing systematic and stochastic components

- If one is bounded, so is the other
- If the stochastic component is bounded, the systematic component must be globally nonlinear (though possibly locally linear)
- All modeling decisions are about the data generation process — how the information made its way from the world (including how the world produced the data) to your data set

Choosing systematic and stochastic components

- If one is bounded, so is the other
- If the stochastic component is bounded, the systematic component must be globally nonlinear (though possibly locally linear)
- All modeling decisions are about the data generation process — how the information made its way from the world (including how the world produced the data) to your data set
- What if we don't know the DGP (& we usually don't)?

Choosing systematic and stochastic components

- If one is bounded, so is the other
- If the stochastic component is bounded, the systematic component must be globally nonlinear (though possibly locally linear)
- All modeling decisions are about the data generation process — how the information made its way from the world (including how the world produced the data) to your data set
- What if we don't know the DGP (& we usually don't)?
  - The problem: model dependence

Choosing systematic and stochastic components

- If one is bounded, so is the other
- If the stochastic component is bounded, the systematic component must be globally nonlinear (though possibly locally linear)
- All modeling decisions are about the data generation process — how the information made its way from the world (including how the world produced the data) to your data set
- What if we don't know the DGP (& we usually don't)?
  - The problem: model dependence
  - Our first approach: make "reasonable" assumptions and check fit (& other observable implications of the assumptions)

Choosing systematic and stochastic components

- If one is bounded, so is the other
- If the stochastic component is bounded, the systematic component must be globally nonlinear (though possibly locally linear)
- All modeling decisions are about the data generation process — how the information made its way from the world (including how the world produced the data) to your data set
- What if we don't know the DGP (& we usually don't)?
  - The problem: model dependence
  - Our first approach: make "reasonable" assumptions and check fit (& other observable implications of the assumptions)
  - Later:

Choosing systematic and stochastic components

- If one is bounded, so is the other
- If the stochastic component is bounded, the systematic component must be globally nonlinear (though possibly locally linear)
- All modeling decisions are about the data generation process — how the information made its way from the world (including how the world produced the data) to your data set
- What if we don't know the DGP (& we usually don't)?
  - The problem: model dependence
  - Our first approach: make "reasonable" assumptions and check fit (& other observable implications of the assumptions)
  - Later:
    - Avoid it: relax assumptions

Choosing systematic and stochastic components

- If one is bounded, so is the other
- If the stochastic component is bounded, the systematic component must be globally nonlinear (though possibly locally linear)
- All modeling decisions are about the data generation process — how the information made its way from the world (including how the world produced the data) to your data set
- What if we don't know the DGP (& we usually don't)?
  - The problem: model dependence
  - Our first approach: make "reasonable" assumptions and check fit (& other observable implications of the assumptions)
  - Later:
    - Avoid it: relax assumptions
    - Detect remaining model dependence

Choosing systematic and stochastic components

- If one is bounded, so is the other
- If the stochastic component is bounded, the systematic component must be globally nonlinear (though possibly locally linear)
- All modeling decisions are about the data generation process — how the information made its way from the world (including how the world produced the data) to your data set
- What if we don't know the DGP (& we usually don't)?
  - The problem: model dependence
  - Our first approach: make "reasonable" assumptions and check fit (& other observable implications of the assumptions)
  - Later:
    - Avoid it: relax assumptions
    - Detect remaining model dependence
    - Remove model dependence: preprocess data

Choosing systematic and stochastic components

- If one is bounded, so is the other
- If the stochastic component is bounded, the systematic component must be globally nonlinear (though possibly locally linear)
- All modeling decisions are about the data generation process — how the information made its way from the world (including how the world produced the data) to your data set
- What if we don't know the DGP (& we usually don't)?
  - The problem: model dependence
  - Our first approach: make "reasonable" assumptions and check fit (& other observable implications of the assumptions)
  - Later:
    - ⋆ Avoid it: relax assumptions
    - ⋆ Detect remaining model dependence
    - ⋆ Remove model dependence: preprocess data
    - ⋆ Characterize behavior under model mispecification

Preview: Generalized Linear Models

Preview: Generalized Linear Models

- The models we will learn this semester are created from a choice of stochastic and systematic components.

Preview: Generalized Linear Models

- The models we will learn this semester are created from a choice of stochastic and systematic components.
- That is Y_i is a (stochastic) function of a systematic component parameterized by a linear predictor, $X_i^\top \beta$

Preview: Generalized Linear Models

- The models we will learn this semester are created from a choice of stochastic and systematic components.
- That is Y_i is a (stochastic) function of a systematic component parameterized by a linear predictor, $X_i^\top \beta$
- They share many nice properties that let us map insights from regression onto new models.

Preview: Generalized Linear Models

- The models we will learn this semester are created from a choice of stochastic and systematic components.
- That is Y_i is a (stochastic) function of a systematic component parameterized by a linear predictor, $X_i^\top \beta$
- They share many nice properties that let us map insights from regression onto new models.
- 3 components of a GLM

Preview: Generalized Linear Models

- The models we will learn this semester are created from a choice of stochastic and systematic components.
- That is Y_i is a (stochastic) function of a systematic component parameterized by a linear predictor, $X_i^\top \beta$
- They share many nice properties that let us map insights from regression onto new models.
- 3 components of a GLM
    1. Systematic component: $X_i^\top \beta$
        - ★ Must be a linear function of X_i

Preview: Generalized Linear Models

- The models we will learn this semester are created from a choice of stochastic and systematic components.
- That is Y_i is a (stochastic) function of a systematic component parameterized by a linear predictor, $X_i^\top \beta$
- They share many nice properties that let us map insights from regression onto new models.
- 3 components of a GLM
    1. Systematic component: $X_i^\top \beta$
        - Must be a linear function of X_i

Preview: Generalized Linear Models

- The models we will learn this semester are created from a choice of stochastic and systematic components.
- That is Y_i is a (stochastic) function of a systematic component parameterized by a linear predictor, $X_i^\top \beta$
- They share many nice properties that let us map insights from regression onto new models.
- 3 components of a GLM
  1. Systematic component: $X_i^\top \beta$
     - ★ Must be a linear function of X_i
  2. Random component: $f(Y; \theta, \alpha)$
     - ★ θ is called the canonical parameter
     - ★ α: is called the dispersion parameter

Preview: Generalized Linear Models

- The models we will learn this semester are created from a choice of stochastic and systematic components.
- That is Y_i is a (stochastic) function of a systematic component parameterized by a linear predictor, $X_i^\top \beta$
- They share many nice properties that let us map insights from regression onto new models.
- 3 components of a GLM
  1. Systematic component: $X_i^\top \beta$
     - ★ Must be a linear function of X_i
  2. Random component: $f(Y; \theta, \alpha)$
     - ★ θ is called the canonical parameter
     - ★ α: is called the dispersion parameter
  3. Link function: $g(\mu_i) = X_i^\top \beta$ where $\mu_i = \mathbb{E}(Y_i \mid X_i)$
     - ★ Must be monotonic and differentiable wrt μ_i

One Step Deeper: Exponential Family

Nearly every distribution we will discuss is in the exponential family. An exponential family distribution has the density of the following form:

$$f_Y(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$$

One Step Deeper: Exponential Family

Nearly every distribution we will discuss is in the exponential family. An exponential family distribution has the density of the following form:

$$f_Y(y; \theta, \phi) = \exp\left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\}$$

Example: Poisson(μ):

$$\Pr(Y_i = y \mid \mu) \; = \; \exp\{y \log \mu - \exp(\log \mu) - \log y!\}$$

One Step Deeper: Exponential Family

Nearly every distribution we will discuss is in the exponential family. An exponential family distribution has the density of the following form:

$$f_Y(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$$

Example: Poisson(μ):

$$\Pr(Y_i = y \mid \mu) \;=\; \exp\left\{y \log\mu - \exp(\log\mu) - \log y!\right\}$$

One Step Deeper: Exponential Family

Nearly every distribution we will discuss is in the exponential family. An exponential family distribution has the density of the following form:

$$f_Y(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$$

Example: Poisson(μ):

$$\Pr(Y_i = y \mid \mu) \;=\; \exp\{y\log\mu - \exp(\log\mu) - \log y!\}$$

One Step Deeper: Exponential Family

Nearly every distribution we will discuss is in the exponential family. An exponential family distribution has the density of the following form:

$$f_Y(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$$

Example: Poisson(μ):

$$\Pr(Y_i = y \mid \mu) \;=\; \exp\{y \log\mu - \exp(\log\mu) - \log y!\}$$

$\Longrightarrow \theta = \log\mu$, $\phi = 1$, $a(\phi) = \phi, b(\theta) = \exp(\theta)$, and $c = -\log y!$

One Step Deeper: Exponential Family

Nearly every distribution we will discuss is in the exponential family. An exponential family distribution has the density of the following form:

$$f_Y(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$$

Example: Poisson(μ):

$$\Pr(Y_i = y \mid \mu) \ = \ \exp\{y \log \mu - \exp(\log \mu) - \log y!\}$$

$\Longrightarrow \theta = \log \mu$, $\phi = 1$, $a(\phi) = \phi, b(\theta) = \exp(\theta)$, and $c = -\log y!$

Many other examples, including: Normal, Bernoulli/binomial, Gamma, multinomial, exponential, negative binomial, beta, uniform, chi-squared, etc.

One Step Deeper: Properties of the Exponential Family

- Mean is a function of θ and given by

$$\mathbb{E}(Y) \equiv \mu = b'(\theta)$$

One Step Deeper: Properties of the Exponential Family

- Mean is a function of θ and given by

$$\mathbb{E}(Y) \equiv \mu = b'(\theta)$$

- Variance is a function of θ and ϕ and given by

$$\mathbb{V}(Y) \equiv V = b''(\theta)a(\phi)$$

One Step Deeper: Properties of the Exponential Family

- Mean is a function of θ and given by

$$\mathbb{E}(Y) \equiv \mu = b'(\theta)$$

- Variance is a function of θ and ϕ and given by

$$\mathbb{V}(Y) \equiv V = b''(\theta)a(\phi)$$

- Common forms of $a(\phi)$: 1 (Poisson, Bernoulli), ϕ (normal, Gamma), and ϕ/ω_i (binomial)

One Step Deeper: Properties of the Exponential Family

- Mean is a function of θ and given by

$$\mathbb{E}(Y) \equiv \mu = b'(\theta)$$

- Variance is a function of θ and ϕ and given by

$$\mathbb{V}(Y) \equiv V = b''(\theta)a(\phi)$$

- Common forms of $a(\phi)$: 1 (Poisson, Bernoulli), ϕ (normal, Gamma), and ϕ/ω_i (binomial)
- $b''(\theta)$ is called the variance function

One Step Deeper: Properties of the Exponential Family

- Mean is a function of θ and given by

$$\mathbb{E}(Y) \equiv \mu = b'(\theta)$$

- Variance is a function of θ and ϕ and given by

$$\mathbb{V}(Y) \equiv V = b''(\theta)a(\phi)$$

- Common forms of $a(\phi)$: 1 (Poisson, Bernoulli), ϕ (normal, Gamma), and ϕ/ω_i (binomial)
- $b''(\theta)$ is called the variance function
- In the Poisson model, $\theta_i = \log \mu_i$, $a(\phi) = 1$ and $b(\theta_i) = \exp(\theta_i)$

One Step Deeper: Properties of the Exponential Family

- Mean is a function of θ and given by

$$\mathbb{E}(Y) \equiv \mu = b'(\theta)$$

- Variance is a function of θ and ϕ and given by

$$\mathbb{V}(Y) \equiv V = b''(\theta)a(\phi)$$

- Common forms of $a(\phi)$: 1 (Poisson, Bernoulli), ϕ (normal, Gamma), and ϕ/ω_i (binomial)
- $b''(\theta)$ is called the variance function
- In the Poisson model, $\theta_i = \log \mu_i$, $a(\phi) = 1$ and $b(\theta_i) = \exp(\theta_i)$
  $\Rightarrow \mathbb{E}(Y_i) = \frac{db(\theta_i)}{d\theta_i} = \exp(\theta_i) = \mu_i$ and $\mathbb{V}(Y_i) = \frac{d^2 b(\theta_i)}{d\theta_i^2} = \exp(\theta_i) = \mu_i$

One Step Deeper: Link Functions

- Link function: $g(\mu_i) = X_i^\top \beta$
- Defines the relationship between $X_i^\top \beta$ and the mean μ_i
- Must map the real line onto the possible range of μ_i

One Step Deeper: Link Functions

- Link function: $g(\mu_i) = X_i^\top \beta$
- Defines the relationship between $X_i^\top \beta$ and the mean μ_i
- Must map the real line onto the possible range of μ_i

- Recall that $\mu_i = b'(\theta_i)$
- Therefore, θ_i is always a (often simple) function of $X_i^\top \beta$
- In particular, when $\theta_i = X_i^\top \beta$, the link is called the canonical link

One Step Deeper: Link Functions

- Link function: $g(\mu_i) = X_i^\top \beta$
- Defines the relationship between $X_i^\top \beta$ and the mean μ_i
- Must map the real line onto the possible range of μ_i
- Recall that $\mu_i = b'(\theta_i)$
- Therefore, θ_i is always a (often simple) function of $X_i^\top \beta$
- In particular, when $\theta_i = X_i^\top \beta$, the link is called the canonical link
  - In Poisson, $\theta_i = \log(\mu_i) =$

One Step Deeper: Link Functions

- Link function: $g(\mu_i) = X_i^\top \beta$
- Defines the relationship between $X_i^\top \beta$ and the mean μ_i
- Must map the real line onto the possible range of μ_i

- Recall that $\mu_i = b'(\theta_i)$
- Therefore, θ_i is always a (often simple) function of $X_i^\top \beta$
- In particular, when $\theta_i = X_i^\top \beta$, the link is called the canonical link
  - In Poisson, $\theta_i = \log(\mu_i) = \log(\exp(X_i^\top \beta)) = X_i^\top \beta$

One Step Deeper: Link Functions

- Link function: $g(\mu_i) = X_i^\top \beta$
- Defines the relationship between $X_i^\top \beta$ and the mean μ_i
- Must map the real line onto the possible range of μ_i

- Recall that $\mu_i = b'(\theta_i)$
- Therefore, θ_i is always a (often simple) function of $X_i^\top \beta$
- In particular, when $\theta_i = X_i^\top \beta$, the link is called the canonical link
  - In Poisson, $\theta_i = \log(\mu_i) = \log(\exp(X_i^\top \beta)) = X_i^\top \beta$
    $\longrightarrow \exp^{-1} = \log$ is the canonical link function

One Step Deeper: Link Functions

- Link function: $g(\mu_i) = X_i^\top \beta$
- Defines the relationship between $X_i^\top \beta$ and the mean μ_i
- Must map the real line onto the possible range of μ_i

- Recall that $\mu_i = b'(\theta_i)$
- Therefore, θ_i is always a (often simple) function of $X_i^\top \beta$
- In particular, when $\theta_i = X_i^\top \beta$, the link is called the canonical link
  - In Poisson, $\theta_i = \log(\mu_i) = \log(\exp(X_i^\top \beta)) = X_i^\top \beta$
    $\longrightarrow \exp^{-1} = \log$ is the canonical link function

- Must be monotonic and differentiable
- This allows us to express the mean function as: $\mu_i = g^{-1}(X_i^\top \beta)$

Welcome Back!