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: 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.

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: 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
  - 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

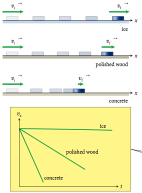
In the preceding two chapters, we developed a mathematical framework for describing motion along a straight line. In this chapter, we continue our study of motion by investigating inertia, a property of objects that affects their motion. The experiments we carry out in studying inertia lead us to discover one of the most fundamental laws in physics—conservation of momentum.

4.1 Friction

Picture a block of wood sitting motionless on a smooth wooden surface. If you give the block a shove, it slides some distance but eventually comes to rest. Depending on the smoothness of the block and the smoothness of the wooden surface, this stopping may happen sooner or it may happen later. If the two surfaces in contact are very smooth and slippery, the block slides for a longer time interval than if the surfaces are rough or sticky. This you know from everyday experience: A hockey puck slides easily on ice but not on a rough road.

Figure 4.1 shows how the velocity of a wooden block decreases on three different surfaces. The slowing down is due to *friction*—the resistance to motion that one surface or object encounters when moving over another. Notice that, during the interval covered by the velocity-versus-time graph, the velocity decrease as the block slides over ice is hardly observable. The block slides easily over ice because there is very little friction between the two surfaces. The effect of friction is to bring two objects to rest with respect to each other—in this case the wooden block and the surface it is sliding on. The less friction there is, the longer it takes for the block to come to rest.

Figure 4.1 Velocity-versus-time graph for a wooden block sliding on three different surfaces. The rougher the surface, the more quickly the velocity decreases.



CONCEPTS

Figure 4.2 Low-friction track and carts used in the experiments described in this chapter.

You may wonder whether it is possible to make surfaces that have no friction at all, such that an object, once given a shove, continues to glide forever. There is no totally frictionless surface over which objects slide forever, but there are ways to minimize friction. You can, for instance, float an object on a cushion of air. This is most easily accomplished with a low-friction track—a track whose surface is dotted with little holes through which pressurized air blows. The air serves as a cushion on which a conveniently shaped object can float, with friction between the object and the track all but eliminated. Alternatively, one can use wheeled carts with low-friction bearings on an ordinary track. Figure 4.2 shows low-friction carts you may have encountered in your lab or class. Although there is still some friction both for low-friction tracks and for the track shown in Figure 4.2, this friction is so small that it can be ignored during an experiment. For example, if the track in Figure 4.2 is horizontal, carts move along its length without slowing down appreciably. In other words:

In the absence of friction, objects moving along a horizontal track keep moving without slowing down.

Another advantage of using such carts is that the track constrains the motion to being along a straight line. We can then use a high-speed camera to record the cart's position at various instants, and from that information determine its speed and acceleration.

4.1 (*a*) Are the accelerations of the motions shown in Figure 4.1 constant? (*b*) For which surface is the acceleration largest in magnitude?

4.2 Inertia

We can discover one of the most fundamental principles of physics by studying how the velocities of two low-friction carts change when the carts collide. Let's first see what happens with two identical carts. We call these *standard carts* because we'll use them as a standard against which to compare the motion of other carts. First we put one standard cart on the low-friction track and make sure it doesn't move. Next we place the second cart on the track some distance from the first one and give the second cart a shove toward the first. The two carts collide, and the collision alters the velocities of both.

ANNOTATION

Alan: I remember, in high school, being amazed at how quickly carts could travel on these tracks - air would blow up through these tiny holes evenly distributed along the length of the track and the cart would essentially float on the air and consequently - the cart would move very quickly with the slightest push.

Bob: Although there is no way to create frictionless surfaces, I find it interesting that we consider experiments "in the absence of friction." In a way, this relates back to Chapter 1.5 where we talked about the importance of having too little or too much information in our representations. In some cases, the friction is so insignificant that we ignore it (simplifying our representation).

Claire: Does this only apply to solid surfaces? I feel as if a substance that floats on water either has negligible or very little friction.

Alan: Why is this? I don't get it.

David: believe this applies to almost every surface, although I'm not sure if water would count more as resistance than friction. Anyways, the best example I could think of would be a surf board. If people who were paddling in the same direction as the waves experienced no resistance, they would continually speed up, and eventually reach very high speeds. However, in reality if they were two stop paddling they'd slow down and only the waves would slowly push them to shore.

Alan: Is it possible to have a surface, in real life, that inflicts NO friction at all?

Erica: Doesn't air resistance factor into this at all? It seems that it is not enough for there to be only an absense of friction for something to keep moving without slowing down. What about some other opposing force - like air resistance? Or is air resistance just another example of friction?

Bob: The key word is "appreciably". In the absence of friction, the cart does not slow down appreciably but still would a little due to air resistance

Alan: a) yes b) concrete has the acceleration of greatest magnitude

Erica: I would think that they are not constant because if we think of the formula F=ma, the force of friction is different in every case so that would change the acceleration value (where mass would stay the same since it's assumed that th object is the same in each situation).

Claire: As a theoretical question about inertia, if an object in motion will stay in motion, but is being affected by friction, will it slow down perpetually but remain in motion, or will it eventually stop completely due to the friction? Just curious.

Alan: With friction everything slows down to a half at one point or another. It is only if an outside force acts on the object if that object will maintain motion after the effects of inertia.

Claire: Standard carts: identical carts in mass, shape, etc. I like this notion of standard carts, it provides a good baseline to compare other motion and to understand the concepts before building on it.

Alan: Great visual representation of friction! It is interesting how this compares the velocity of things on different surfaces

Bob: The rougher the surface, the more friction between the surface and the wooden block, and thus acceleration will be greater.

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

Xinyi Duan[a] and Herrissa Lamothe[a]

[a]Princeton Sociology Department

This manuscript was compiled on May 14, 2016

Scarcity, defined as having less than one needs, has been shown to cause cognitive strain and to decrease cognitive performance [1]. Mani et al. (2013) find experimentally that before and after payday, Indian farmers who were paid once a year for their crops showed worse cognitive performance before payday, a sign of scarcity. Carvalho et al. conduct similar experiments in the U.S. context and conclude that scarcity's effect on cognition and decision making for the most part does not hold in the U.S. [2]. We reanalyze the Carvalho et al. data and find that identification and design issues may have attenuated or muddled the effect. We also conduct two online studies more specifically examining both inflow and outflow of financial resources, and their impact on scarcity and cognitive performance. We find that changes in U.S. participants' financial resources are multifaceted and hence scarcity does not consistently change with any one aspects of payday, debt, income or expenditures. Rather, we find scarcity and cognitive strain occurred after recent experiences of late or penalized bill payment. We argue that scarcity in the U.S. context is induced by unmanaged mismatches between two types of volatility: income and bill payment volatility[3], that result in a penalty. This finding holds important implications for scarcity, cognitive strain, and quality of economic decision-making in the U.S. context.

Scarcity | Income Volatility | Expense Volatility | Economic Decision Making | Penalty| Cognition

Scarcity, defined as having less than one needs, has been shown to impose a cognitive strain that impedes decision-making. Particularly in the case of poverty, cognitive strain is hypothesized to result from the cognitively costly process of "managin[ing] sporadic income, juggl[ing] expenses, and making[ing] difficult tradeoff" [1](p. 976). Key to this process is the idea that the cognitive task of juggling various types of volatility in inflow and outflow is itself a source of additional preoccupation and cognitive strain that leaves fewer cognitive resources for other cognitive tasks– including decision-making. While the impact of economic scarcity on decision-making has been directly observed in international contexts[1],it is not until recently that this analysis has been applied to the U.S. context[2].

Mani et al. (2013) find that Indian sugar-cane farmers demonstrate diminished cognitive performance before their annual payday when compared to performance after payday. Carvalho et al. (2016)'s experimental adaptation of this pre-post payday design to the U.S. context finds no such differences in cognitive performance. However, the average low income U.S. context is very different from the Indian farmer context. U.S. participants face frequent paydays– often weekly or bi-weekly, volatility in the timing and amount of these paydays, debt payments and due dates, and the power of financial institutions regulating bill payments. These are key features of the U.S. context without which we would not be able to understand the full picture of the state of financial resources in the American experience. An analysis of cognitive strain and decision-making in the U.S. must take these factors into account.

Scarcity and Cognitive Strain in the U.S. Context. In this paper, we re-analyze Carvalho et al.(2016)'s study results, and refute the argument that scarcity in the U.S. context does not induce cognitive strain. Despite design, compliance, and identification issues that bias findings towards null results; data from the Carvalho et al.(2016) study suggest that scarcity in the U.S. context negatively affects cognitive performance and impairs decision-making. However, counter to Carvalho et al.'s argument, being in the period before payday does not seem to cause a difference in scarcity mindset in the U.S. Nor is increased expenditure a sign of general economic well-being and decrease in scarcity. Instead, participants' subjective feeling of scarcity increases as expenditure increases. Moreover, participants earning less than \$20,000 a year feel more scarce than participants earning more than that amount in the Carvalho et al.(2016) data.

Having identified a key link between scarcity and cognitive strain in the U.S. context, we conduct two complementary studies to identify the mechanisms inducing scarcity in the U.S. context. Our first study deals with an identification issue in the Carvalho et al.(2016) study regarding what was considered a payday, and compliance issues that may have attenuated results. The second study consists of an experiment that manipulated participants' level of preoccupation about bill payment to test the hypothesis that scarcity in the U.S. context is a function of bill payment expenditures and not having enough financial resources to pay bills. Interestingly, we find that neither receiving payment on payday, nor the ability to pay bill payment amounts in full were associated with higher feelings of scarcity and cognitive strain in the U.S. context. Instead we find the structural mechanism of being late on a bill and incurring late fees or a penalty were associated with higher levels of scarcity. This suggests a very different

Significance Statement

Scarcity– having less than what one feels one needs– negatively impacted U.S. participants' cognitive reaction time and decision quality. Scarcity in our U.S. sample occurred not when participants had not yet been paid, or when they had not paid a bill in full; instead scarcity occurred when participants were late and penalized on a recent bill. We suggest these episodes bring participants' concerns about their financial situation to the foreground. We hypothesize that such instances of penalty occur when income/inflow and bill payment/outflow volatilities are mismatched and result in feelings of scarcity and signs of cognitive strain.

Please provide details of author contributions here.

[1]Xinyi Duan and Herrissa Lamothe contributed equally to this work.

www.pnas.org/cgi/doi/10.1073/pnas.XXXXXXXXXX

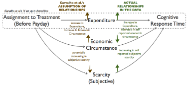


Fig. 1. Causal Paths in Carvalho et al.'s IV design.

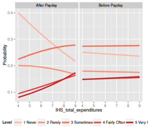


Fig. 2. Interaction of Expenditure and Treatment on Being Worried about Finances Carvalho et al.'s IV design.

reported expenditure measure (expenditure within the last seven days), on the outcome of interest (reaction time). The construction required the following assumptions: 1) those in the before payday assigned group have less expenditure than those in the after group, 2) that less expenditure means people are more scarce– perhaps because they have, or feel that they have, fewer resources and hence spend less, 3) and that less expenditure should negatively impact reaction time as predicted by scarcity theory. Not very assumption relationship held in their data (Figure 1).

The first glaring problem is one of IV's exclusion restriction. By assumption, expenditure stands in for both better economic circumstances and scarcity. The two variables are related but are not the same concept, nor are they analogous with expenditure. Consequently, within the model proposed by Carvalho et al., we already have alternative paths through which intent to treat may impact cognitive response time that would fail the IV exclusion requirement. The second major problem is that assumptions 2 and 3 do not hold in their own data. While the authors theorize that expenditure means better economic circumstances and less scarcity, their own data show that increasing expenditure actually aggravates the feeling of scarcity. For one of their self reported measures of scarcity, "in the last 24 hours, how often did you worry about

[0]We show in our appendix that assumption 1 only holds for Study 2. Conditional on those earning less \$20,000 a year, there is no significant difference in expenditure before and after payday for Study 1

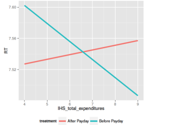


Fig. 3. Interaction of Expenditure and Treatment on Predicted Reaction Time from Carvalho et al. data

having enough money to make ends meet?", controlling for treatment (before payday), increasing expenditure actually increases people's likelihood of being worried. More intriguingly, there is an interaction with treatment such that in the after payday group, the more they spend, the more worried they seem to be, while before payday group shows the same level of worry across expenditure (Figure 2).

This interaction indicates that the simple main effect of treatment may not have cleanly identified periods of scarcity as Carvalho et al. have posited, and that treatment is interacting with other indicators of economic circumstance which actually induces scarcity. Examining the interaction of expenditure with treatment on reaction time, we see that the before payday group is indeed significantly slower than after payday, when controlling for expenditure and the interaction of expenditure and treatment. More importantly, the interaction between treatment and expenditure is also significant (Figure 3). An ANOVA test shows that the interaction model is a significantly better model than a model with just treatment, at $p < 0.0001$ level. Upon examination of the expenditure question, we hypothesize that perhaps expenditure captured bill payments as the question asked explicitly about bills, which may be why those in the after payday group feel strained as they spend more because they were paying more on bills, as after payday often is a time for paying off debt.

In light of these analyses, which revealed unaccounted interactions in Carvalho et al.'s data, we conclude that scarcity in the American context does not seem to hold at the simple before and after payday line; and what mechanisms driving scarcity are more complex. Some groups feel more subjectively scarce constantly (those making less than \$20,000 in income compared to those who make more), while some feel more scarce as they manage more outflows after they receive income. Both of these groups who report feeling more scarcity do indeed have slower reaction time or less accuracy on the Stroop task.

[0]We examined the other self reported scarcity measures as well for the rest of the analyses and found similar patterns

[0]We also conduct this analysis with income less than 20k as an additional covariate. We find that the the having lower income also account for a chunk of variation in people's feeling of scarcity, along with treatment and expenditure, showing the heterogeneous effects for the less than 20k income group again.

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)

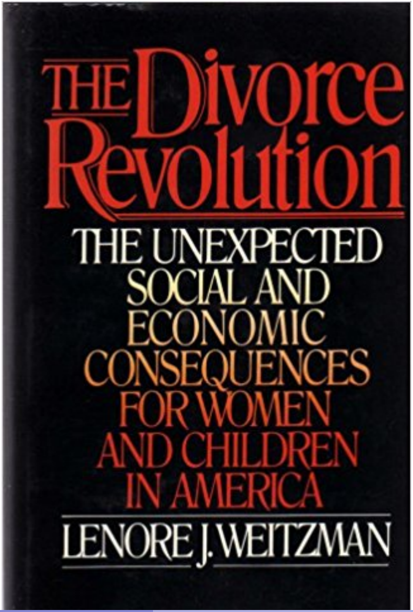
Why a Replication Project?

- 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?

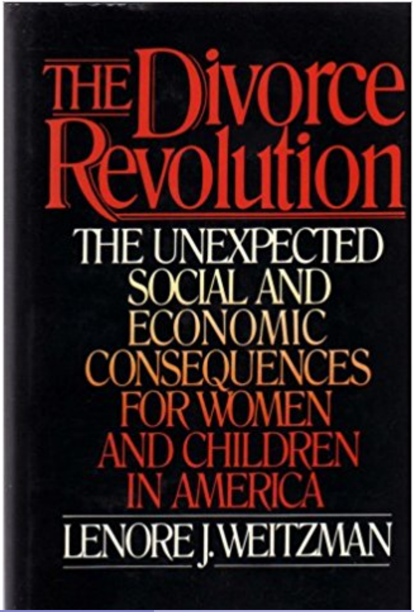
Why Replicability?

- Changes in living standard after divorce



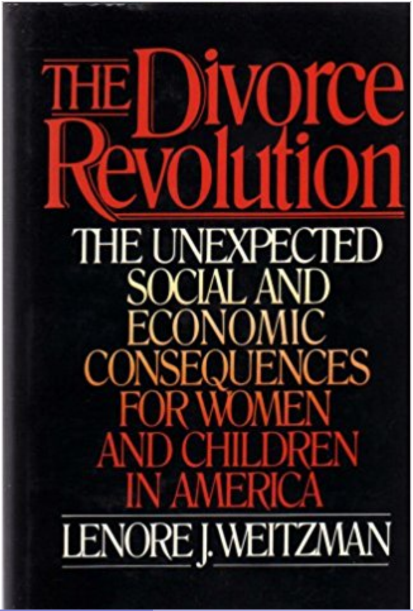
Why Replicability?

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



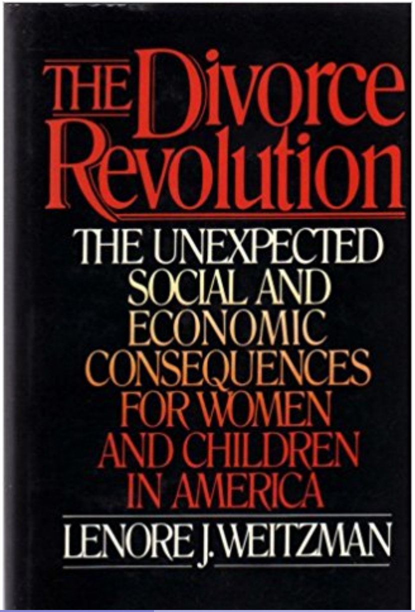
Why Replicability?

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



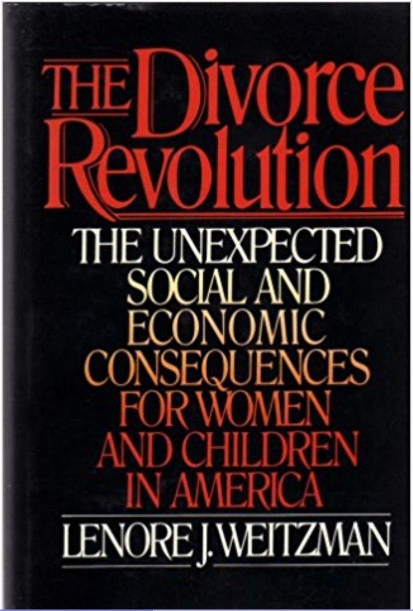
Why Replicability?

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



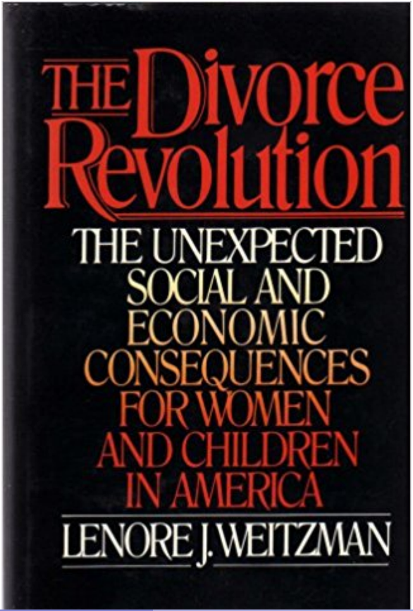
Why Replicability?

- 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



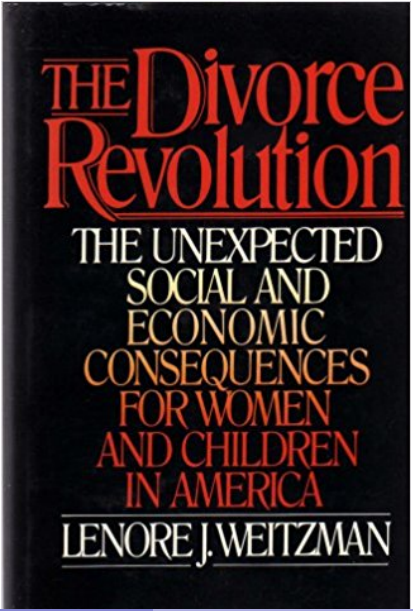
Why Replicability?

- 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)

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

Equivalent Linear Regression Notation

- Standard version (last semester)

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

Equivalent Linear Regression Notation

- Standard version (last semester)

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

- Alternative version

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

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)

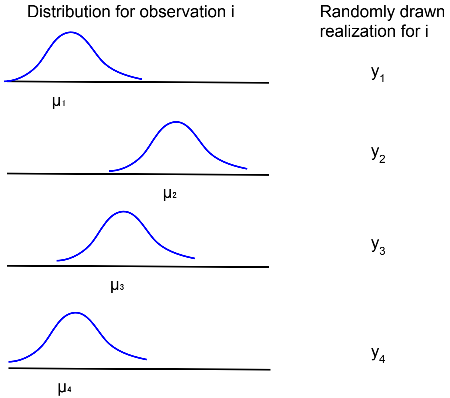
$$\begin{aligned} Y_i &= x_i\beta + \epsilon_i && \text{systematic} + \text{stochastic} \\ \epsilon_i &\sim f_N(0, \sigma^2) \end{aligned}$$

- Alternative version

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

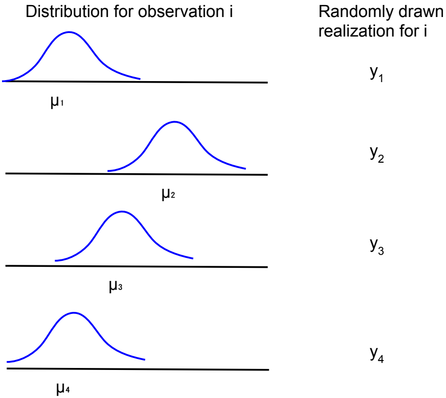
Understanding the Alternative Regression Notation

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

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

Generalized Alternative Notation

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

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

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

where

Y_i random outcome variable

$f(\cdot)$ probability density

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

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

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

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.

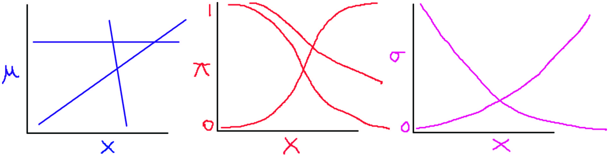
Forms of Uncertainty

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

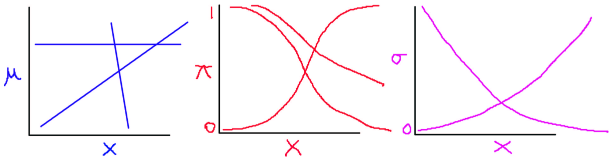
- Estimation uncertainty: Lack of knowledge of β and α. Vanishes as n gets larger.
- Fundamental uncertainty: Represented by the stochastic component. Exists no matter what the researcher does; no matter how large n is.
- (A Test: If you know the model, is $R^2 = 1$?)

Systematic Components: Examples

Systematic Components: Examples

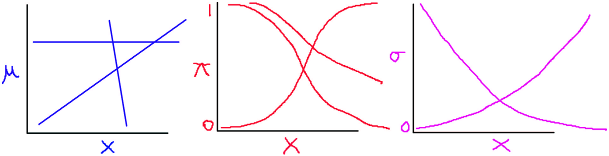


Systematic Components: Examples



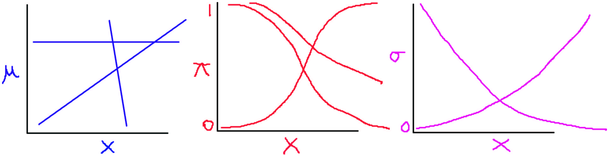
- $E(Y_i) \equiv \mu_i = X_i\beta = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}$

Systematic Components: Examples



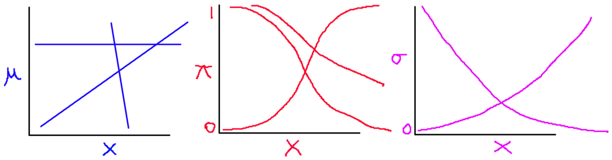
- $E(Y_i) \equiv \mu_i = X_i\beta = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}$
- $\Pr(Y_i = 1) \equiv \pi_i = \frac{1}{1+e^{-x_i\beta}}$

Systematic Components: Examples



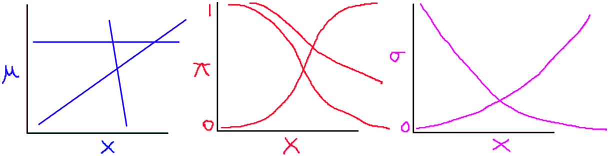
- $E(Y_i) \equiv \mu_i = X_i\beta = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}$
- $\Pr(Y_i = 1) \equiv \pi_i = \frac{1}{1+e^{-x_i\beta}}$
- $V(Y_i) \equiv \sigma_i^2 = e^{x_i\beta}$

Systematic Components: Examples



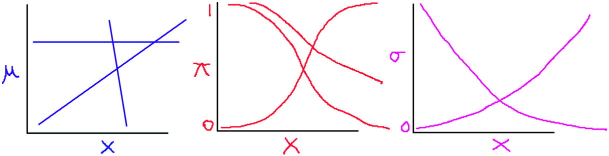
- $E(Y_i) \equiv \mu_i = X_i\beta = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}$
- $\Pr(Y_i = 1) \equiv \pi_i = \frac{1}{1+e^{-x_i\beta}}$
- $V(Y_i) \equiv \sigma_i^2 = e^{x_i\beta}$
- Interpretation:

Systematic Components: Examples



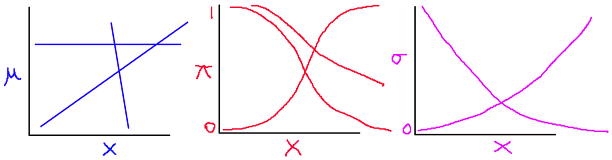
- $E(Y_i) \equiv \mu_i = X_i\beta = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}$
- $\Pr(Y_i = 1) \equiv \pi_i = \frac{1}{1+e^{-x_i\beta}}$
- $V(Y_i) \equiv \sigma_i^2 = e^{x_i\beta}$
- Interpretation:
  - Each is a class of functional forms

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

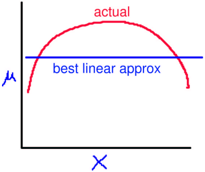
- We will:
  - Assume a class of functional forms (called link functions) from theory (each form is flexible and maps out many potential relationships)
  - Estimate β from data to choose a member of the class
  - Be uncertain (Because of sampling, measurement, & fundamental uncertainties) about:
   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

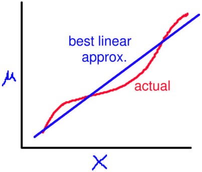
Systematic Components: Examples

- 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

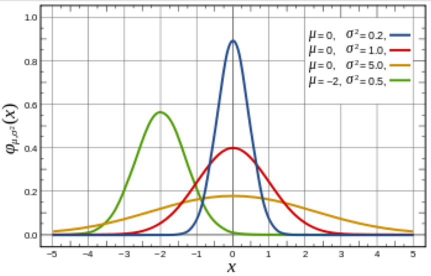
- We will:
  - Assume a class of functional forms (called link functions) from theory (each form is flexible and maps out many potential relationships)
  - Estimate β from data to choose a member of the class
  - Be uncertain (Because of sampling, measurement, & fundamental uncertainties) about:
   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





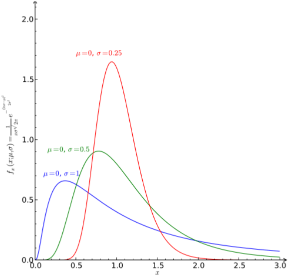
Overview of Stochastic Components

Overview of Stochastic Components



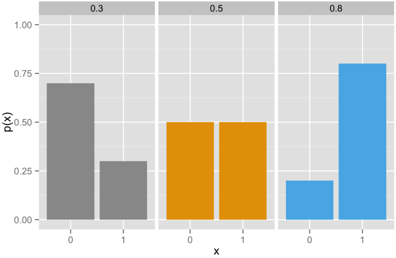
- Normal — continuous, unimodal, symmetric, unbounded

Overview of Stochastic Components



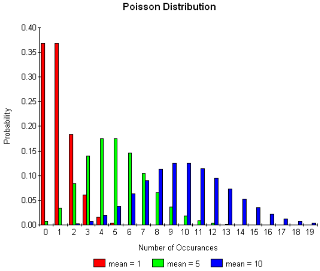
- Normal — continuous, unimodal, symmetric, unbounded
- Log-normal — continuous, unimodal, skewed, bounded from below by zero

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:
    - ⋆ <u>Avoid</u> it: relax assumptions
    - ⋆ <u>Detect</u> 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):

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

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):

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

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!