SPIA586a: Machine Learning for Policy Analysis
Week 2: Discovery and Measurement

Brandon Stewart[1]

Princeton

February 3–5, 2025

[1]Huge thanks to Ian Lundberg and Matt Salganik for materials around the Fragile Families Challenge.

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

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

- Last Week
  - course structure
  - core ideas of the course
  - what is machine learning
  - seven tools
  - review of regression

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

- Last Week
  - course structure
  - core ideas of the course
  - what is machine learning
  - seven tools
  - review of regression
- This Week
  - discovery & measurement
  - regularization, GAMs

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

- Last Week
  - course structure
  - core ideas of the course
  - what is machine learning
  - seven tools
  - review of regression
- This Week
  - discovery & measurement
  - regularization, GAMs
- Next Week
  - prediction within population

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

- Last Week
  - course structure
  - core ideas of the course
  - what is machine learning
  - seven tools
  - review of regression
- This Week
  - discovery & measurement
  - regularization, GAMs
- Next Week
  - prediction within population
- Long Run
  - target tasks $\rightarrow$ data sources $\rightarrow$ guiding values

1. Review
2. Frameworks for Discovery and Measurement
3. Example Application: Political Emails
4. Frameworks for Measurement
5. A Failure of Measurement
6. Introduction to the Fragile Families Challenge
7. Regression with Regularization
8. Logistic Regression
9. Generalized Additive Models

1. Review
2. Frameworks for Discovery and Measurement
3. Example Application: Political Emails
4. Frameworks for Measurement
5. A Failure of Measurement
6. Introduction to the Fragile Families Challenge
7. Regression with Regularization
8. Logistic Regression
9. Generalized Additive Models

Review: Breaking Down Sources of Error

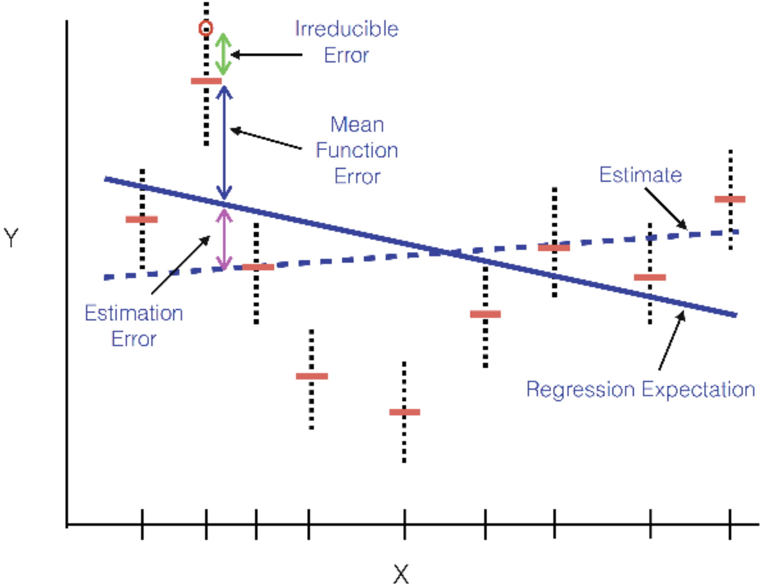


Image Credit: Berk (2020) *Statistical learning from a regression perspective*

1. Review
2. Frameworks for Discovery and Measurement
3. Example Application: Political Emails
4. Frameworks for Measurement
5. A Failure of Measurement
6. Introduction to the Fragile Families Challenge
7. Regression with Regularization
8. Logistic Regression
9. Generalized Additive Models

1. Review
2. Frameworks for Discovery and Measurement
3. Example Application: Political Emails
4. Frameworks for Measurement
5. A Failure of Measurement
6. Introduction to the Fragile Families Challenge
7. Regression with Regularization
8. Logistic Regression
9. Generalized Additive Models

Discovery and Measurement

Discovery and Measurement

- "Concept formation lies at the heart of all social science endeavors. . . Concepts are integral to every argument for they address the most basic question of social science research: what are we talking about?" (Gerring, 2012, pg 112).

Discovery and Measurement

- "Concept formation lies at the heart of all social science endeavors... Concepts are integral to every argument for they address the most basic question of social science research: what are we talking about?" (Gerring, 2012, pg 112).
- Concepts are not really right or wrong, just more or less useful.

Discovery and Measurement

- "Concept formation lies at the heart of all social science endeavors... Concepts are integral to every argument for they address the most basic question of social science research: what are we talking about?" (Gerring, 2012, pg 112).
- Concepts are not really right or wrong, just more or less useful.
- Roughly speaking concepts are useful when they consistently distinguish one consistent type of phenomena from others in a way that can be cleanly articulated.

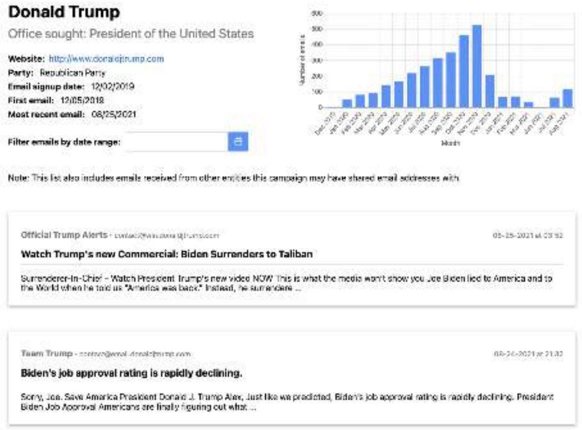
1. Review
2. Frameworks for Discovery and Measurement
3. Example Application: Political Emails
4. Frameworks for Measurement
5. A Failure of Measurement
6. Introduction to the Fragile Families Challenge
7. Regression with Regularization
8. Logistic Regression
9. Generalized Additive Models

1. Review
2. Frameworks for Discovery and Measurement
3. Example Application: Political Emails
4. Frameworks for Measurement
5. A Failure of Measurement
6. Introduction to the Fragile Families Challenge
7. Regression with Regularization
8. Logistic Regression
9. Generalized Additive Models

Tracking Manipulative Tactics in Political Email

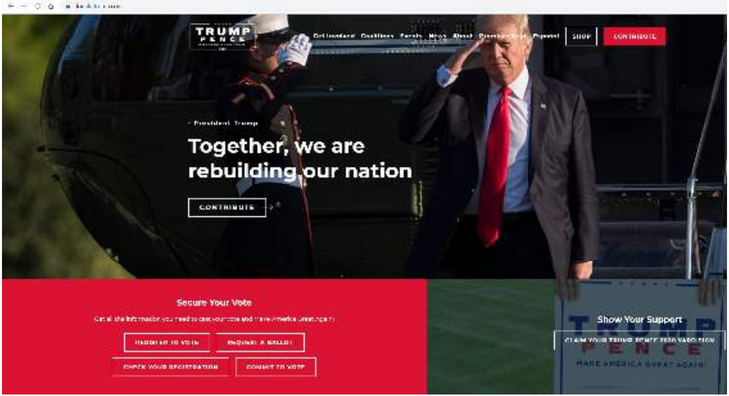
Mathur, Wang, Schwemmer, Hamin, Stewart and Narayanan

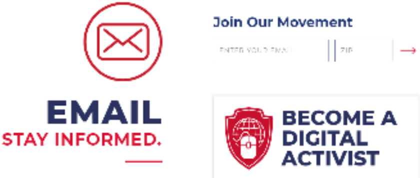
"Manipulative tactics are the norm in political emails: Evidence from 300K emails from the 2020 U.S. election cycle", *Big Data & Society*, 2023.



electionemails2020.org

Campaign website example





Email Example

This is about more than $25

Joe Biden

Office sought: President of the United States

From name: Joseph R. Biden Jr.
From address: info@joebiden.com
Email date: 10/22/2020
Email time: 15:32

Folks,

Tonight, I am debating Donald Trump for the final time.

I hope that tonight the whole country will get to hear about our vision for how we're going to build this country back better than it was before. I'll be fighting for you up there, and I hope I make you proud.

But after the debate, the next step will be up to you -- because we'll only have 11 days until Election Day, and I need to know if you'll be standing with me tonight and until the last ballot is counted. And we rely on nights like this for fundraising. Without those added resources, we can't win.

This is not only about your first donation. This is about you standing by my side when it matters the most. Can I count on you to donate $25 before I take the stage tonight to make today our biggest yet?

If you've saved payment information with ActBlue Express, your donation will go through immediately:

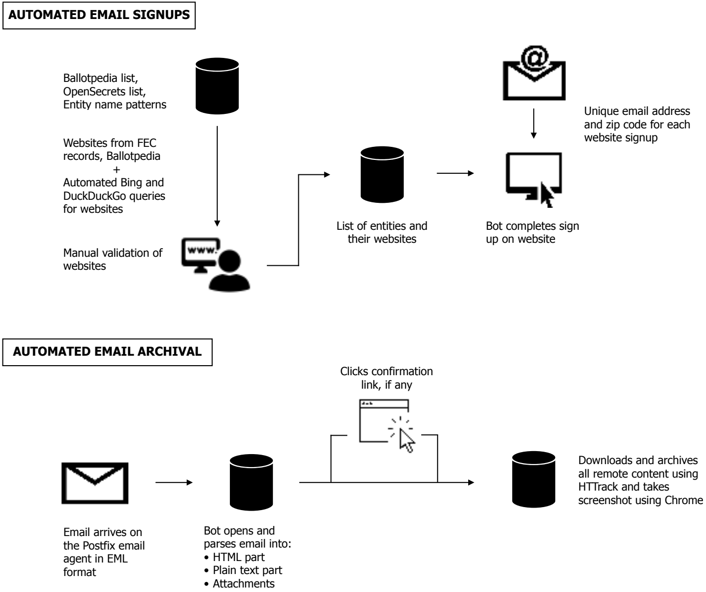
Click here to express donate $25 >>

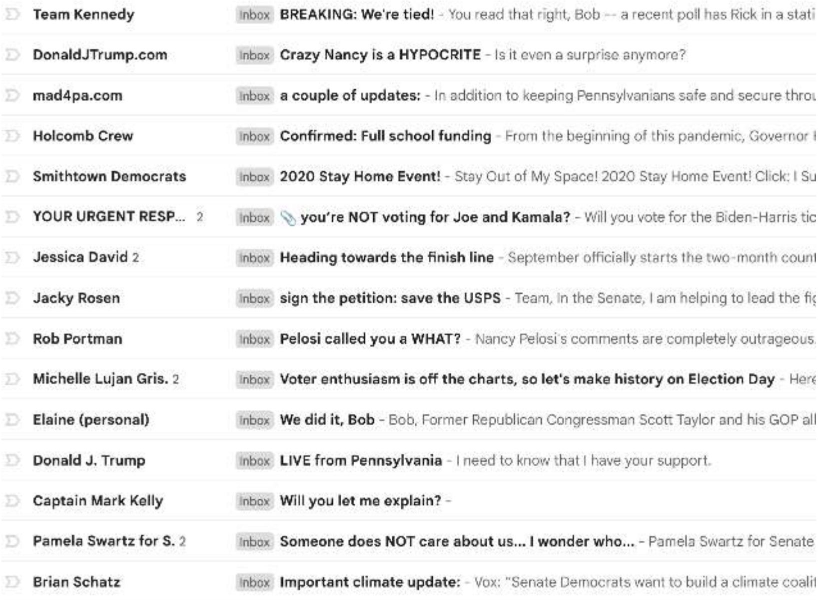
Click here to express donate $50 >>

Click here to express donate $100 >>

- emails contain HTML elements such as links & images, but mostly text
- three parts: header information, body and any outgoing sites

Bot-Based Signup Pipeline




Team Kennedy Inbox BREAKING: We're tied! - You read that right, Bob -- a recent poll has Rick in a stati
DonaldJTrump.com Inbox Crazy Nancy is a HYPOCRITE - Is it even a surprise anymore?
mad4pa.com Inbox a couple of updates: - In addition to keeping Pennsylvanians safe and secure throu
Holcomb Crew Inbox Confirmed: Full school funding - From the beginning of this pandemic, Governor
Smithtown Democrats Inbox 2020 Stay Home Event! - Stay Out of My Space! 2020 Stay Home Event! Click: I Su
YOUR URGENT RESP... 2 Inbox you're NOT voting for Joe and Kamala? - Will you vote for the Biden-Harris tic
Jessica David 2 Inbox Heading towards the finish line - September officially starts the two-month count
Jacky Rosen Inbox sign the petition: save the USPS - Team, In the Senate, I am helping to lead the fig
Rob Portman Inbox Pelosi called you a WHAT? - Nancy Pelosi's comments are completely outrageous.
Michelle Lujan Gris. 2 Inbox Voter enthusiasm is off the charts, so let's make history on Election Day - Here
Elaine (personal) Inbox We did it, Bob - Bob, Former Republican Congressman Scott Taylor and his GOP all
Donald J. Trump Inbox LIVE from Pennsylvania - I need to know that I have your support.
Captain Mark Kelly Inbox Will you let me explain? -
Pamela Swartz for S. 2 Inbox Someone does NOT care about us... I wonder who... - Pamela Swartz for Senate
Brian Schatz Inbox Important climate update: - Vox: "Senate Democrats want to build a climate coalit


30%
20%
10%
0%
Percentage
Political Campaigns & Events
Political Issues
Fundraising
Misc. / Other
Voter Mobilization
Ceremonial / Niceties

Fundraising


The Washington Post
Democracy Dies in Darkness

Outlook • Perspective

Trump's campaign lures donors with absurd financial promises — and insults

Its fundraising efforts take a bizarre, aggressive tone.

Campaign manager Bill Stepien, right, listens as President Trump speaks at a campaign stop in Old Forge, Pa. (Evan Vucci/AP)

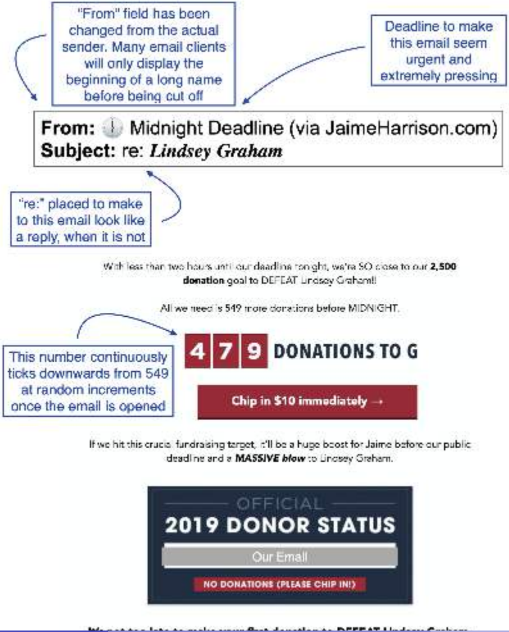
Ben Adler, October 9 2020

"I've decided to activate an exclusive 3-Months-Out 700%-MATCH just for YOU. This offer is only available for the NEXT HOUR."

"Why haven't you entered to win a chance to meet President Trump yet?"

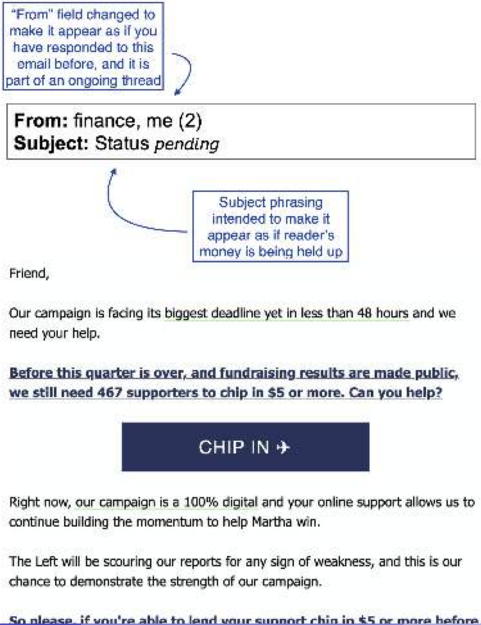
"The President saw the list of Patriots who have already entered & he noticed that your name was MISSING."

Example 1



"From" field has been changed from the actual sender. Many email clients will only display the beginning of a long name before being cut off
Deadline to make this email seem urgent and extremely pressing
From: Midnight Deadline (via JaimeHarrison.com)
Subject: re: Lindsey Graham
"re:" placed to make to this email look like a reply, when it is not
With less than two hours until our deadline tonight, we're SO close to our 2,500 donation goal to DEFEAT Lindsey Graham!!
All we need is 549 more donations before MIDNIGHT.
This number continuously ticks downwards from 549 at random increments once the email is opened
4 7 9 DONATIONS TO G
Chip in $10 immediately →
If we hit this crucial fundraising target, it'll be a huge boost for Jaime before our public deadline and a MASSIVE blow to Lindsey Graham.
OFFICIAL
2019 DONOR STATUS
Our Email
NO DONATIONS (PLEASE CHIP IN!)

Example 2


"From" field changed to make it appear as if you have responded to this email before, and it is part of an ongoing thread
From: finance, me (2)
Subject: Status *pending*
Subject phrasing intended to make it appear as if reader's money is being held up
Friend,
Our campaign is facing its biggest deadline yet in less than 48 hours and we need your help.
Before this quarter is over, and fundraising results are made public, we still need 467 supporters to chip in $5 or more. Can you help?
CHIP IN
Right now, our campaign is a 100% digital and your online support allows us to continue building the momentum to help Martha win.
The Left will be scouring our reports for any sign of weakness, and this is our chance to demonstrate the strength of our campaign.
So please, if you're able to lend your support chip in $5 or more before

Manipulative Practices

Open rate for political emails is only about 20%.

Manipulative Practices

Open rate for political emails is only about 20%.

We found two manipulative ways within sender & subject fields to manipulate recipients into opening emails:

Manipulative Practices

Open rate for political emails is only about 20%.

We found two manipulative ways within sender & subject fields to manipulate recipients into opening emails:

- Clickbait:
  - forward referencing/info. withholding
  - sensationalism
  - urgency / time scarcity

Manipulative Practices

Open rate for political emails is only about 20%.

We found two manipulative ways within sender & subject fields to manipulate recipients into opening emails:

- Clickbait:
  - forward referencing/info. withholding
  - sensationalism
  - urgency / time scarcity
- User-interface manipulation:
  - obscured name
  - ongoing thread
  - re: / fwd:

Manipulative Practices

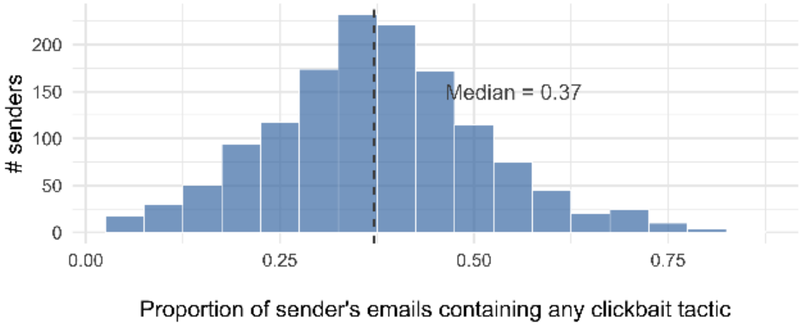
Open rate for political emails is only about 20%.

We found two manipulative ways within sender & subject fields to manipulate recipients into opening emails:

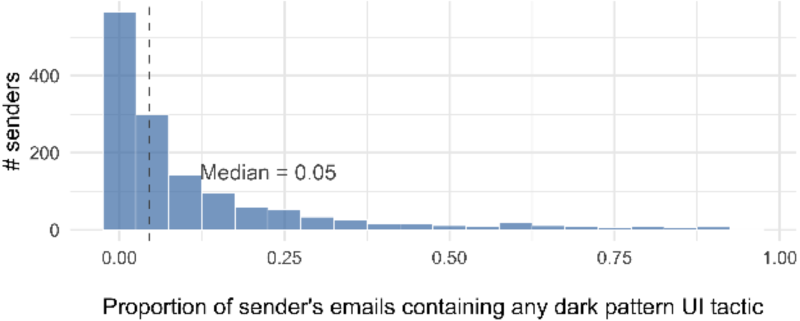
- Clickbait:
  - forward referencing/info. withholding
  - sensationalism
  - urgency / time scarcity
- User-interface manipulation:
  - obscured name
  - ongoing thread
  - re: / fwd:

Likely impacts are highest for low digital literacy (highly correlated with age).

Manipulative Practices — Distribution



Manipulative Practices — Distribution



Princeton Corpus of Political Emails

This corpus contains emails from candidates running for state and federal office, political parties, and other political organizations like PACs.

As of Tue Sep 7 06:57:16 EDT 2021, the corpus contains **435,436 emails** from **3,129 senders**.

Search the corpus below or view the corpus organized by senders instead

Search the corpus (e.g., 'Trump' or 'Biden') Search

PRINCETON UNIVERSITY

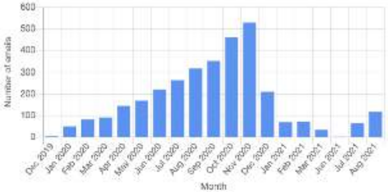
This project is led by researchers at Princeton University, affiliated with its Center for Information Technology Policy (CITP). Our contact details are listed below.

Donald Trump

Office sought: President of the United States

Website: http://www.donaldjtrump.com
Party: Republican Party
Email signup date: 12/02/2019
First email: 12/05/2019
Most recent email: 08/25/2021

Filter emails by date range:



Note: This list also includes emails received from other entities this campaign may have shared email addresses with.

Official Trump Alerts · contact@win.donaldjtrump.com 08-25-2021 at 03:52

Watch Trump's new Commercial: Biden Surrenders to Taliban

Surrenderer-In-Chief - Watch President Trump's new video NOW This is what the media won't show you Joe Biden lied to America and to the World when he told us "America was back." Instead, he surrendere ...

Team Trump · contact@email.donaldjtrump.com 08-24-2021 at 21:32

Biden's job approval rating is rapidly declining.

Sorry, Joe. Save America President Donald J. Trump Alex, Just like we predicted, Biden's job approval rating is rapidly declining. President Biden Job Approval Americans are finally figuring out what ...

Studying Political Emails

- We could have studied any facet of political emails—we happened to choose manipulative practices.

Studying Political Emails

- We could have studied any facet of political emails—we happened to choose manipulative practices.
- Given that we chose manipulate practices we had to define each practice apart from the data itself.

Studying Political Emails

- We could have studied any facet of political emails—we happened to choose manipulative practices.
- Given that we chose manipulate practices we had to define each practice apart from the data itself.
- Then we needed a way to measure the practices at scale.

Studying Political Emails

- We could have studied any facet of political emails—we happened to choose manipulative practices.
- Given that we chose manipulate practices we had to define each practice apart from the data itself.
- Then we needed a way to measure the practices at scale.
- All of these steps are complicated and come with many caveats!

1. Review
2. Frameworks for Discovery and Measurement
3. Example Application: Political Emails
4. Frameworks for Measurement
5. A Failure of Measurement
6. Introduction to the Fragile Families Challenge
7. Regression with Regularization
8. Logistic Regression
9. Generalized Additive Models

1 Review

2 Frameworks for Discovery and Measurement

3 Example Application: Political Emails

4 Frameworks for Measurement

5 A Failure of Measurement

6 Introduction to the Fragile Families Challenge

7 Regression with Regularization

8 Logistic Regression

9 Generalized Additive Models

Measurement

- Roughly speaking, measurement is the process of linking units to concepts by assigning them some kind of label.

Measurement

- Roughly speaking, measurement is the process of linking units to concepts by assigning them some kind of label.
- What is easily measurable isn't always interesting and what is interesting isn't always easily measurable.

Measurement

- Roughly speaking, measurement is the process of linking units to concepts by assigning them some kind of label.
- What is easily measurable isn't always interesting and what is interesting isn't always easily measurable.
- Many useful social concepts are latent properties like power, social capital, legitimacy, culture and justice.

Measurement

- Roughly speaking, measurement is the process of linking units to concepts by assigning them some kind of label.
- What is easily measurable isn't always interesting and what is interesting isn't always easily measurable.
- Many useful social concepts are latent properties like power, social capital, legitimacy, culture and justice.
- Social phenomena are particularly difficult because the unit can react to the act of being measured or preemptively adjust to the possibility of being measured.

Measurement

- Roughly speaking, measurement is the process of linking units to concepts by assigning them some kind of label.
- What is easily measurable isn't always interesting and what is interesting isn't always easily measurable.
- Many useful social concepts are latent properties like power, social capital, legitimacy, culture and justice.
- Social phenomena are particularly difficult because the unit can react to the act of being measured or preemptively adjust to the possibility of being measured.
- Observationally equivalent phenomena can be different based on context (e.g. a 'wink' vs a 'twitch' to use Geertz's famous example).

Measurement

- Roughly speaking, measurement is the process of linking units to concepts by assigning them some kind of label.
- What is easily measurable isn't always interesting and what is interesting isn't always easily measurable.
- Many useful social concepts are latent properties like power, social capital, legitimacy, culture and justice.
- Social phenomena are particularly difficult because the unit can react to the act of being measured or preemptively adjust to the possibility of being measured.
- Observationally equivalent phenomena can be different based on context (e.g. a 'wink' vs a 'twitch' to use Geertz's famous example).
- Many ways to describe a good measurement but they often involve notions of reliability and validity

The Realities of Measurement

The Realities of Measurement

- Challenge of measurement is that a measure can't be all things to all people, it has to be specific to a particular purpose.

The Realities of Measurement

- Challenge of measurement is that a measure can't be all things to all people, it has to be specific to a particular purpose.
- Yet the narrower the purpose, the harder it is to justify extensive validation.

The Realities of Measurement

- Challenge of measurement is that a measure can't be all things to all people, it has to be specific to a particular purpose.
- Yet the narrower the purpose, the harder it is to justify extensive validation.
- Quantitative turn in 20th century social science driven by large surveys which activitely solicit information from people (e.g. the General Social Survey, American National Election Survey).

The Realities of Measurement

- Challenge of measurement is that a measure can't be all things to all people, it has to be specific to a particular purpose.
- Yet the narrower the purpose, the harder it is to justify extensive validation.
- Quantitative turn in 20th century social science driven by large surveys which activitely solicit information from people (e.g. the General Social Survey, American National Election Survey).
- Now: massive increase in data $\rightsquigarrow$ communities leaving digitized footprints which we can passively measure.

The Realities of Measurement

- Challenge of measurement is that a measure can't be all things to all people, it has to be specific to a particular purpose.
- Yet the narrower the purpose, the harder it is to justify extensive validation.
- Quantitative turn in 20th century social science driven by large surveys which activitely solicit information from people (e.g. the General Social Survey, American National Election Survey).
- Now: massive increase in data $\rightsquigarrow$ communities leaving digitized footprints which we can passively measure.
- Thus we are entering an age of custom measurement where a single researcher can more straightforwardly develop their own one-off measurement.

The Realities of Measurement

- Challenge of measurement is that a measure can't be all things to all people, it has to be specific to a particular purpose.
- Yet the narrower the purpose, the harder it is to justify extensive validation.
- Quantitative turn in 20th century social science driven by large surveys which activitely solicit information from people (e.g. the General Social Survey, American National Election Survey).
- Now: massive increase in data $\rightsquigarrow$ communities leaving digitized footprints which we can passively measure.
- Thus we are entering an age of custom measurement where a single researcher can more straightforwardly develop their own one-off measurement.
- Validation has thus moved from a community effort to an individual effort.

Measurement and Machine Learning

- Measurement is often difficult to think about with the tools of machine learning because the gold standard often isn't available.

Measurement and Machine Learning

- Measurement is often difficult to think about with the tools of machine learning because the gold standard often isn't available.
- In practice, researchers generally don't agree on the concept until the measurement is pinned down and even then it can be tricky.

Measurement and Machine Learning

- Measurement is often difficult to think about with the tools of machine learning because the gold standard often isn't available.
- In practice, researchers generally don't agree on the concept until the measurement is pinned down and even then it can be tricky.
- Thus, it is hard to imagine what a 'shared task' competition would even look like for measurement.

Measurement and Machine Learning

- Measurement is often difficult to think about with the tools of machine learning because the gold standard often isn't available.
- In practice, researchers generally don't agree on the concept until the measurement is pinned down and even then it can be tricky.
- Thus, it is hard to imagine what a 'shared task' competition would even look like for measurement.
- In practice though, many issues with applications of machine learning come from not taking seriously enough the question of how something is measured and whether the quantity being predicted is really the quantity of interest.

1. Review
2. Frameworks for Discovery and Measurement
3. Example Application: Political Emails
4. Frameworks for Measurement
5. A Failure of Measurement
6. Introduction to the Fragile Families Challenge
7. Regression with Regularization
8. Logistic Regression
9. Generalized Additive Models

1 Review
2 Frameworks for Discovery and Measurement
3 Example Application: Political Emails
4 Frameworks for Measurement
5 A Failure of Measurement
6 Introduction to the Fragile Families Challenge
7 Regression with Regularization
8 Logistic Regression
9 Generalized Additive Models

RESEARCH ARTICLE

ECONOMICS

Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer[1,2]*, Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5]*†

Health systems rely on commercial prediction algorithms to identify and help patients with complex health needs. We show that a widely used algorithm, typical of this industry-wide approach and affecting millions of patients, exhibits significant racial bias: At a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses. Remedying this disparity would increase the percentage of Black patients receiving additional help from 17.7 to 46.5%. The bias arises because the algorithm predicts health care costs rather than illness, but unequal access to care means that we spend less money caring for Black patients than for White patients. Thus, despite health care cost appearing to be an effective proxy for health by some measures of predictive accuracy, large racial biases arise. We suggest that the choice of convenient, seemingly effective proxies for ground truth can be an important source of algorithmic bias in many contexts.

Obermeyer et. al. 2019

Obermeyer et. al. 2019

- Industry Goal: target patients for ‘high-risk care management.’ A broader case of algorithmic prioritization.

Obermeyer et. al. 2019

- Industry Goal: target patients for ‘high-risk care management.’ A broader case of algorithmic prioritization.
- Who would be helped most? Difficult causal inference problem $\rightsquigarrow$ simplifying assumption: those with greatest care needs will benefit most.

Obermeyer et. al. 2019

- Industry Goal: target patients for 'high-risk care management.' A broader case of algorithmic prioritization.
- Who would be helped most? Difficult causal inference problem $\rightsquigarrow$ simplifying assumption: those with greatest care needs will benefit most.
- Algorithm designers use health-care costs as a proxy for 'greatest care needs.'

Obermeyer et. al. 2019

- Industry Goal: target patients for 'high-risk care management.' A broader case of algorithmic prioritization.
- Who would be helped most? Difficult causal inference problem $\rightsquigarrow$ simplifying assumption: those with greatest care needs will benefit most.
- Algorithm designers use health-care costs as a proxy for 'greatest care needs.'
- Obermeyer et. al. investigate, but "Algorithms deployed on large scales are typically proprietary, making it difficult for independent researchers to dissect them."

Obermeyer et. al. 2019

- Industry Goal: target patients for 'high-risk care management.' A broader case of algorithmic prioritization.
- Who would be helped most? Difficult causal inference problem $\rightsquigarrow$ simplifying assumption: those with greatest care needs will benefit most.
- Algorithm designers use health-care costs as a proxy for 'greatest care needs.'
- Obermeyer et. al. investigate, but "Algorithms deployed on large scales are typically proprietary, making it difficult for independent researchers to dissect them."
- Here they are able to get necessary data and measure health with ICD codes and lab results. (this is still a measurement call!)

Obermeyer et. al. 2019

- Industry Goal: target patients for 'high-risk care management.' A broader case of algorithmic prioritization.
- Who would be helped most? Difficult causal inference problem $\rightsquigarrow$ simplifying assumption: those with greatest care needs will benefit most.
- Algorithm designers use health-care costs as a proxy for 'greatest care needs.'
- Obermeyer et. al. investigate, but "Algorithms deployed on large scales are typically proprietary, making it difficult for independent researchers to dissect them."
- Here they are able to get necessary data and measure health with ICD codes and lab results. (this is still a measurement call!)

Do White patients and Black patients with the same risk score have the same average health?

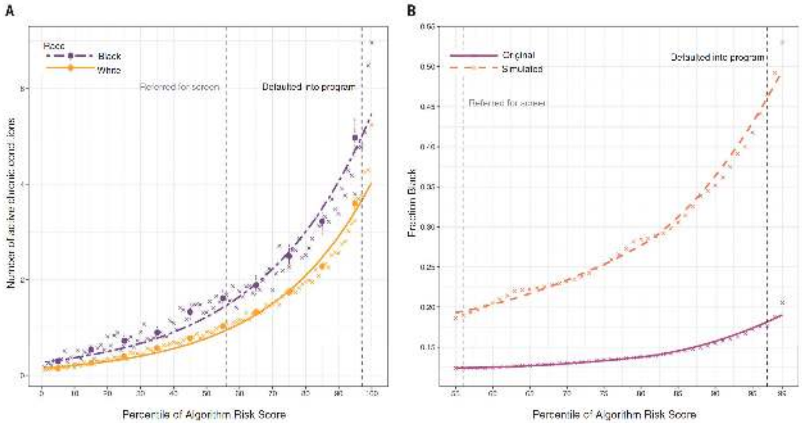


Fig. 1. Number of chronic illnesses versus algorithm-predicted risk, by race. (**A**) Mean number of chronic conditions by race, plotted against algorithm risk score. (**B**) Fraction of Black patients at or above a given risk score for the original algorithm ("original") and for a simulated scenario that removes algorithmic bias ("simulated": at each threshold of risk, defined at a given percentile on the x axis, healthier Whites above the threshold are replaced with less healthy Blacks below the threshold, until the marginal patient is equally healthy). The × symbols show risk percentiles by race; circles show risk deciles with 95% confidence intervals clustered by patient. The dashed vertical lines show the auto-identification threshold (the black line, which denotes the 97th percentile) and the screening threshold (the gray line, which denotes the 55th percentile).

Image Credit: Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. "Dissecting racial bias in an algorithm used to manage the health of populations." Science 366, no. 6464: 447-453.

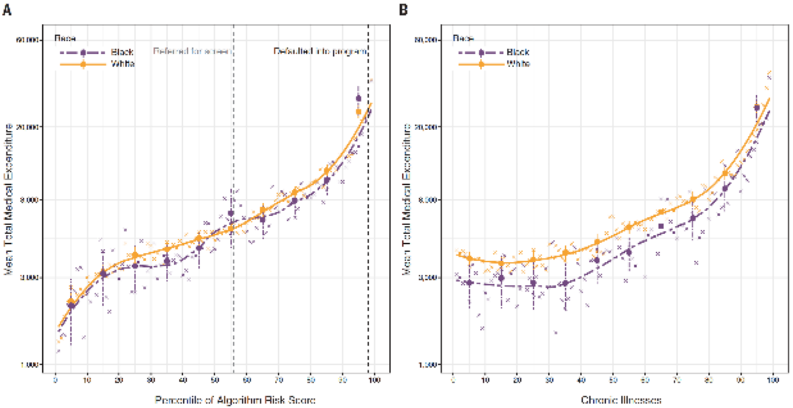


Fig. 3. Costs versus algorithm-predicted risk, and costs versus health, by race. **(A)** Total medical expenditures by race, conditional on algorithm risk score. The dashed vertical lines show the auto-identification threshold (black line: 97th percentile) and the screening threshold (gray line: 55th percentile). **(B)** Total medical expenditures by race, conditional on number of chronic conditions. The × symbols show risk percentiles; circles show risk deciles with 95% confidence intervals clustered by patient. The y axis uses a log scale.

Image Credit: Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. "Dissecting racial bias in an algorithm used to manage the health of populations." Science 366, no. 6464: 447-453.

SOCIAL SCIENCE

Assessing risk, automating racism

A health care algorithm reflects underlying racial bias in society

By **Ruha Benjamin**

As more organizations and industries adopt digital tools to identify risk and allocate resources, the automation of racial discrimination is a growing concern. Social scientists have been at the forefront of studying the historical, political, economic, and ethical dimensions of such tools (*1–3*). But most analysts do not have access to widely

era, the intention to deepen racial inequities was more explicit, today coded inequity is perpetuated precisely because those who design and adopt such tools are not thinking carefully about systemic racism.

Obermeyer *et al.* gained access to the training data, algorithm, and contextual data for one of the largest commercial tools used by health insurers to assess the health profiles for millions of patients. The purpose of the tool is to identify a subset of patients who re-

beyond the algorithm developers by constructing a more fine-grained measure of health outcomes, by extracting and cleaning data from electronic health records to determine the severity, not just the number, of conditions. Crucially, they found that so long as the tool remains effective at predicting costs, the outputs will continue to be racially biased by design, even as they may not explicitly attempt to take race into account. For this reason, Obermeyer *et al.*

Upshot

Upshot

- Black patients substantially sicker than same risk White patients.

Upshot

- Black patients substantially sicker than same risk White patients.
- Changes would yield an increase in Black patients receiving help from 17.7% to 46.5%!

Upshot

- Black patients substantially sicker than same risk White patients.
- Changes would yield an increase in Black patients receiving help from 17.7% to 46.5%!
- Note: machine learning won't fix this because at a technical level it is a measurement and/or research design issue.

Upshot

- Black patients substantially sicker than same risk White patients.
- Changes would yield an increase in Black patients receiving help from 17.7% to 46.5%!
- Note: machine learning won't fix this because at a technical level it is a measurement and/or research design issue.
- Many core causes including: systemic racism driving data patterns and lack of diversity leads to nobody checking for bias.

Upshot

- Black patients substantially sicker than same risk White patients.
- Changes would yield an increase in Black patients receiving help from 17.7% to 46.5%!
- Note: machine learning won't fix this because at a technical level it is a measurement and/or research design issue.
- Many core causes including: systemic racism driving data patterns and lack of diversity leads to nobody checking for bias.

> *Labels matter greatly, not only in algorithm design but also in algorithm analysis. Black patients do not "cost less," so much as they are valued less.*
>
> *Benjamin, 2019*

1. Review
2. Frameworks for Discovery and Measurement
3. Example Application: Political Emails
4. Frameworks for Measurement
5. A Failure of Measurement
6. Introduction to the Fragile Families Challenge
7. Regression with Regularization
8. Logistic Regression
9. Generalized Additive Models

1 Review

2 Frameworks for Discovery and Measurement

3 Example Application: Political Emails

4 Frameworks for Measurement

5 A Failure of Measurement

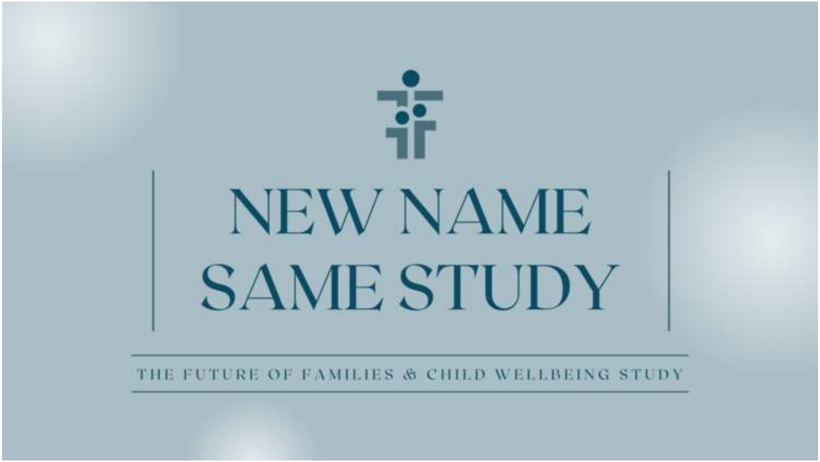
6 Introduction to the Fragile Families Challenge

7 Regression with Regularization

8 Logistic Regression

9 Generalized Additive Models

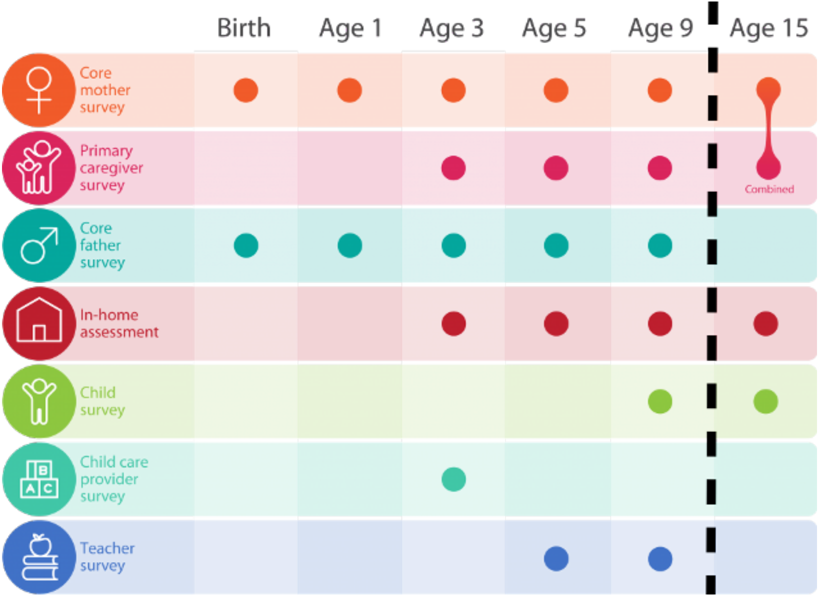
Fragile Families
& Child Wellbeing Study
PRINCETON | COLUMBIA

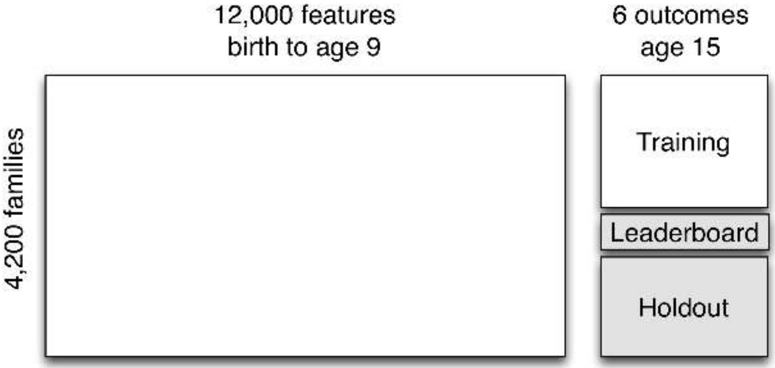

NEW NAME
SAME STUDY
THE FUTURE OF FAMILIES & CHILD WELLBEING STUDY

Future of Families

& Child Wellbeing Study

PRINCETON | COLUMBIA


Birth
Age 1
Age 3
Age 5
Age 9
Age 15
Core mother survey
Primary caregiver survey
Combined
Core father survey
In-home assessment
Child survey
Child care provider survey
Teacher survey


12,000 features
birth to age 9
6 outcomes
age 15
4,200 families
Training
Leaderboard
Holdout

Six Outcomes of Fragile Families Challenge

Outcomes measured at age 15.

- GPA
- Material Hardship
- Grit
- Evicted
- Job training
- Job loss

GPA

IAN LUNDBERG

MARCH 6, 2017

GPA

Uncategorized   No comments

GPA measures academic achievement.

We want to know:

- What helps disadvantaged children to beat the odds and succeed academically?
- What derails children so that they perform unexpectedly poorly?

SURVEY QUESTION

B20. At the {most recent grading period/last grading period in the spring} what was your grade in ...

		A	B	C	D OR LOWER	NO GRADE OR PASS/FAIL	REF	DK	N/A HOMESCHOOLED
B20A	English or language arts? ..	1	2	3	4	5	-1	-2	7 → GO TO B22A
B20B	Math?	1	2	3	4	5	-1	-2	7 → GO TO B22A
B20C	History or social studies? ..	1	2	3	4	5	-1	-2	7 → GO TO B22A
B20D	Science?	1	2	3	4	5	-1	-2	7 → GO TO B22A

HOW WE CLEANED THE DATA

Our measure of GPA is self-reported by the child at approximately age 15. We marked as NA the

Eviction

IAN LUNDBERG
MARCH 6, 2017

EVICTION

Uncategorized · No comments

Eviction is a traumatic experience in which families are forced from their homes for not paying the rent or mortgage.

We want to know: As children transition into adulthood, does eviction cause negative outcomes?

SURVEY QUESTION

When children were about 15 years old, each child's primary caregiver was asked the following question:

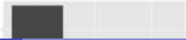
J51. Since {MONTH AND YEAR COHORT CITY FIELDED IN YR 9}, were you evicted from your home or apartment for not paying the rent or mortgage?

YES..1
NO...2
REFUSED................................-1
DON'T KNOW..........................-2

HOW WE CLEANED THE DATA

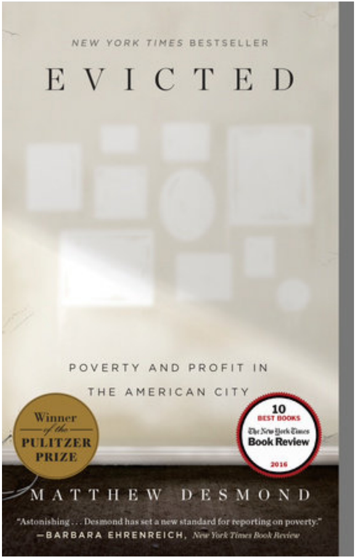
Those who did not participate in the age 15 interview, as well as those who refused (-1) or didn't know (-2), were coded as NA. Those who responded "Yes" were coded 1, and those who responded "No" were coded 0. We additionally coded as 1 a small group of respondents who answered in a previous question that they were evicted in the past year, and thus were skipped over this question.

DISTRIBUTION IN THE TRAINING SET



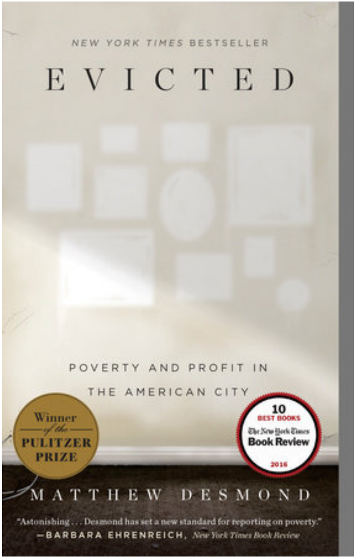
Considerations in Measuring Eviction

- 'Have you ever been evicted?' type questions seem very straightforward.



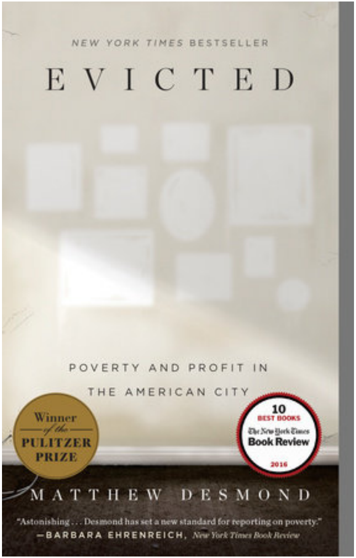
Considerations in Measuring Eviction

- 'Have you ever been evicted?' type questions seem very straightforward.
- People may only answer this question affirmatively if a sheriff shows up with a court order.



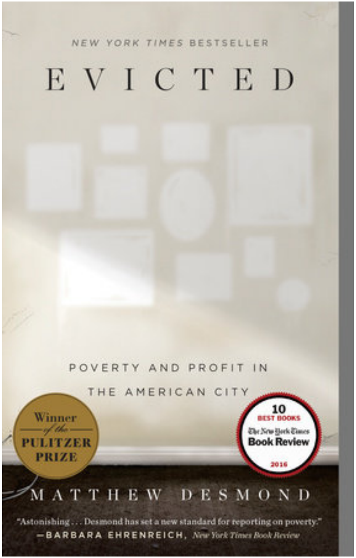
Considerations in Measuring Eviction

- 'Have you ever been evicted?' type questions seem very straightforward.
- People may only answer this question affirmatively if a sheriff shows up with a court order.
- Work from Desmond's MARS survey has shown that informal evictions are twice as common as formal evictions.



Considerations in Measuring Eviction

- 'Have you ever been evicted?' type questions seem very straightforward.
- People may only answer this question affirmatively if a sheriff shows up with a court order.
- Work from Desmond's MARS survey has shown that informal evictions are twice as common as formal evictions.
- MARS survey asks a series of six questions.



MARS Survey Questions

involuntary removals and ending with voluntary moves. These questions are asked of each move that respondents experienced during the previous two years:

(1) An eviction is when your landlord forces you to move when you don't want to. Were you, or a person you were staying with, evicted?
(2) Did you, or a person you were staying with, [leave after receiving] an eviction notice?
(3) Did you move away from this place because your landlord told you, or a person you were staying with, to leave?
(4) Did you move away from this place because you, or a person you were staying with, missed a rent payment and thought that if you didn't move, you would be evicted?
(5) Did you move away from this place because the city condemned the property and forced you to leave?
(6) Did you move away from this place because (a) the landlord raised the rent? (b) the neighborhood was dangerous? (c) the landlord wouldn't fix anything, and your place was getting run down? (d) the landlord went into foreclosure?

Reasoning with Imperfect Data

- Do you have to throw out a measure because it isn't perfect?

Reasoning with Imperfect Data

- Do you have to throw out a measure because it isn't perfect?
- No! However it is important to understand how your measure deviates from the underlying construct so you can reason correctly.

Reasoning with Imperfect Data

- Do you have to throw out a measure because it isn't perfect?
- No! However it is important to understand how your measure deviates from the underlying construct so you can reason correctly.
- Arguably the Fragile Families question provides an approximate lower bound on the prevalence of eviction. Using data from studies like MARS we can even guess by how much.

Reasoning with Imperfect Data

- Do you have to throw out a measure because it isn't perfect?
- No! However it is important to understand how your measure deviates from the underlying construct so you can reason correctly.
- Arguably the Fragile Families question provides an approximate lower bound on the prevalence of eviction. Using data from studies like MARS we can even guess by how much.
- No study is perfect—even designed data can have flaws (more in Weeks 7–9).

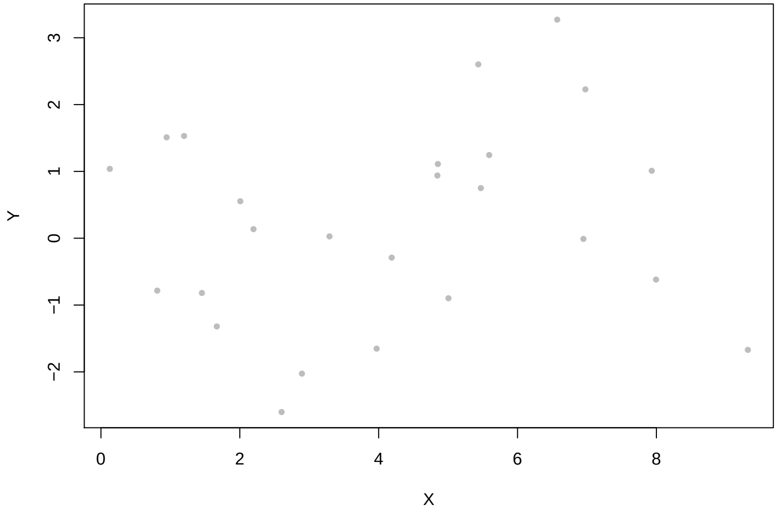
Reasoning with Imperfect Data

- Do you have to throw out a measure because it isn't perfect?
- No! However it is important to understand how your measure deviates from the underlying construct so you can reason correctly.
- Arguably the Fragile Families question provides an approximate lower bound on the prevalence of eviction. Using data from studies like MARS we can even guess by how much.
- No study is perfect—even designed data can have flaws (more in Weeks 7–9).
- However, beware the risk of unintended consequences

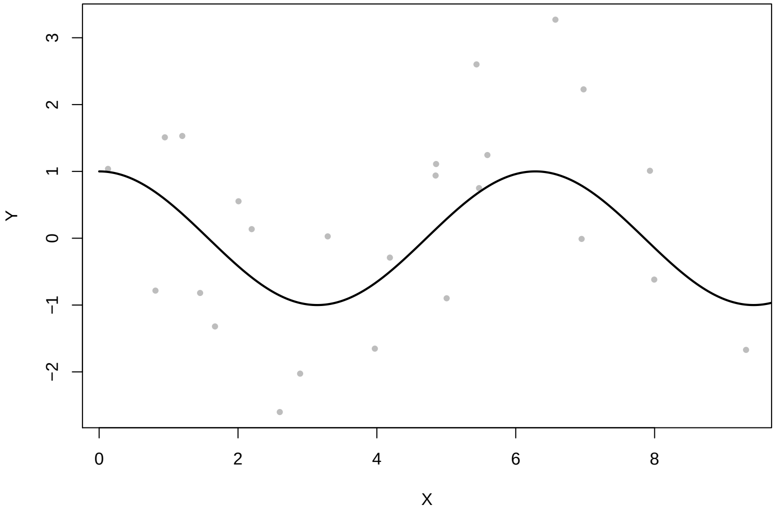
1. Review
2. Frameworks for Discovery and Measurement
3. Example Application: Political Emails
4. Frameworks for Measurement
5. A Failure of Measurement
6. Introduction to the Fragile Families Challenge
7. Regression with Regularization
8. Logistic Regression
9. Generalized Additive Models

1. Review
2. Frameworks for Discovery and Measurement
3. Example Application: Political Emails
4. Frameworks for Measurement
5. A Failure of Measurement
6. Introduction to the Fragile Families Challenge
7. **Regression with Regularization**
8. Logistic Regression
9. Generalized Additive Models

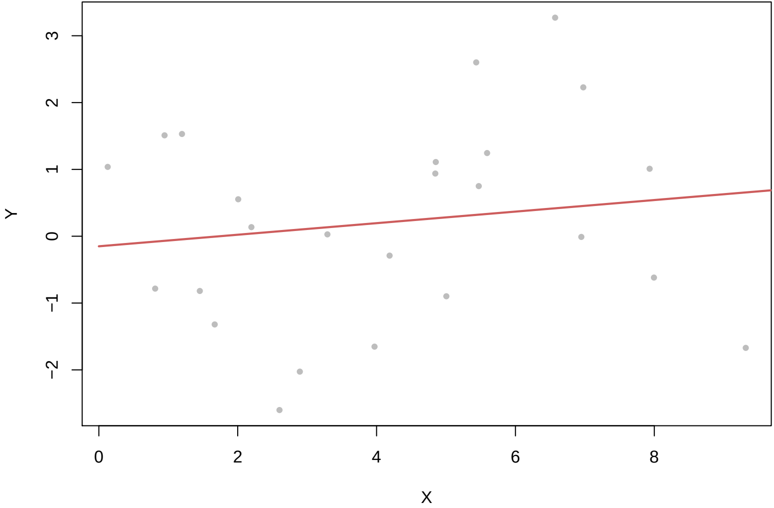
Visual Intuition



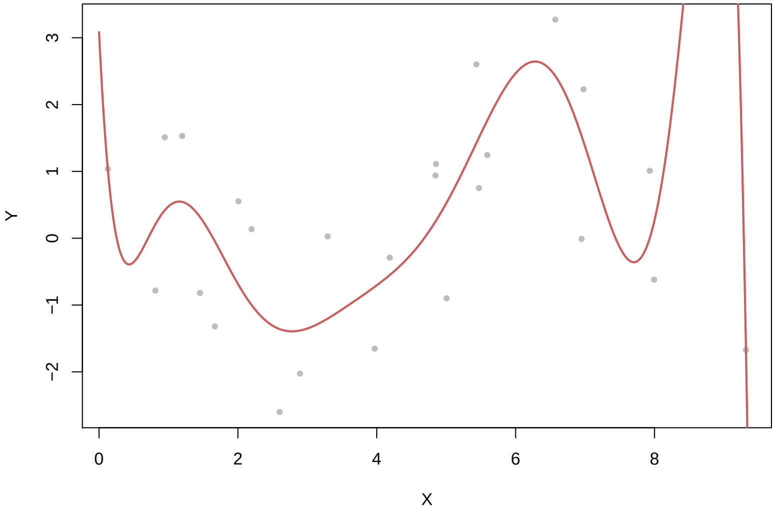
Visual Intuition



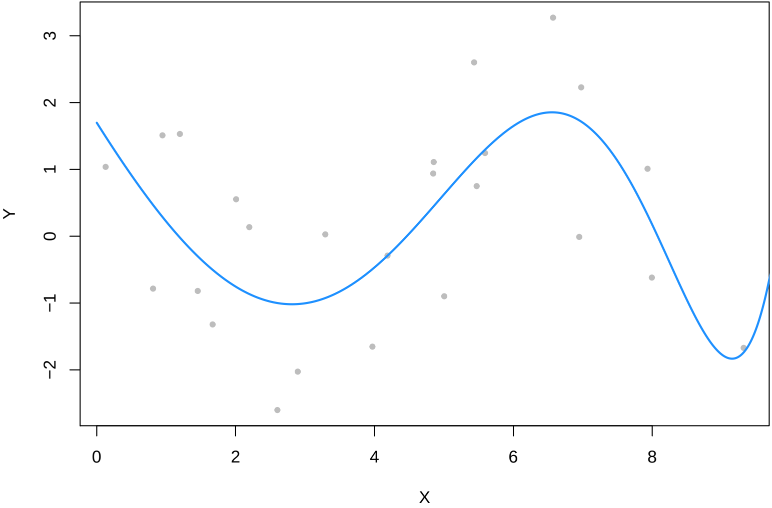
Visual Intuition



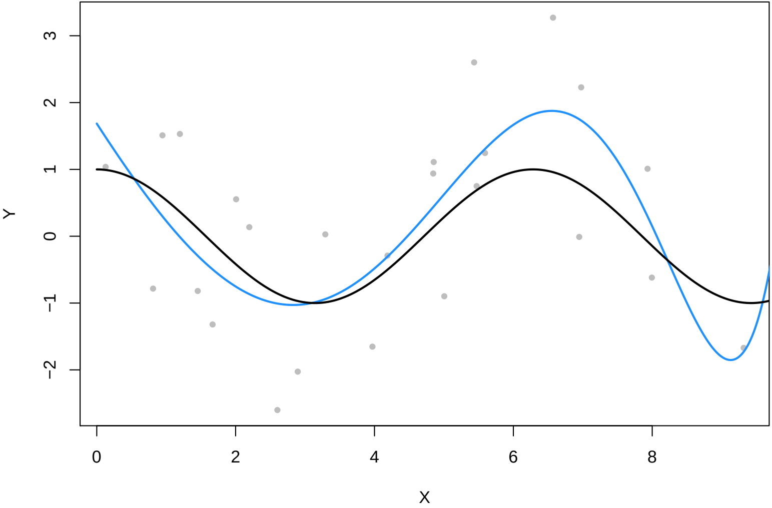
Visual Intuition



Visual Intuition



Visual Intuition



Mathematical Form

- In standard linear regression we minimize the following criterion:

$$\sum_{i=1}^{n}\left(y_i - \left(\beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}\right)\right)^2$$

Mathematical Form

- In standard linear regression we minimize the following criterion:

$$\sum_{i=1}^{n}\left(y_i - \left(\beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}\right)\right)^2$$

- In regularization we add a penalty such that we want to minimize:

$$\sum_{i=1}^{n}\left(y_i - \left(\beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}\right)\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|^q$$

  for some positive regularization penalty λ and a value q which determines the type of regularizer

Mathematical Form

- In standard linear regression we minimize the following criterion:

$$\sum_{i=1}^{n}\left(y_i - \left(\beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}\right)\right)^2$$

- In regularization we add a penalty such that we want to minimize:

$$\sum_{i=1}^{n}\left(y_i - \left(\beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}\right)\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|^q$$

  for some positive regularization penalty λ and a value q which determines the type of regularizer

- As the coefficients get larger, the overall penalty increases.

Mathematical Form

- In standard linear regression we minimize the following criterion:

$$\sum_{i=1}^{n}\left(y_i-\left(\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}\right)\right)^2$$

- In regularization we add a penalty such that we want to minimize:

$$\sum_{i=1}^{n}\left(y_i-\left(\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}\right)\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|^q$$

  for some positive regularization penalty λ and a value q which determines the type of regularizer
- As the coefficients get larger, the overall penalty increases.
- The math here is for linear regression but works for logistic regression and other models.

Two Canonical Regularizers

Two Canonical Regularizers

- There are many regularizers.

Two Canonical Regularizers

- There are many regularizers.
- They have really great names too:

Two Canonical Regularizers

- There are many regularizers.
- They have really great names too: tikhonov regularization, ridge regression,

Two Canonical Regularizers

- There are many regularizers.
- They have really great names too: tikhonov regularization, ridge regression, least absolute shrinkage and selection operator (LASSO),

Two Canonical Regularizers

- There are many regularizers.
- They have really great names too: tikhonov regularization, ridge regression, least absolute shrinkage and selection operator (LASSO), elastic net,

Two Canonical Regularizers

- There are many regularizers.
- They have really great names too: tikhonov regularization, ridge regression, least absolute shrinkage and selection operator (LASSO), elastic net, grouped lasso, fused lasso, adaptive lasso, gamma lasso, Bayesian lasso, square-root lasso, hierarchical adaptive lasso,

Two Canonical Regularizers

- There are many regularizers.
- They have really great names too: tikhonov regularization, ridge regression, least absolute shrinkage and selection operator (LASSO), elastic net, grouped lasso, fused lasso, adaptive lasso, gamma lasso, Bayesian lasso, square-root lasso, hierarchical adaptive lasso, smoothly clipped absolute deviation, horseshoe, bridge regression ...

Two Canonical Regularizers

- There are many regularizers.
- They have really great names too: tikhonov regularization, ridge regression, least absolute shrinkage and selection operator (LASSO), elastic net, grouped lasso, fused lasso, adaptive lasso, gamma lasso, Bayesian lasso, square-root lasso, hierarchical adaptive lasso, smoothly clipped absolute deviation, horseshoe, bridge regression ...
- We will cover two which come up frequently ridge regression ($q = 2$) and LASSO ($q = 1$)

Ridge and Lasso Quick Comparison

Ridge and Lasso Quick Comparison

- Ridge Regression

- Lasso Regression

Ridge and Lasso Quick Comparison

- Ridge Regression
  - minimizes with penalty $\lambda \sum_{j=1}^{p} |\beta_j|^{2}$
- Lasso Regression
  - minimizes with penalty $\lambda \sum_{j=1}^{p} |\beta_j|^{1}$

Ridge and Lasso Quick Comparison

- Ridge Regression
  - minimizes with penalty $\lambda \sum_{j=1}^{p} |\beta_j|^2$
  - draws coefficients smoothly towards zero

    simplified update: $\beta_j^{\text{Ridge}} = \frac{\beta_j^{OLS}}{1+\lambda}$
- Lasso Regression
  - minimizes with penalty $\lambda \sum_{j=1}^{p} |\beta_j|$
  - drops coefficients exactly to zero (induces sparsity)

    simplified update: $\beta_j^{\text{LASSO}} = \text{sign}\left(\beta_j^{OLS}\right)\left(|\beta_j^{OLS}| - \lambda\right)_+$
    - ★ where $\text{sign}(\cdot) \rightsquigarrow 1$ or -1
    - ★ $\left(|\tilde{\beta}_j| - \lambda\right)_+ = \max(|\tilde{\beta}_j| - \lambda, 0)$

Ridge and Lasso Quick Comparison

- Ridge Regression
  - minimizes with penalty $\lambda \sum_{j=1}^{p} |\beta_j|^2$
  - draws coefficients smoothly towards zero

    simplified update: $\beta_j^{\text{Ridge}} = \frac{\beta_j^{OLS}}{1+\lambda}$
  - useful for shrinking functions towards simpler forms
- Lasso Regression
  - minimizes with penalty $\lambda \sum_{j=1}^{p} |\beta_j|$
  - drops coefficients exactly to zero (induces sparsity)

    simplified update: $\beta_j^{\text{LASSO}} = \text{sign}\left(\beta_j^{OLS}\right)\left(|\beta_j^{OLS}| - \lambda\right)_+$
    - where $\text{sign}(\cdot) \rightsquigarrow 1$ or -1
    - $\left(|\tilde{\beta}_j| - \lambda\right)_+ = \max(|\tilde{\beta}_j| - \lambda, 0)$
  - useful for selecting among many coefficients

Constraint Interpretation

We can rewrite penalty as placing a constraint on our optimization problem.

$$\underset{\tilde{\beta}}{\arg\min} \sum_{i=1}^{n} \left(y_i - \left(\tilde{\beta}_0 + \sum_{j=1}^{p} \tilde{\beta}_j x_{ij} \right) \right)^2 \text{ subject to } \sum_{j=1}^{p} \tilde{\beta}_j^2 \leq s$$

or

$$\underset{\tilde{\beta}}{\arg\min} \sum_{i=1}^{n} \left(y_i - \left(\tilde{\beta}_0 + \sum_{j=1}^{p} \tilde{\beta}_j x_{ij} \right) \right)^2 \text{ subject to } \sum_{j=1}^{p} |\tilde{\beta}_j| \leq s$$

where there is a one-to-one correspondence between s and λ.

Constraint Visualization

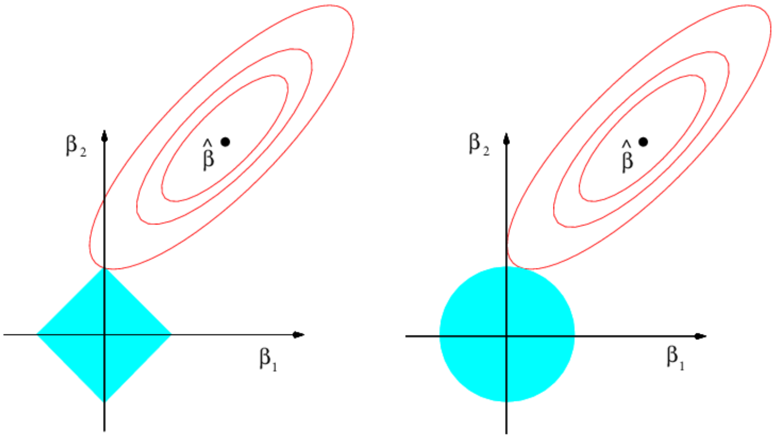


Image Credit: James et al. 2013 *An Introduction to Statistical Learning, with applications in R.*

How to Set λ

- The key parameter here is λ which controls the tradeoff between fitting the loss function and the penalty.

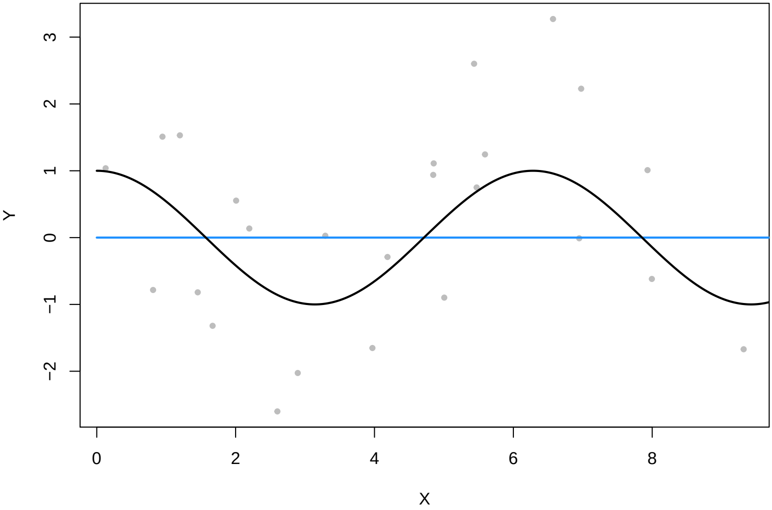
How to Set λ

- The key parameter here is λ which controls the tradeoff between fitting the loss function and the penalty.
- Software like `glmnet` will calculate the whole regularization path—a whole sequence of values of λ from maximal penalization to no penalization.

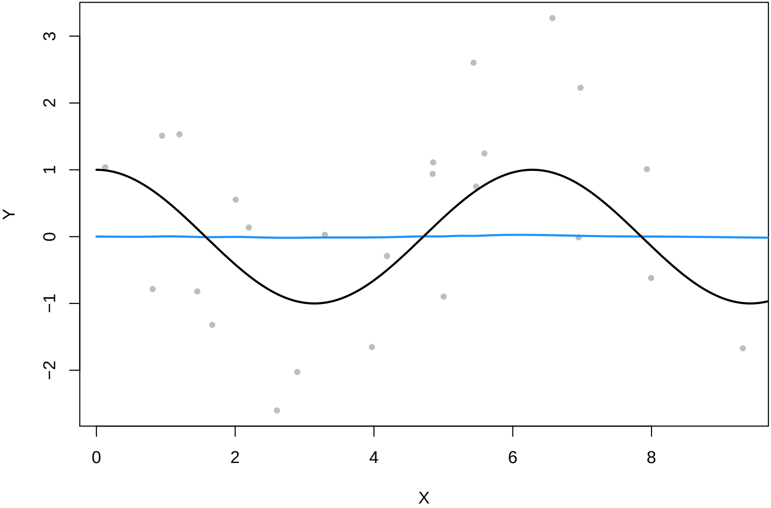
How to Set λ

- The key parameter here is λ which controls the tradeoff between fitting the loss function and the penalty.
- Software like `glmnet` will calculate the whole regularization path—a whole sequence of values of λ from maximal penalization to no penalization.
- We can choose a single value based on cross-validation or a validation set (`glmnet` will even implement this for us with `cv.glmnet`)

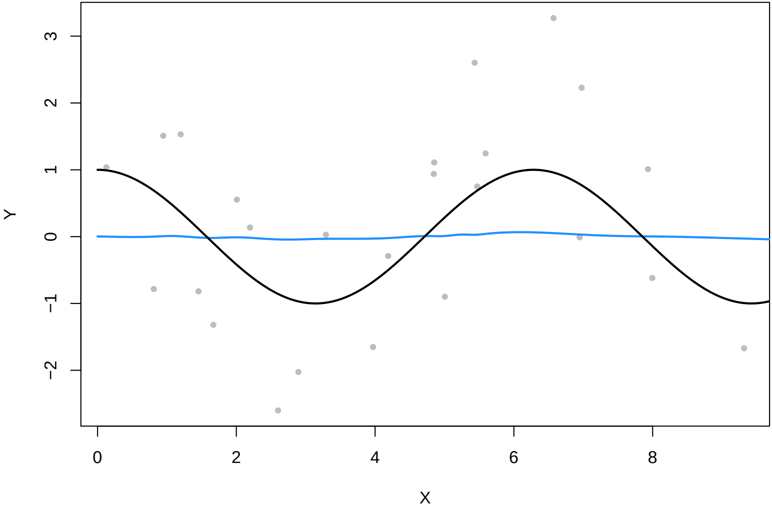
Regularization Path



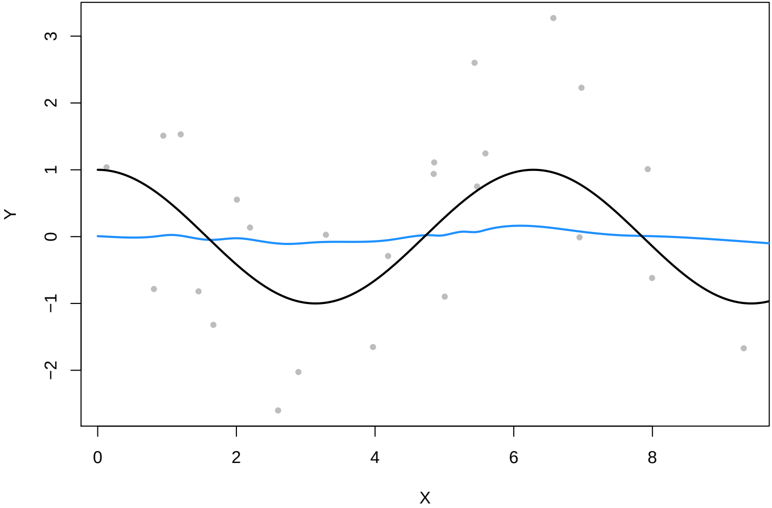
Regularization Path



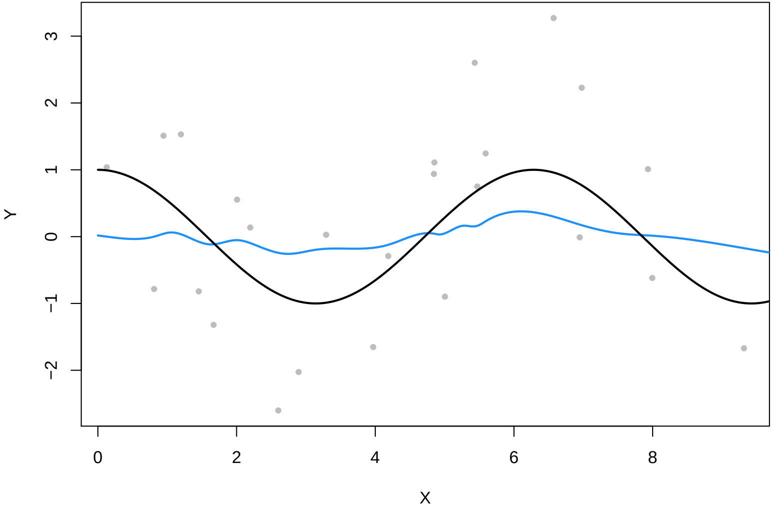
Regularization Path



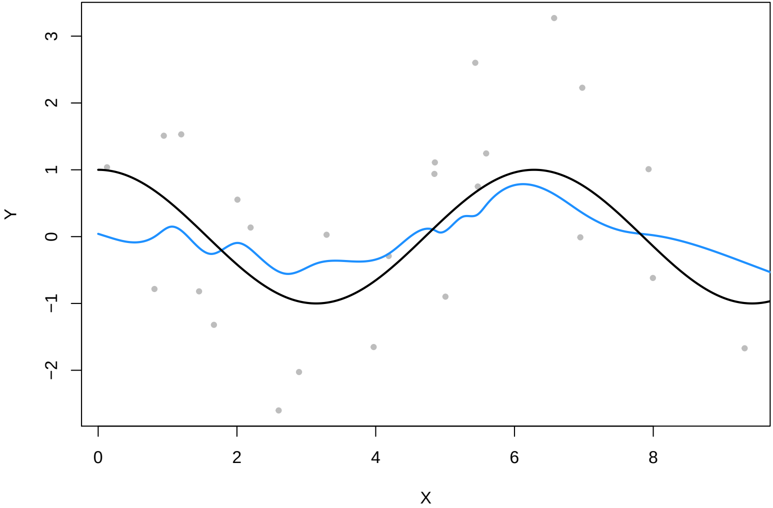
Regularization Path



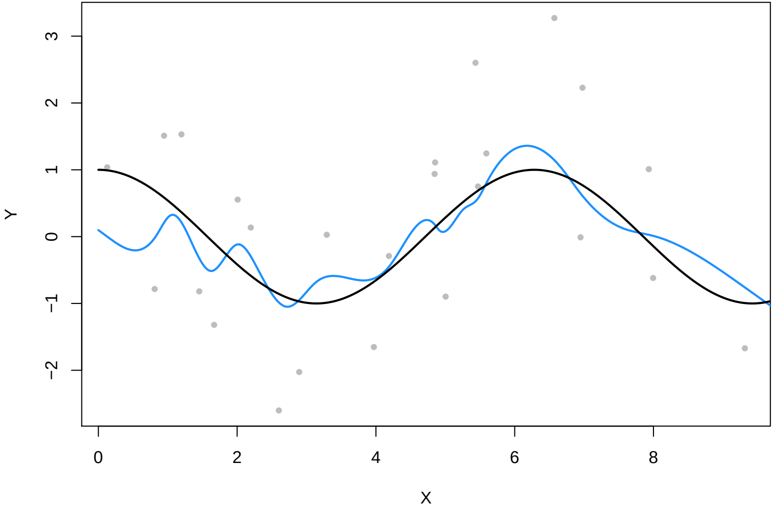
Regularization Path



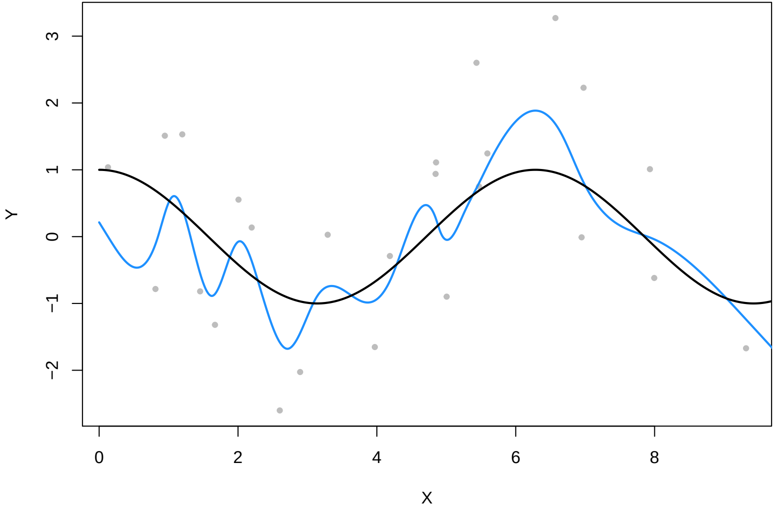
Regularization Path



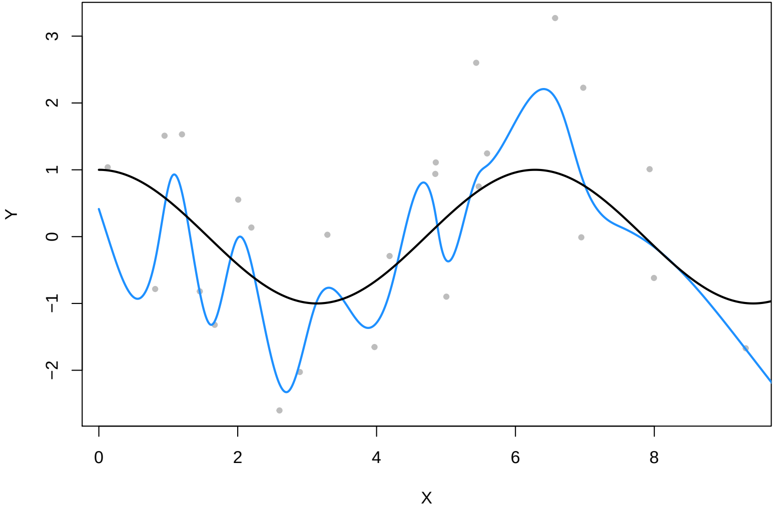
Regularization Path



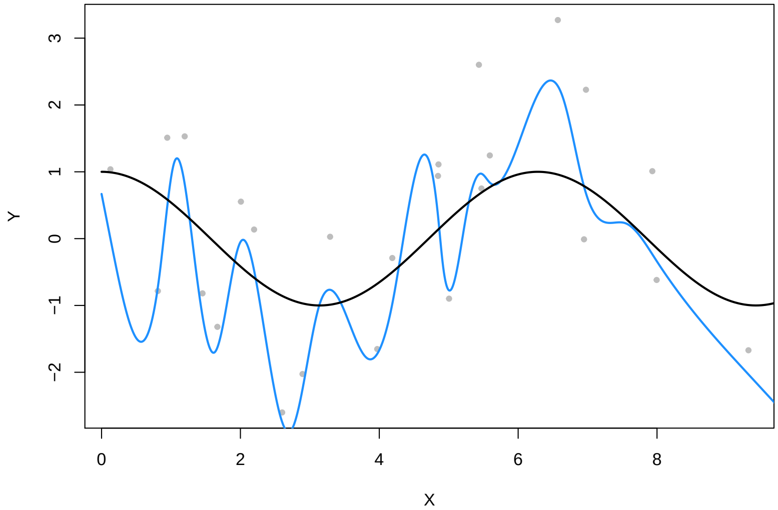
Regularization Path



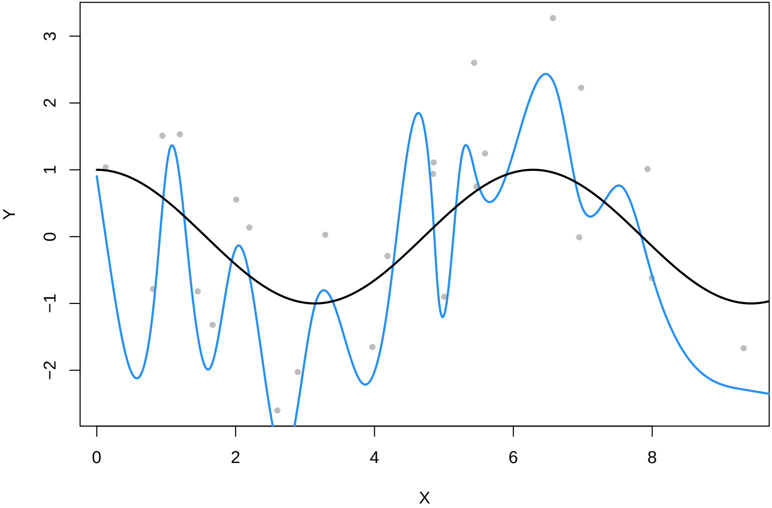
Regularization Path



Regularization Path



Regularization Path



Subtleties to Be Aware Of

- Analyses are now dependent on the scale of the predictors in a way they weren't before.

Subtleties to Be Aware Of

- Analyses are now dependent on the scale of the predictors in a way they weren't before.
- Packages like `glmnet` will standardize the predictors for you but this can also cause complications.

Subtleties to Be Aware Of

- Analyses are now dependent on the scale of the predictors in a way they weren't before.
- Packages like `glmnet` will standardize the predictors for you but this can also cause complications.
- This means that (e.g.) raw and orthogonalized polynomials will no longer yield the same predictions.

Subtleties to Be Aware Of

- Analyses are now dependent on the scale of the predictors in a way they weren't before.
- Packages like `glmnet` will standardize the predictors for you but this can also cause complications.
- This means that (e.g.) raw and orthogonalized polynomials will no longer yield the same predictions.
- We have shown only one kind of regularization (explicity penalization in the objective), but there are many of other techniques that perform similar regularization.

Subtleties to Be Aware Of

- Analyses are now dependent on the scale of the predictors in a way they weren't before.
- Packages like `glmnet` will standardize the predictors for you but this can also cause complications.
- This means that (e.g.) raw and orthogonalized polynomials will no longer yield the same predictions.
- We have shown only one kind of regularization (explicity penalization in the objective), but there are many of other techniques that perform similar regularization.
- Simple linear models effectively provide infinite regularization of non-linear terms by shrinking them to zero—restricting the model class is one extreme way of regularizing.

Summary

- Regularization is a way to smoothly restrict model complexity by penalizing more complex models.

Summary

- Regularization is a way to smoothly restrict model complexity by penalizing more complex models.
- `glmnet` is a nice package for performing lasso regularization (and ridge regression too although more on that next time).

Summary

- Regularization is a way to smoothly restrict model complexity by penalizing more complex models.
- `glmnet` is a nice package for performing lasso regularization (and ridge regression too although more on that next time).
- Lasso yields a sparse solution (exact zeroes for many coefficients) while ridge just shrinks them down.

Summary

- Regularization is a way to smoothly restrict model complexity by penalizing more complex models.
- `glmnet` is a nice package for performing lasso regularization (and ridge regression too although more on that next time).
- Lasso yields a sparse solution (exact zeroes for many coefficients) while ridge just shrinks them down.
- Lasso can be particularly helpful when trying to do regression in the $p > n$ setting.

Summary

- Regularization is a way to smoothly restrict model complexity by penalizing more complex models.
- `glmnet` is a nice package for performing lasso regularization (and ridge regression too although more on that next time).
- Lasso yields a sparse solution (exact zeroes for many coefficients) while ridge just shrinks them down.
- Lasso can be particularly helpful when trying to do regression in the $p > n$ setting.

Going Deeper:

James, Witten, Hastie and Tibshirani (2021) *An Introduction to Statistical Learning with Applications in R, 2nd Edition*. Chapters 6.0–6.2, 6.4
Free online at `www.statlearning.com`

1. Review
2. Frameworks for Discovery and Measurement
3. Example Application: Political Emails
4. Frameworks for Measurement
5. A Failure of Measurement
6. Introduction to the Fragile Families Challenge
7. Regression with Regularization
8. Logistic Regression
9. Generalized Additive Models

1. Review
2. Frameworks for Discovery and Measurement
3. Example Application: Political Emails
4. Frameworks for Measurement
5. A Failure of Measurement
6. Introduction to the Fragile Families Challenge
7. Regression with Regularization
8. Logistic Regression
9. Generalized Additive Models

Last Time: Linear Regression

$$\hat{\beta} = \arg\min_{\tilde{\beta}} \sum_i (y_i - \hat{y}_i)^2$$

$$\hat{y}_i = \tilde{\beta}_0 + \sum_{j=1}^{p} X_{i,j}\tilde{\beta}_j$$

What Happens with Binary Variables?

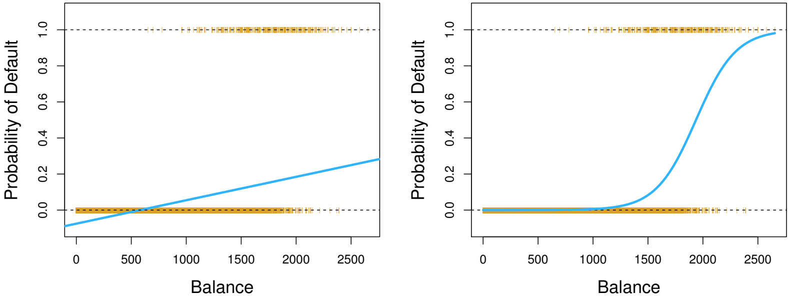


Image Credit: James et al. 2013 *An Introduction to Statistical Learning, with applications in R.*

Logistic Regression

Logistic Regression

- Core Problem: for unbounded inputs, linear prediction $\hat{y}_i$ can always span $-\infty$ to ∞

Logistic Regression

- Core Problem: for unbounded inputs, linear prediction $\hat{y}_i$ can always span $-\infty$ to ∞
- Key idea of logistic regression $\rightarrow$ take the linear form and squish it down to run the range of 0 to 1.

Logistic Regression

- Core Problem: for unbounded inputs, linear prediction $\hat{y}_i$ can always span $-\infty$ to ∞
- Key idea of logistic regression $\rightarrow$ take the linear form and squish it down to run the range of 0 to 1.
- Two equivalent ways of thinking about this:

Logistic Regression

- Core Problem: for unbounded inputs, linear prediction $\hat{y}_i$ can always span $-\infty$ to ∞
- Key idea of logistic regression $\rightarrow$ take the linear form and squish it down to run the range of 0 to 1.
- Two equivalent ways of thinking about this:
    1) linear model of log odds

Logistic Regression

- Core Problem: for unbounded inputs, linear prediction $\hat{y}_i$ can always span $-\infty$ to ∞
- Key idea of logistic regression $\rightarrow$ take the linear form and squish it down to run the range of 0 to 1.
- Two equivalent ways of thinking about this:
  1) linear model of log odds
  2) project line into 0 to 1 space using a convenient function

(1) Log Odds Version

(1) Log Odds Version

- We want something that ranges $-\infty$ to $+\infty$

(1) Log Odds Version

- We want something that ranges $-\infty$ to $+\infty$
- Denote probability $p(Y_i = 1|X_i) = \pi_i$. π_i ranges [0,1].

(1) Log Odds Version

- We want something that ranges $-\infty$ to $+\infty$
- Denote probability $p(Y_i = 1|X_i) = \pi_i$. π_i ranges [0,1].
- The odds are $\frac{\pi_i}{1-\pi_i}$ and range $(0, \infty)$

(1) Log Odds Version

- We want something that ranges $-\infty$ to $+\infty$
- Denote probability $p(Y_i = 1|X_i) = \pi_i$. π_i ranges [0,1].
- The odds are $\frac{\pi_i}{1-\pi_i}$ and range $(0, \infty)$
- The log function takes a value that is in range $(0, \infty)$ and converts to $(-\infty, \infty)$

(1) Log Odds Version

- We want something that ranges $-\infty$ to $+\infty$
- Denote probability $p(Y_i = 1|X_i) = \pi_i$. π_i ranges [0,1].
- The odds are $\frac{\pi_i}{1-\pi_i}$ and range $(0, \infty)$
- The log function takes a value that is in range $(0, \infty)$ and converts to $(-\infty, \infty)$

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{j=1}^{p} X_{i,j}\beta_j$$

$$\pi_i = \frac{1}{1 + e^{-(\beta_0 + \sum_{j=1}^{p} X_{i,j}\beta_j)}}$$

(2) Convenient Function Version

(2) Convenient Function Version

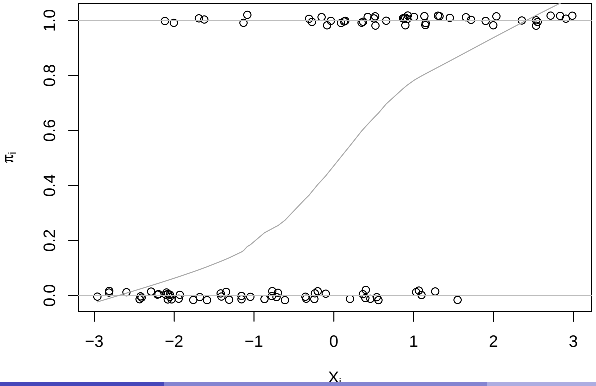
- Another way to think about this: find a function that converts $\beta_0 + \sum_{j=1}^{p} X_{i,j}\beta_j$ from range $-\infty, \infty$ to 0 to 1.

(2) Convenient Function Version

- Another way to think about this: find a function that converts $\beta_0 + \sum_{j=1}^{p} X_{i,j}\beta_j$ from range $-\infty, \infty$ to 0 to 1.
- Different functions lead to (slightly) different models.

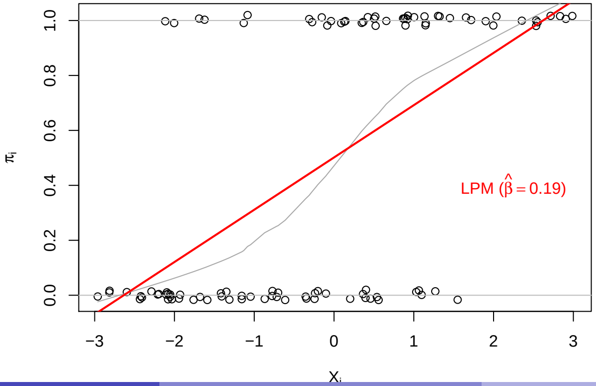
(2) Convenient Function Version

- Another way to think about this: find a function that converts $\beta_0 + \sum_{j=1}^{p} X_{i,j}\beta_j$ from range $-\infty, \infty$ to 0 to 1.
- Different functions lead to (slightly) different models.



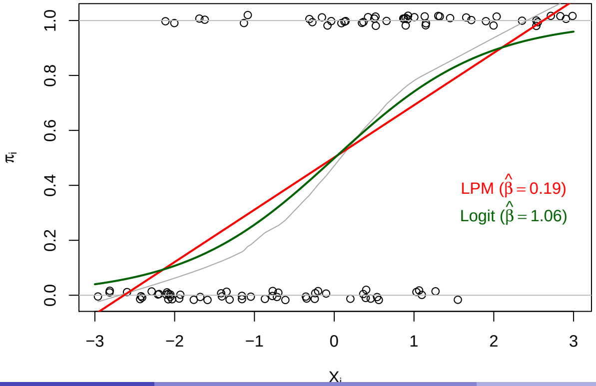
(2) Convenient Function Version

- Another way to think about this: find a function that converts $\beta_0 + \sum_{j=1}^{p} X_{i,j}\beta_j$ from range $-\infty, \infty$ to 0 to 1.
- Different functions lead to (slightly) different models.



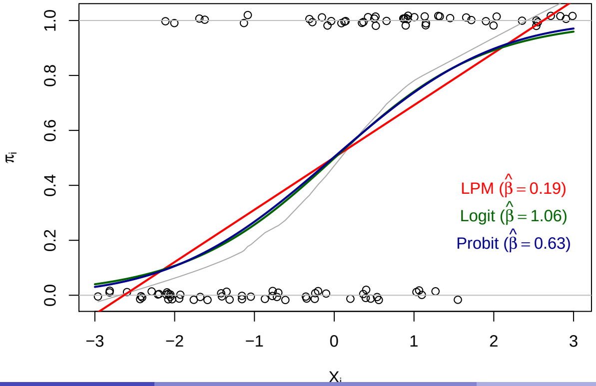
(2) Convenient Function Version

- Another way to think about this: find a function that converts $\beta_0 + \sum_{j=1}^{p} X_{i,j}\beta_j$ from range $-\infty, \infty$ to 0 to 1.
- Different functions lead to (slightly) different models.



(2) Convenient Function Version

- Another way to think about this: find a function that converts $\beta_0 + \sum_{j=1}^{p} X_{i,j}\beta_j$ from range $-\infty, \infty$ to 0 to 1.
- Different functions lead to (slightly) different models.



Loss Function

- We now have a model that connects the probability π_i to the linear predictor $\beta_0 + \sum_{j=1}^{p} X_{i,j}\beta_j$

Loss Function

- We now have a model that connects the probability π_i to the linear predictor $\beta_0 + \sum_{j=1}^{p} X_{i,j}\beta_j$
- To get the loss function we will use maximum likelihood estimation.

Loss Function

- We now have a model that connects the probability π_i to the linear predictor $\beta_0 + \sum_{j=1}^{p} X_{i,j}\beta_j$
- To get the loss function we will use maximum likelihood estimation.
- Core Idea: we are modeling the probability of seeing the data we observe. We want that probability to be as high as possible.

Loss Function

- We now have a model that connects the probability π_i to the linear predictor $\beta_0 + \sum_{j=1}^{p} X_{i,j}\beta_j$
- To get the loss function we will use maximum likelihood estimation.
- Core Idea: we are modeling the probability of seeing the data we observe. We want that probability to be as high as possible.
- In the case of categorical data this leads to what is called the cross entropy loss in machine learning.

Loss Function

- We now have a model that connects the probability π_i to the linear predictor $\beta_0 + \sum_{j=1}^{p} X_{i,j}\beta_j$
- To get the loss function we will use maximum likelihood estimation.
- Core Idea: we are modeling the probability of seeing the data we observe. We want that probability to be as high as possible.
- In the case of categorical data this leads to what is called the cross entropy loss in machine learning.

$$\arg\max \quad \prod_{Y_i=1} \pi_i \prod_{Y_i=0} (1 - \pi_i)$$

Loss Function

- We now have a model that connects the probability π_i to the linear predictor $\beta_0 + \sum_{j=1}^{p} X_{i,j}\beta_j$
- To get the loss function we will use maximum likelihood estimation.
- Core Idea: we are modeling the probability of seeing the data we observe. We want that probability to be as high as possible.
- In the case of categorical data this leads to what is called the cross entropy loss in machine learning.

$$\arg\max \quad \prod_{Y_i=1} \pi_i \prod_{Y_i=0} (1-\pi_i)$$

$$\arg\min \quad -\left(\sum_{Y_i=1} \log\, \pi_i\right) - \left(\sum_{Y_i=0} \log\, (1-\pi_i)\right)$$

Logistic Regression

$$\hat{\beta} = \arg\min_{\tilde{\beta}} - \left(\sum_{Y_i = 1} \log \pi_i \right) - \left(\sum_{Y_i = 0} \log (1 - \pi_i) \right) \tag{1}$$

$$\pi_i = \frac{1}{1 + e^{-(\tilde{\beta}_0 + \sum_{j=1}^{p} X_{i,j} \tilde{\beta}_j)}} \tag{2}$$

Logistic Regression

$$\hat{\beta} = \arg\min_{\tilde{\beta}} - \left(\sum_{Y_i=1} \log\, \pi_i \right) - \left(\sum_{Y_i=0} \log\, (1 - \pi_i) \right) \tag{1}$$

$$\pi_i = \frac{1}{1 + e^{-(\tilde{\beta}_0 + \sum_{j=1}^{p} X_{i,j} \tilde{\beta}_j)}} \tag{2}$$

Key Notes:

- Just as linear regression always makes a line, this will always make a line in log-odds (or more generally an S-shaped curve).

Logistic Regression

$$\hat{\beta} = \arg\min_{\tilde{\beta}} - \left(\sum_{Y_i=1} \log \pi_i \right) - \left(\sum_{Y_i=0} \log (1 - \pi_i) \right) \tag{1}$$

$$\pi_i = \frac{1}{1 + e^{-(\tilde{\beta}_0 + \sum_{j=1}^{p} X_{i,j} \tilde{\beta}_j)}} \tag{2}$$

Key Notes:

- Just as linear regression always makes a line, this will always make a line in log-odds (or more generally an S-shaped curve).
- We can generalize this in the same way we did for linear regression by making the line more flexible with basis functions.

Logistic Regression

$$\hat{\beta} = \arg\min_{\tilde{\beta}} - \left(\sum_{Y_i=1} \log \pi_i \right) - \left(\sum_{Y_i=0} \log (1 - \pi_i) \right) \qquad (1)$$

$$\pi_i = \frac{1}{1 + e^{-(\tilde{\beta}_0 + \sum_{j=1}^{p} X_{i,j} \tilde{\beta}_j)}} \qquad (2)$$

Key Notes:

- Just as linear regression always makes a line, this will always make a line in log-odds (or more generally an S-shaped curve).
- We can generalize this in the same way we did for linear regression by making the line more flexible with basis functions.
- Predictions are in the form of probability that the outcome is 1.

Generalization to Many Outcome Categories

- This strategy works in the general setting of more than two outcome categories.
  - Social Science Name: multinomial logistic regression
  - Machine Learning Name: softmax classifier

Generalization to Many Outcome Categories

- This strategy works in the general setting of more than two outcome categories.
  - Social Science Name: multinomial logistic regression
  - Machine Learning Name: softmax classifier
- New setup:

$$\pi_{ik} = \Pr(Y_i = k \mid X_i) = \frac{e^{\beta_{0,k} + \sum_{j=1}^{p} X_{i,j}\beta_{j,k}}}{\sum_{k'=1}^{K} e^{\beta_{0,k'} + \sum_{j=1}^{p} X_{i,j}\beta_{j,k'}}}$$

Generalization to Many Outcome Categories

- This strategy works in the general setting of more than two outcome categories.
  - Social Science Name: multinomial logistic regression
  - Machine Learning Name: softmax classifier
- New setup:

$$\pi_{ik} = \Pr(Y_i = k \mid X_i) = \frac{e^{\beta_{0,k} + \sum_{j=1}^{p} X_{i,j}\beta_{j,k}}}{\sum_{k'=1}^{K} e^{\beta_{0,k'} + \sum_{j=1}^{p} X_{i,j}\beta_{j,k'}}}$$

- Predictions give a probability of falling in each outcome category.

Subtleties to Be Aware Of

Subtleties to Be Aware Of

- If some variable predicts the outcome perfectly in the training set you can get a problem called perfect separation where one or more coefficient pushes towards ∞ or $-\infty$.

Subtleties to Be Aware Of

- If some variable predicts the outcome perfectly in the training set you can get a problem called perfect separation where one or more coefficient pushes towards ∞ or $-\infty$.
- The uncertainty of logistic regression is connected to the rarest class.

Subtleties to Be Aware Of

- If some variable predicts the outcome perfectly in the training set you can get a problem called perfect separation where one or more coefficient pushes towards ∞ or $-\infty$.
- The uncertainty of logistic regression is connected to the rarest class.
- The coefficients do not have the same interpretation as regression coefficients.

Summary of Logistic Regression

Summary of Logistic Regression

- Works for binary or categorical outcomes

Summary of Logistic Regression

- Works for binary or categorical outcomes
- Provides predictions in probabilities

Summary of Logistic Regression

- Works for binary or categorical outcomes
- Provides predictions in probabilities

Going Deeper:

James, Witten, Hastie and Tibshirani (2021)
An Introduction to Statistical Learning with Applications in R. Chapters 4.0–4.3
Free online at `www.statlearning.com`

1. Review
2. Frameworks for Discovery and Measurement
3. Example Application: Political Emails
4. Frameworks for Measurement
5. A Failure of Measurement
6. Introduction to the Fragile Families Challenge
7. Regression with Regularization
8. Logistic Regression
9. Generalized Additive Models

1. Review
2. Frameworks for Discovery and Measurement
3. Example Application: Political Emails
4. Frameworks for Measurement
5. A Failure of Measurement
6. Introduction to the Fragile Families Challenge
7. Regression with Regularization
8. Logistic Regression
9. Generalized Additive Models

Generalized Additive Models (GAM)

Recall the linear model,

$$\hat{y}_i = \beta_0 + X_{1i}\beta_1 + X_{2i}\beta_2 + X_{3i}\beta_3 + \ldots$$

Generalized Additive Models (GAM)

Recall the linear model,

$$\hat{y}_i = \beta_0 + X_{1i}\beta_1 + X_{2i}\beta_2 + X_{3i}\beta_3 + \dots$$

For GAMs, we maintain additivity, but instead of imposing termwise linearity we allow flexible functional forms for each explanatory variable, where $s_1(\cdot), s_2(\cdot),$ and $s_3(\cdot)$ are smooth functions that are estimated from the data:

Generalized Additive Models (GAM)

Recall the linear model,

$$\hat{y}_i = \beta_0 + X_{1i}\beta_1 + X_{2i}\beta_2 + X_{3i}\beta_3 + \dots$$

For GAMs, we maintain additivity, but instead of imposing termwise linearity we allow flexible functional forms for each explanatory variable, where $s_1(\cdot), s_2(\cdot),$ and $s_3(\cdot)$ are smooth functions that are estimated from the data:

$$\hat{y}_i = \beta_0 + s_1(X_{1i}) + s_2(X_{2i}) + s_3(X_{3i}) + \dots$$

Generalized Additive Models (GAM)

$$\hat{y}_i = \beta_0 + s_1(X_{1i}) + s_2(X_{2i}) + s_3(X_{3i}) + \dots$$

- $s_j(\cdot)$ are usually estimated with locally weighted regression smoothers or cubic smoothing splines (but many approaches are possible)

Generalized Additive Models (GAM)

$$\hat{y}_i = \beta_0 + s_1(X_{1i}) + s_2(X_{2i}) + s_3(X_{3i}) + \dots$$

- $s_j(\cdot)$ are usually estimated with locally weighted regression smoothers or cubic smoothing splines (but many approaches are possible)
- Theory and estimation are somewhat involved, but they are easy to use:
  - ```
    gam.out <- gam(y~s(x1)+s(x2)+x3)
    plot(gam.out)
    ```
  - Multiple functions but I recommend `mgcv` package
    ```

Generalized Additive Models (GAM)

The GAM approach can be extended to allow interactions ($s_{12}(\cdot)$) between explanatory variables, but this eats up degrees of freedom so you need a lot of data.

$$\hat{y}_i = \beta_0 + s_{12}(X_{1i}, X_{2i}) + s_3(X_{3i})$$

Generalized Additive Models (GAM)

The GAM approach can be extended to allow interactions ($s_{12}(\cdot)$) between explanatory variables, but this eats up degrees of freedom so you need a lot of data.

$$\hat{y}_i = \beta_0 + s_{12}(X_{1i}, X_{2i}) + s_3(X_{3i})$$

It can also be used for hybrid models where we model some variables as linear and others with a flexible function:

$$\hat{y}_i = \beta_0 + \beta_1 X_{1i} + s_2(X_{2i}) + s_3(X_{3i})$$

GAM Fit

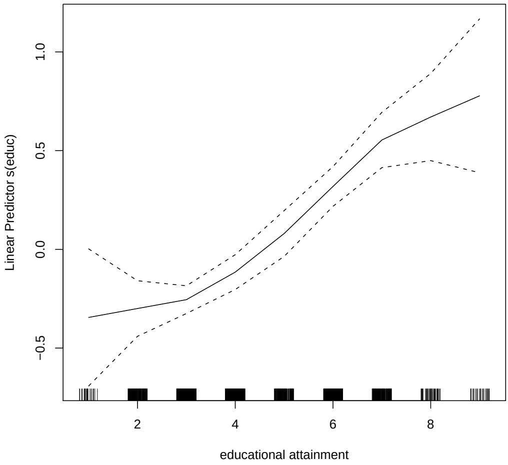


Image Credit: Jens Hainmueller

GAM Fit

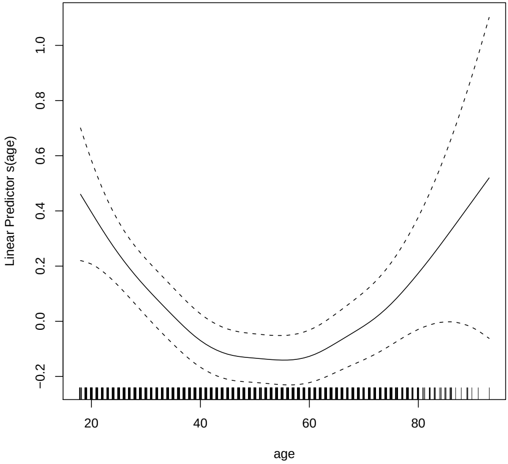


Image Credit: Jens Hainmueller

GAM Fit

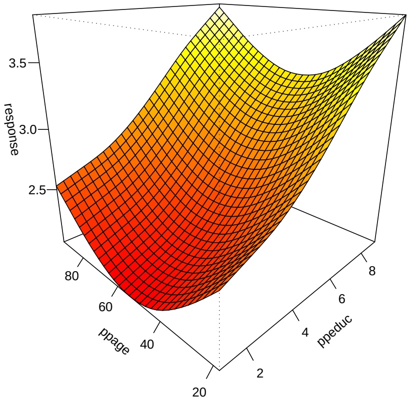


Image Credit: Jens Hainmueller

GAMs using mgcv

- There is an avalanche of small, consequential decisions when doing machine learning. Many strategies that seem perfectly feasible on paper work poorly in practice.

GAMs using `mgcv`

- There is an avalanche of small, consequential decisions when doing machine learning. Many strategies that seem perfectly feasible on paper work poorly in practice.
- Simon Wood and colleagues have spent years getting this right for GAMs and implementing in `mgcv`.

GAMs using `mgcv`

- There is an avalanche of small, consequential decisions when doing machine learning. Many strategies that seem perfectly feasible on paper work poorly in practice.
- Simon Wood and colleagues have spent years getting this right for GAMs and implementing in `mgcv`.
- This includes careful choices of tuning parameters, extensions to all kinds of generalized linear models, visualizations, confidence intervals etc.

GAMs using `mgcv`

- There is an avalanche of small, consequential decisions when doing machine learning. Many strategies that seem perfectly feasible on paper work poorly in practice.
- Simon Wood and colleagues have spent years getting this right for GAMs and implementing in `mgcv`.
- This includes careful choices of tuning parameters, extensions to all kinds of generalized linear models, visualizations, confidence intervals etc.
- The big limitation is limited interactivity, but in many cases that might be completely fine.

Summary

- GAMs with `mgcv` offer a nice turnkey solution to doing additive smoothing.

Summary

- GAMs with `mgcv` offer a nice turnkey solution to doing additive smoothing.
- Additivity makes it easy to interpret what is going on and `mgcv` comes with nice plotting options.

Summary

- GAMs with `mgcv` offer a nice turnkey solution to doing additive smoothing.
- Additivity makes it easy to interpret what is going on and `mgcv` comes with nice plotting options.
- We have skimmed over many, many details...

Going Deeper:

Wood, Simon (2017) *Generalized Additive Models An Introduction with R, Second Edition*.

This Week in Review

This Week in Review

- Discovery and Measurement!
- An Introduction to Future of Families and the Fragile Families Challenge!
- Methods: regularization and GAMs

To go deeper: Gerring, John. 2012. *Social Science Methodology: A Unified Framework, Second Edition*. Chapters 5–7.

This Week in Review

- Discovery and Measurement!
- An Introduction to Future of Families and the Fragile Families Challenge!
- Methods: regularization and GAMs

To go deeper: Gerring, John. 2012. *Social Science Methodology: A Unified Framework, Second Edition*. Chapters 5–7.

Reading Preview for Prediction within a Population:

- Salganik et. al. 2020. "Measuring the predictability of life outcomes with a scientific mass collaboration" *Proceedings of the National Academy of Sciences*.
- Lundberg, et al. 2024 "The origins of unpredictability in life outcome prediction tasks." *Proceedings of the National Academy of Sciences*
- Blumenstock et al. 2015. Predicting Poverty and Wealth from Mobile Phone Metadata. *Science.*
- JWHT Chapter 8