Week 9: Diagnostics

Brandon Stewart[1]

Princeton

November 5–7, 2024

[1]These slides are heavily influenced by Matt Blackwell, Adam Glynn, Jens Hainmueller, Erin Hartman, Kevin Quinn, and Matt Salganik.

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

- Last Week
  - multiple regression

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

- Last Week
  - multiple regression
- This Week
  - peeking under the hood of the linear models
  - non-linearity
  - unusual and influential data

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

- Last Week
  - multiple regression
- This Week
  - peeking under the hood of the linear models
  - non-linearity
  - unusual and influential data
- Next Week
  - adjusting for observed confounding

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

- Last Week
  - multiple regression
- This Week
  - peeking under the hood of the linear models
  - non-linearity
  - unusual and influential data
- Next Week
  - adjusting for observed confounding
- Long Run
  - probability $\rightarrow$ inference $\rightarrow$ regression $\rightarrow$ causal inference

Pulling Back the Curtain

Pulling Back the Curtain

- There is a pedagogical tension between teaching you to be conversant in the field as it is now and preparing you for methods of the future.

Pulling Back the Curtain

- There is a pedagogical tension between teaching you to be conversant in the field as it is now and preparing you for methods of the future.
- We skew a bit more towards the latter, because most people won't continue to study methods long term—I want your knowledge to stay current as long as possible!

Pulling Back the Curtain

- There is a pedagogical tension between teaching you to be conversant in the field as it is now and preparing you for methods of the future.
- We skew a bit more towards the latter, because most people won't continue to study methods long term—I want your knowledge to stay current as long as possible!
- Towards the beginning of the course there is a lot of material that is simply foundational material you simply need to know, but we are starting to get into the practice of statistics and research which means fewer black and white answers.

Pulling Back the Curtain

- There is a pedagogical tension between teaching you to be conversant in the field as it is now and preparing you for methods of the future.
- We skew a bit more towards the latter, because most people won't continue to study methods long term—I want your knowledge to stay current as long as possible!
- Towards the beginning of the course there is a lot of material that is simply foundational material you simply need to know, but we are starting to get into the practice of statistics and research which means fewer black and white answers.
- This week we will talk about how to diagnose and fix problems which is naturally a pretty context-specific activity.

Four Themes

Four Themes

(1) Clearly define your goal.
(2) Kick the tires.
(3) Diagnostics are a double-edged sword.
(4) Re-examine defaults.

(1) Clearly Define Your Goal

(1) Clearly Define Your Goal

- Best practices are rarely best for all applications, people generally have a distribution of applications in mind when they give advice so best to be specific about your application.

(1) Clearly Define Your Goal

- Best practices are rarely best for all applications, people generally have a distribution of applications in mind when they give advice so best to be specific about your application.
- When thinking about whether an estimator is biased, always think 'biased with respect to what?'

(1) Clearly Define Your Goal

- Best practices are rarely best for all applications, people generally have a distribution of applications in mind when they give advice so best to be specific about your application.
- When thinking about whether an estimator is biased, always think 'biased with respect to what?'
- Predictive generalization is one unifying way to solve questions about what is best (particularly in machine learning).

(1) Clearly Define Your Goal

- Best practices are rarely best for all applications, people generally have a distribution of applications in mind when they give advice so best to be specific about your application.
- When thinking about whether an estimator is biased, always think 'biased with respect to what?'
- Predictive generalization is one unifying way to solve questions about what is best (particularly in machine learning).
- You may not know how to precisely define your goal yet. That's okay—we will talk more next week!

(2) Kick the Tires

(2) Kick the Tires

Residuals are important. Look at them.

(2) Kick the Tires

(2) Kick the Tires

- There are so many ways that something can go wrong, particularly when you don't know every nuance of the data.

(2) Kick the Tires

- There are so many ways that something can go wrong, particularly when you don't know every nuance of the data.
- Examining the model helps us detect cases where assumptions have failed to hold or data is contaminated.

(2) Kick the Tires

- There are so many ways that something can go wrong, particularly when you don't know every nuance of the data.
- Examining the model helps us detect cases where assumptions have failed to hold or data is contaminated.
- When diagnosing problems it can often be hard to tell how consequential they are.

(2) Kick the Tires

- There are so many ways that something can go wrong, particularly when you don't know every nuance of the data.
- Examining the model helps us detect cases where assumptions have failed to hold or data is contaminated.
- When diagnosing problems it can often be hard to tell how consequential they are.
- One strategy is diagnosis through treatment, i.e. we just use a procedure that would address the problem and see if it changes our conclusions rather than diagnosing in the first place.

(3) Diagnostics are a Double-edged Sword

(3) Diagnostics are a Double-edged Sword

- Most diagnostics aren't full proof tests—if an assumption could truly be checked it would probably be incorporated in the procedure.

(3) Diagnostics are a Double-edged Sword

- Most diagnostics aren't full proof tests—if an assumption could truly be checked it would probably be incorporated in the procedure.
- All diagnostics have a care tradeoff: if I follow a procedure conditional on passing a test, I have a new procedure which might have different properties.

(3) Diagnostics are a Double-edged Sword

- Most diagnostics aren't full proof tests—if an assumption could truly be checked it would probably be incorporated in the procedure.
- All diagnostics have a care tradeoff: if I follow a procedure conditional on passing a test, I have a new procedure which might have different properties.
- Double checking that that things haven't gone horribly wrong is most likely okay, but there is always a slippery slope argument that leads to what Gelman calls the 'garden of forking paths.' A formal way out of this is train-test splits.

(3) Diagnostics are a Double-edged Sword

- Most diagnostics aren't full proof tests—if an assumption could truly be checked it would probably be incorporated in the procedure.
- All diagnostics have a care tradeoff: if I follow a procedure conditional on passing a test, I have a new procedure which might have different properties.
- Double checking that that things haven't gone horribly wrong is most likely okay, but there is always a slippery slope argument that leads to what Gelman calls the 'garden of forking paths.' A formal way out of this is train-test splits.
- Caution: diagnostics typically check one thing and often (only) work when everything else is perfect.

(4) Re-Examine Defaults

(4) Re-Examine Defaults

- In many ways statistics is the science of defaults.

(4) Re-Examine Defaults

- In many ways statistics is the science of defaults.
- Defaults are hugely important because people rarely change defaults.

(4) Re-Examine Defaults

- In many ways statistics is the science of defaults.
- Defaults are hugely important because people rarely change defaults.
- For example, you see a lot of classical standard errors because that's what `lm()` produces. This is why packages like `estimatr` use robust standard errors by default.

(4) Re-Examine Defaults

- In many ways statistics is the science of defaults.
- Defaults are hugely important because people rarely change defaults.
- For example, you see a lot of classical standard errors because that's what `lm()` produces. This is why packages like `estimatr` use robust standard errors by default.
- But remember, different solutions are better for different settings—defaults are a function of their time and context.

(4) Re-Examine Defaults

- In many ways statistics is the science of defaults.
- Defaults are hugely important because people rarely change defaults.
- For example, you see a lot of classical standard errors because that's what `lm()` produces. This is why packages like `estimatr` use robust standard errors by default.
- But remember, different solutions are better for different settings—defaults are a function of their time and context.
- The history of 20th century social science is defined by scarcity—data was hard to find, surveys were costly to field, computing resources were expensive and difficult to access. The methods we use are a consequence of that scarcity.

(4) Re-Examine Defaults

- In many ways statistics is the science of defaults.
- Defaults are hugely important because people rarely change defaults.
- For example, you see a lot of classical standard errors because that's what `lm()` produces. This is why packages like `estimatr` use robust standard errors by default.
- But remember, different solutions are better for different settings—defaults are a function of their time and context.
- The history of 20th century social science is defined by scarcity—data was hard to find, surveys were costly to field, computing resources were expensive and difficult to access. The methods we use are a consequence of that scarcity.
- But now we have abundance—new forms of data, cheap surveys, huge computation! This should change the methods we consider.

The Most Important Lesson: Check Your Data

The Most Important Lesson: Check Your Data

"Do not attempt to build a model on a set of poor data! In human surveys, one often finds 14-inch men, 1000-pound women, students with 'no' lungs, and so on. In manufacturing data, one can find 10,000 pounds of material in a 100 pound capacity barrel, and similar obvious errors.

All the planning, and training in the world will not eliminate these sorts of problems. In our decades of experience with 'messy data,' we have yet to find a large data set completely free of such quality problems."

Draper and Smith (1981, p. 418)

The Most Important Lesson: Check Your Data

"Do not attempt to build a model on a set of poor data! In human surveys, one often finds 14-inch men, 1000-pound women, students with 'no' lungs, and so on. In manufacturing data, one can find 10,000 pounds of material in a 100 pound capacity barrel, and similar obvious errors.

All the planning, and training in the world will not eliminate these sorts of problems. In our decades of experience with 'messy data,' we have yet to find a large data set completely free of such quality problems."

Draper and Smith (1981, p. 418)

Carefully Examine the Data First!!

It is easy to make a mistake.

Example 1

Does Diversity Pay?: Race, Gender, and the Business Case for Diversity

Cedric Herring
University of Illinois at Chicago

The value-in-diversity perspective argues that a diverse workforce, relative to a homogeneous one, is generally beneficial for business, including but not limited to corporate profits and earnings. This is in contrast to other accounts that view diversity as either nonconsequential to business success or actually detrimental by creating conflict, undermining cohesion, and thus decreasing productivity. Using data from the 1996 to 1997 National Organizations Survey, a national sample of for-profit business organizations, this article tests eight hypotheses derived from the value-in-diversity thesis. The results support seven of these hypotheses: racial diversity is associated with increased sales revenue, more customers, greater market share, and greater relative profits. Gender diversity is associated with increased sales revenue, more customers, and greater relative profits. I discuss the implications of these findings relative to alternative views of diversity in the workplace.

https://doi.org/10.1177/0003122409074002O3

Example 1

Comment on Herring, ASR, April 2009

American Sociological Review
2017, Vol. 82(4) 657–667

DOI: 10.1177/0003122417714422
journals.sagepub.com/home/asr

Does Diversity Pay? A Replication of Herring (2009)

Dragana Stojmenovska,[a] Thijs Bol,[a] and Thomas Leopold[a]

Abstract

In an influential article published in the *American Sociological Review* in 2009, Herring finds that diverse workforces are beneficial for business. His analysis supports seven out of eight hypotheses on the positive effects of gender and racial diversity on sales revenue, number of customers, perceived relative market share, and perceived relative profitability. This comment points out that Herring's analysis contains two errors. First, missing codes on the outcome variables are treated as substantive codes. Second, two control variables—company size and establishment size—are highly skewed, and this skew obscures their positive associations with the predictor and outcome variables. We replicate Herring's analysis correcting for both errors. The findings support only one of the original eight hypotheses, suggesting that diversity is nonconsequential, rather than beneficial, to business success.

https://doi.org/10.1177/0003122417714422

Example 1

In our correspondence with Herring, he did not offer a definitive explanation for these discrepancies, but indicated that he may have treated all codes other than "not applicable" (−999) as substantive codes. Given (1) the large difference between his sample size and the number of valid observations in the NOS, and (2) the large number of missing values due to reasons other than "not applicable"—in particular for sales revenue and number of customers—this coding error appears likely to account for much of the discrepancies. This means, for example, that 206 business organizations in which the sales revenue was unknown were treated as if they had sales of 88,888,888,888 US Dollars. Yet, even when we replicated this error (i.e., keeping all organizations with missing values other than –999 in our sample), we were unable to recover Herring's sample sizes, although the differences were smaller.

https://doi.org/10.1177/0003122417714422

Example 1

American Sociological Review
Volume 82, Issue 4, August 2017, Pages 868-877

https://doi.org/10.1177/0003122417716611

SAGE journals

Comment and Reply

Is Diversity Still a Good Thing?

Cedric Herring

Abstract

In this reply to Stojmenovska, Bol, and Leopold's comment on my 2009 *ASR* article, "Does Diversity Pay?" I offer an updated analysis of the relationship between racial and gender diversity in business organizations and such business outcomes as sales revenue, number of customers, market share, and relative profits. Despite the claims in their comment, an updated analysis supports the business case for diversity perspective. Racial and gender diversity are related to business outcomes in the direction predicted. I conclude that diversity is still a good thing, not just because it is related to business outcomes, but also because it reinforces the belief that everyone deserves an equal opportunity.

https://doi.org/10.1177/0003122417716611

Example 2

MARITAL COITAL FREQUENCY AND THE PASSAGE OF TIME: ESTIMATING THE SEPARATE EFFECTS OF SPOUSES' AGES AND MARITAL DURATION, BIRTH AND MARRIAGE COHORTS, AND PERIOD INFLUENCES*

GUILLERMINA JASSO
University of Minnesota

To the extent that marital coital frequency is linked to both within-couple and societal fertility and sex ratio, it may be implicated in a wide range of behavioral and social phenomena. Variation in marital coital frequency over the life course as well as across cohorts may thus affect many aspects of the social life. This paper reports estimates of the ceteris paribus cohort-free effects of spouses' ages, marital duration, and contemporaneous period influences on coital frequency, as well as of the correlations between coital frequency and birth- and marriage-cohort factors. These estimates are obtained by (a) using the properties of the fixed-effects statistical model in order to separate the effects of cohort influences from the age/duration and period effects and to control for the operation of couple-specific unobservables, and (b) using strictly monotonic nonlinear transformations in order to separate the effects of wife's age, husband's age and marital duration.

https://www.jstor.org/stable/2095411

Example 2

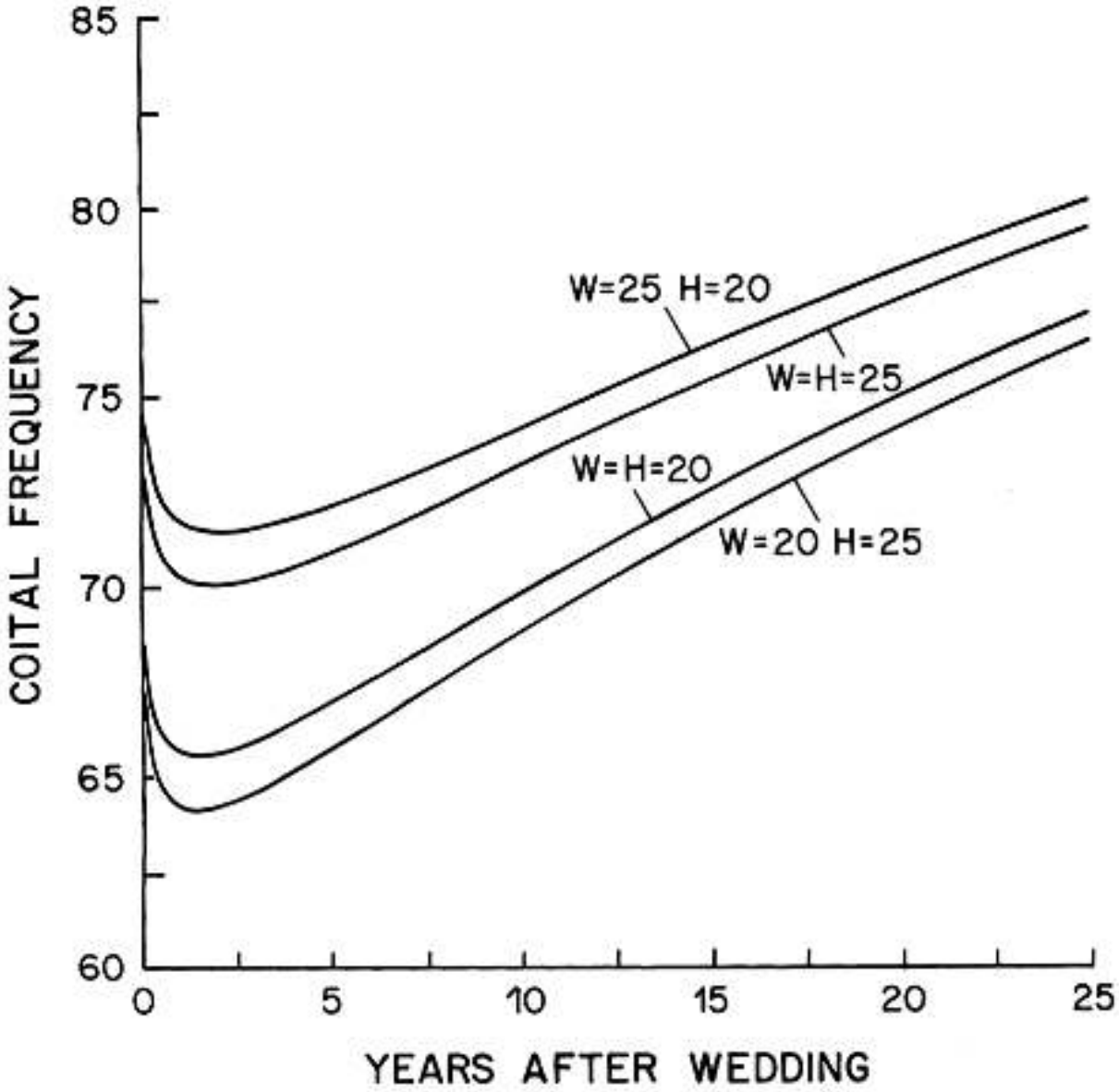


Figure 2. Combined Estimated Effects of Spouses' Ages and Marital Duration on Coital Frequency over the Course of the Marriage

Example 2

MARITAL COITAL FREQUENCY: UNNOTICED OUTLIERS AND UNSPECIFIED INTERACTIONS LEAD TO ERRONEOUS CONCLUSIONS*

JOAN R. KAHN J. RICHARD UDRY

Carolina Population Center, The University of North Carolina at Chapel Hill

https://www.jstor.org/stable/2095496

Example 2

In a recent application of fixed-effects modeling, Jasso estimates age and period effects on marital coital frequency that take into account cohort effects. Although the application is an innovation for researchers interested in identifying net age and period effects, the results are in some respects problematic. Jasso claims that over the period 1970–1975 (1) net of cohort and age effects, average marital coital frequency declined significantly by 50 percent ($p < .01$), and (2) net of period and cohort effects, coital frequency increased with wife's age ($p < .05$).

Both of these findings are difficult to reconcile with previous findings in the literature on sexual behavior. The negative period effect is inconsistent

https://www.jstor.org/stable/2095496

Example 2

results (see column 2 of Table 1). However, after applying standard regression diagnostic techniques, we found serious problems with outliers and specification error. First, in the 1970 wave, four observations were coded as 88, even though the missing code was specified as 99 in the codebook. We are fairly certain that the 88s should have been coded as missing since no other values was greater than 63. In fact, 99.5 percent of the sample had values less than 40. Furthermore, all of the cases with 88 in 1970 had coital frequencies of less than or equal to 10 in 1975, indicating that these couples were probably not highly active.

Deleting the four 88-codes reduced the significance level on the wife's age effect from $p < .05$ to $p < .10$, and reversed the sign on the husband's age effect (see column 3 in Table 1). The fit of the model improved substantially from an R^2 of .047 to .057. This in itself is not damning evidence against Jasso's analysis, although it indicates that her results are sensitive to specific cases. To determine the potential influence of the other cases, we examined the studentized residuals[1] (Belsley et al., 1980; Bollen and Jackman, 1985; Weisberg, 1980) and found a handful of cases with extremely large residuals. The 88s had by far the largest studentized residuals, ranging between 10 and 12. Four additional cases had studentized residuals between 3 and 7. An examination of the DFBETAs[2] showed that the omission of these observations resulted in large changes in the parameter estimates. These observations correspond to the four observations with coital frequencies of greater than 40 (i.e., they are at the extreme high end of the coital frequency distribution).[3]

https://www.jstor.org/stable/2095496

Example 2

IS IT OUTLIER DELETION OR IS IT SAMPLE TRUNCATION? NOTES ON SCIENCE AND SEXUALITY*

(Reply to Kahn and Udry, *ASR*, October, 1986)

GUILLERMINA JASSO
University of Minnesota

https://www.jstor.org/stable/2095497

Example 2

I argue that the estimates of the effects of spouses' ages, marital duration, and contemporaneous period influences on martial coital frequency reported in Jasso (1985) are superior to the outlier-deleting estimates reported in Kahn and Udry (1986)—for chiefly one reason: If the outlying observations in question are indeed erroneous, as claimed by Kahn and Udry, then (assuming correct specification) both the (full-sample) Kahn and Udry (1986) and the Jasso (1985) estimates are unbiased estimates of the underlying parameters; in contrast, if the outlying observations are correct, then only the Jasso estimates are unbiased, the Kahn and Udry estimates now reflecting biases associated with truncation of the dependent variable. The Kahn and Udry estimates on partitions of the sample, even ignoring the truncation bias, cannot be evaluated since the equations estimated do not represent the process proposed in their verbal argument (interactions with marital duration) but instead something quite different (interactions with date of marriage), for which no rationale is offered.

https://www.jstor.org/stable/2095497

Example 3

American Sociological Review
2015, Vol. 80(3) 469–495

DOI: 10.1177/0003122415579389
http://asr.sagepub.com

Her Support, His Support: Money, Masculinity, and Marital Infidelity

Christin L. Munsch[a]

Abstract

Recent years have seen great interest in the relationship between relative earnings and marital outcomes. Using data from the 1997 National Longitudinal Survey of Youth, I examine the effect of relative earnings on infidelity, a marital outcome that has received little attention. Theories of social exchange predict that the greater one's relative income, the more likely one will be to engage in infidelity. Yet, emerging literature raises questions about the utility of gender-neutral exchange approaches, particularly when men are economically dependent and women are breadwinners. I find that, for men, breadwinning increases infidelity. For women, breadwinning decreases infidelity. I argue that by remaining faithful, breadwinning women neutralize their gender deviance and keep potentially strained relationships intact. I also find that, for both men and women, economic dependency is associated with a higher likelihood of engaging in infidelity; but, the influence of dependency on men's infidelity is greater than the influence of dependency on women's infidelity. For economically dependent persons, infidelity may be an attempt to restore relationship equity; however, for men, dependence may be particularly threatening. Infidelity may allow economically dependent men to engage in compensatory behavior while simultaneously distancing themselves from breadwinning spouses.

https://doi.org/10.1177/0003122418780369

Example 3

American Sociological Review
Volume 83, Issue 4, August 2018, Pages 833–838
© American Sociological Association 2018. Article Reuse Guidelines
https://doi.org/10.1177/0003122418780369

SAGE journals

Correction

Correction: "Her Support, His Support: Money, Masculinity, and Marital Infidelity" *American Sociological Review* 80(3):469–95

Christin L. Munsch

University of Connecticut

I made several errors related to the coding of missing data in my June 2015 *ASR* article, "Her Support, His Support: Money, Masculinity, and Marital Infidelity." Upon discovery, I contacted the editors who asked me to prepare a short memo explaining the errors and to replicate the analyses exactly as presented in the original article. It is worth noting, however, that modeling decisions that may have been appropriate using the previous version of the data may not be appropriate using the corrected version. Additional analyses using new procedures can be found on my website (http://www.christinlmunsch.com/wp-content/uploads/2018/05/asrcorrection.pdf).

https://doi.org/10.1177/0003122418780369

Example 3

Errors

First, I made an error in the coding of the dependent variable, infidelity. Persons with the same spouse in consecutive years, who reported more than one sex partner since the date of the last interview, were defined as unfaithful. However, because Stata treats missing values as positive infinity, I inadvertently coded 246 cases in which the respondent was missing a sex partner count as adulterous.[1] Unfortunately,

https://doi.org/10.1177/0003122418780369

1. Opening Themes
2. Preliminaries: The Hat Matrix
3. Fun With (Misleading) Visualization
4. Extreme Values
   - Outliers
   - Leverage Points
   - Influence Points
   - Fun with Outliers
5. Robust Regression Methods
   - Appendix: Robustness
6. Nonlinearity
   - Linear Basis Function Models
   - Generalized Additive Models

Prereq: Defining the Hat Matrix

Prereq: Defining the Hat Matrix

- We want to figure out how the residuals relate to the true errors.

Prereq: Defining the Hat Matrix

- We want to figure out how the residuals relate to the true errors.
- To get there let's write the residual $\widehat{\mathbf{u}}$ in terms of y, then we will be able to replace $\mathbf{y}$ with $\mathbf{X}\boldsymbol{\beta} + \mathbf{u}$.

Prereq: Defining the Hat Matrix

- We want to figure out how the residuals relate to the true errors.
- To get there let's write the residual $\widehat{\mathbf{u}}$ in terms of y, then we will be able to replace $\mathbf{y}$ with $\mathbf{X}\boldsymbol{\beta} + \mathbf{u}$.

$$\widehat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$$

Prereq: Defining the Hat Matrix

- We want to figure out how the residuals relate to the true errors.
- To get there let's write the residual $\widehat{\mathbf{u}}$ in terms of y, then we will be able to replace $\mathbf{y}$ with $\mathbf{X}\boldsymbol{\beta} + \mathbf{u}$.

$$\begin{aligned}\widehat{\mathbf{u}} &= \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} \\ &= \mathbf{y} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}\end{aligned}$$

Prereq: Defining the Hat Matrix

- We want to figure out how the residuals relate to the true errors.
- To get there let's write the residual $\widehat{\mathbf{u}}$ in terms of y, then we will be able to replace $\mathbf{y}$ with $\mathbf{X}\boldsymbol{\beta} + \mathbf{u}$.

$$\begin{aligned}\widehat{\mathbf{u}} &= \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} \\ &= \mathbf{y} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \\ &\equiv \mathbf{y} - \mathbf{H}\mathbf{y}\end{aligned}$$

Prereq: Defining the Hat Matrix

- We want to figure out how the residuals relate to the true errors.
- To get there let's write the residual $\widehat{\mathbf{u}}$ in terms of y, then we will be able to replace $\mathbf{y}$ with $\mathbf{X}\boldsymbol{\beta} + \mathbf{u}$.

$$\begin{aligned}
\widehat{\mathbf{u}} &= \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} \\
&= \mathbf{y} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \\
&\equiv \mathbf{y} - \mathbf{H}\mathbf{y} \\
&= (\mathbf{I} - \mathbf{H})\mathbf{y}
\end{aligned}$$

Prereq: Defining the Hat Matrix

- We want to figure out how the residuals relate to the true errors.
- To get there let's write the residual $\widehat{\mathbf{u}}$ in terms of y, then we will be able to replace $\mathbf{y}$ with $\mathbf{X}\boldsymbol{\beta} + \mathbf{u}$.

$$\begin{aligned}
\widehat{\mathbf{u}} &= \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} \\
&= \mathbf{y} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \\
&\equiv \mathbf{y} - \mathbf{H}\mathbf{y} \\
&= (\mathbf{I} - \mathbf{H})\mathbf{y}
\end{aligned}$$

- $\mathbf{H} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$ is the hat matrix because it puts the "hat" on $\mathbf{y}$:

$$\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}$$

Prereq: Defining the Hat Matrix

- We want to figure out how the residuals relate to the true errors.
- To get there let's write the residual $\widehat{\mathbf{u}}$ in terms of y, then we will be able to replace $\mathbf{y}$ with $\mathbf{X}\boldsymbol{\beta} + \mathbf{u}$.

$$\begin{aligned}\widehat{\mathbf{u}} &= \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} \\ &= \mathbf{y} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \\ &\equiv \mathbf{y} - \mathbf{H}\mathbf{y} \\ &= (\mathbf{I} - \mathbf{H})\mathbf{y}\end{aligned}$$

- $\mathbf{H} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$ is the hat matrix because it puts the "hat" on $\mathbf{y}$:

$$\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}$$

it has a few nice properties:

Prereq: Defining the Hat Matrix

- We want to figure out how the residuals relate to the true errors.
- To get there let's write the residual $\widehat{\mathbf{u}}$ in terms of y, then we will be able to replace $\mathbf{y}$ with $\mathbf{X}\boldsymbol{\beta} + \mathbf{u}$.

$$\begin{aligned}\widehat{\mathbf{u}} &= \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} \\ &= \mathbf{y} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \\ &\equiv \mathbf{y} - \mathbf{H}\mathbf{y} \\ &= (\mathbf{I} - \mathbf{H})\mathbf{y}\end{aligned}$$

- $\mathbf{H} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$ is the hat matrix because it puts the "hat" on $\mathbf{y}$:

$$\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}$$

it has a few nice properties:

  - $\mathbf{H}$ is an $n \times n$ symmetric matrix

Prereq: Defining the Hat Matrix

- We want to figure out how the residuals relate to the true errors.
- To get there let's write the residual $\widehat{\mathbf{u}}$ in terms of y, then we will be able to replace $\mathbf{y}$ with $\mathbf{X}\boldsymbol{\beta} + \mathbf{u}$.

$$\begin{aligned}\widehat{\mathbf{u}} &= \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} \\ &= \mathbf{y} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \\ &\equiv \mathbf{y} - \mathbf{H}\mathbf{y} \\ &= (\mathbf{I} - \mathbf{H})\mathbf{y}\end{aligned}$$

- $\mathbf{H} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$ is the hat matrix because it puts the "hat" on $\mathbf{y}$:

$$\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}$$

it has a few nice properties:

  - $\mathbf{H}$ is an $n \times n$ symmetric matrix
  - $\mathbf{H}$ is idempotent: $\mathbf{H}\mathbf{H} = \mathbf{H}$

Relating the Residuals to the Errors

With the hat matrix, we are ready to relate the residuals to the errors.

$$\widehat{\mathbf{u}} = (\mathbf{I} - \mathbf{H})(\mathbf{y})$$

Relating the Residuals to the Errors

With the hat matrix, we are ready to relate the residuals to the errors.

$$\begin{aligned}\widehat{\mathbf{u}} &= (\mathbf{I} - \mathbf{H})(\mathbf{y}) \\ &= (\mathbf{I} - \mathbf{H})(\mathbf{X}\boldsymbol{\beta} + \mathbf{u})\end{aligned}$$

Relating the Residuals to the Errors

With the hat matrix, we are ready to relate the residuals to the errors.

$$\begin{aligned}\widehat{\mathbf{u}} &= (\mathbf{I} - \mathbf{H})(\mathbf{y}) \\ &= (\mathbf{I} - \mathbf{H})(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) \\ &= (\mathbf{I} - \mathbf{H})\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \mathbf{H})\mathbf{u}\end{aligned}$$

Relating the Residuals to the Errors

With the hat matrix, we are ready to relate the residuals to the errors.

$$\begin{aligned}
\widehat{\mathbf{u}} &= (\mathbf{I} - \mathbf{H})(\mathbf{y}) \\
&= (\mathbf{I} - \mathbf{H})(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) \\
&= (\mathbf{I} - \mathbf{H})\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \mathbf{H})\mathbf{u} \\
&= \mathbf{I}\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \mathbf{H})\mathbf{u}
\end{aligned}$$

Relating the Residuals to the Errors

With the hat matrix, we are ready to relate the residuals to the errors.

$$
\begin{aligned}
\widehat{\mathbf{u}} &= (\mathbf{I} - \mathbf{H})(\mathbf{y}) \\
&= (\mathbf{I} - \mathbf{H})(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) \\
&= (\mathbf{I} - \mathbf{H})\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \mathbf{H})\mathbf{u} \\
&= \mathbf{I}\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \mathbf{H})\mathbf{u} \\
&= \mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \mathbf{H})\mathbf{u}
\end{aligned}
$$

Relating the Residuals to the Errors

With the hat matrix, we are ready to relate the residuals to the errors.

$$
\begin{aligned}
\widehat{\mathbf{u}} &= (\mathbf{I} - \mathbf{H})(\mathbf{y}) \\
&= (\mathbf{I} - \mathbf{H})(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) \\
&= (\mathbf{I} - \mathbf{H})\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \mathbf{H})\mathbf{u} \\
&= \mathbf{I}\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \mathbf{H})\mathbf{u} \\
&= \mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \mathbf{H})\mathbf{u} \\
&= (\mathbf{I} - \mathbf{H})\mathbf{u}
\end{aligned}
$$

Relating the Residuals to the Errors

With the hat matrix, we are ready to relate the residuals to the errors.

$$
\begin{aligned}
\widehat{\mathbf{u}} &= (\mathbf{I} - \mathbf{H})(\mathbf{y}) \\
&= (\mathbf{I} - \mathbf{H})(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) \\
&= (\mathbf{I} - \mathbf{H})\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \mathbf{H})\mathbf{u} \\
&= \mathbf{I}\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \mathbf{H})\mathbf{u} \\
&= \mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \mathbf{H})\mathbf{u} \\
&= (\mathbf{I} - \mathbf{H})\mathbf{u}
\end{aligned}
$$

- Residuals $\widehat{\mathbf{u}}$ are a linear function of the errors, $\mathbf{u}$.

Relating the Residuals to the Errors

With the hat matrix, we are ready to relate the residuals to the errors.

$$\begin{aligned}
\widehat{\mathbf{u}} &= (\mathbf{I} - \mathbf{H})(\mathbf{y}) \\
&= (\mathbf{I} - \mathbf{H})(\mathbf{X}\beta + \mathbf{u}) \\
&= (\mathbf{I} - \mathbf{H})\mathbf{X}\beta + (\mathbf{I} - \mathbf{H})\mathbf{u} \\
&= \mathbf{I}\mathbf{X}\beta - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\beta + (\mathbf{I} - \mathbf{H})\mathbf{u} \\
&= \mathbf{X}\beta - \mathbf{X}\beta + (\mathbf{I} - \mathbf{H})\mathbf{u} \\
&= (\mathbf{I} - \mathbf{H})\mathbf{u}
\end{aligned}$$

- Residuals $\widehat{\mathbf{u}}$ are a linear function of the errors, $\mathbf{u}$.
- For instance,

$$\widehat{u}_1 = (1 - h_{11})u_1 - \sum_{i=2}^{n} h_{1i}u_i$$

Relating the Residuals to the Errors

With the hat matrix, we are ready to relate the residuals to the errors.

$$\begin{aligned}
\widehat{\mathbf{u}} &= (\mathbf{I} - \mathbf{H})(\mathbf{y}) \\
&= (\mathbf{I} - \mathbf{H})(\mathbf{X}\beta + \mathbf{u}) \\
&= (\mathbf{I} - \mathbf{H})\mathbf{X}\beta + (\mathbf{I} - \mathbf{H})\mathbf{u} \\
&= \mathbf{I}\mathbf{X}\beta - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\beta + (\mathbf{I} - \mathbf{H})\mathbf{u} \\
&= \mathbf{X}\beta - \mathbf{X}\beta + (\mathbf{I} - \mathbf{H})\mathbf{u} \\
&= (\mathbf{I} - \mathbf{H})\mathbf{u}
\end{aligned}$$

- Residuals $\widehat{\mathbf{u}}$ are a linear function of the errors, $\mathbf{u}$.
- For instance,

$$\widehat{u}_1 = (1 - h_{11})u_1 - \sum_{i=2}^{n} h_{1i}u_i$$

- Note that each residual is a function of all of the errors.

Characterizing the Distribution of the Residuals

What can we say about the distribution of the residuals now that we have the expression: $\widehat{\mathbf{u}} = (\mathbf{I} - \mathbf{H})\mathbf{u}$.

Characterizing the Distribution of the Residuals

What can we say about the distribution of the residuals now that we have the expression: $\widehat{\mathbf{u}} = (\mathbf{I} - \mathbf{H})\mathbf{u}$.

$$E[\widehat{\mathbf{u}}] =$$

Characterizing the Distribution of the Residuals

What can we say about the distribution of the residuals now that we have the expression: $\widehat{\mathbf{u}} = (\mathbf{I} - \mathbf{H})\mathbf{u}$.

$$E[\widehat{\mathbf{u}}] = (\mathbf{I} - \mathbf{H})E[\mathbf{u}] =$$

Characterizing the Distribution of the Residuals

What can we say about the distribution of the residuals now that we have the expression: $\widehat{\mathbf{u}} = (\mathbf{I} - \mathbf{H})\mathbf{u}$.

$$E[\widehat{\mathbf{u}}] = (\mathbf{I} - \mathbf{H})E[\mathbf{u}] = \mathbf{0}$$

Characterizing the Distribution of the Residuals

What can we say about the distribution of the residuals now that we have the expression: $\widehat{\mathbf{u}} = (\mathbf{I} - \mathbf{H})\mathbf{u}$.

$$E[\widehat{\mathbf{u}}] = (\mathbf{I} - \mathbf{H})E[\mathbf{u}] = \mathbf{0}$$
$$V[\hat{\mathbf{u}}] =$$

Characterizing the Distribution of the Residuals

What can we say about the distribution of the residuals now that we have the expression: $\widehat{\mathbf{u}} = (\mathbf{I} - \mathbf{H})\mathbf{u}$.

$$E[\widehat{\mathbf{u}}] = (\mathbf{I} - \mathbf{H})E[\mathbf{u}] = \mathbf{0}$$
$$V[\widehat{\mathbf{u}}] = \sigma_u^2(\mathbf{I} - \mathbf{H})$$

Characterizing the Distribution of the Residuals

What can we say about the distribution of the residuals now that we have the expression: $\widehat{\mathbf{u}} = (\mathbf{I} - \mathbf{H})\mathbf{u}$.

$$E[\widehat{\mathbf{u}}] = (\mathbf{I} - \mathbf{H})E[\mathbf{u}] = \mathbf{0}$$
$$V[\hat{\mathbf{u}}] = \sigma_u^2(\mathbf{I} - \mathbf{H})$$

The variance of the ith residual $\hat{u}_i$ is $V[\hat{u}_i] = \sigma^2(1 - h_{ii})$, where h_{ii} is the ith diagonal element of the matrix $\mathbf{H}$ (called the hat value).

Properties of the Distribution of Residuals

Properties of the Distribution of Residuals

Notice that even when the unobserved errors are Normal and homoskedastic, the estimated residuals have some different properties. They

Properties of the Distribution of Residuals

Notice that even when the unobserved errors are Normal and homoskedastic, the estimated residuals have some different properties. They

1. are not independent
   (because they must satisfy the two constraints $\sum_{i=1}^{n} \widehat{u}_i = 0$ and $\sum_{i=1}^{n} \widehat{u}_i x_i = 0$)

Properties of the Distribution of Residuals

Notice that even when the unobserved errors are Normal and homoskedastic, the estimated residuals have some different properties. They

1. are not independent
   (because they must satisfy the two constraints $\sum_{i=1}^{n} \widehat{u}_i = 0$ and $\sum_{i=1}^{n} \widehat{u}_i x_i = 0$)
2. do not have the same variance.
   The variance of the residuals varies across data points
   $V[\hat{u}_i] = \sigma^2(1 - h_{ii})$, even though the unobserved errors all have the same variance σ^2

Properties of the Distribution of Residuals

Notice that even when the unobserved errors are Normal and homoskedastic, the estimated residuals have some different properties. They

1. are not independent
   (because they must satisfy the two constraints $\sum_{i=1}^{n} \widehat{u}_i = 0$ and $\sum_{i=1}^{n} \widehat{u}_i x_i = 0$)
2. do not have the same variance.
   The variance of the residuals varies across data points $V[\hat{u}_i] = \sigma^2(1 - h_{ii})$, even though the unobserved errors all have the same variance σ^2

These properties make it hard to learn about the errors (which our assumptions are about and we don't have access to) from our residuals (which we have estimated and can examine).

Properties of the Distribution of Residuals

Notice that even when the unobserved errors are Normal and homoskedastic, the estimated residuals have some different properties. They

1. are not independent
   (because they must satisfy the two constraints $\sum_{i=1}^{n} \widehat{u}_i = 0$ and $\sum_{i=1}^{n} \widehat{u}_i x_i = 0$)
2. do not have the same variance.
   The variance of the residuals varies across data points $V[\hat{u}_i] = \sigma^2(1 - h_{ii})$, even though the unobserved errors all have the same variance σ^2

These properties make it hard to learn about the errors (which our assumptions are about and we don't have access to) from our residuals (which we have estimated and can examine).
This is inconvenient for diagnostics.

Properties of the Distribution of Residuals

Notice that even when the unobserved errors are Normal and homoskedastic, the estimated residuals have some different properties. They

1. are not independent
   (because they must satisfy the two constraints $\sum_{i=1}^{n} \widehat{u}_i = 0$ and $\sum_{i=1}^{n} \widehat{u}_i x_i = 0$)
2. do not have the same variance.
   The variance of the residuals varies across data points $V[\hat{u}_i] = \sigma^2(1 - h_{ii})$, even though the unobserved errors all have the same variance σ^2

These properties make it hard to learn about the errors (which our assumptions are about and we don't have access to) from our residuals (which we have estimated and can examine).
This is inconvenient for diagnostics.

What if we could transform the residuals to address the two issues above?

Standardized Residuals

Standardized Residuals

Let's address the second problem (unequal variances) by standardizing $\hat{u}_i$, i.e., dividing by unit i's estimated standard deviation.

Standardized Residuals

Let's address the second problem (unequal variances) by standardizing $\hat{u}_i$, i.e., dividing by unit i's estimated standard deviation.

This produces standardized (or "internally studentized") residuals:

$$\hat{u}_i' = \frac{\hat{u}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

where $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$ and $\hat{\sigma}^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n-(k+1)}$ is our usual estimate of the error variance.

Standardized Residuals

Let's address the second problem (unequal variances) by standardizing $\hat{u}_i$, i.e., dividing by unit i's estimated standard deviation.

This produces standardized (or "internally studentized") residuals:

$$\hat{u}_i' = \frac{\hat{u}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

where $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$ and $\hat{\sigma}^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n-(k+1)}$ is our usual estimate of the error variance.

The standardized residuals are still not ideal, since the numerator and denominator of $\hat{u}_i'$ are not independent. This makes the distribution of $\hat{u}_i'$ nonstandard. If the distribution is non-standard, we can't easily check for violations.

Studentized Residuals

Studentized Residuals

If we remove observation i from the estimation of σ^2, then we can eliminate the dependence and the result will have a standard distribution.

Studentized Residuals

If we remove observation i from the estimation of σ^2, then we can eliminate the dependence and the result will have a standard distribution.

- estimate error variance without residual i:

$$\widehat{\sigma}^2_{-i} = \frac{\widehat{\mathbf{u}}'\widehat{\mathbf{u}} - \widehat{u}_i^2/(1-h_{ii})}{n-k-2}$$

Studentized Residuals

If we remove observation i from the estimation of σ^2, then we can eliminate the dependence and the result will have a standard distribution.

- estimate error variance without residual i:

$$\widehat{\sigma}^2_{-i} = \frac{\widehat{\mathbf{u}}'\widehat{\mathbf{u}} - \widehat{u}_i^2/(1-h_{ii})}{n-k-2}$$

- Use this i-free estimate to standardize, which creates the studentized residuals:

$$\widehat{u}_i^* = \frac{\widehat{u}_i}{\widehat{\sigma}_{-i}\sqrt{1-h_{ii}}}$$

Studentized Residuals

If we remove observation i from the estimation of σ^2, then we can eliminate the dependence and the result will have a standard distribution.

- estimate error variance without residual i:

$$\widehat{\sigma}^2_{-i} = \frac{\widehat{\mathbf{u}}'\widehat{\mathbf{u}} - \widehat{u}_i^2/(1-h_{ii})}{n-k-2}$$

- Use this i-free estimate to standardize, which creates the studentized residuals:

$$\widehat{u}_i^* = \frac{\widehat{u}_i}{\widehat{\sigma}_{-i}\sqrt{1-h_{ii}}}$$

- If the errors are Normal, the studentized residuals, $\widehat{u}_i^*$, follow a t distribution with $(n-k-2)$ degrees of freedom.

Studentized Residuals

If we remove observation i from the estimation of σ^2, then we can eliminate the dependence and the result will have a standard distribution.

- estimate error variance without residual i:

$$\widehat{\sigma}^2_{-i} = \frac{\widehat{\mathbf{u}}'\widehat{\mathbf{u}} - \widehat{u}_i^2/(1-h_{ii})}{n-k-2}$$

- Use this i-free estimate to standardize, which creates the studentized residuals:

$$\widehat{u}_i^* = \frac{\widehat{u}_i}{\widehat{\sigma}_{-i}\sqrt{1-h_{ii}}}$$

- If the errors are Normal, the studentized residuals, $\widehat{u}_i^*$, follow a t distribution with $(n-k-2)$ degrees of freedom.
- Deviations from this t distribution of the residuals imply violation of Normality in the errors.

Example: Buchanan votes in Florida, 2000

Example: Buchanan votes in Florida, 2000

Wand et al. show that the ballot caused 2,000 Democratic voters to vote by mistake for Buchanan, a number more than enough to have tipped the vote in FL from Bush to Gore, thus giving him FL's 25 electoral votes and the presidency.

American Political Science Review Vol. 95, No. 4 December 2001

The Butterfly Did It: The Aberrant Vote for Buchanan in Palm Beach County, Florida

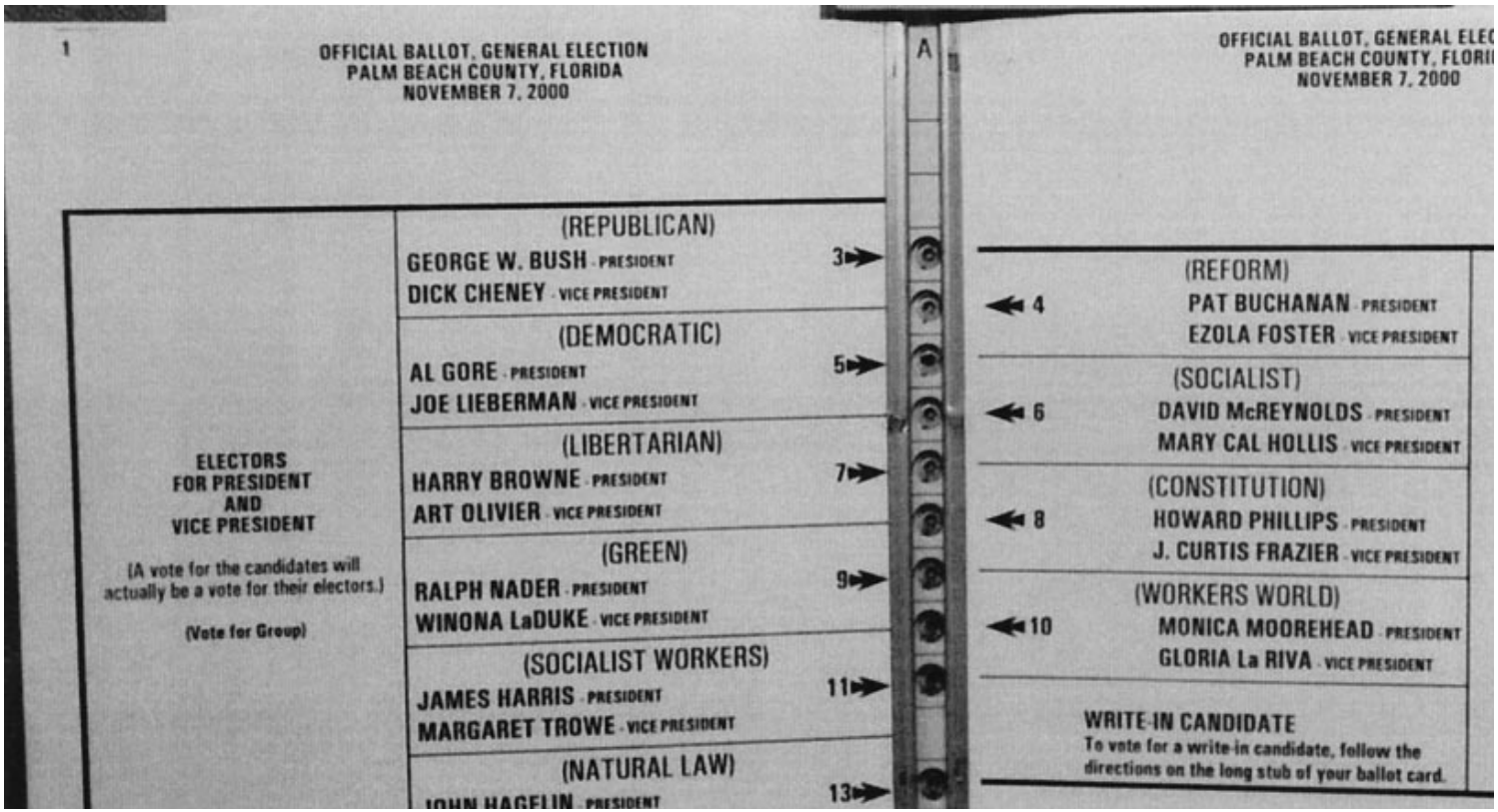
JONATHAN N. WAND *Cornell University*
KENNETH W. SHOTTS *Northwestern University*
JASJEET S. SEKHON *Harvard University*
WALTER R. MEBANE, JR. *Cornell University*
MICHAEL C. HERRON *Northwestern University*
HENRY E. BRADY *University of California, Berkeley*

We show that the butterfly ballot used in Palm Beach County, Florida, in the 2000 presidential election caused more than 2,000 Democratic voters to vote by mistake for Reform candidate Pat Buchanan, a number larger than George W. Bush's certified margin of victory in Florida. We use multiple methods and several kinds of data to rule out alternative explanations for the votes Buchanan received in Palm Beach County. Among 3,053 U.S. counties where Buchanan was on the ballot, Palm Beach County has the most anomalous excess of votes for him. In Palm Beach County, Buchanan's proportion of the vote on election-day ballots is four times larger than his proportion on absentee (nonbutterfly) ballots, but Buchanan's proportion does not differ significantly between election-day and absentee ballots in any other Florida county. Unlike other Reform candidates in Palm Beach County, Buchanan tended to receive election-day votes in Democratic precincts and from individuals who voted for the Democratic U.S. Senate candidate. Robust estimation of overdispersed binomial regression models underpins much of the analysis.

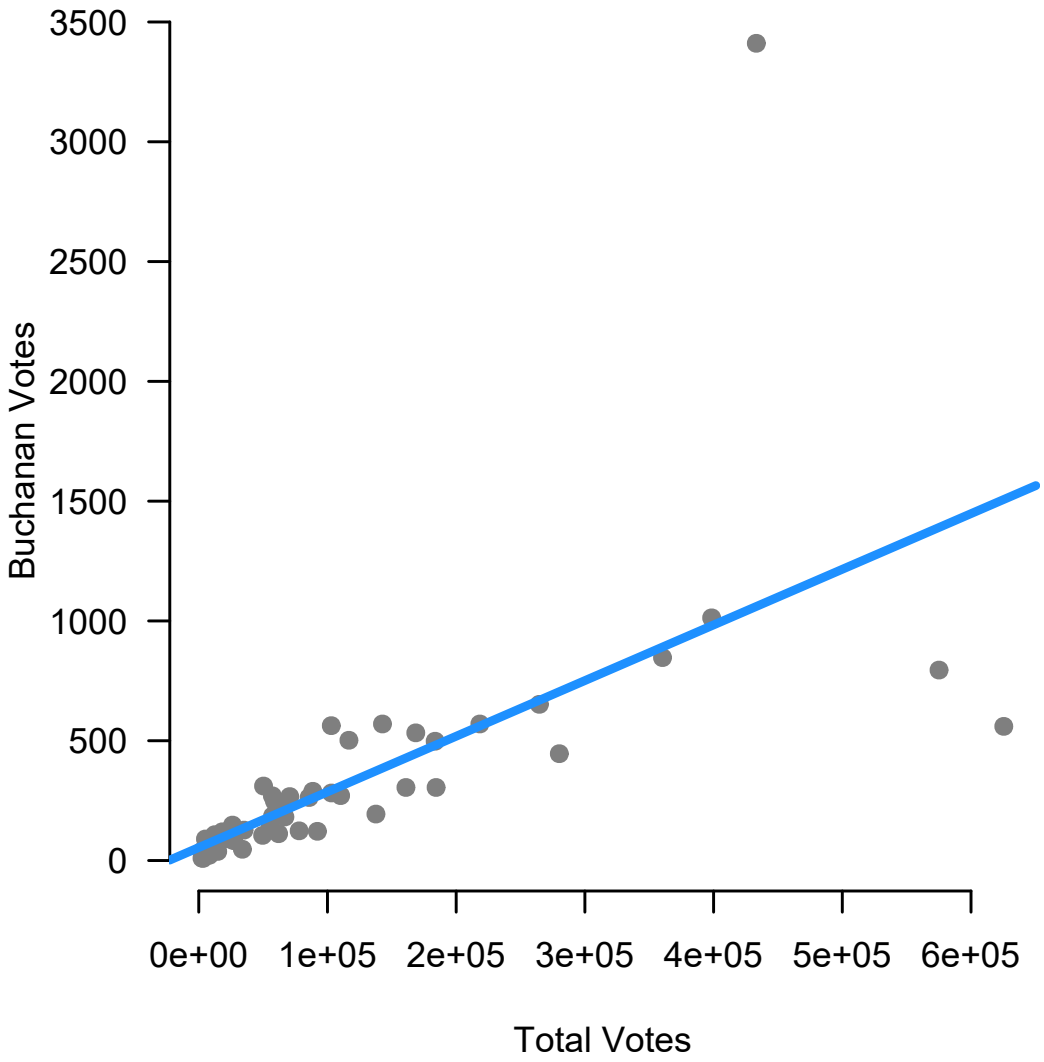
Example: Buchanan votes in Florida, 2000

Wand et al. show that the ballot caused 2,000 Democratic voters to vote by mistake for Buchanan, a number more than enough to have tipped the vote in FL from Bush to Gore, thus giving him FL's 25 electoral votes and the presidency.

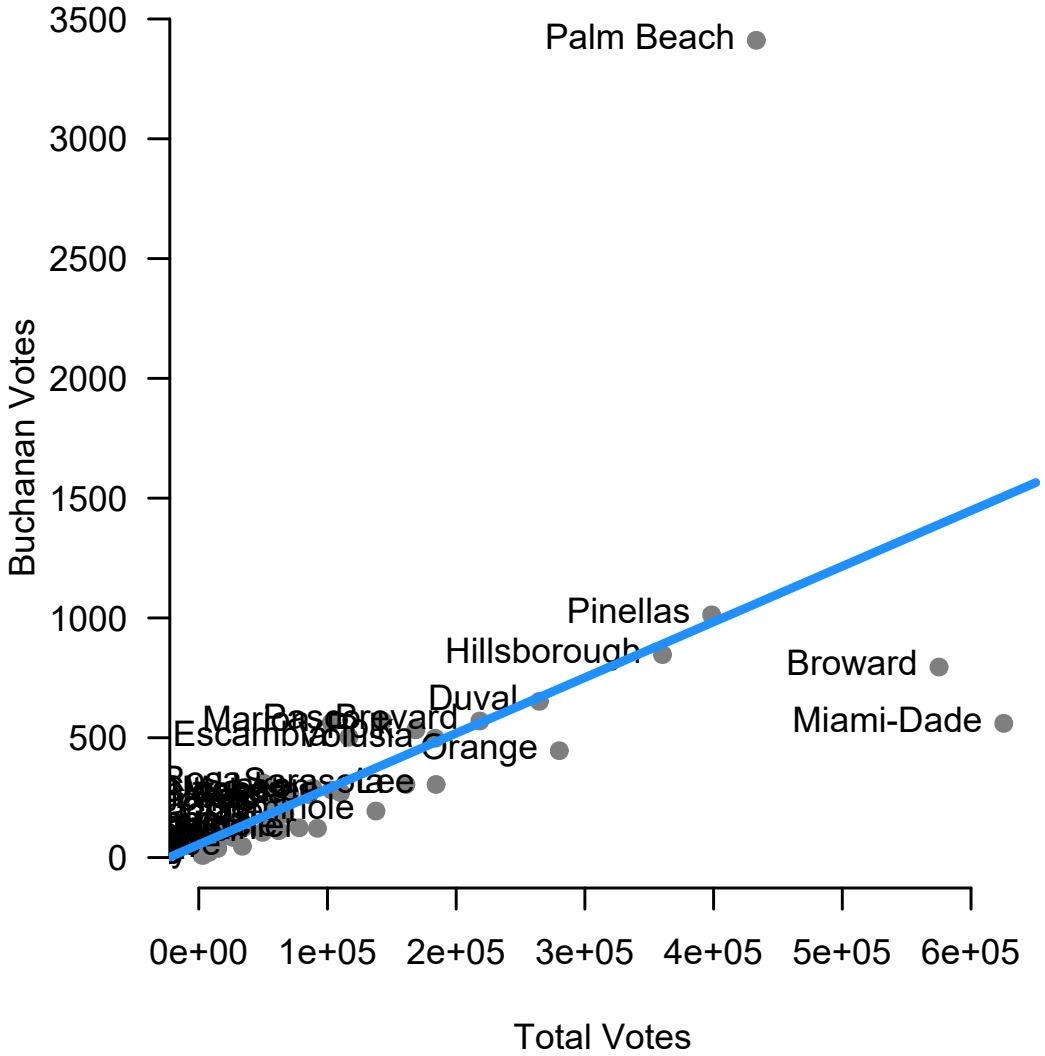
FIGURE 1. The Palm Beach County Bufferfly Ballot



Example: Buchanan votes in Florida, 2000



Example: Buchanan votes in Florida, 2000



Example: Buchanan Votes in Florida

Example: Buchanan Votes in Florida

- Now that our studentized residuals follow a known standard distribution (if A1-A6 hold!), we can proceed with diagnostic analysis for the nonnormal errors.

Example: Buchanan Votes in Florida

- Now that our studentized residuals follow a known standard distribution (if A1-A6 hold!), we can proceed with diagnostic analysis for the nonnormal errors.

- We examine data from the 2000 presidential election in Florida used in Wand et al. (2001).

Example: Buchanan Votes in Florida

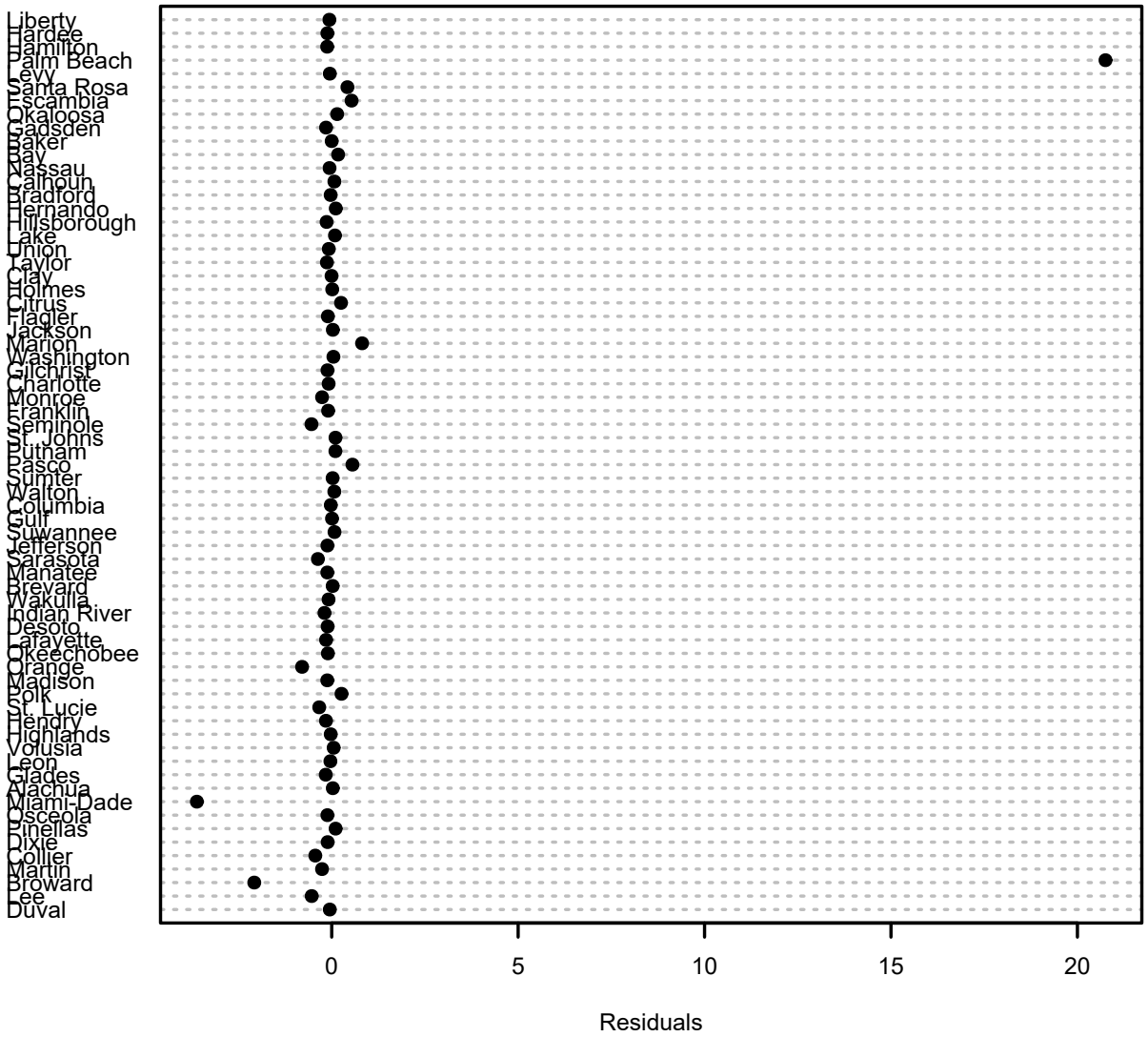
- Now that our studentized residuals follow a known standard distribution (if A1-A6 hold!), we can proceed with diagnostic analysis for the nonnormal errors.
- We examine data from the 2000 presidential election in Florida used in Wand et al. (2001).
- Our analysis takes place at the county level and we will regress the number of Buchanan votes in each county on the total number of votes in each county.

Buchanan Votes and Total Votes

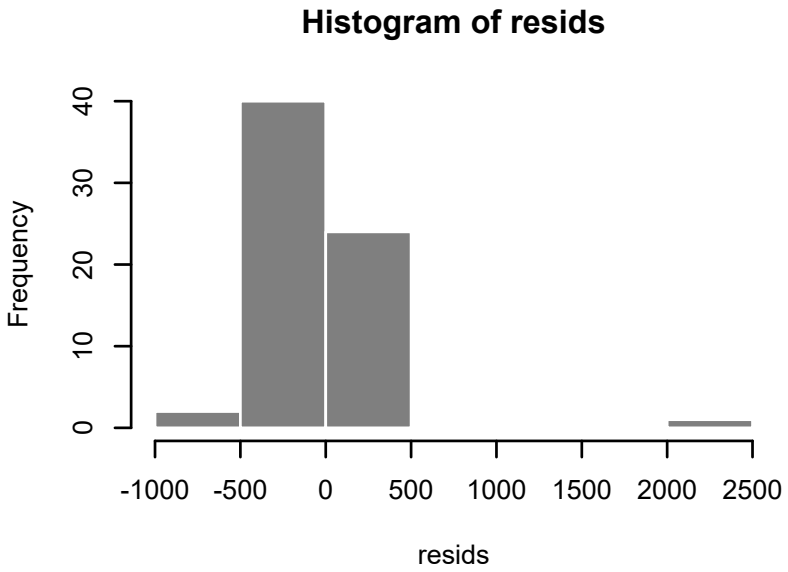
R Code

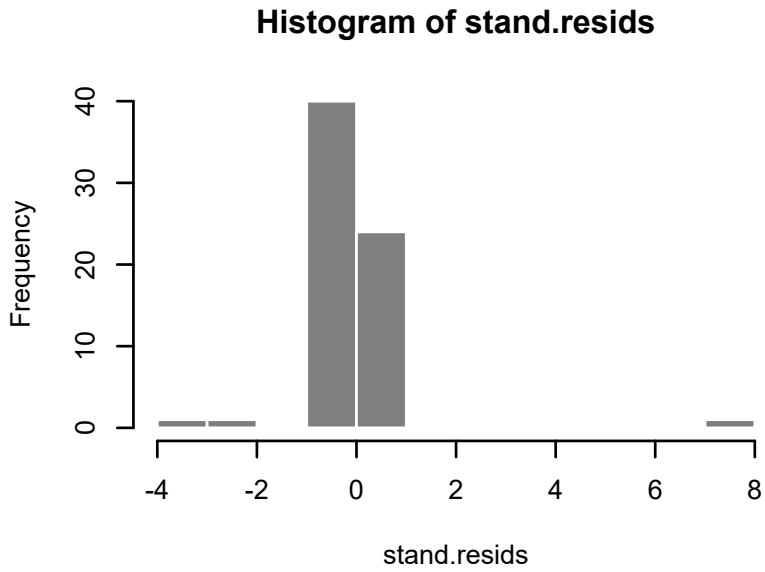
```
> mod1 <- lm(buchanan00~TotalVotes00,data=dta)
> summary(mod1)
Residuals:
    Min      1Q  Median      3Q     Max
-947.05  -41.74  -19.47   20.20 2350.54

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.423e+01  4.914e+01   1.104    0.274
TotalVotes00 2.323e-03  3.104e-04   7.483 2.42e-10 ***
---
Residual standard error: 332.7 on 65 degrees of freedom
Multiple R-squared: 0.4628,        Adjusted R-squared: 0.4545
F-statistic:    56 on 1 and 65 DF,  p-value: 2.417e-10

> residuals              <- resid(mod1)
> standardized_residuals <- rstandard(mod1)
> studentized_residuals  <- rstudent(mod1)
> dotchart(residuals,dta$name,cex=.7,xlab="Residuals")
```

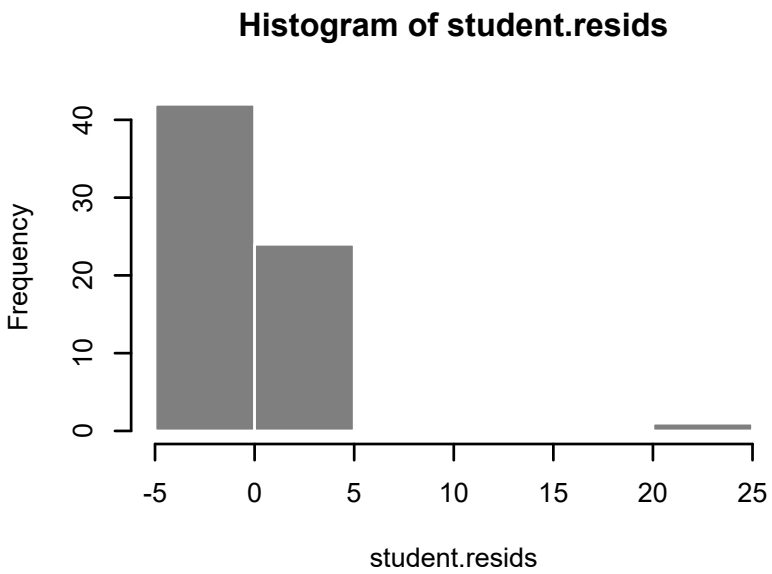
Plotting the residuals



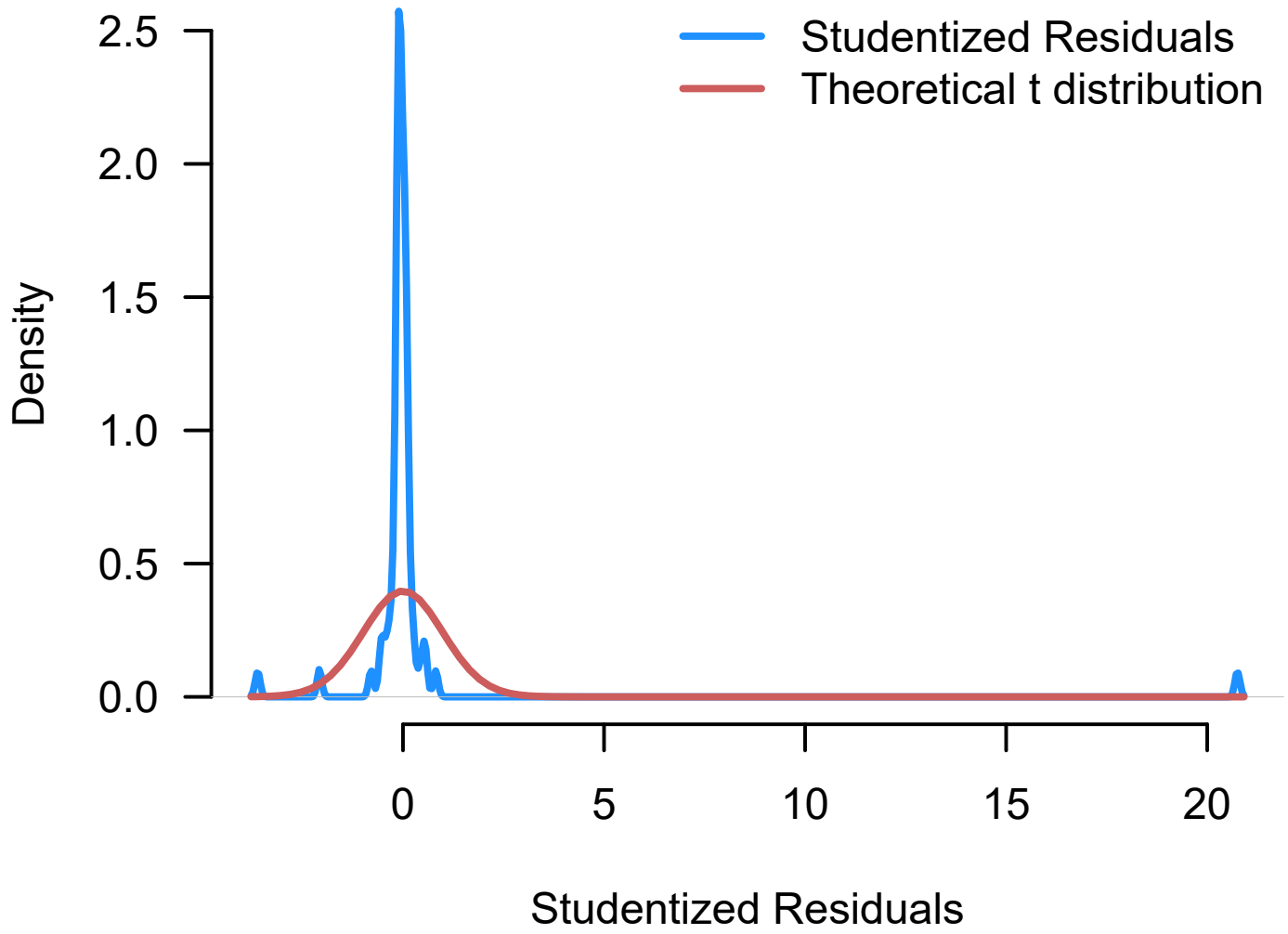
Plotting the residuals







Plotting the residuals



Quantile-Quantile plots

- How can we easily compare our actual distribution of residuals to the theoretical distribution?

Quantile-Quantile plots

- How can we easily compare our actual distribution of residuals to the theoretical distribution?
- Quantile-quantile plot or QQ-plot is useful for comparing distributions

Quantile-Quantile plots

- How can we easily compare our actual distribution of residuals to the theoretical distribution?
- Quantile-quantile plot or QQ-plot is useful for comparing distributions
- Plots the quantiles of one distribution against those of another distribution

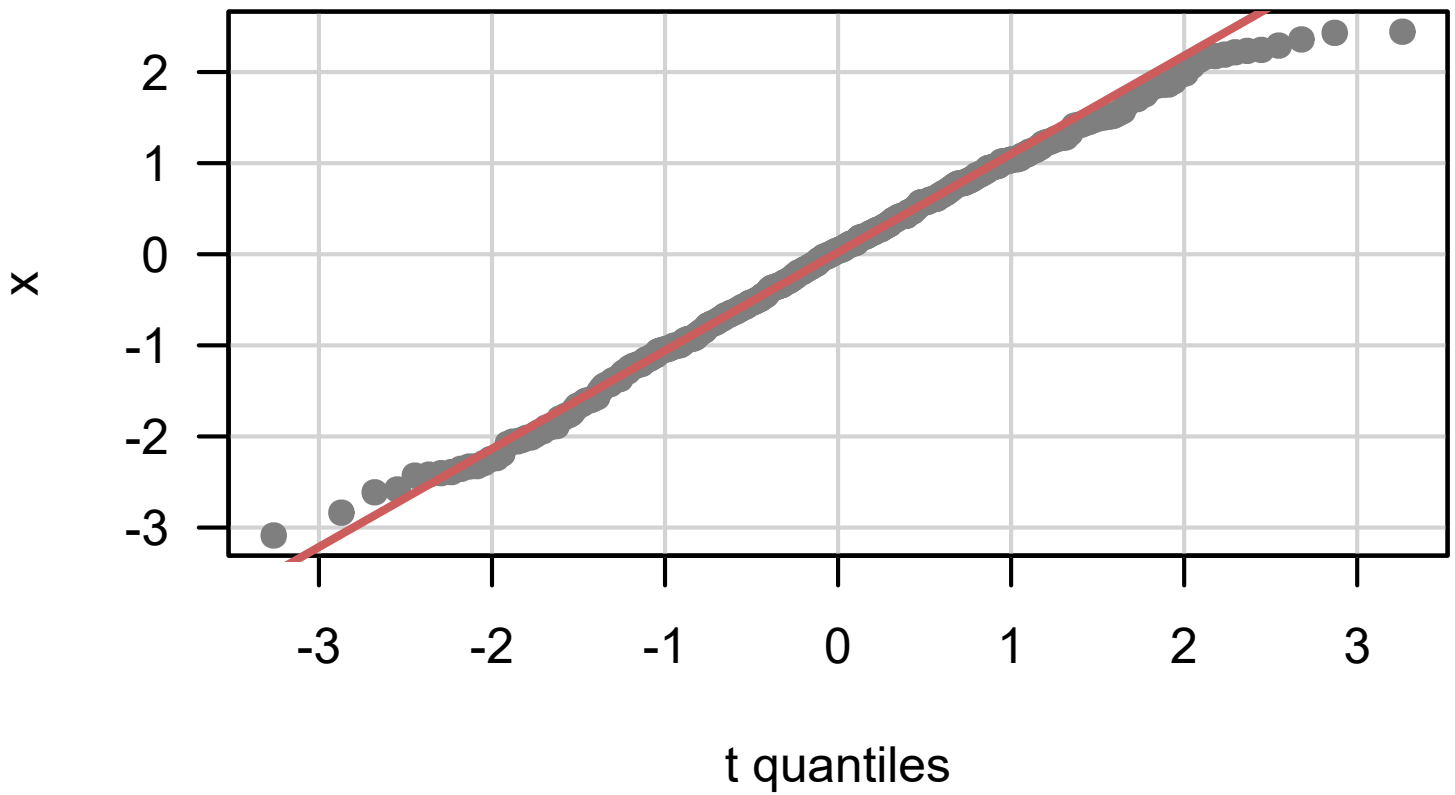
Quantile-Quantile plots

- How can we easily compare our actual distribution of residuals to the theoretical distribution?
- Quantile-quantile plot or QQ-plot is useful for comparing distributions
- Plots the quantiles of one distribution against those of another distribution
- For example, one point is the (m_x, m_y) where m_x is the median of the x distribution and m_y is the median for the y distribution

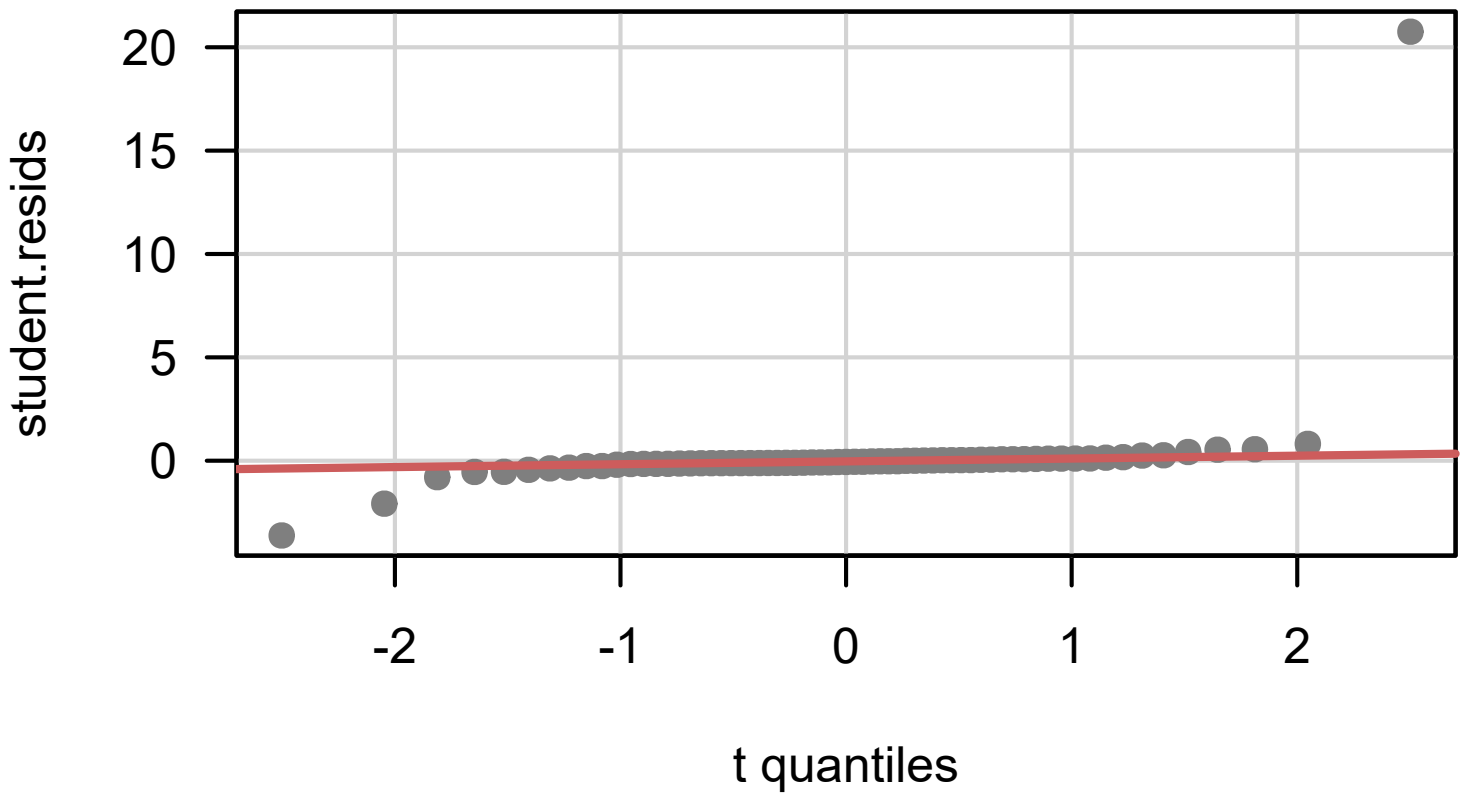
Quantile-Quantile plots

- How can we easily compare our actual distribution of residuals to the theoretical distribution?
- Quantile-quantile plot or QQ-plot is useful for comparing distributions
- Plots the quantiles of one distribution against those of another distribution
- For example, one point is the (m_x, m_y) where m_x is the median of the x distribution and m_y is the median for the y distribution
- If distributions are equal $\implies$ 45 degree line

Good QQ-plot



Buchanan QQ-plot

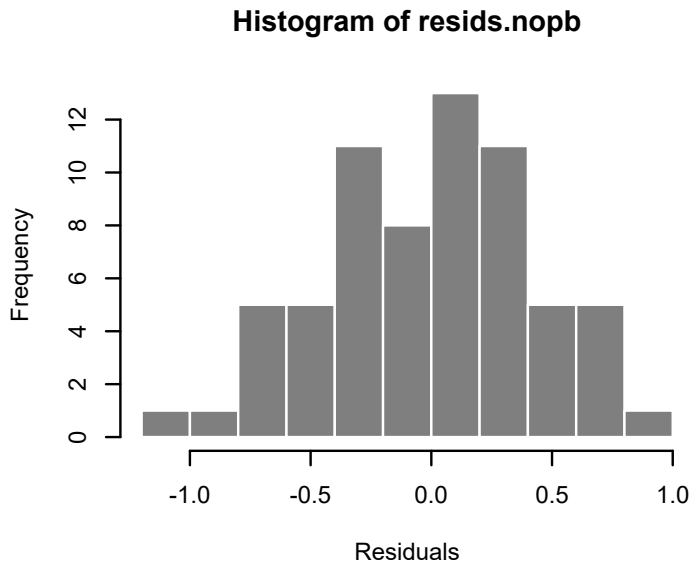


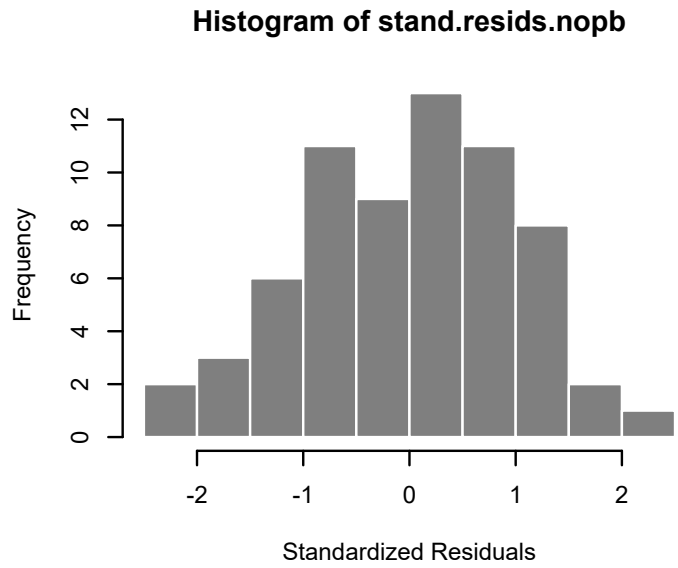
Buchanan Revisited

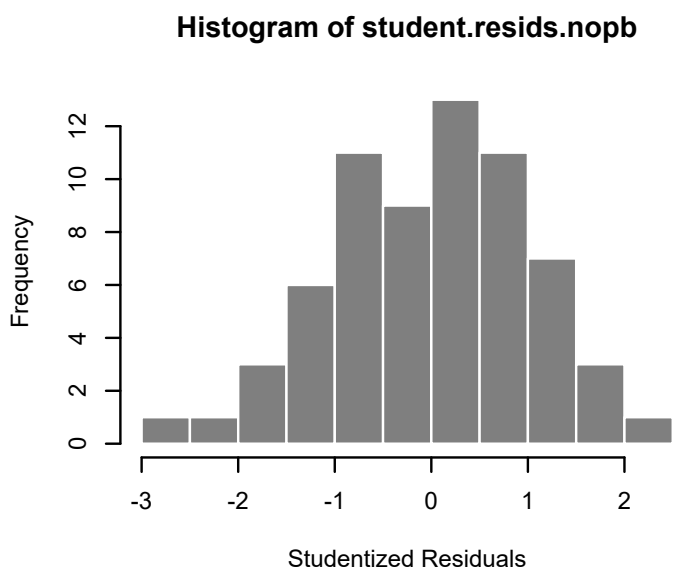
Let's delete Palm Beach and also use log transformations for both variables

```
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -2.48597    0.37889  -6.561 1.09e-08 ***
## log(edaytotal)  0.70311    0.03621  19.417  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4362 on 64 degrees of freedom
## Multiple R-squared:  0.8549, Adjusted R-squared:  0.8526
## F-statistic:   377 on 1 and 64 DF,  p-value: < 2.2e-16
```

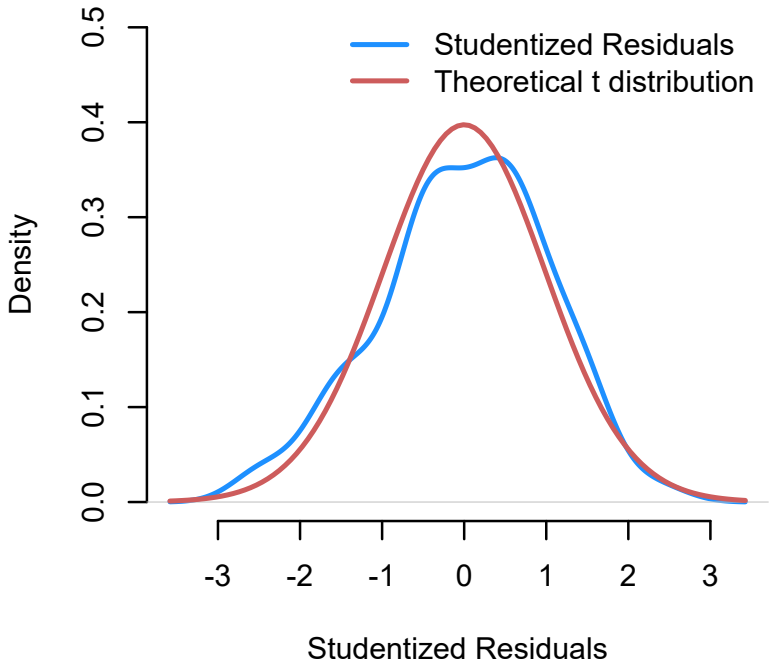
Buchanan Revisited

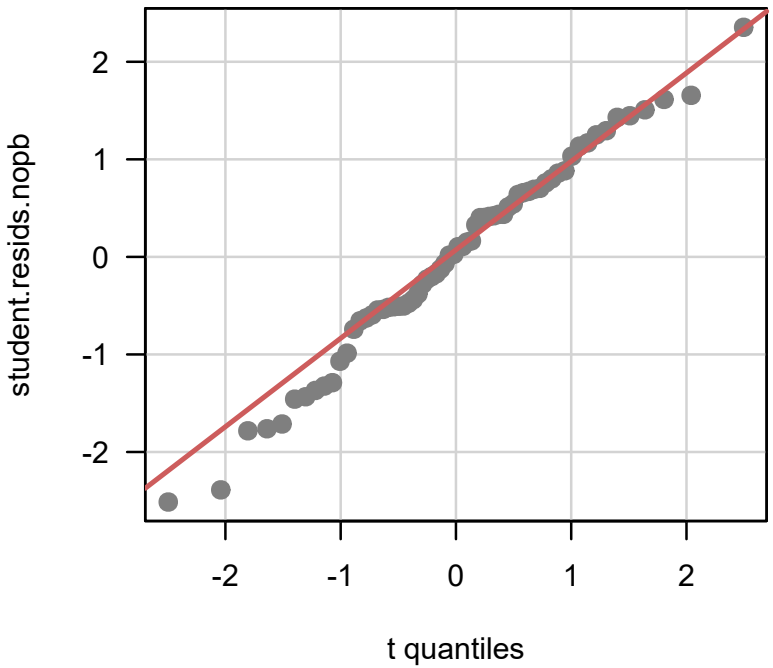






Buchanan Revisited





Summary

- The Hat matrix is useful for lots of things (including the extreme values section this week).

Summary

- The Hat matrix is useful for lots of things (including the extreme values section this week).
- The standardized/studentized residuals presume classical linear model assumptions, but they will come up from time to time.

Summary

- The Hat matrix is useful for lots of things (including the extreme values section this week).
- The standardized/studentized residuals presume classical linear model assumptions, but they will come up from time to time.
- QQ plots are often presented (as here) as diagnostics for normality of the errors, but they are just generically useful for comparing distributions.

1. Opening Themes
2. Preliminaries: The Hat Matrix
3. Fun With (Misleading) Visualization
4. Extreme Values
   - Outliers
   - Leverage Points
   - Influence Points
   - Fun with Outliers
5. Robust Regression Methods
   - Appendix: Robustness
6. Nonlinearity
   - Linear Basis Function Models
   - Generalized Additive Models

Fun With (Misleading) Visualization in the New York Times

AMERICAS

How Stable Are Democracies? 'Warning Signs Are Flashing

The Interpreter

By AMANDA TAUB NOV. 29, 2016

WASHINGTON — Yascha Mounk is used to being the most pessimistic person in the room. Mr. Mounk, a lecturer in government at Harvard, has spent the past few years challenging one of the bedrock assumptions of Western politics: that once a country becomes a liberal democracy, it will stay that way.

His research suggests something quite different: that liberal democracies around the world may be at serious risk of decline.

Mr. Mounk's interest in the topic began rather unusually. In 2014, he published a book, "Stranger in My Own Country." It started as a memoir of his experiences growing up as a Jew in Germany, but became a broader

Fun With (Misleading) Visualization in the New York Times

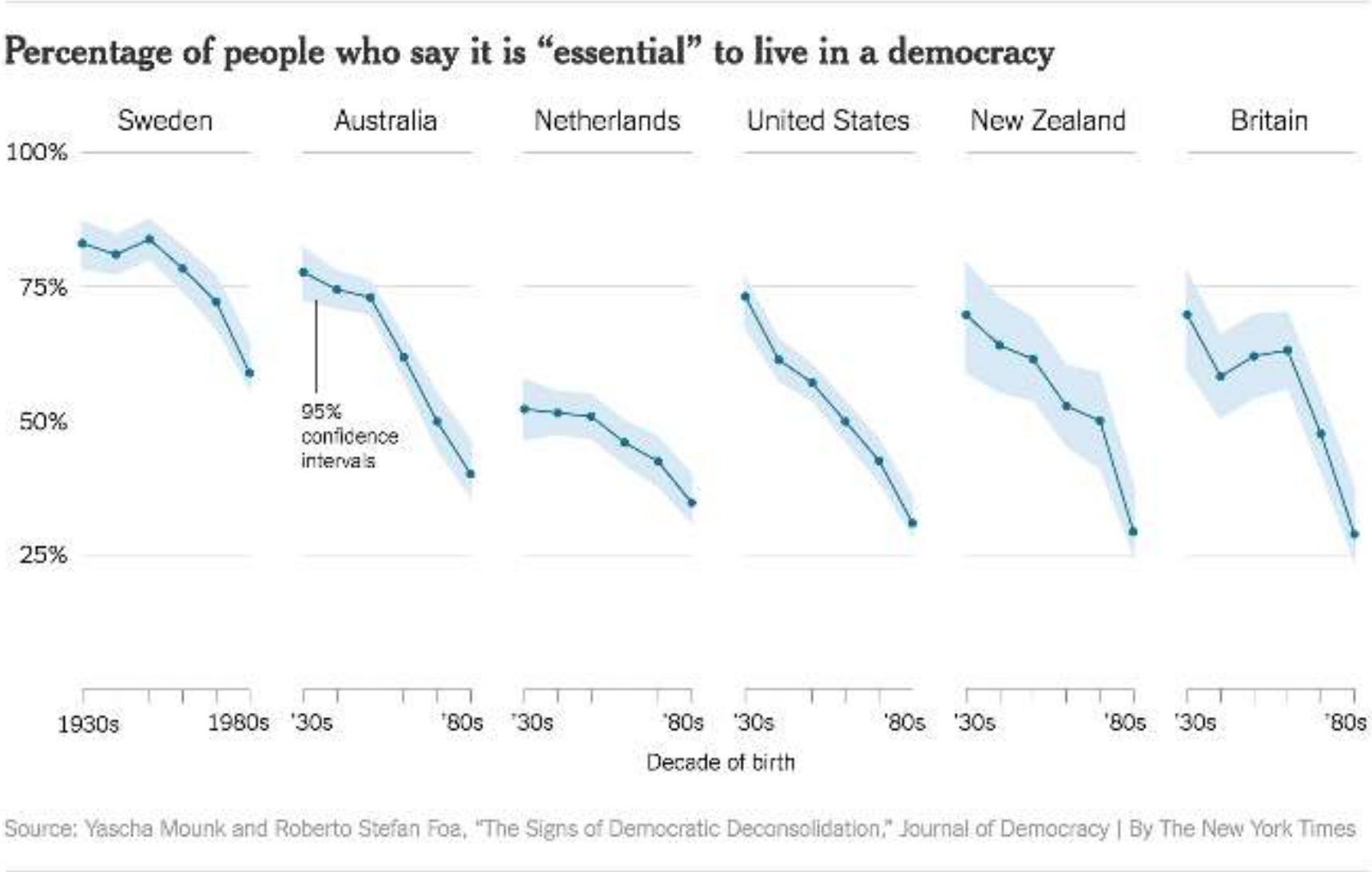
The Danger of Deconsolidation

THE DEMOCRATIC DISCONNECT

Roberto Stefan Foa and Yascha Mounk

Roberto Stefan Foa *is a principal investigator of the World Values Survey and fellow of the Laboratory for Comparative Social Research. His writing has appeared in a wide range of journals, books, and publications by the UN, OECD, and World Bank.* ***Yascha Mounk*** *is a lecturer on political theory in Harvard University's Government Department and a Carnegie Fellow at New America, a Washington, D.C.–based think tank. His dissertation on the role of personal responsibility in contemporary politics and philosophy will be published by Harvard University Press, and his essays have appeared in* Foreign Affairs, *the* New York Times, *and the* Wall Street Journal.

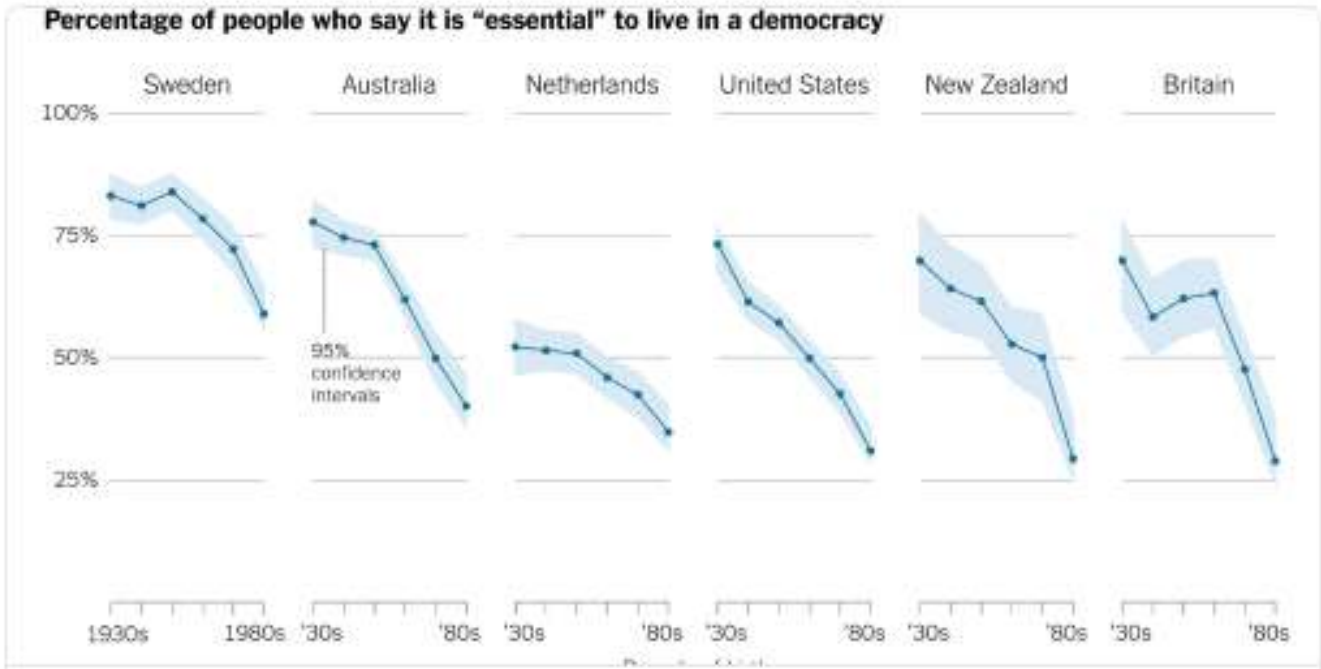
Fun With (Misleading) Visualization in the New York Times



Fun With (Misleading) Visualization in the New York Times

Ryan D. Enos @RyanDEnos · 19h

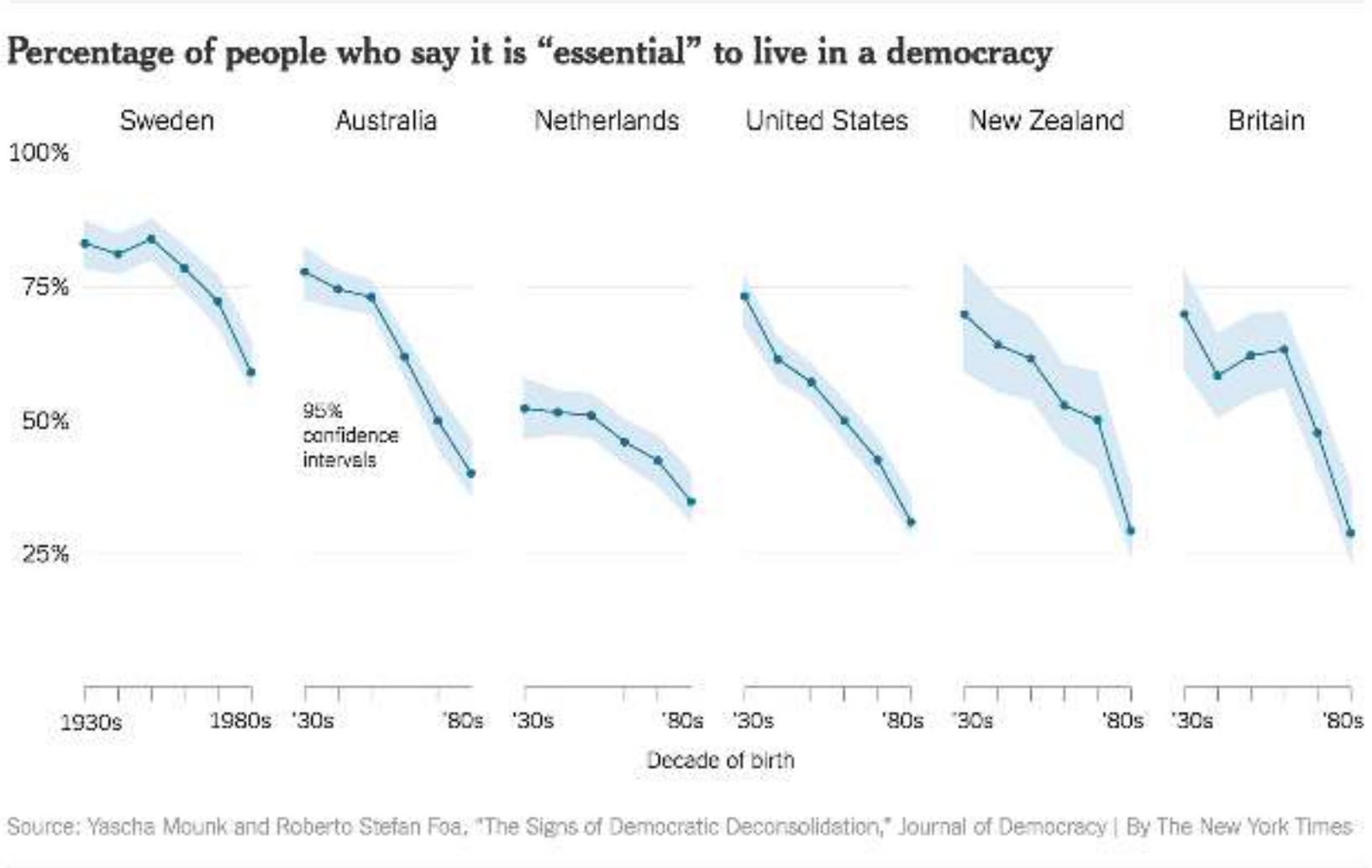
Lots of worried chatter a/b @amandataub article on work of @Yascha_Mounk. Important, but want to raise cautions 1/



How Stable Are Democracies? 'Warning Signs Are Flashing Red'

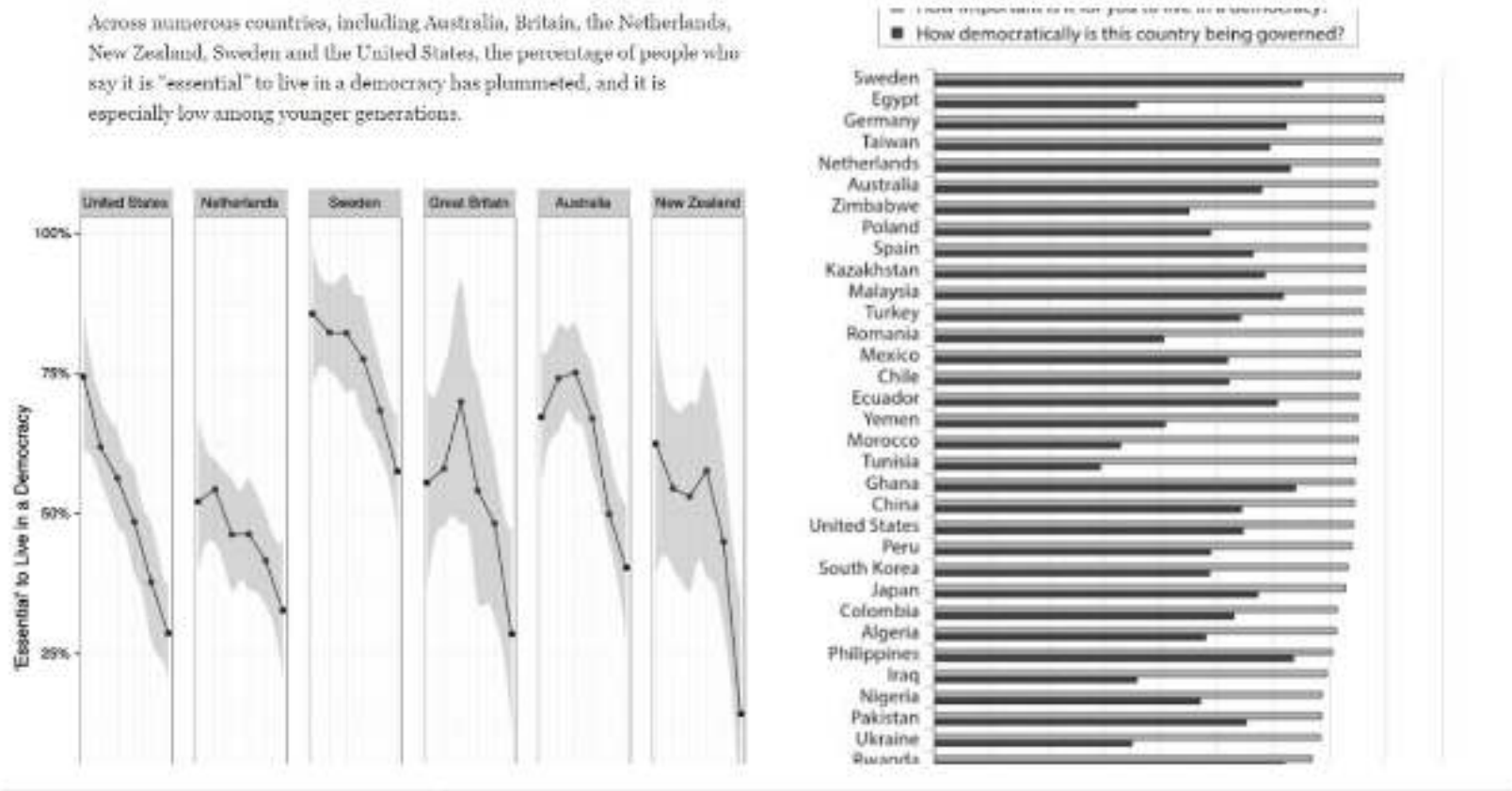
New research tries to spot the collapse of liberal democracies before they happen, and it suggests that Western democracy may be seriously ill.

Alternate Graphs



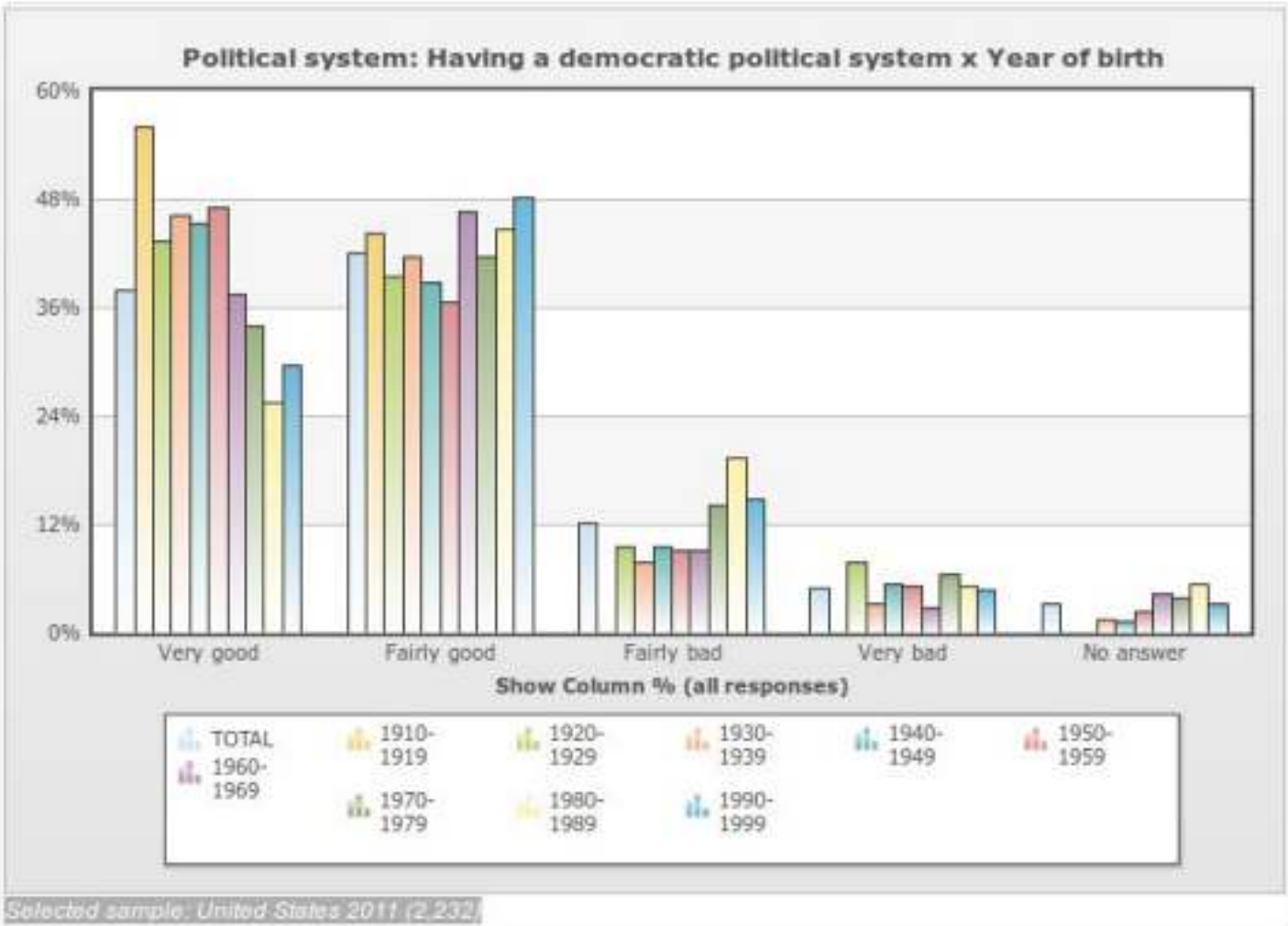
Alternate Graphs

.@RyanDEnos Compare NYT/JoD (left) to the very same data analysed differently by Bartels and Achen (2016) (right). Extreme score vs means.

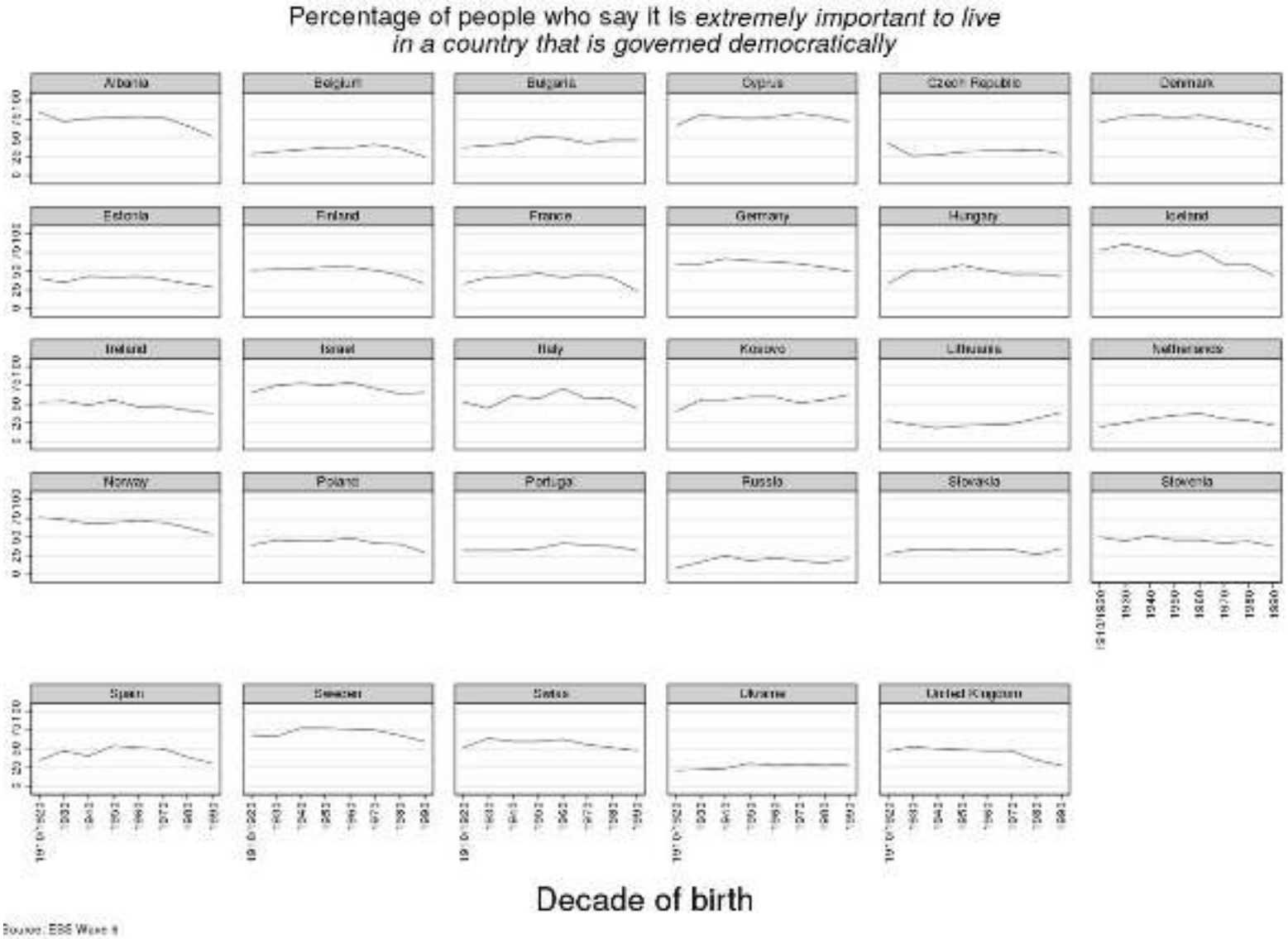


Alternate Graphs

@RyanDEnos They also stop at the 80s cohort. The data has the 90's as well. I wonder why they would stop there...



Alternate Graphs

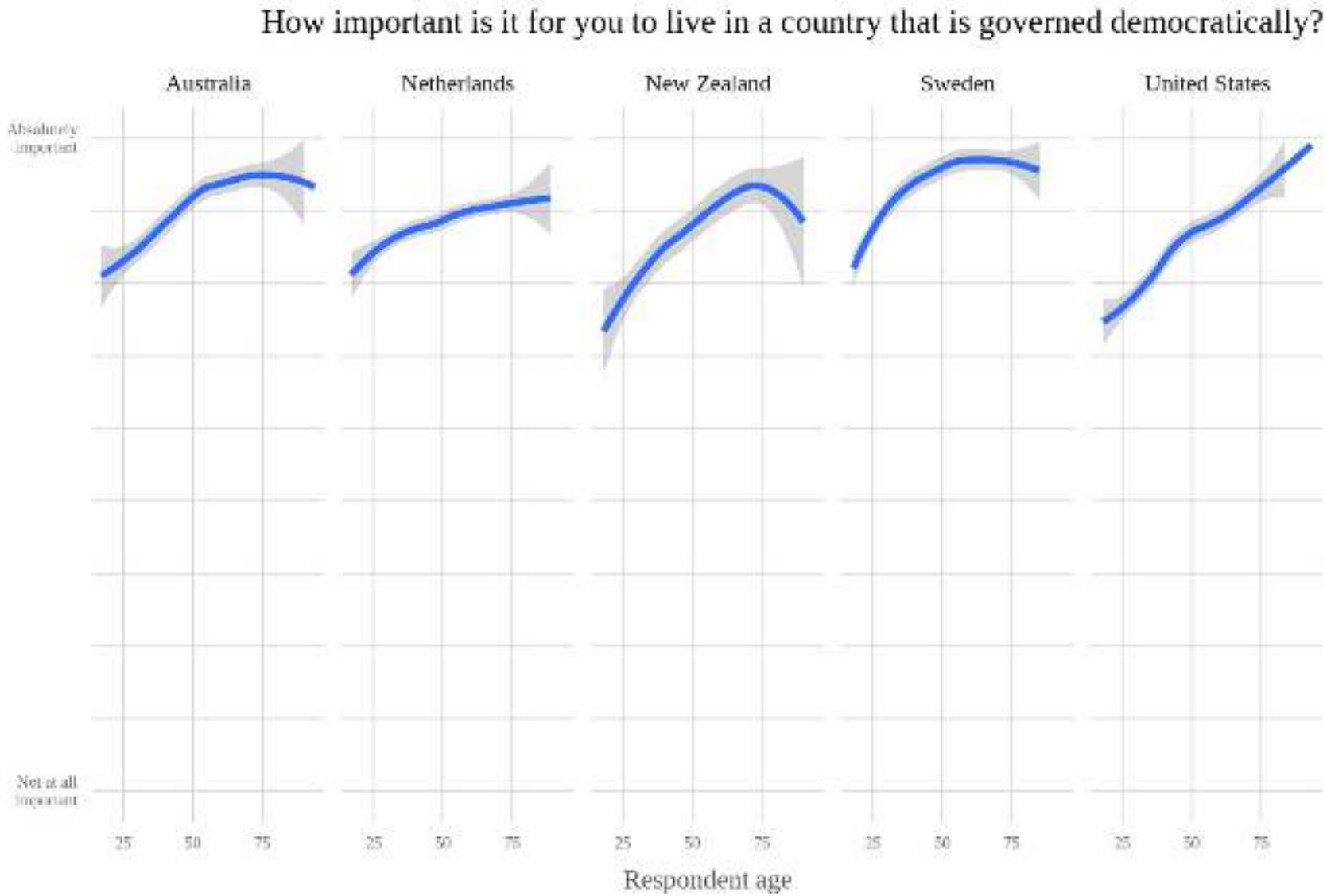


In reply to Ryan D. Enos

Benjamin Sack @bcsack · 15h

@RyanDEnos Same analysis strategy with comparable data from @ESS_Survey (similar item, 0-10 scale) shows slightly different pattern, too.

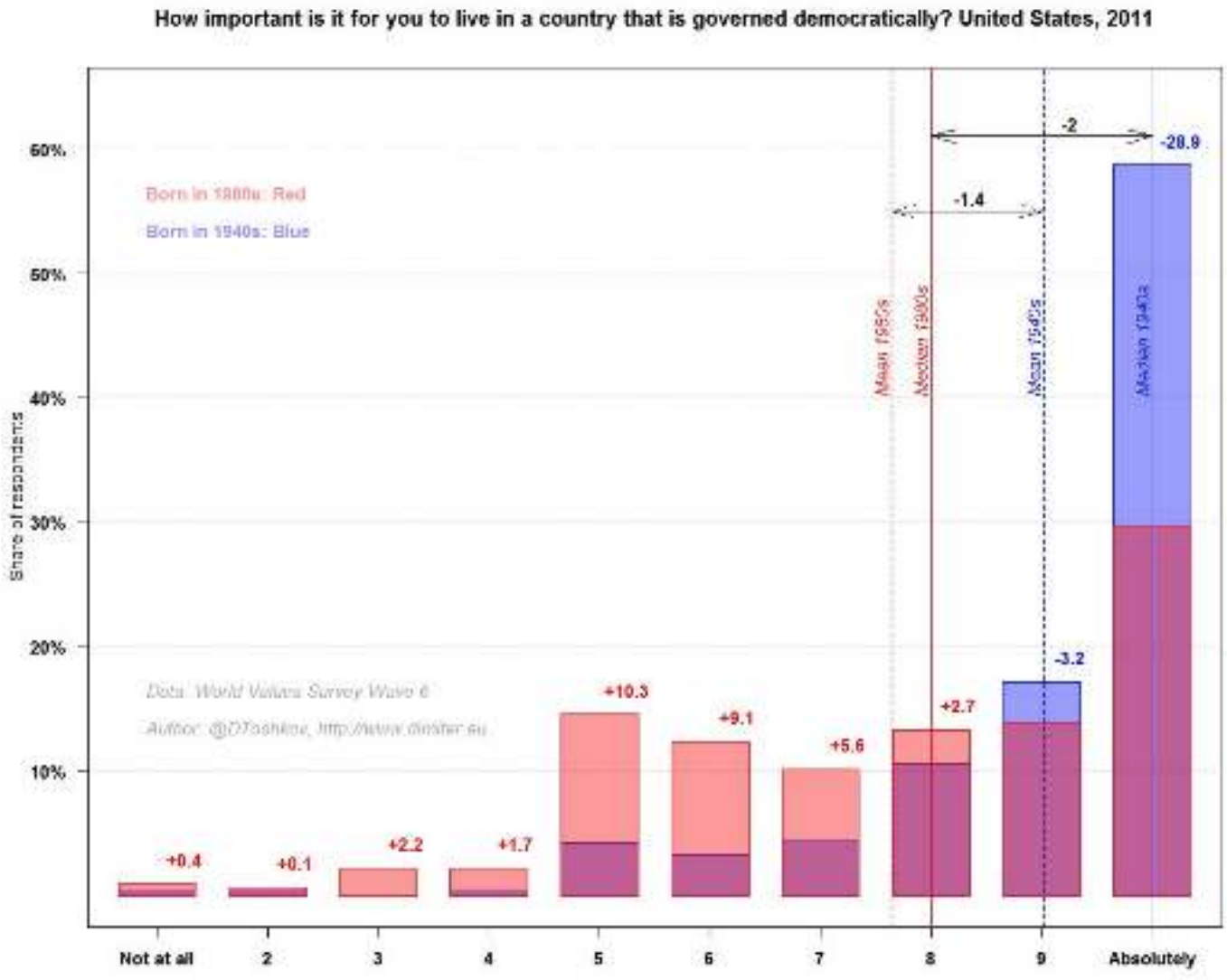
Alternate Graphs



614 Bantam @jpbnch · 16h

@RyanDEnos @bshor @natallemjb @TomWCvdMeer this is a "quick and dirty" plot I did with WVS wave 6. Not quite so terrifying.

Alternate Graphs



Dimiter Toshkov @DToshkov · 31m

my take on the democratic deconsolidation graph that scared everyone yesterday. Blue is 1940s cohort, red is 1980s. First, United States

The Trouble with Norway

The Trouble with Norway

- Lange and Garrett (1985): organizational and political power of labor interact to improve economic growth

The Trouble with Norway

- Lange and Garrett (1985): organizational and political power of labor interact to improve economic growth
- Jackman (1987): relationship just due to North Sea Oil?

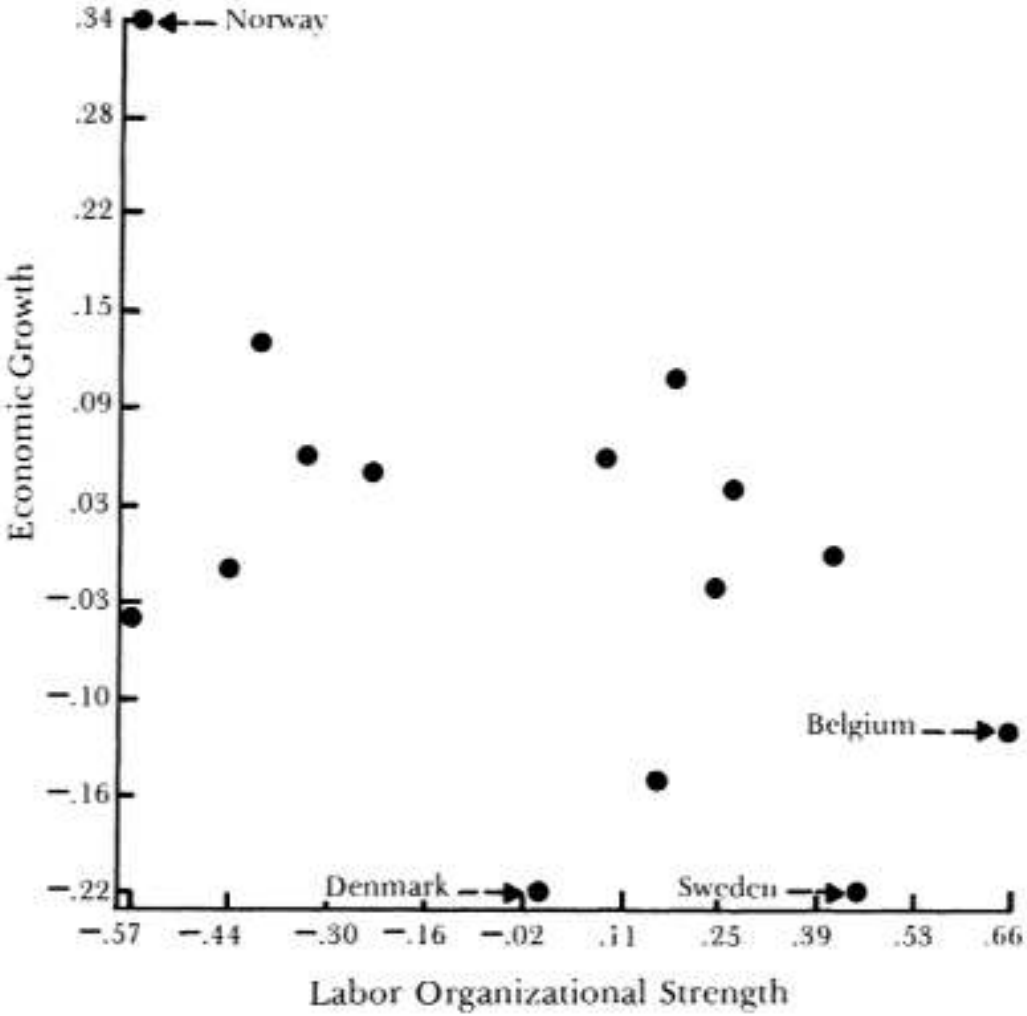
The Trouble with Norway

- Lange and Garrett (1985): organizational and political power of labor interact to improve economic growth
- Jackman (1987): relationship just due to North Sea Oil?
- Table guide:
  - x_1 = organizational power of labor
  - x_2 = political power of labor
  - Parentheses contain t-statistics

The Trouble with Norway

- Lange and Garrett (1985): organizational and political power of labor interact to improve economic growth
- Jackman (1987): relationship just due to North Sea Oil?
- Table guide:
  - x_1 = organizational power of labor
  - x_2 = political power of labor
  - Parentheses contain t-statistics

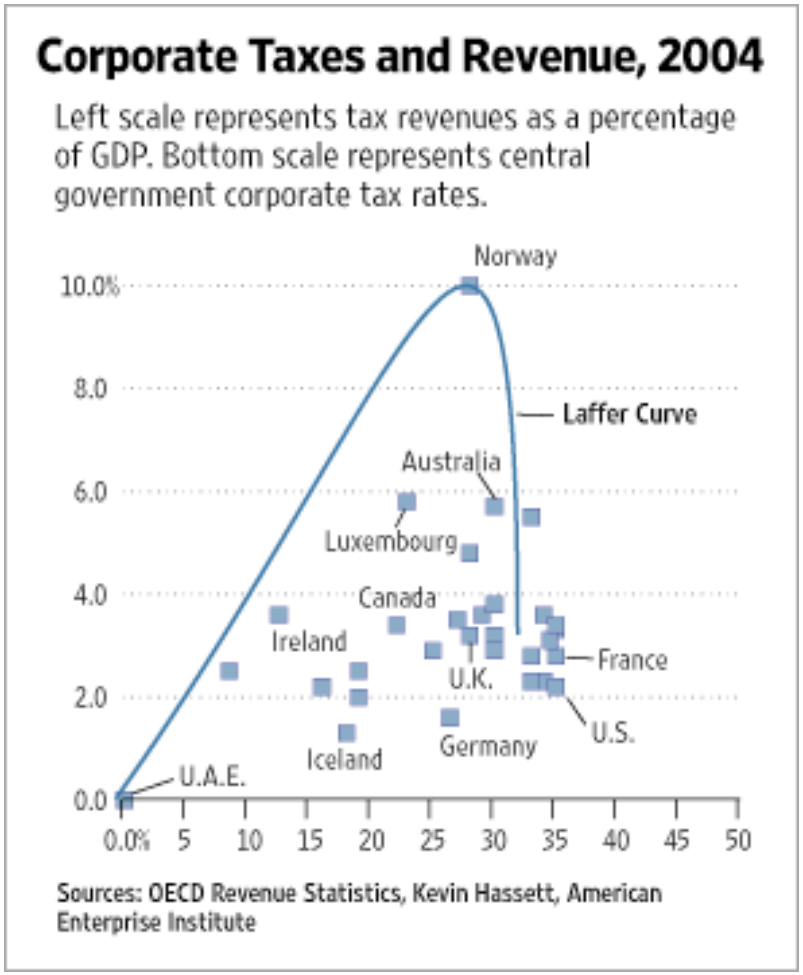
	Constant	x_1	x_2	$x_1 \cdot x_2$
Norway Obs Included	.814	-.192	-.278	.137
	(4.7)	(2.0)	(2.4)	(2.9)
Norway Obs Excluded	.641	-.068	-.138	.054
	(4.8)	(0.9)	(1.5)	(1.3)



Source: Jackman (1987), Fig 1 Norway looks different. But should we exclude Norway? What about Belgium?

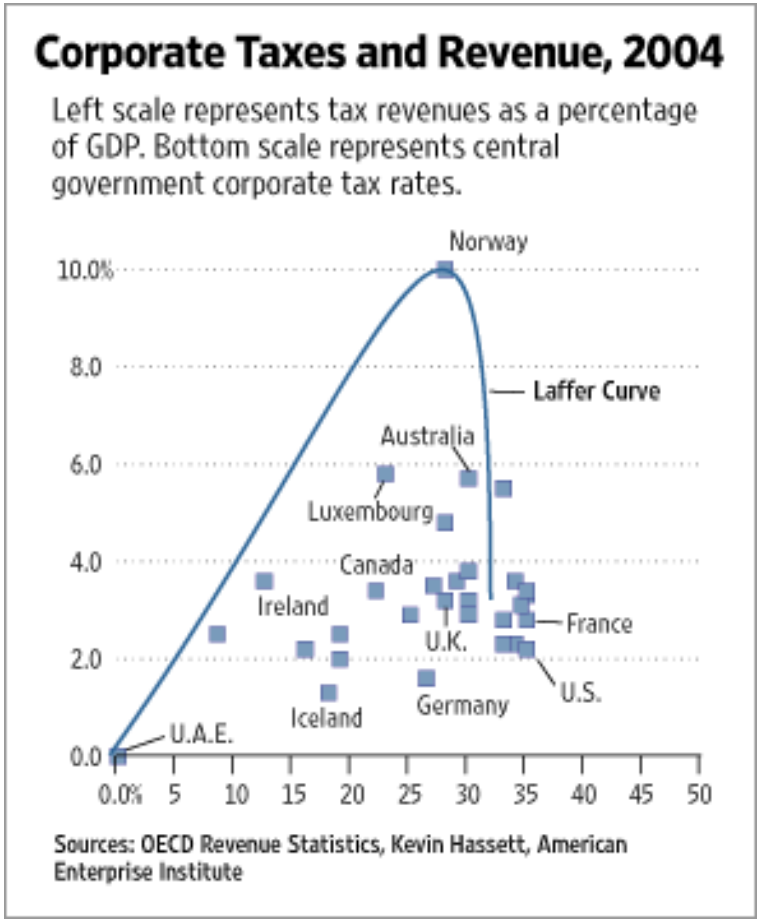
Creative Curve Fitting with Norway

Creative Curve Fitting with Norway

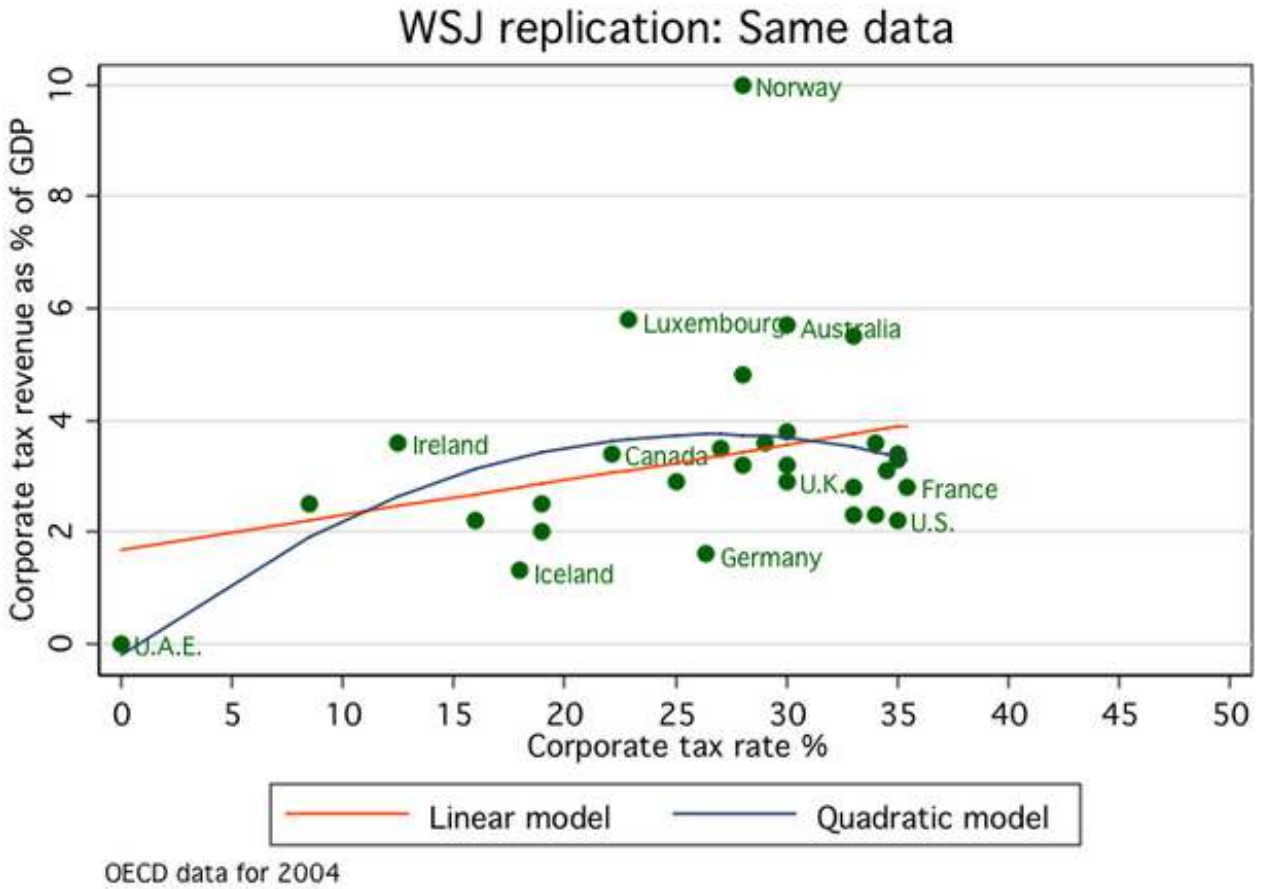


Laffer curve: relationship between rates of taxation and the resulting levels of the government's tax revenue.

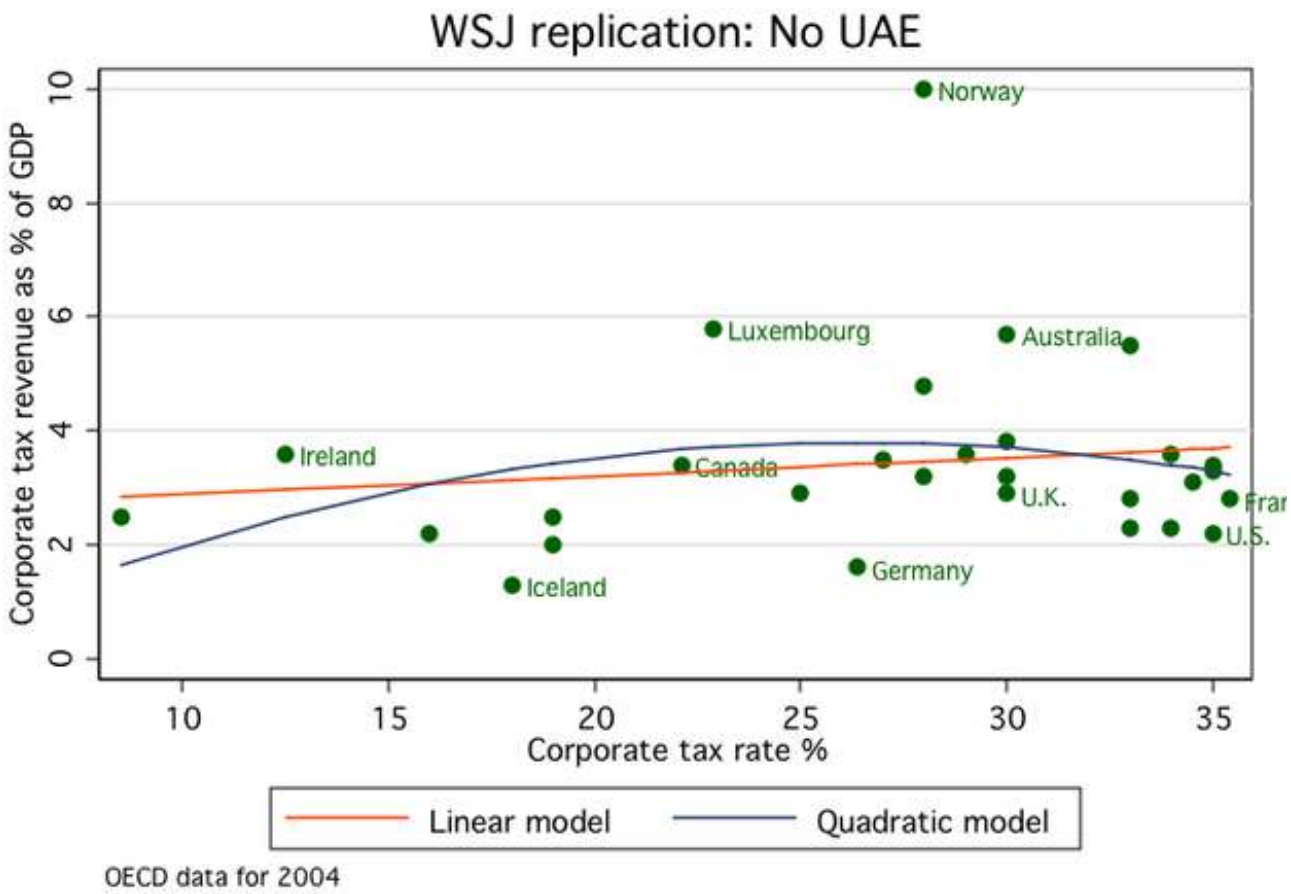
Laffer curve: relationship between rates of taxation and the resulting levels of the government's tax revenue.



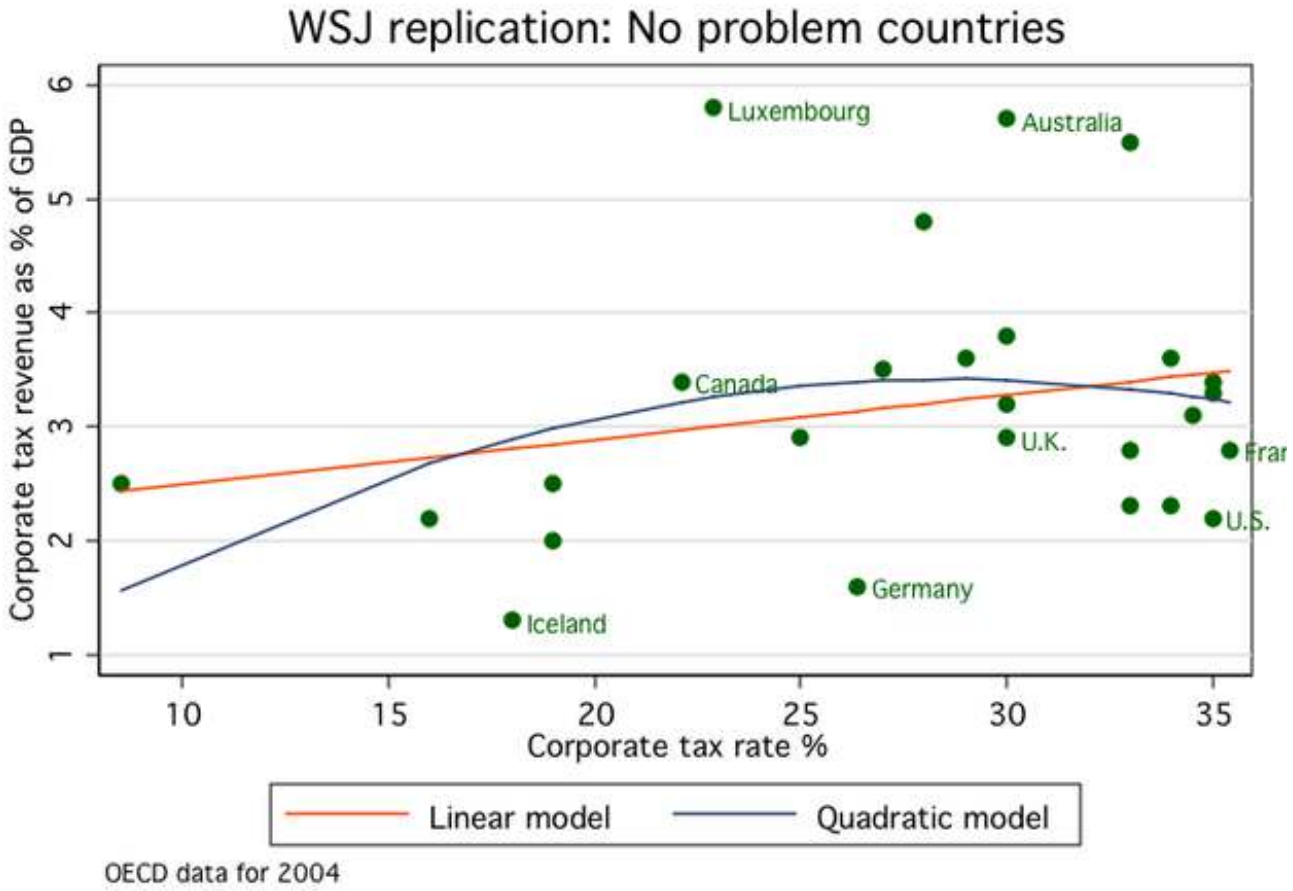
Source: `https://www.brendan-nyhan.com/blog/2007/08/replicating-the.html` (hat tip to Matt Salganik)



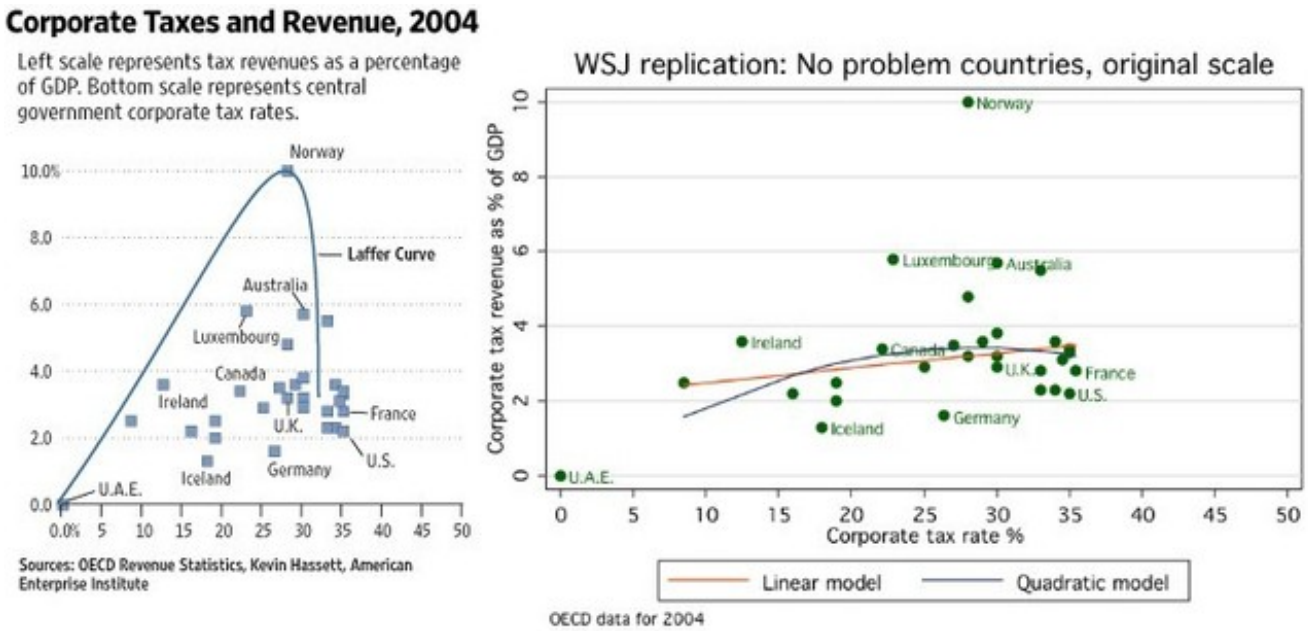
Replication with same data



Replication with same data minus UAE (which is not in OECD)



Replication with same data minus UAE (which is not in OECD) and "problem countries": "Ireland (a noted tax haven), Norway (a country with unusual oil revenues), and Switzerland (a country with significant internal variation in taxation),"



Original vs modified replication

This example showed:

- non-linearity (maybe) and influential points (maybe)
- Some things seem pretty clear (drop UAE from a plot of OECD countries) and some don't (drop Iceland, Norway, and Switzerland)

1. Opening Themes
2. Preliminaries: The Hat Matrix
3. Fun With (Misleading) Visualization
4. Extreme Values
   - Outliers
   - Leverage Points
   - Influence Points
   - Fun with Outliers
5. Robust Regression Methods
   - Appendix: Robustness
6. Nonlinearity
   - Linear Basis Function Models
   - Generalized Additive Models

Three Types of Extreme Values

Three Types of Extreme Values

1. Outlier: extreme in the y direction

Three Types of Extreme Values

1. Outlier: extreme in the y direction
2. Leverage point: extreme in one x direction

Three Types of Extreme Values

1. Outlier: extreme in the y direction
2. Leverage point: extreme in one x direction
3. Influence point: extreme in both directions

Three Types of Extreme Values

1. Outlier: extreme in the y direction
2. Leverage point: extreme in one x direction
3. Influence point: extreme in both directions

- Not all of these are problematic

Three Types of Extreme Values

1. Outlier: extreme in the y direction
2. Leverage point: extreme in one x direction
3. Influence point: extreme in both directions

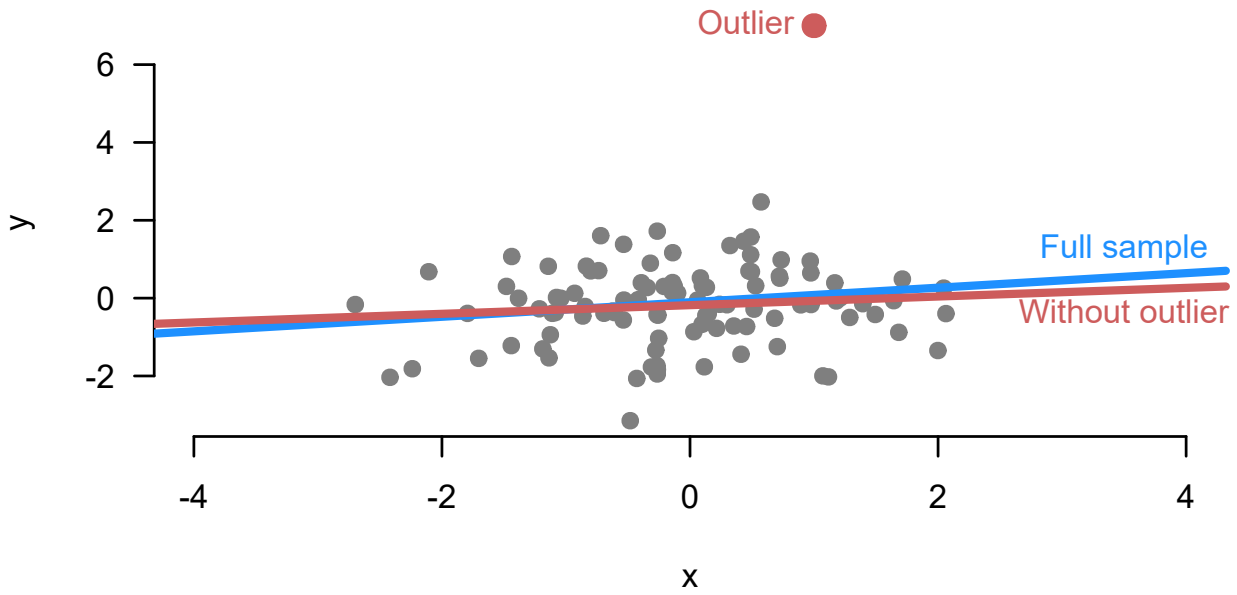
- Not all of these are problematic
- If the data are truly "contaminated" (come from a different distribution), can cause inefficiency and possibly bias

Three Types of Extreme Values

1. Outlier: extreme in the y direction
2. Leverage point: extreme in one x direction
3. Influence point: extreme in both directions

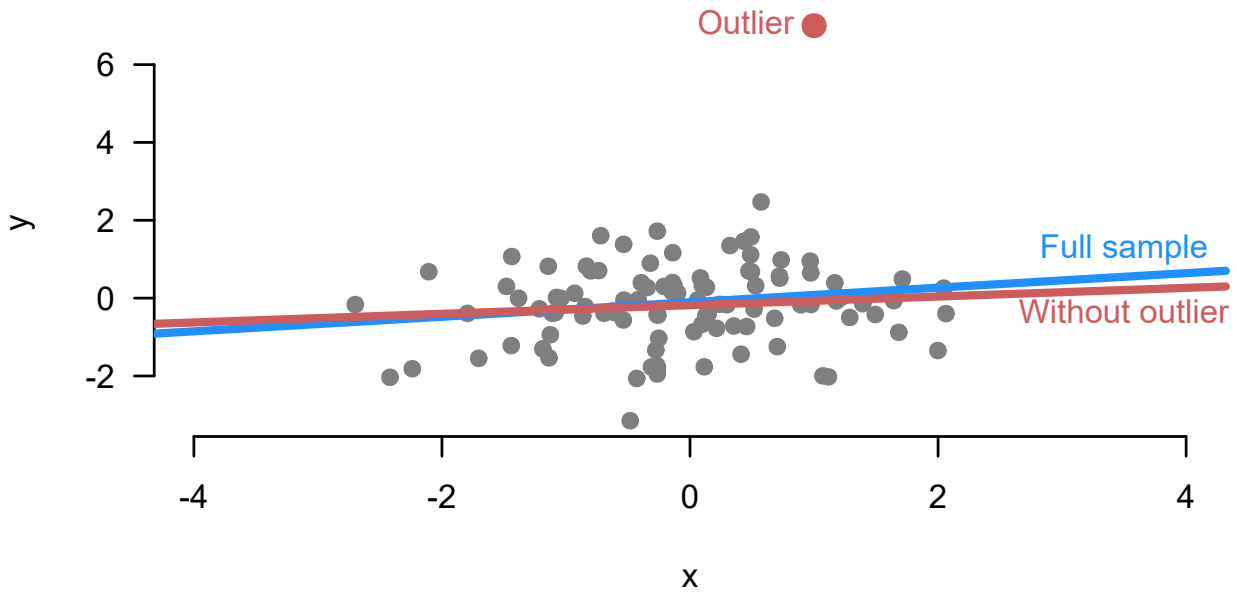
- Not all of these are problematic
- If the data are truly "contaminated" (come from a different distribution), can cause inefficiency and possibly bias
- Can be a violation of iid (not identically distributed)

Outlier Definition



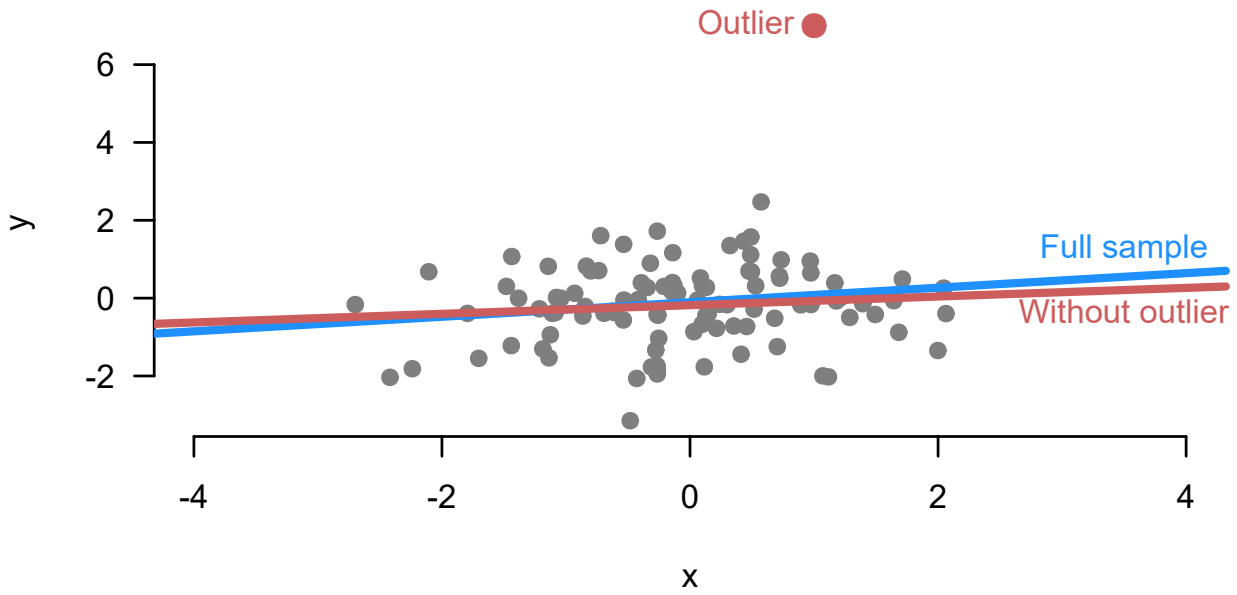
- An **outlier** is a data point with very large regression errors, u_i

Outlier Definition



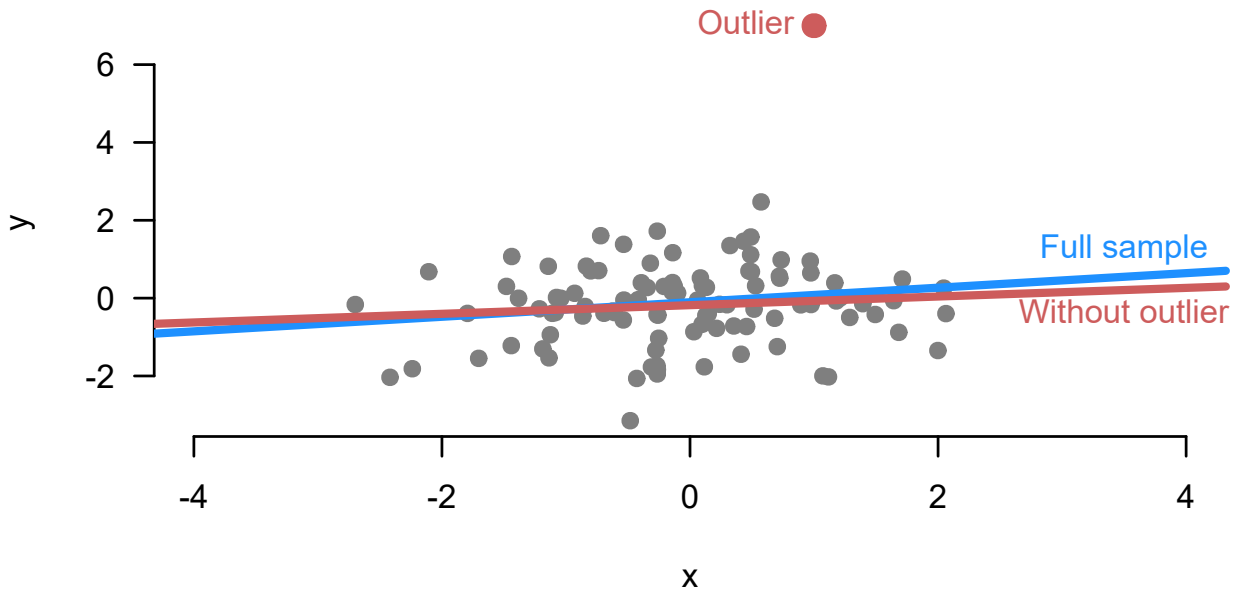
- An **outlier** is a data point with very large regression errors, u_i
- Very distant from the rest of the data in the y-dimension

Outlier Definition



- An **outlier** is a data point with very large regression errors, u_i
- Very distant from the rest of the data in the y-dimension
- Increases standard errors (by increasing $\widehat{\sigma}^2$)

Outlier Definition



- An **outlier** is a data point with very large regression errors, u_i
- Very distant from the rest of the data in the y-dimension
- Increases standard errors (by increasing $\widehat{\sigma}^2$)
- No bias if typical in the x's

Detecting Outliers

- Look at standardized residuals, $\widehat{u}_i'$?

Detecting Outliers

- Look at standardized residuals, $\widehat{u}_i'$?
  - but $\widehat{\sigma}^2$ could be biased upwards by the large residual from the outlier

Detecting Outliers

- Look at standardized residuals, $\widehat{u}_i'$?
  - but $\widehat{\sigma}^2$ could be biased upwards by the large residual from the outlier
  - Makes detecting residuals harder

Detecting Outliers

- Look at standardized residuals, $\widehat{u}_i'$?
  - but $\widehat{\sigma}^2$ could be biased upwards by the large residual from the outlier
  - Makes detecting residuals harder
- Possible solution: use studentized residuals

$$\widehat{u}_i^* = \frac{\widehat{u}_i}{\widehat{\sigma}_{-i}\sqrt{1-h_i}}$$

Detecting Outliers

- Look at standardized residuals, $\widehat{u}_i'$?
  - but $\widehat{\sigma}^2$ could be biased upwards by the large residual from the outlier
  - Makes detecting residuals harder
- Possible solution: use studentized residuals

$$\widehat{u}_i^* = \frac{\widehat{u}_i}{\widehat{\sigma}_{-i}\sqrt{1-h_i}}$$

- $\widehat{\sigma} > \widehat{\sigma}_{-i}$ because we drop the large residual from the outlier, and so $\widehat{u}_i' < \widehat{u}_i^*$

Cutoff Rules for Outliers

- The studentized residuals follow a t distribution, $u_i^* \sim t_{n-k-2}$, when $u_i \sim N(0, \sigma^2)$

Cutoff Rules for Outliers

- The studentized residuals follow a t distribution, $u_i^* \sim t_{n-k-2}$, when $u_i \sim N(0, \sigma^2)$
- Rule of thumb: $|\widehat{u_i^*}| > 2$ will be relatively rare

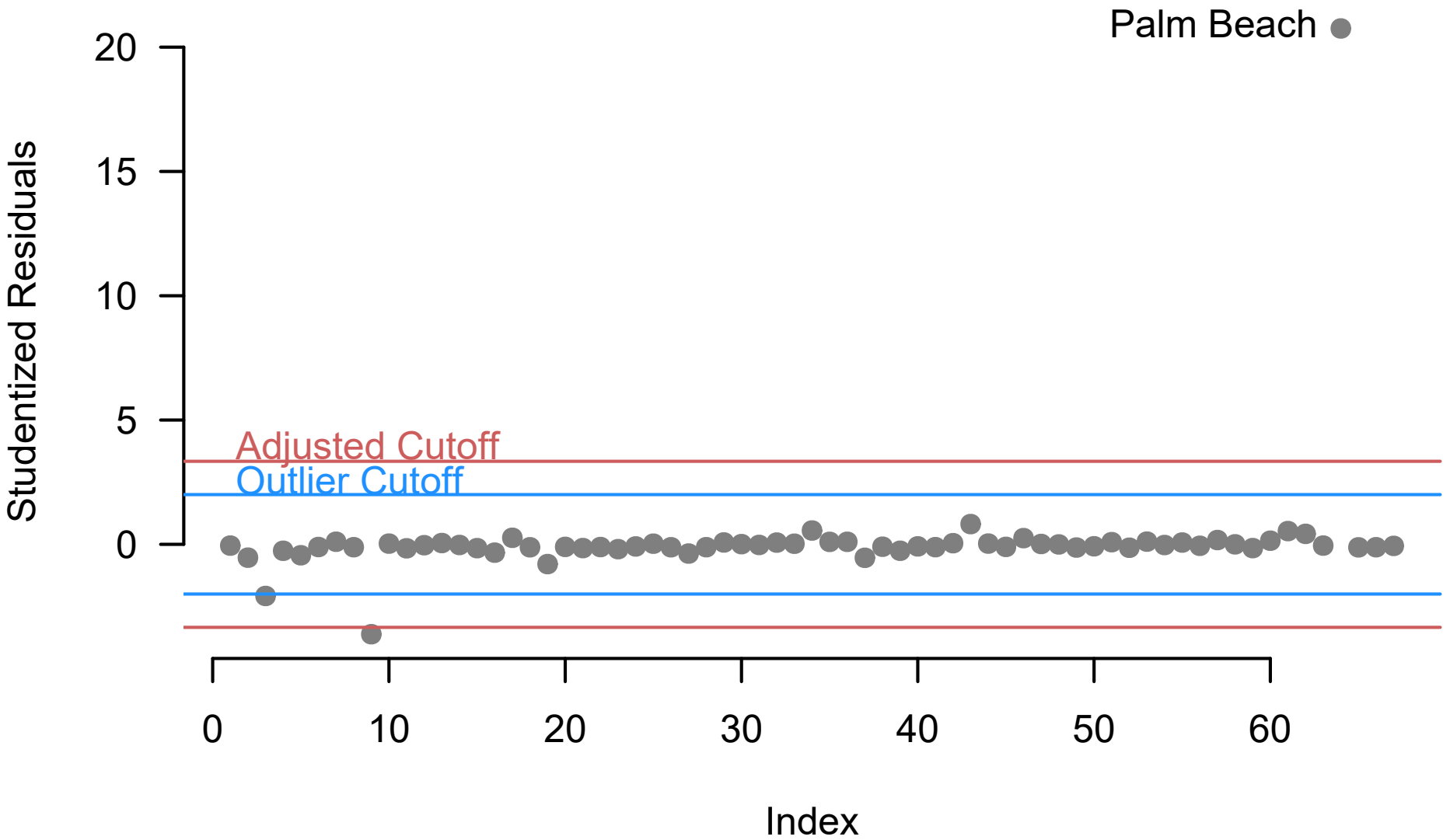
Cutoff Rules for Outliers

- The studentized residuals follow a t distribution, $u_i^* \sim t_{n-k-2}$, when $u_i \sim N(0, \sigma^2)$
- Rule of thumb: $|\widehat{u_i^*}| > 2$ will be relatively rare
- Extreme outliers, $|\widehat{u_i^*}| > 4 - 5$ are much less likely

Cutoff Rules for Outliers

- The studentized residuals follow a t distribution, $u_i^* \sim t_{n-k-2}$, when $u_i \sim N(0, \sigma^2)$
- Rule of thumb: $|\widehat{u_i^*}| > 2$ will be relatively rare
- Extreme outliers, $|\widehat{u_i^*}| > 4 - 5$ are much less likely
- People usually adjust cutoff for multiple testing

Buchanan outliers



What to do about outliers

- Is the data corrupted?

What to do about outliers

- Is the data corrupted?
  - Fix the observation (obvious data entry errors)

What to do about outliers

- Is the data corrupted?
  - Fix the observation (obvious data entry errors)
  - Remove the observation

What to do about outliers

- Is the data corrupted?
  - Fix the observation (obvious data entry errors)
  - Remove the observation
  - Be transparent either way

What to do about outliers

- Is the data corrupted?
  - Fix the observation (obvious data entry errors)
  - Remove the observation
  - Be transparent either way
- Is the outlier part of the data generating process?

What to do about outliers

- Is the data corrupted?
  - Fix the observation (obvious data entry errors)
  - Remove the observation
  - Be transparent either way
- Is the outlier part of the data generating process?
  - Transform the dependent variable ($\log(y)$)

What to do about outliers

- Is the data corrupted?
  - Fix the observation (obvious data entry errors)
  - Remove the observation
  - Be transparent either way
- Is the outlier part of the data generating process?
  - Transform the dependent variable ($\log(y)$)
  - Use a method that is robust to outliers (robust regression)

What to do about outliers

- Is the data corrupted?
  - Fix the observation (obvious data entry errors)
  - Remove the observation
  - Be transparent either way
- Is the outlier part of the data generating process?
  - Transform the dependent variable ($\log(y)$)
  - Use a method that is robust to outliers (robust regression)
- Key question is what is the goal? If you want to estimate the expectation of a distribution and a property of that distribution is extreme observations, that's just part of the story.

Stylized Anecdote: The "Discovery" of the Ozone Hole

Stylized Anecdote: The "Discovery" of the Ozone Hole

- In the late 70s, NASA used an automated data processing program on satellite measurements of atmospheric data to track changes in atmospheric variables such as ozone.

Stylized Anecdote: The "Discovery" of the Ozone Hole

- In the late 70s, NASA used an automated data processing program on satellite measurements of atmospheric data to track changes in atmospheric variables such as ozone.

- This data "quality control" algorithm rejected abnormally low readings of ozone over the Antarctic as unreasonable.

Stylized Anecdote: The "Discovery" of the Ozone Hole

- In the late 70s, NASA used an automated data processing program on satellite measurements of atmospheric data to track changes in atmospheric variables such as ozone.
- This data "quality control" algorithm rejected abnormally low readings of ozone over the Antarctic as unreasonable.
- This (arguably) delayed the detection of the ozone hole by several years until British Antarctic Survey scientists discovered it based on analysis of their own observations (*Nature*, May 1985).

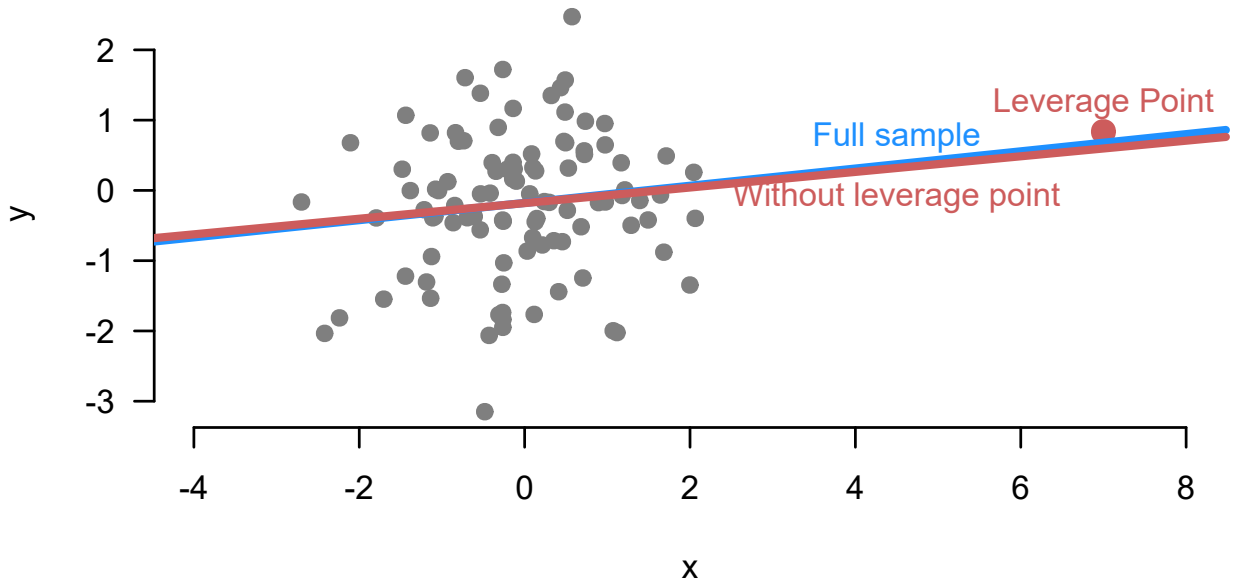
Stylized Anecdote: The "Discovery" of the Ozone Hole

- In the late 70s, NASA used an automated data processing program on satellite measurements of atmospheric data to track changes in atmospheric variables such as ozone.
- This data "quality control" algorithm rejected abnormally low readings of ozone over the Antarctic as unreasonable.
- This (arguably) delayed the detection of the ozone hole by several years until British Antarctic Survey scientists discovered it based on analysis of their own observations (*Nature*, May 1985).
- The ozone hole was detected in satellite data only when the raw data was reprocessed. When the software was rerun without the pre-processing flags, the ozone hole was seen as far back as 1976.

Stylized Anecdote: The "Discovery" of the Ozone Hole

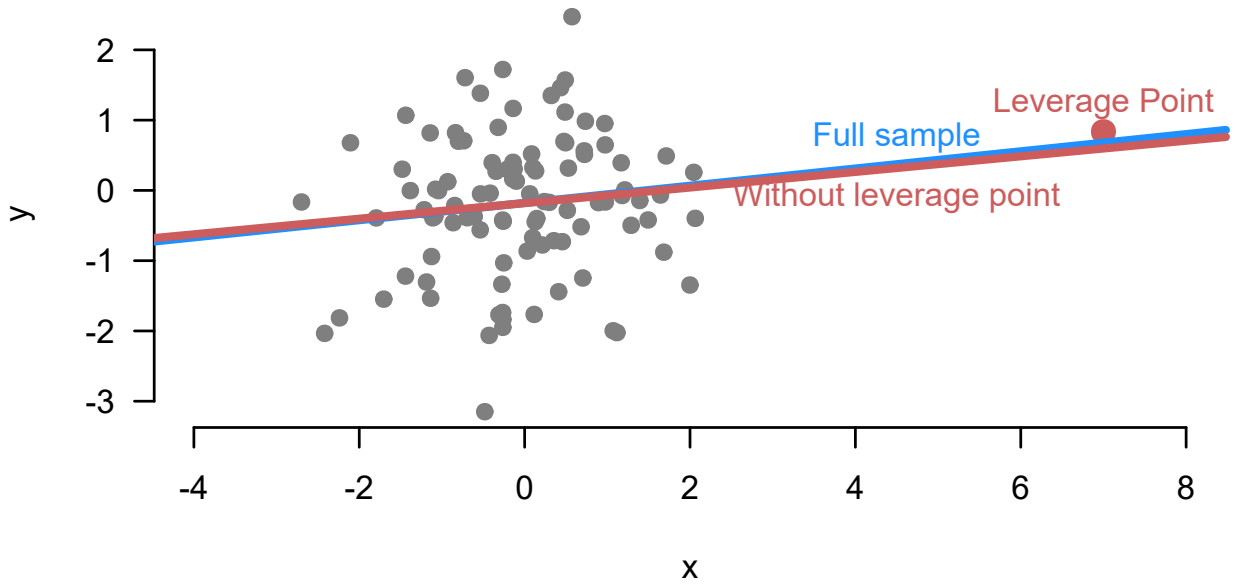
- In the late 70s, NASA used an automated data processing program on satellite measurements of atmospheric data to track changes in atmospheric variables such as ozone.
- This data "quality control" algorithm rejected abnormally low readings of ozone over the Antarctic as unreasonable.
- This (arguably) delayed the detection of the ozone hole by several years until British Antarctic Survey scientists discovered it based on analysis of their own observations (*Nature*, May 1985).
- The ozone hole was detected in satellite data only when the raw data was reprocessed. When the software was rerun without the pre-processing flags, the ozone hole was seen as far back as 1976.
- NB: This stylized story is only partially true. More info: `https://robjhyndman.com/hyndsight/ozone-hole-anomaly.html`

Leverage Point Definition



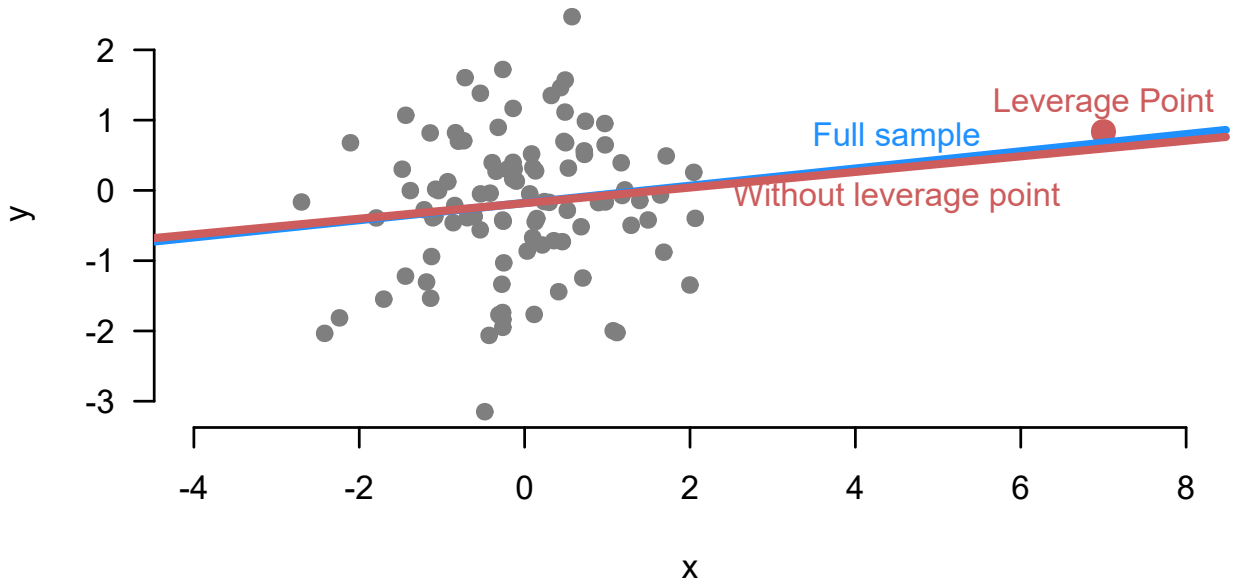
- Values that are extreme in the x direction

Leverage Point Definition



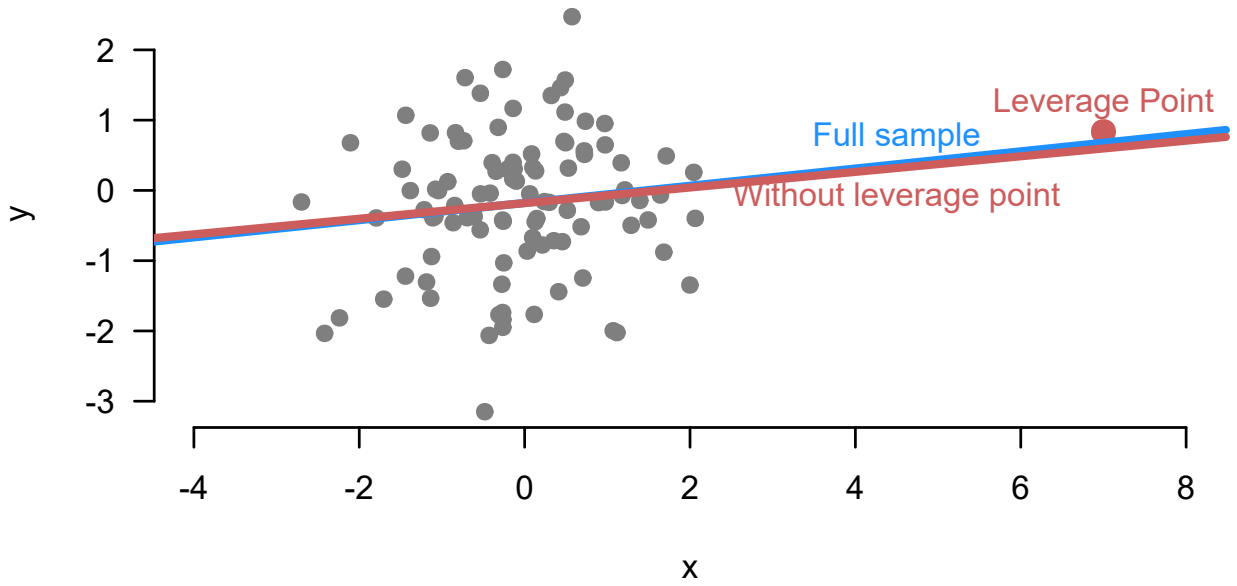
- Values that are extreme in the x direction
- That is, values far from the center of the covariate distribution

Leverage Point Definition



- Values that are extreme in the x direction
- That is, values far from the center of the covariate distribution
- Decrease SEs (more X variation)

Leverage Point Definition



- Values that are extreme in the x direction
- That is, values far from the center of the covariate distribution
- Decrease SEs (more X variation)
- No bias if typical in y dimension

Leverage Points: Hat values

To measure leverage in multivariate data we will go back to the hat matrix **H**:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$$

H is $n \times n$, symmetric, and idempotent. It generates fitted values as follows:

$$\hat{y}_i = \mathbf{h}_i'\mathbf{y} = \begin{bmatrix} h_{i,1} & h_{i,2} & \cdots & h_{i,n} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \sum_{j=1}^{n} h_{i,j} y_j$$

Leverage Points: Hat values

To measure leverage in multivariate data we will go back to the hat matrix $\mathbf{H}$:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$$

$\mathbf{H}$ is $n \times n$, symmetric, and idempotent. It generates fitted values as follows:

$$\hat{y}_i = \mathbf{h}_i'\mathbf{y} = \begin{bmatrix} h_{i,1} & h_{i,2} & \cdots & h_{i,n} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \sum_{j=1}^{n} h_{i,j} y_j$$

Therefore,

- h_{ij} dictates how important y_j is for the fitted value $\hat{y}_i$ (regardless of the actual value of y_j, since $\mathbf{H}$ depends only on $\mathbf{X}$)

Leverage Points: Hat values

To measure leverage in multivariate data we will go back to the hat matrix $\mathbf{H}$:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$$

$\mathbf{H}$ is $n \times n$, symmetric, and idempotent. It generates fitted values as follows:

$$\hat{y}_i = \mathbf{h}_i'\mathbf{y} = \begin{bmatrix} h_{i,1} & h_{i,2} & \cdots & h_{i,n} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \sum_{j=1}^{n} h_{i,j} y_j$$

Therefore,

- h_{ij} dictates how important y_j is for the fitted value $\hat{y}_i$ (regardless of the actual value of y_j, since $\mathbf{H}$ depends only on $\mathbf{X}$)
- The diagonal entries $h_{ii} = \sum_{j=1}^{n} h_{ij}^2$, so they summarize how important y_i is for all the fitted values. We call them the hat values or leverages and a single subscript notation is used: $h_i = h_{ii}$

Leverage Points: Hat values

To measure leverage in multivariate data we will go back to the hat matrix **H**:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$$

H is $n \times n$, symmetric, and idempotent. It generates fitted values as follows:

$$\hat{y}_i = \mathbf{h}_i'\mathbf{y} = \begin{bmatrix} h_{i,1} & h_{i,2} & \cdots & h_{i,n} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \sum_{j=1}^{n} h_{i,j} y_j$$

Therefore,

- h_{ij} dictates how important y_j is for the fitted value $\hat{y}_i$ (regardless of the actual value of y_j, since **H** depends only on **X**)
- The diagonal entries $h_{ii} = \sum_{j=1}^{n} h_{ij}^2$, so they summarize how important y_i is for all the fitted values. We call them the hat values or leverages and a single subscript notation is used: $h_i = h_{ii}$
- Intuitively, the hat values measure how far a unit's vector of characteristics $\mathbf{x}_i$ is from the vector of means of **X**

Leverage Points: Hat values

To measure leverage in multivariate data we will go back to the hat matrix **H**:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$$

H is $n \times n$, symmetric, and idempotent. It generates fitted values as follows:

$$\hat{y}_i = \mathbf{h}_i'\mathbf{y} = \begin{bmatrix} h_{i,1} & h_{i,2} & \cdots & h_{i,n} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \sum_{j=1}^{n} h_{i,j} y_j$$

Therefore,

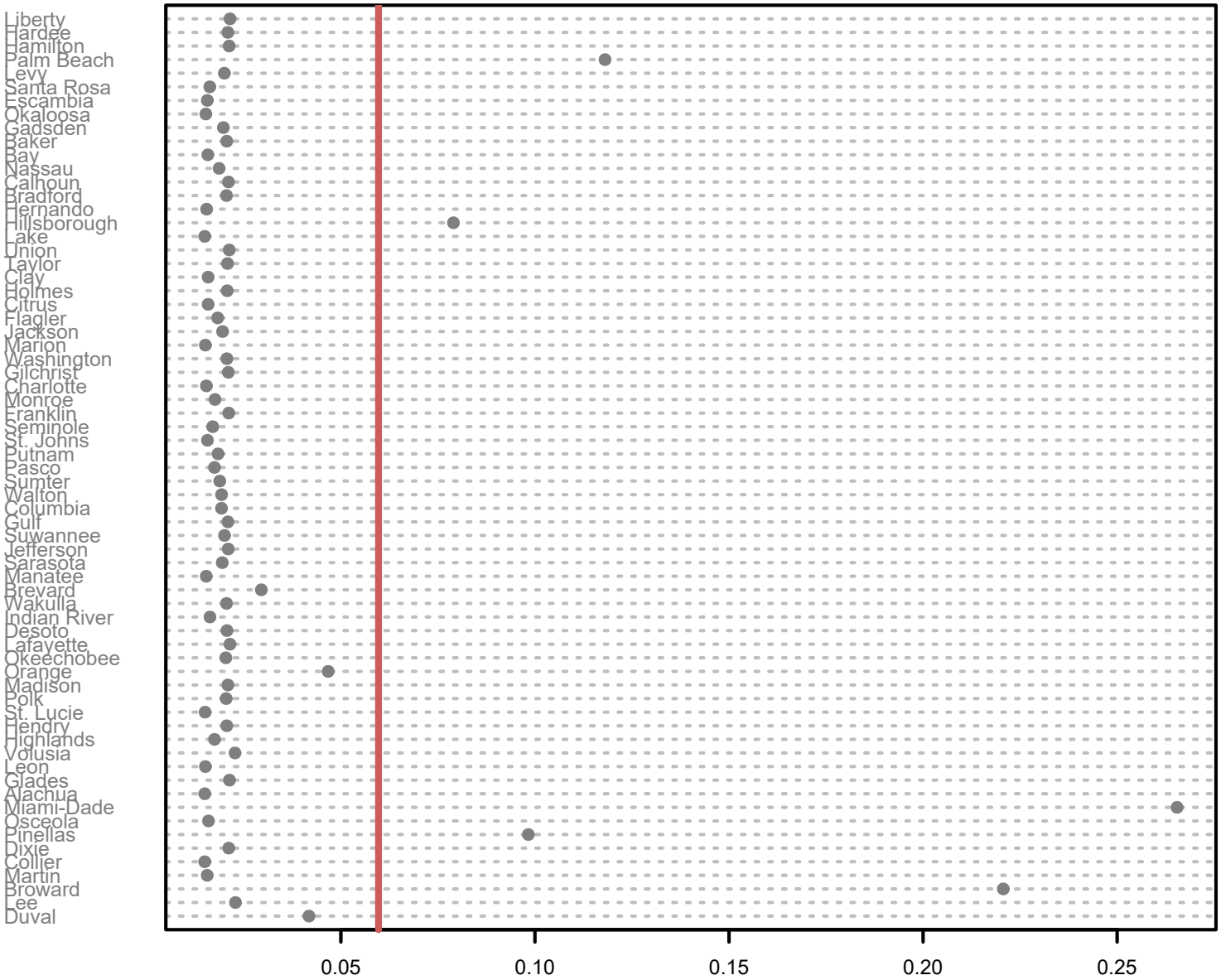
- h_{ij} dictates how important y_j is for the fitted value $\hat{y}_i$ (regardless of the actual value of y_j, since **H** depends only on **X**)
- The diagonal entries $h_{ii} = \sum_{j=1}^{n} h_{ij}^2$, so they summarize how important y_i is for all the fitted values. We call them the hat values or leverages and a single subscript notation is used: $h_i = h_{ii}$
- Intuitively, the hat values measure how far a unit's vector of characteristics $\mathbf{x}_i$ is from the vector of means of **X**
- Rule of thumb: examine hat values greater than $2(k+1)/n$

Appendix: Facts about Hat Values

- $\sum_{i=1}^{n} h_i = k + 1$
- $1/n \leq h_i \leq 1$ for all i
- $\text{Var}[\widehat{u}_i] = (1 - h_i)\sigma^2$
- With a simple linear regression, we have

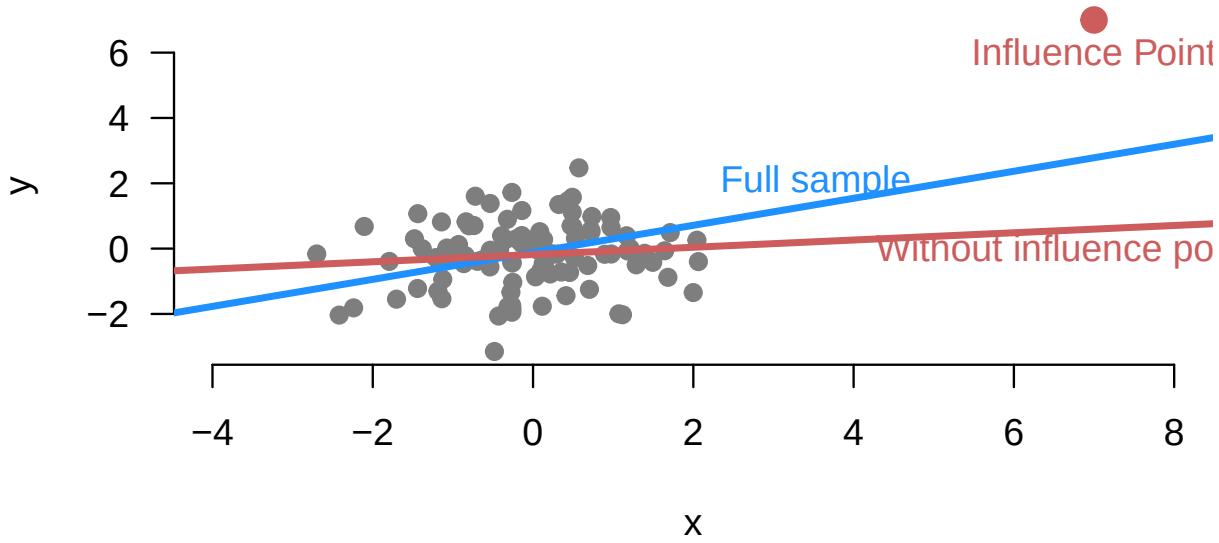
$$h_i = \frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum_{j=1}^{n}(X_j - \overline{X})^2}$$

Buchanan hats

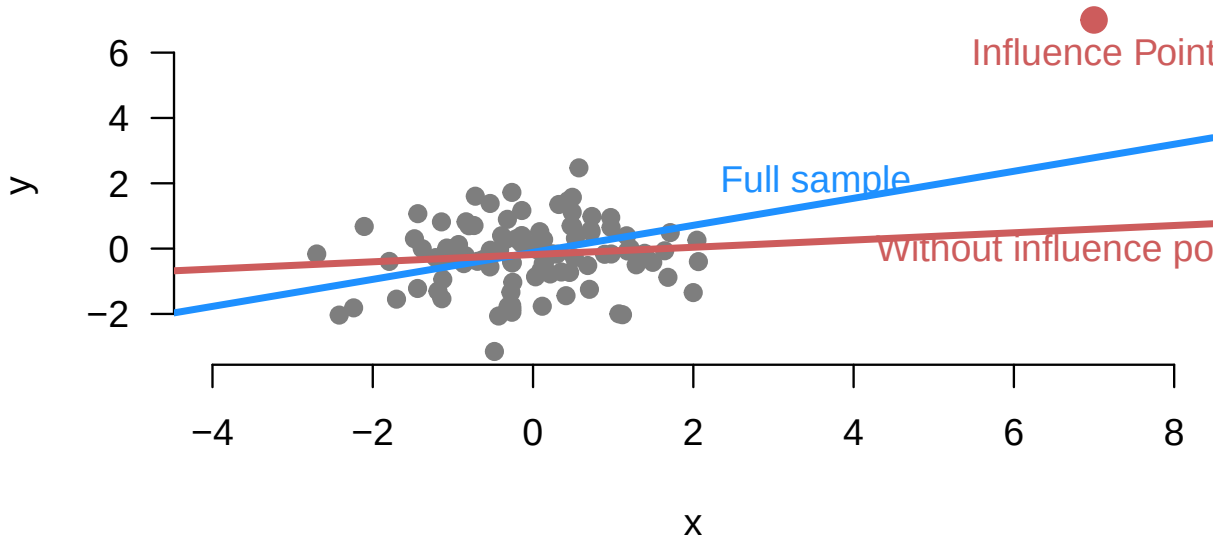


Influence points

Influence points

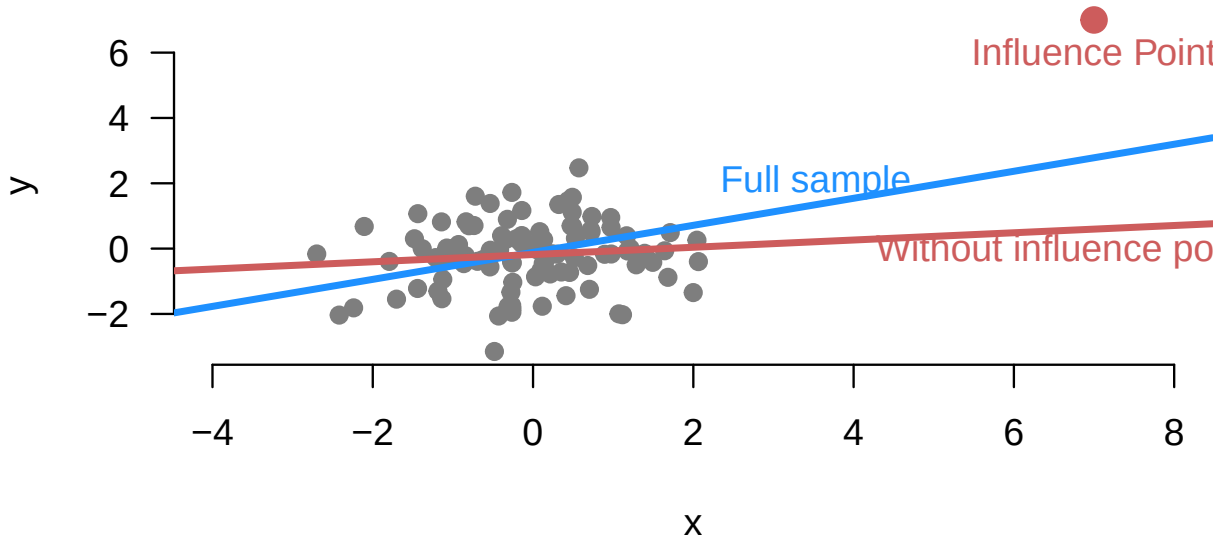


Influence points



- An influence point is one that is both an outlier (extreme in Y) and a leverage point (extreme in X).

Influence points



- An influence point is one that is both an outlier (extreme in Y) and a leverage point (extreme in X).
- Causes the regression line to move toward it.

Detecting Influence Points/Bad Leverage Points

- Influence Points:
  Influence on coefficients = Leverage $\times$ Outlyingness

Detecting Influence Points/Bad Leverage Points

- Influence Points:
  Influence on coefficients = Leverage $\times$ Outlyingness

- More formally: Measure the change that occurs in the slope estimates when an observation is removed from the data set. Let

$$D_{ij} = \hat{\beta}_j - \hat{\beta}_{j(-i)}, \quad i = 1, \ldots, n, \quad j = 0, \ldots, k$$

  where $\hat{\beta}_{j(-i)}$ is the estimate of the jth coefficient from the same regression once observation i has been removed from the data set.

Detecting Influence Points/Bad Leverage Points

- Influence Points:
  Influence on coefficients = Leverage $\times$ Outlyingness

- More formally: Measure the change that occurs in the slope estimates when an observation is removed from the data set. Let

$$D_{ij} = \hat{\beta}_j - \hat{\beta}_{j(-i)}, \quad i = 1, \ldots, n, \quad j = 0, \ldots, k$$

  where $\hat{\beta}_{j(-i)}$ is the estimate of the jth coefficient from the same regression once observation i has been removed from the data set.

- D_{ij} is called the DFbeta, which measures the influence of observation i on the estimated coefficient for the jth explanatory variable.

Standardized Influence

To make comparisons across coefficients, it is helpful to scale D_{ij} by the estimated standard error of the coefficients:

$$D_{ij}^{*} = \frac{\hat{\beta}_j - \hat{\beta}_{j(-i)}}{\hat{SE}_{-i}(\hat{\beta}_j)}$$

where D_{ij}^{*} is called DFbetaS.

Standardized Influence

To make comparisons across coefficients, it is helpful to scale D_{ij} by the estimated standard error of the coefficients:

$$D_{ij}^{*} = \frac{\hat{\beta}_j - \hat{\beta}_{j(-i)}}{\hat{SE}_{-i}(\hat{\beta}_j)}$$

where D_{ij}^{*} is called DFbetaS.

- $D_{ij}^{*} > 0$ implies that removing observation i decreases the estimate of $\beta_j \rightarrow$ obs i has a positive influence on β_j.

Standardized Influence

To make comparisons across coefficients, it is helpful to scale D_{ij} by the estimated standard error of the coefficients:

$$D_{ij}^{*} = \frac{\hat{\beta}_j - \hat{\beta}_{j(-i)}}{\hat{SE}_{-i}(\hat{\beta}_j)}$$

where D_{ij}^{*} is called DFbetaS.

- $D_{ij}^{*} > 0$ implies that removing observation i decreases the estimate of $\beta_j \rightarrow$ obs i has a positive influence on β_j.
- $D_{ij}^{*} < 0$ implies that removing observation i increases the estimate of $\beta_j \rightarrow$ obs i has a negative influence on β_j.

Standardized Influence

To make comparisons across coefficients, it is helpful to scale D_{ij} by the estimated standard error of the coefficients:

$$D_{ij}^{*} = \frac{\hat{\beta}_j - \hat{\beta}_{j(-i)}}{\hat{SE}_{-i}(\hat{\beta}_j)}$$

where D_{ij}^{*} is called DFbetaS.

- $D_{ij}^{*} > 0$ implies that removing observation i decreases the estimate of $\beta_j \rightarrow$ obs i has a positive influence on β_j.
- $D_{ij}^{*} < 0$ implies that removing observation i increases the estimate of $\beta_j \rightarrow$ obs i has a negative influence on β_j.
- Values of $|D_{ij}^{*}| > 2/\sqrt{n}$ are an indication of high influence.

Standardized Influence

To make comparisons across coefficients, it is helpful to scale D_{ij} by the estimated standard error of the coefficients:

$$D_{ij}^{*} = \frac{\hat{\beta}_j - \hat{\beta}_{j(-i)}}{\hat{SE}_{-i}(\hat{\beta}_j)}$$

where D_{ij}^{*} is called DFbetaS.

- $D_{ij}^{*} > 0$ implies that removing observation i decreases the estimate of $\beta_j \rightarrow$ obs i has a positive influence on β_j.
- $D_{ij}^{*} < 0$ implies that removing observation i increases the estimate of $\beta_j \rightarrow$ obs i has a negative influence on β_j.
- Values of $|D_{ij}^{*}| > 2/\sqrt{n}$ are an indication of high influence.
- In R: `dfbetas(model)`

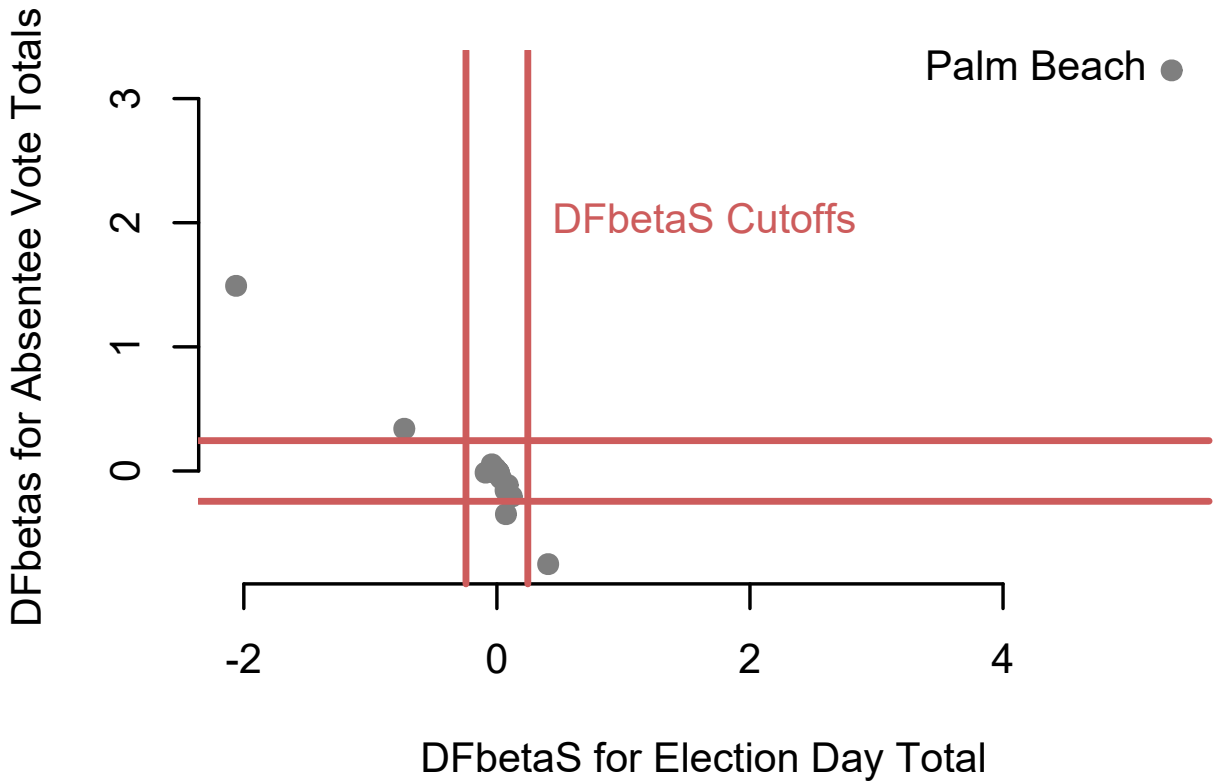
Buchanan influence

```
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.935e+01  5.520e+01  -0.532  0.59686
## edaytotal     1.100e-03  4.797e-04   2.293  0.02529 *
## absnbuchanan  6.895e+00  2.129e+00   3.238  0.00195 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 317.2 on 61 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.5361, Adjusted R-squared:  0.5209
## F-statistic: 35.24 on 2 and 61 DF,  p-value: 6.711e-11
```

Buchanan influence

```
##      (Intercept)     edaytotal  absnbuchanan
## 1  0.3454475146  0.4050504921 -0.7505222758
## 2 -0.0234266617 -0.0241000045 -0.0131672181
## 3  0.0650795039 -0.7319311820  0.3401669862
## 4 -0.0333980968  0.0133802934 -0.0087505576
## 5 -0.0397626659 -0.0073746223  0.0096551713
## 6 -0.0009277798  0.0001505476  0.0002210247
```

Buchanan influence



- Palm Beach county moves each of the coefficients by more than 3 standard errors!

Summarizing Influence across All Coefficients

- Leverage tells us how much one data point affects a single coefficient.

Summarizing Influence across All Coefficients

- Leverage tells us how much one data point affects a single coefficient.
- A number of summary measures exist for influence of data points across all coefficients, all involving both leverage and outlyingness.

Summarizing Influence across All Coefficients

- Leverage tells us how much one data point affects a single coefficient.
- A number of summary measures exist for influence of data points across all coefficients, all involving both leverage and outlyingness.
- A popular measure is Cook's distance:

$$D_i = \underbrace{\frac{\hat{u}_i'^2}{k+1}}_{\text{outlyingness}} \times \underbrace{\frac{h_i}{1-h_i}}_{\text{leverage}}$$

where $\hat{u}_i'$ is the standardized residual and h_i is the hat value.

Summarizing Influence across All Coefficients

- Leverage tells us how much one data point affects a single coefficient.
- A number of summary measures exist for influence of data points across all coefficients, all involving both leverage and outlyingness.
- A popular measure is Cook's distance:

$$D_i = \underbrace{\frac{\hat{u}_i'^2}{k+1}}_{\text{outlyingness}} \times \underbrace{\frac{h_i}{1-h_i}}_{\text{leverage}}$$

where $\hat{u}_i'$ is the standardized residual and h_i is the hat value.

  - It can be shown that D_i is a weighted sum of $k+1$ DFbetaS's for observation i

Summarizing Influence across All Coefficients

- Leverage tells us how much one data point affects a single coefficient.
- A number of summary measures exist for influence of data points across all coefficients, all involving both leverage and outlyingness.
- A popular measure is Cook's distance:

$$D_i = \underbrace{\frac{\hat{u}_i'^2}{k+1}}_{\text{outlyingness}} \times \underbrace{\frac{h_i}{1-h_i}}_{\text{leverage}}$$

where $\hat{u}_i'$ is the standardized residual and h_i is the hat value.

  - It can be shown that D_i is a weighted sum of $k+1$ DFbetaS's for observation i
  - In R, `cooks.distance(model)`

Summarizing Influence across All Coefficients

- Leverage tells us how much one data point affects a single coefficient.
- A number of summary measures exist for influence of data points across all coefficients, all involving both leverage and outlyingness.
- A popular measure is Cook's distance:

$$D_i = \underbrace{\frac{\hat{u}_i'^2}{k+1}}_{\text{outlyingness}} \times \underbrace{\frac{h_i}{1-h_i}}_{\text{leverage}}$$

where $\hat{u}_i'$ is the standardized residual and h_i is the hat value.

  - It can be shown that D_i is a weighted sum of $k+1$ DFbetaS's for observation i
  - In R, `cooks.distance(model)`
  - $D > 4/(n-k-1)$ is commonly considered large

Summarizing Influence across All Coefficients

- Leverage tells us how much one data point affects a single coefficient.
- A number of summary measures exist for influence of data points across all coefficients, all involving both leverage and outlyingness.
- A popular measure is Cook's distance:

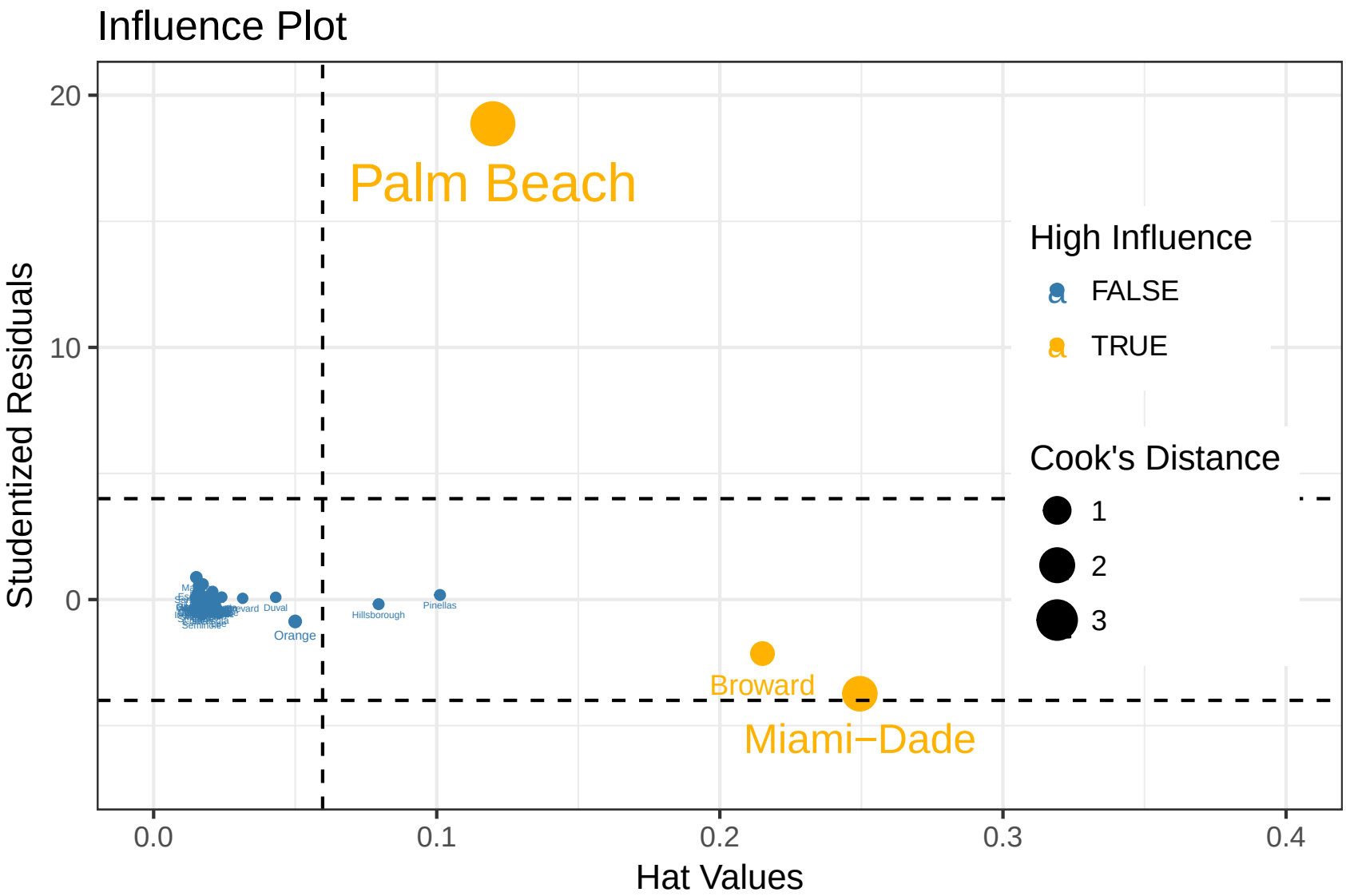
$$D_i = \underbrace{\frac{\hat{u}_i'^2}{k+1}}_{\text{outlyingness}} \times \underbrace{\frac{h_i}{1-h_i}}_{\text{leverage}}$$

where $\hat{u}_i'$ is the standardized residual and h_i is the hat value.

  - It can be shown that D_i is a weighted sum of $k+1$ DFbetaS's for observation i
  - In R, `cooks.distance(model)`
  - $D > 4/(n-k-1)$ is commonly considered large

- The influence plot: the studentized residuals plotted against the hat values, size of points proportional to Cook's distance.

Influence Plot Buchanan



Courtesy of Erin Hartman

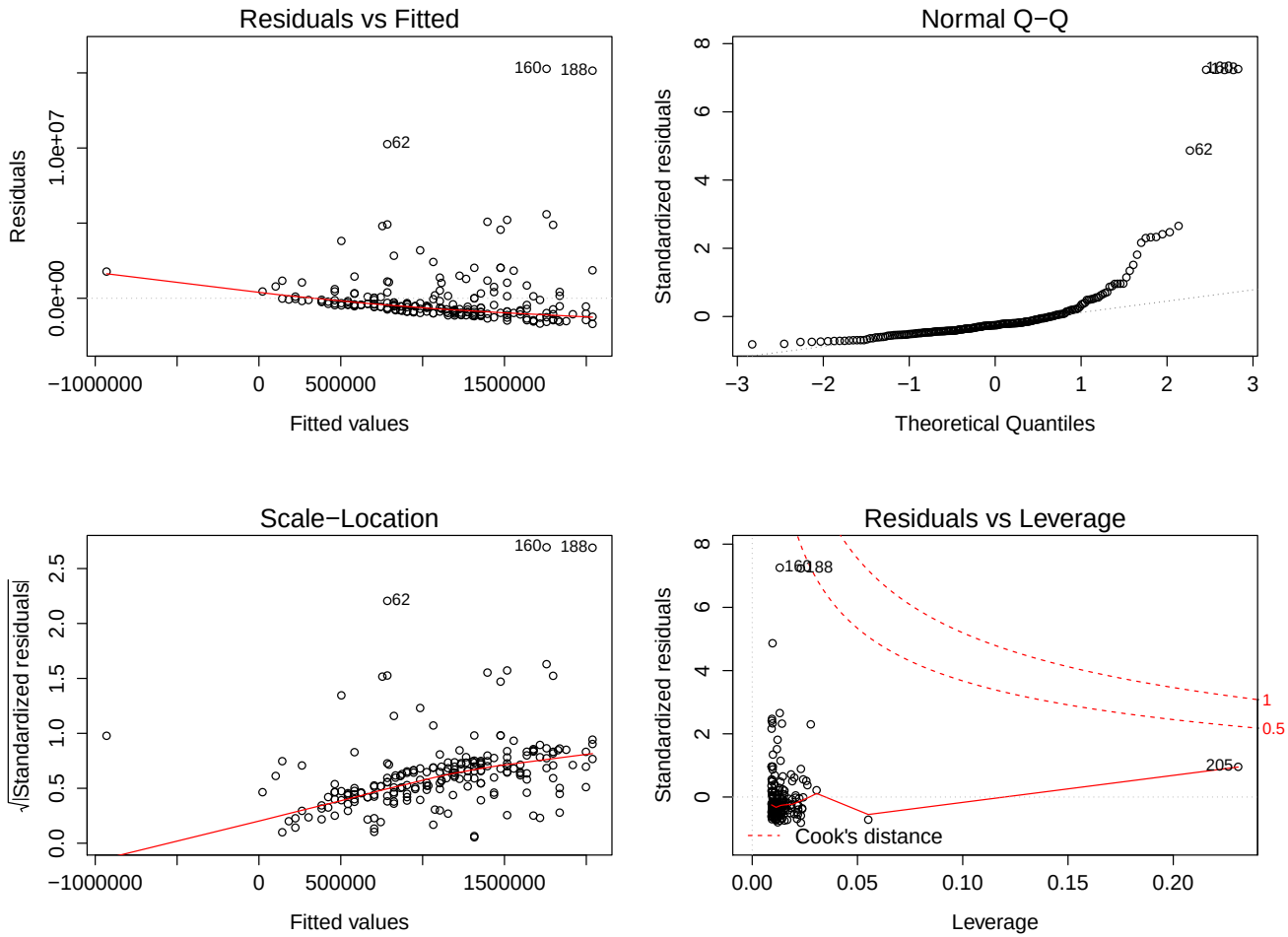
Code for Influence Plot

```
ggplot(fl_lm, aes(x = .hat, y = rstudent(fl_lm),
size = .cooksd,
col = .cooksd > 4/(nrow(fl_data) - 1 - 1),
label = fl_data$county)) +
geom_point() + geom_text(vjust = 2) +
xlab("Hat Values") + ylab("Studentized Residuals") +
geom_vline(xintercept = 2 * (fl_lm$rank - 1 + 1)/nrow(fl_data)
, linetype = 2) +
geom_hline(yintercept = c(-4, 4), linetype = 2) +
scale_color_manual("High Influence",
values = c("TRUE" = ucla_gold,
"FALSE" = ucla_blue)) +
scale_size("Cook's Distance") + theme_bw() +
theme(legend.position = c(0.9, 0.5)) + ylim(c(-7, 20)) +
xlim(c(0, 0.4)) + ggtitle("Influence Plot")
```

A Quick Function for Standard Diagnostic Plots

R Code

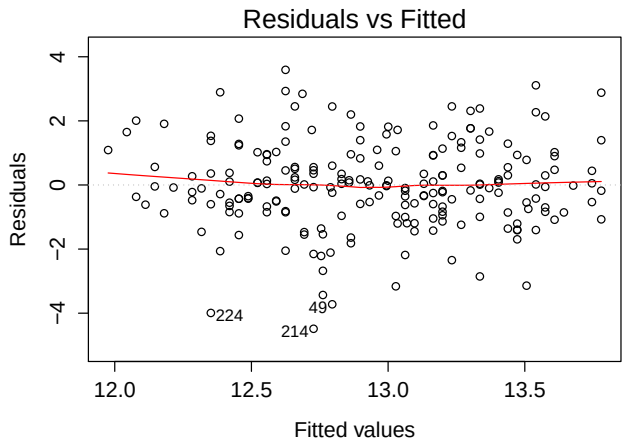
```
> par(mfrow=c(2,2))
> plot(mod1)
```

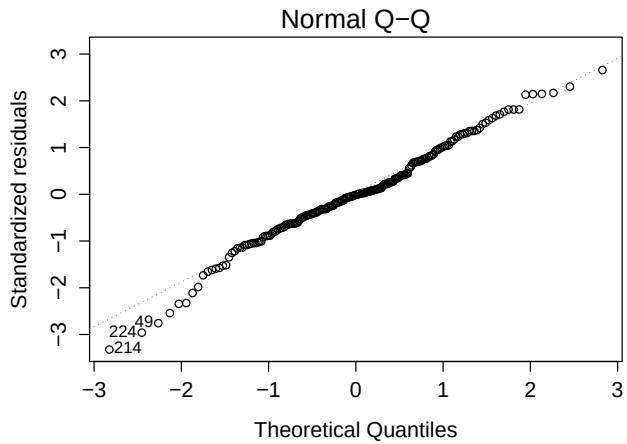


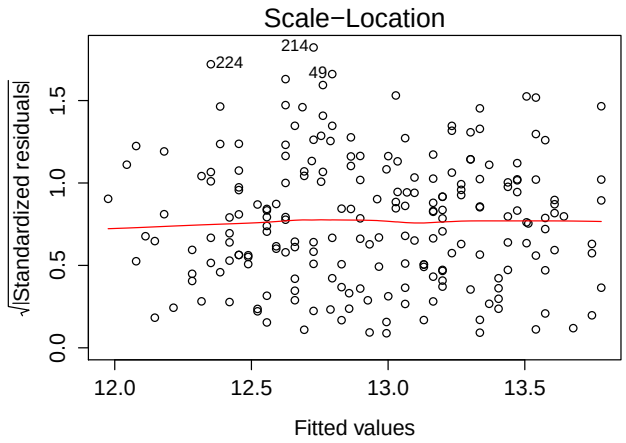
The Improved Model

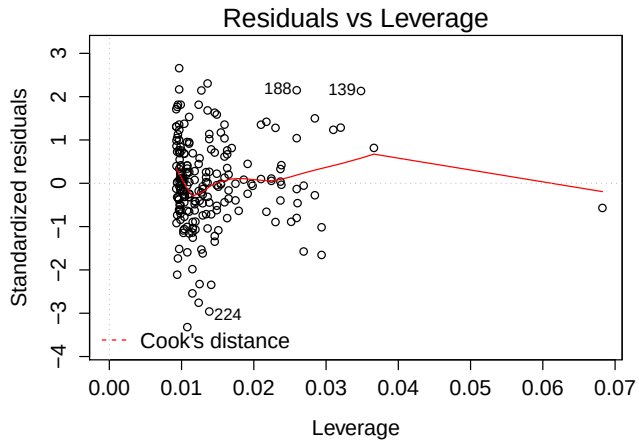
R Code

```
> par(mfrow=c(2,2))
> plot(mod2)
```









'Fun With Outliers'! (via FiveThirtyEight)

NOV. 9, 2018, AT 12:20 PM

Something Looks Weird In Broward County. Here's What We Know About A Possible Florida Recount.

By Nathaniel Rakich

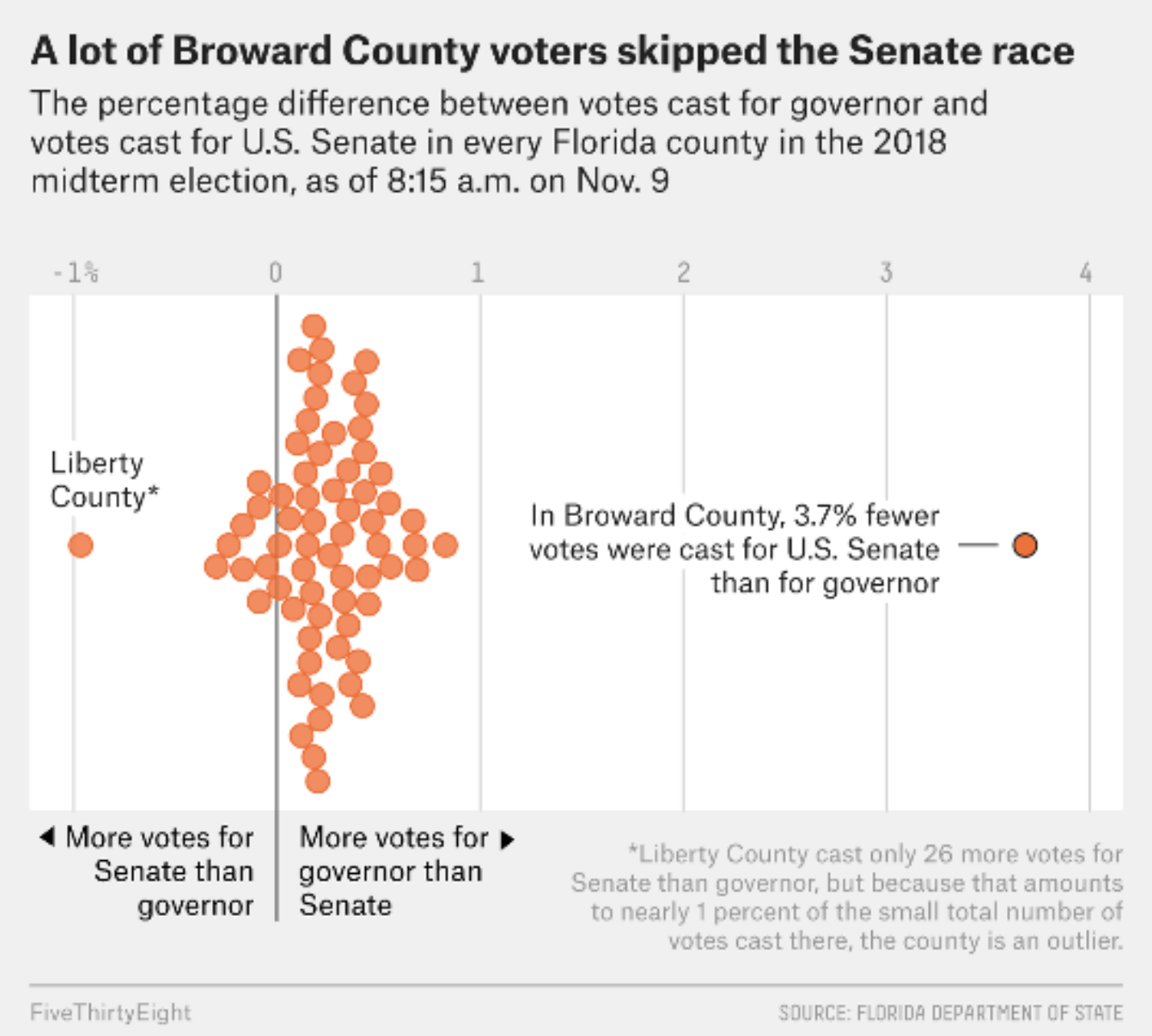
Filed under 2018 Election

ILLUSTRATION BY FIVETHIRTYEIGHT

'Fun With Outliers'! (via FiveThirtyEight)

The Florida U.S. Senate race is still too close to call. According to unofficial results on the Florida Department of State website at 11:45 a.m. Eastern on Friday, Nov. 9, Republican Gov. Rick Scott led Democratic Sen. Bill Nelson by 15,046 votes — or 0.18 percentage points. We're watching that margin closely because if it stays about that small, it will trigger a recount. It's already narrowed since election night, when Scott initially declared victory with a 56,000-vote lead.

‘Fun With Outliers’! (via FiveThirtyEight)

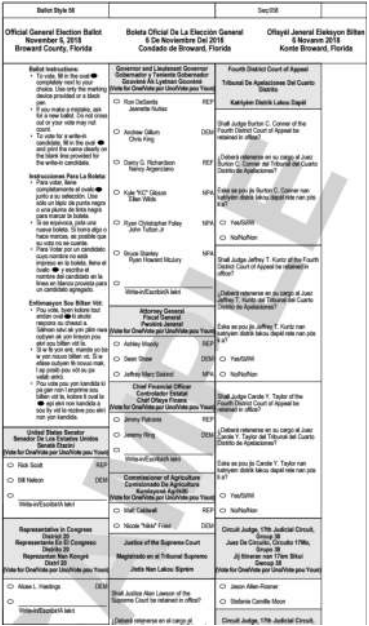


'Fun With Outliers'! (via FiveThirtyEight)

Broward County's undervote rate is *way* out of line with every other county in Florida, which exhibited, at most, a 0.8-percent difference. (There is one outlier — the sparsely populated Liberty County — where votes cast in the Senate race were 1 percent higher than in the governor race, but there we're talking about a difference of 26 votes, not more than 26,000, as is the case in Broward.)

To put in perspective what an eye-popping number of undervotes that is, more Broward County residents voted for the down-ballot constitutional offices of chief financial officer and state agriculture commissioner than U.S. Senate — an extremely high-profile election in which $181 million was spent. Generally, the higher the elected office, the less likely voters are to skip it on their ballots. Something sure does seem off in Broward County; we just don't know what yet.

'Fun With Outliers'! (via FiveThirtyEight)



Ballot Style 58

Sec 058

Official General Election Ballot
November 6, 2018
Broward County, Florida

Boleta Oficial De La Elección General
6 De Noviembre Del 2018
Condado de Broward, Florida

Ofisyèl Jeneral Eleksyon Bilten
6 Novanm 2018
Konte Broward, Florida

Ballot Instructions:

Instrucciones Para La Boleta:

Enfòmasyon Sou Bilten Vòt:

United States Senator
Senador De Los Estados Unidos
Senatè Etazini
Vote for One/Vote por Uno/Vote pou Youn

- Rick Scott REP
- Bill Nelson DEM
- Write-in/Escrito/A lekri

Representative in Congress
District 20
Representante En El Congreso
Distrito 20
Reprezantan Nan Kongrè
Distri 20
Vote for One/Vote por Uno/Vote pou Youn

- Alcee L. Hastings DEM
- Write-in/Escrito/A lekri

Governor and Lieutenant Governor
Gobernador y Teniente Gobernador
Gouvènè Ak Lyetnan Gouvènè
Vote for One/Vote por Uno/Vote pou Youn

- Ron DeSantis / Jeanette Nuñez REP
- Andrew Gillum / Chris King DEM
- Darcy G. Richardson / Nancy Argenziano REF
- Kyle "KC" Gibson / Ellen Wilds NPA
- Ryan Christopher Foley / John Tutton Jr NPA
- Bruce Stanley / Ryan Howard McJury NPA
- Write-in/Escrito/A lekri

Attorney General
Fiscal General
Pwokirè Jeneral
Vote for One/Vote por Uno/Vote pou Youn

- Ashley Moody REP
- Sean Shaw DEM
- Jeffrey Marc Siskind NPA

Chief Financial Officer
Contralor Estatal
Chèf Ofisye Finans
Vote for One/Vote por Uno/Vote pou Youn

- Jimmy Patronis REP
- Jeremy Ring DEM
- Write-in/Escrito/A lekri

Commissioner of Agriculture
Comisionado De Agricultura
Komisyonè Agrikilti
Vote for One/Vote por Uno/Vote pou Youn

- Matt Caldwell REP
- Nicole "Nikki" Fried DEM

Justice of the Supreme Court
Magistrado en el Tribunal Supremo
Jij Nan Lakou Siprèm

Shall Justice Alan Lawson of the Supreme Court be retained in office?

Fourth District Court of Appeal
Tribunal De Apelaciones Del Cuarto Distrito
Katriyèm Distrik Lakou Dapèl

Shall Judge Burton C. Conner of the Fourth District Court of Appeal be retained in office?

- Yes/Sí/Wi
- No/No/Non

Shall Judge Jeffrey T. Kuntz of the Fourth District Court of Appeal be retained in office?

- Yes/Sí/Wi
- No/No/Non

Shall Judge Carole Y. Taylor of the Fourth District Court of Appeal be retained in office?

- Yes/Sí/Wi
- No/No/Non

Circuit Judge, 17th Judicial Circuit, Group 38
Vote for One/Vote por Uno/Vote pou Youn

- Jason Allen Fosner
- Stefanie Camille Moon

Circuit Judge, 17th Judicial Circuit,

'Fun With Outliers'! (via FiveThirtyEight)

Sun Sentinel reporters talked with a ballot expert, who said that some voters may not have noticed the Senate race (perhaps thinking it was just part of the ballot instructions) and started filling out their ballot with the governor race instead. That theory is supported by a data consultant who's worked for several political campaigns in Florida, who found that the parts of Broward County that fall in the 24th Congressional District did see higher levels of undervoting than other parts of the county. That might be because the 24th District was uncontested, which according to Florida law means that the congressional race did not appear on the ballot at all. As you can see in the sample ballot above, the congressional race would also appear in the lower-left corner on many ballots, along with the Senate race. In districts where there was no congressional race on the ballot, however, that corner would have looked even emptier, perhaps making it easier for voters to inadvertently skip over the Senate race.

'Fun With Outliers'! (via FiveThirtyEight)

One possible reason for the discrepancy is poor ballot design. Broward County ballots listed the U.S. Senate race first, right after the ballot instructions. But that pushed the U.S. Senate race to the far bottom left of the ballot, where voters may have skimmed over it, while the governor's race appears at the top of the ballot's center column, immediately to the right of the instructions.

'Fun With Outliers'! (via FiveThirtyEight)

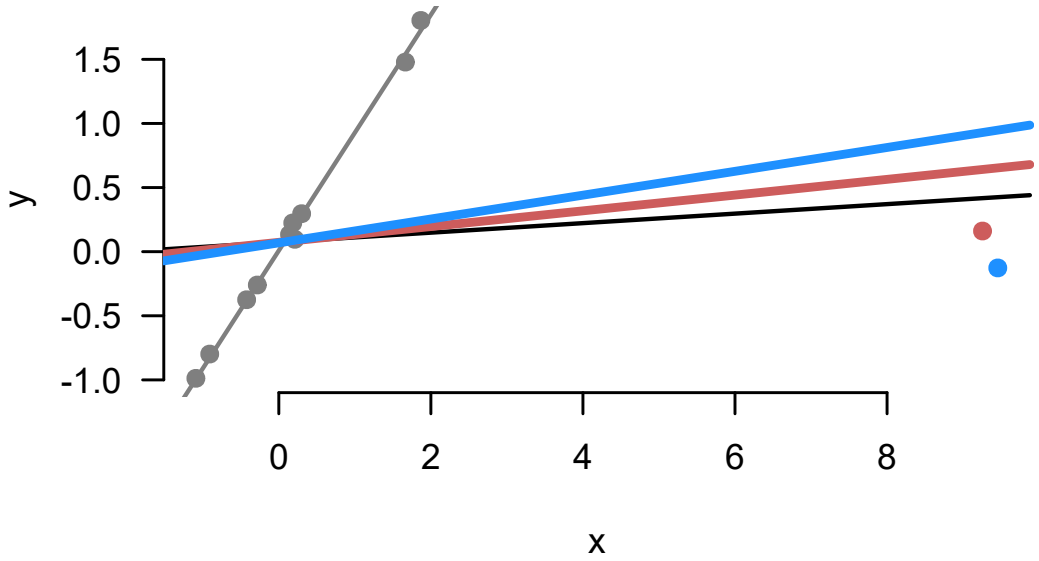
An alternative explanation is that an error with the vote-tabulating machines in Broward County caused them to sometimes not read people's votes for U.S. Senate. If that's true, we would probably only find out if there is a manual recount. According to Florida law, any election that's within half a percentage point (as this one currently is) triggers a machine recount; then, after the machine recount, if the race is within a *quarter* of a percentage point, it goes to a much more complex manual recount — a.k.a. each ballot is recounted by hand. As long as the machine recount doesn't change the Senate results too much (barring a surprise in the remaining ballots in Broward and Palm Beach), it looks like that's where we're headed. In addition, Republican former Rep. Ron DeSantis and Democratic Tallahassee Mayor Andrew Gillum are separated by just 0.44 points in the governor's race, so that could go to a machine recount, too.

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

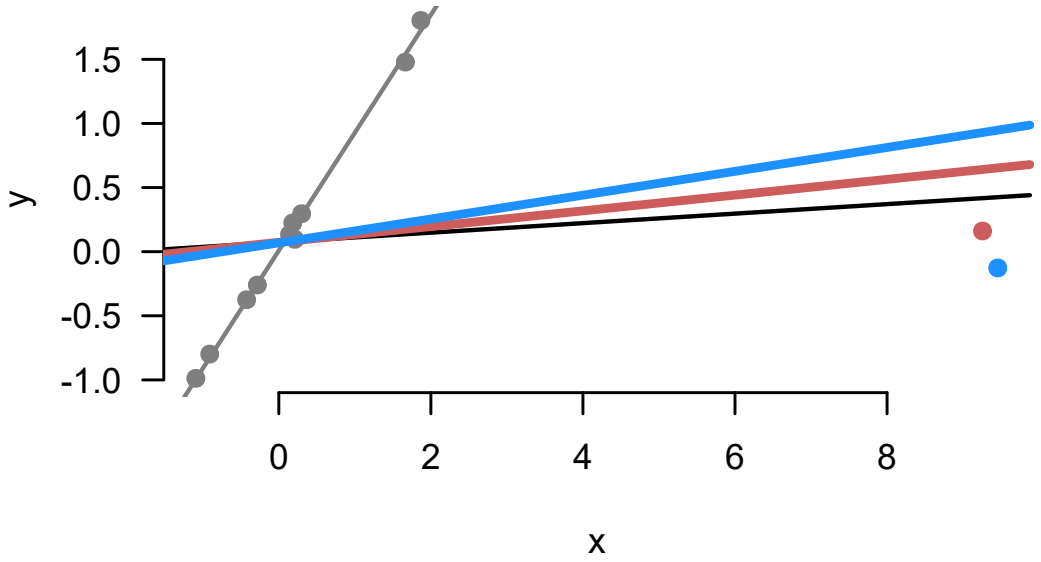
- Last Week
  - multiple regression
- This Week
  - diagnosing problems and troubleshooting the linear model
  - unusual and influential data $\rightarrow$ robust estimation
  - non-linearity $\rightarrow$ generalized additive models
  - unusual errors $\rightarrow$ sandwich SEs
- Next Week
  - frameworks for causal inference
- Long Run
  - probability $\rightarrow$ inference $\rightarrow$ regression $\rightarrow$ causal inference

1. Opening Themes
2. Preliminaries: The Hat Matrix
3. Fun With (Misleading) Visualization
4. Extreme Values
   - Outliers
   - Leverage Points
   - Influence Points
   - Fun with Outliers
5. Robust Regression Methods
   - Appendix: Robustness
6. Nonlinearity
   - Linear Basis Function Models
   - Generalized Additive Models

Limitations of the Standard Tools

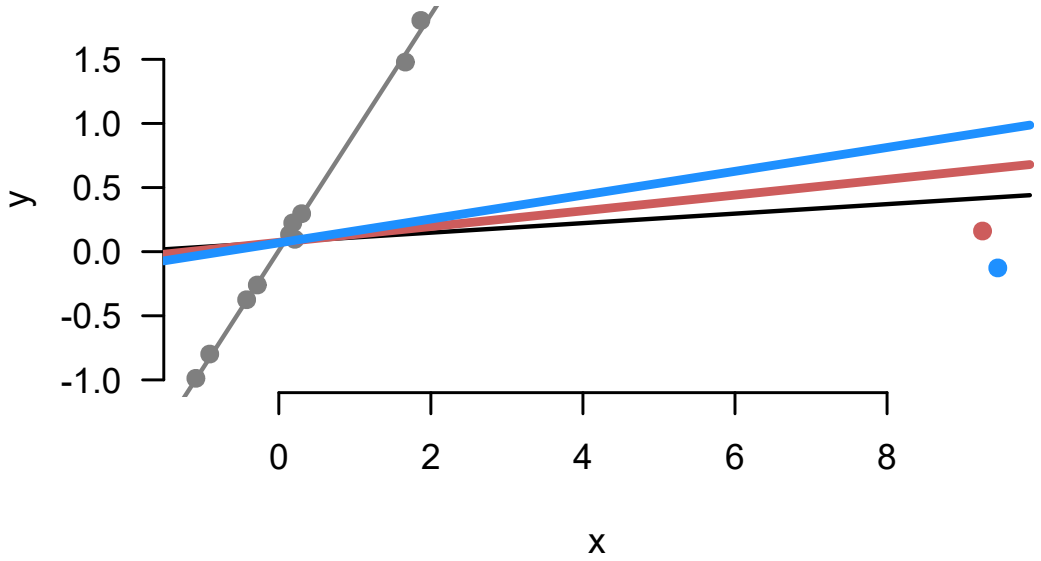


Limitations of the Standard Tools



- What happens when there are two influence points?
- Red line drops the red influence point
- Blue line drops the blue influence point

Limitations of the Standard Tools



- What happens when there are two influence points?
- Red line drops the red influence point
- Blue line drops the blue influence point
- Neither of the "leave-one-out" approaches helps recover the line

The Idea of Robustness

- We have and will cover a few ideas in robust statistics (much of which is due directly or indirectly to Peter Huber).

The Idea of Robustness

- We have and will cover a few ideas in robust statistics (much of which is due directly or indirectly to Peter Huber).
- Robust methods are procedures that are designed to continue to provide 'reasonable' answers in the presence of violation of some assumptions.

The Idea of Robustness

- We have and will cover a few ideas in robust statistics (much of which is due directly or indirectly to Peter Huber).
- Robust methods are procedures that are designed to continue to provide 'reasonable' answers in the presence of violation of some assumptions.
- A lot of social scientists use robust standard errors but far fewer use robust regression tools.

The Idea of Robustness

- We have and will cover a few ideas in robust statistics (much of which is due directly or indirectly to Peter Huber).
- Robust methods are procedures that are designed to continue to provide 'reasonable' answers in the presence of violation of some assumptions.
- A lot of social scientists use robust standard errors but far fewer use robust regression tools.
- These methods used to be computationally prohibitive but haven't been for the last 10-15 years

The Idea of Robustness

- We have and will cover a few ideas in robust statistics (much of which is due directly or indirectly to Peter Huber).
- Robust methods are procedures that are designed to continue to provide 'reasonable' answers in the presence of violation of some assumptions.
- A lot of social scientists use robust standard errors but far fewer use robust regression tools.
- These methods used to be computationally prohibitive but haven't been for the last 10-15 years

But What About Gauss-Markov and BLUE?

- One argument here is that even without normality, we know that Gauss-Markov is the Best Linear Unbiased Estimator (BLUE)

But What About Gauss-Markov and BLUE?

- One argument here is that even without normality, we know that Gauss-Markov is the Best Linear Unbiased Estimator (BLUE)
- How comforting should this be?

But What About Gauss-Markov and BLUE?

- One argument here is that even without normality, we know that Gauss-Markov is the Best Linear Unbiased Estimator (BLUE)
- How comforting should this be? Not very.

But What About Gauss-Markov and BLUE?

- One argument here is that even without normality, we know that Gauss-Markov is the Best Linear Unbiased Estimator (BLUE)
- How comforting should this be? Not very.
- The Linear point is an artificial restriction. It means the estimator has to be of the form $\hat{\beta} = \mathbf{W}y$ but why only use those?

But What About Gauss-Markov and BLUE?

- One argument here is that even without normality, we know that Gauss-Markov is the Best Linear Unbiased Estimator (BLUE)
- How comforting should this be? Not very.
- The Linear point is an artificial restriction. It means the estimator has to be of the form $\hat{\beta} = \mathbf{W}y$ but why only use those?
- With normality assumption we get Best Unbiased Estimator (BUE) which is quite comforting when $n \gg p$ (number of observations much larger than number of variables).

This Point is Not Obvious

This flies in the face of most conventional wisdom in textbooks.

This Point is Not Obvious

This flies in the face of most conventional wisdom in textbooks.

> *"[Even without normally distributed errors] OLS coefficient estimators remain unbiased and efficient."*
>
> *- Berry (1993)*

Quotes from Rainey and Baissa (2015) presentation

This Point is Not Obvious

This flies in the face of most conventional wisdom in textbooks.

> *"[The Gauss-Markov theorem] justifies the use of the OLS method rather than using a variety of competing estimators"*
>
> *- Wooldridge (2013)*

Quotes from Rainey and Baissa (2015) presentation

This Point is Not Obvious

This flies in the face of most conventional wisdom in textbooks.

> *"We need not look for another linear unbiased estimator, for we will not find such an estimator whose variance is smaller than the OLS estimator"*
>
> *- Gujarati (2004)*

Quotes from Rainey and Baissa (2015) presentation

This Point is Not Obvious

This flies in the face of most conventional wisdom in textbooks.

> *"The Gauss-Markov theorem allows us to have considerable confidence in the least squares estimators."*
>
> *- Berry and Feldman (1993)*

Quotes from Rainey and Baissa (2015) presentation

This Point is Not Obvious

This flies in the face of most conventional wisdom in textbooks.

> *The Gauss-Markov theorem has convinced researchers in political science that as long as . . . the Gauss-Markov assumptions are met, the distribution of the errors is unimportant. But the distribution of the errors is crucial to a linear regression analysis. Deviations from normality, especially large deviations commonly found in regression models in political science, can devastate the performance of least squares compared to alternative estimators*
>
> *- Baissa and Rainey (2018)*

Robustly Estimating a Location

- Let's simplify- what if we want to estimate the center of a symmetric distribution.

Robustly Estimating a Location

- Let's simplify- what if we want to estimate the center of a symmetric distribution.
- Two options (of many): mean and median

Robustly Estimating a Location

- Let's simplify- what if we want to estimate the center of a symmetric distribution.
- Two options (of many): mean and median
- Characteristics to consider: efficiency when assumptions hold, sensitivity to assumption violation.

Robustly Estimating a Location

- Let's simplify- what if we want to estimate the center of a symmetric distribution.
- Two options (of many): mean and median
- Characteristics to consider: efficiency when assumptions hold, sensitivity to assumption violation.
- For normal data $y_i \sim \mathcal{N}(\mu, \sigma^2)$, median is less efficient:
  - $V(\hat{\mu}_{\text{mean}}) = \frac{\sigma^2}{n}$
  - $V(\hat{\mu}_{\text{median}}) = \frac{\pi\sigma^2}{2n}$
  - Median is $\frac{\pi}{2}$ times larger (i.e. less efficient)

Robustly Estimating a Location

- Let's simplify- what if we want to estimate the center of a symmetric distribution.
- Two options (of many): mean and median
- Characteristics to consider: efficiency when assumptions hold, sensitivity to assumption violation.
- For normal data $y_i \sim \mathcal{N}(\mu, \sigma^2)$, median is less efficient:
  - $V(\hat{\mu}_{\text{mean}}) = \frac{\sigma^2}{n}$
  - $V(\hat{\mu}_{\text{median}}) = \frac{\pi\sigma^2}{2n}$
  - Median is $\frac{\pi}{2}$ times larger (i.e. less efficient)
- We can measure sensitivity with the influence function which measures change in estimator based on corruption in one datapoint.

Influence Function

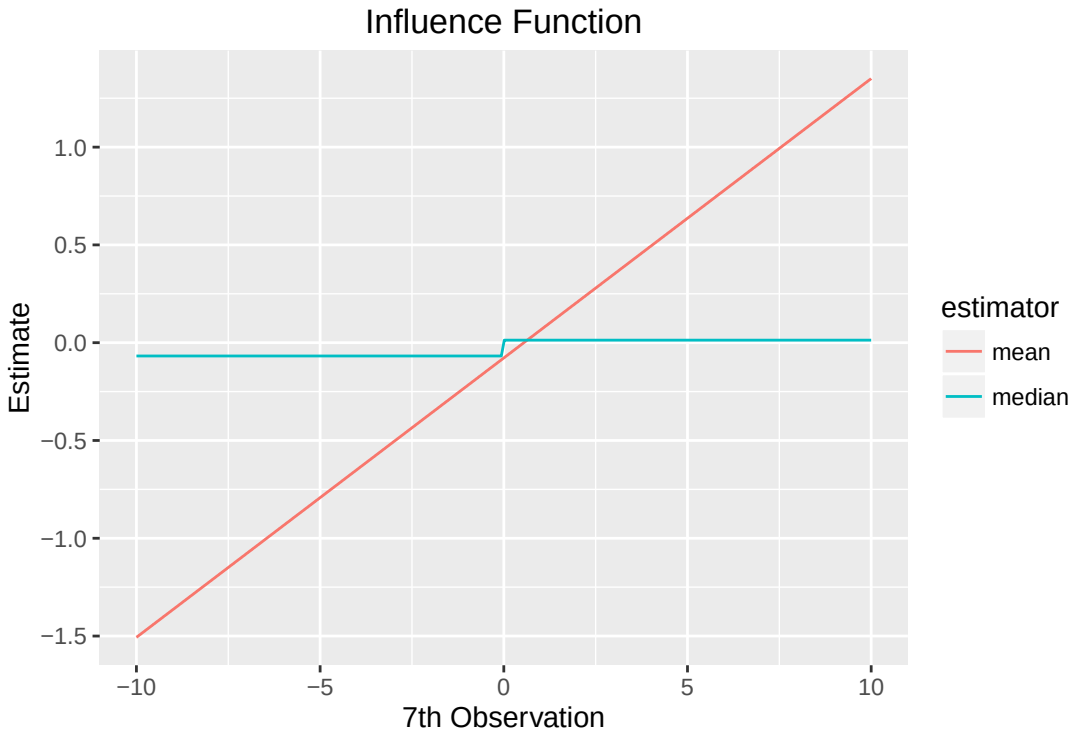
- Imagine that we had a sample Y from a standard normal: -0.068, -1.282, 0.013, 0.141, -0.980, 1.63. $\bar{Y} = -1.52$

Influence Function

- Imagine that we had a sample Y from a standard normal: -0.068, -1.282, 0.013, 0.141, -0.980, 1.63. $\bar{Y} = -1.52$
- Now imagine we add a contaminated 7th observation which could range from -10 to +10. How would the estimator change for the median and mean?

Influence Function

- Imagine that we had a sample Y from a standard normal: -0.068, -1.282, 0.013, 0.141, -0.980, 1.63. $\bar{Y} = -1.52$
- Now imagine we add a contaminated 7th observation which could range from -10 to +10. How would the estimator change for the median and mean?



Example from Fox

Breakdown Point

- The influence function showed us how one aberrant point can change the resulting estimate.

Breakdown Point

- The influence function showed us how one aberrant point can change the resulting estimate.
- We also want to characterize the breakdown point which is the fraction of arbitrarily bad data that the estimator can tolerate without being affected to an arbitrarily large extent

Breakdown Point

- The influence function showed us how one aberrant point can change the resulting estimate.
- We also want to characterize the breakdown point which is the fraction of arbitrarily bad data that the estimator can tolerate without being affected to an arbitrarily large extent
- The breakdown point of the mean is 0 because (as we have seen) a single bad data point can change things a lot.

Breakdown Point

- The influence function showed us how one aberrant point can change the resulting estimate.
- We also want to characterize the breakdown point which is the fraction of arbitrarily bad data that the estimator can tolerate without being affected to an arbitrarily large extent
- The breakdown point of the mean is 0 because (as we have seen) a single bad data point can change things a lot.
- The median has a breakdown point of 50% because half the data can be bad without causing the median to become completely unstuck.

M-estimators

- We can phrase this more generally than the mean or the median which will allow us to extend the ideas to regression via M-estimation

M-estimators

- We can phrase this more generally than the mean or the median which will allow us to extend the ideas to regression via M-estimation
- M-estimators minimize a sum over an objective function $\sum_i^n \rho(E)$ where E is $Y_i - \hat{\mu}$

M-estimators

- We can phrase this more generally than the mean or the median which will allow us to extend the ideas to regression via M-estimation
- M-estimators minimize a sum over an objective function $\sum_i^n \rho(E)$ where E is $Y_i - \hat{\mu}$
  - The mean has $\sum_i \rho(E) = \sum_i (Y_i - \hat{\mu})^2$
  - The median has $\sum_i \rho(E) = \sum_i |(Y_i - \hat{\mu})|$

M-estimators

- We can phrase this more generally than the mean or the median which will allow us to extend the ideas to regression via M-estimation
- M-estimators minimize a sum over an objective function $\sum_i^n \rho(E)$ where E is $Y_i - \hat{\mu}$
  - The mean has $\sum_i \rho(E) = \sum_i (Y_i - \hat{\mu})^2$
  - The median has $\sum_i \rho(E) = \sum_i |(Y_i - \hat{\mu})|$
- The shape of the influence function is determined by the derivative of the objective function with respect to E.

M-estimators

- We can phrase this more generally than the mean or the median which will allow us to extend the ideas to regression via M-estimation
- M-estimators minimize a sum over an objective function $\sum_i^n \rho(E)$ where E is $Y_i - \hat{\mu}$
  - The mean has $\sum_i \rho(E) = \sum_i (Y_i - \hat{\mu})^2$
  - The median has $\sum_i \rho(E) = \sum_i |(Y_i - \hat{\mu})|$
- The shape of the influence function is determined by the derivative of the objective function with respect to E.
- Other objectives include the Huber objective and Tukey's biweight objective which have different properties.

M-estimators

- We can phrase this more generally than the mean or the median which will allow us to extend the ideas to regression via M-estimation
- M-estimators minimize a sum over an objective function $\sum_i^n \rho(E)$ where E is $Y_i - \hat{\mu}$
  - The mean has $\sum_i \rho(E) = \sum_i (Y_i - \hat{\mu})^2$
  - The median has $\sum_i \rho(E) = \sum_i |(Y_i - \hat{\mu})|$
- The shape of the influence function is determined by the derivative of the objective function with respect to E.
- Other objectives include the Huber objective and Tukey's biweight objective which have different properties.
- Calculating robust M estimators often requires an iterative procedure and a careful initialization.

M-estimation for Regression

- We can apply this to regression fairly straightforwardly. In robust M-estimators we choose $\rho()$ so that observations with large residuals get less weight.

M-estimation for Regression

- We can apply this to regression fairly straightforwardly. In robust M-estimators we choose $\rho()$ so that observations with large residuals get less weight.
- Can be very robust to outliers in the Y space (less so in the X space usually)

M-estimation for Regression

- We can apply this to regression fairly straightforwardly. In robust M-estimators we choose $\rho()$ so that observations with large residuals get less weight.
- Can be very robust to outliers in the Y space (less so in the X space usually)
- Some options:
  - Least Median Squares: choose $\hat{\beta}$ to minimize $\text{median}\left\{(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{\text{LMS}})^2\right\}_{i=1}^{n}$. Very high breakdown point, but very inefficient.

M-estimation for Regression

- We can apply this to regression fairly straightforwardly. In robust M-estimators we choose $\rho()$ so that observations with large residuals get less weight.
- Can be very robust to outliers in the Y space (less so in the X space usually)
- Some options:
  - Least Median Squares: choose $\hat{\beta}$ to minimize $\text{median}\left\{(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{\text{LMS}})^2\right\}_{i=1}^{n}$. Very high breakdown point, but very inefficient.
  - Least Trimmed Squares: choose $\hat{\beta}$ to minimize the sum of the p smallest elements of $\left\{(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{\text{LTS}})^2\right\}_{i=1}^{n}$. High breakdown point and more efficient, still not as efficient as some.

M-estimation for Regression

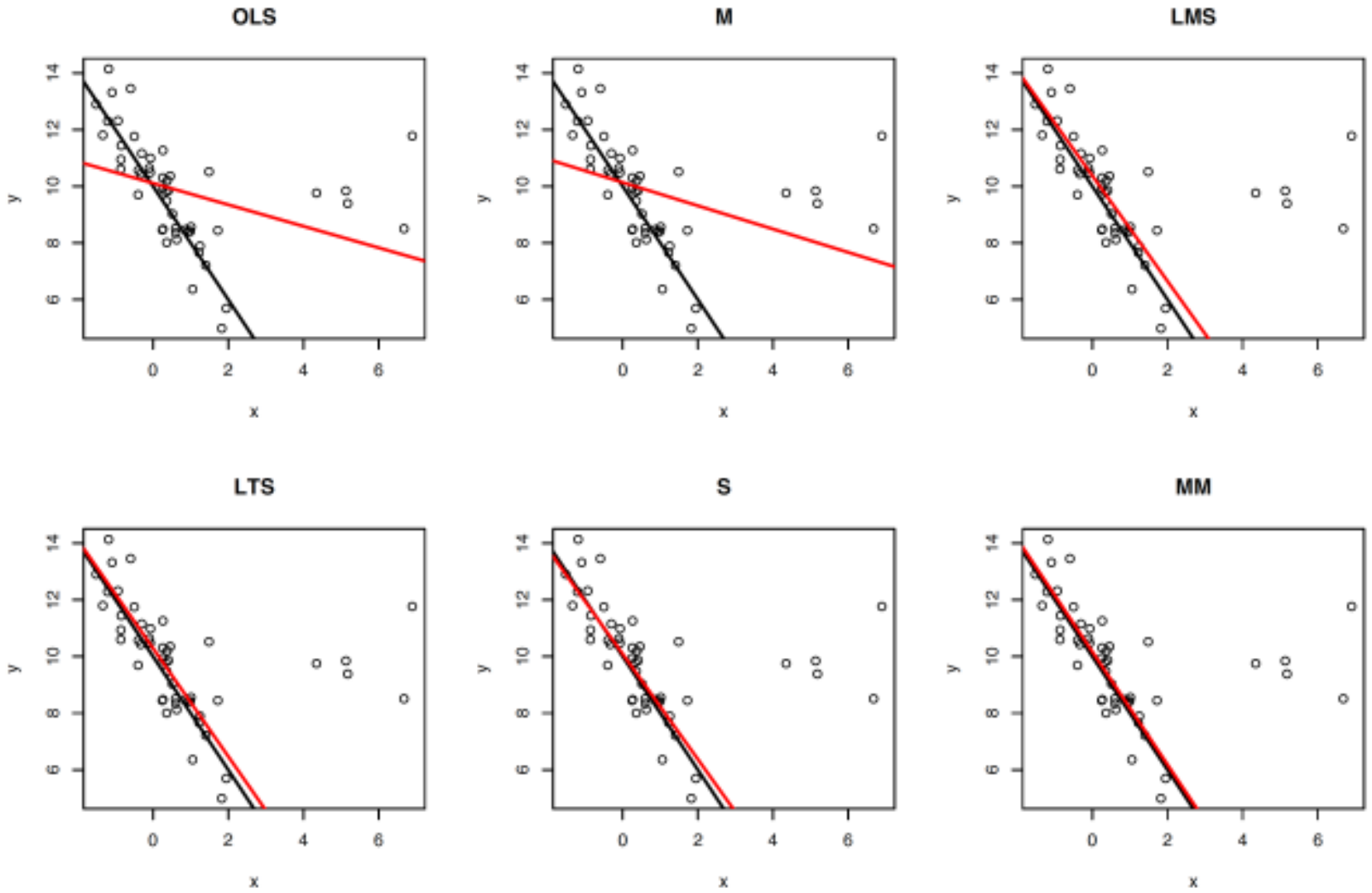
- We can apply this to regression fairly straightforwardly. In robust M-estimators we choose $\rho()$ so that observations with large residuals get less weight.
- Can be very robust to outliers in the Y space (less so in the X space usually)
- Some options:
  - Least Median Squares: choose $\hat{\beta}$ to minimize $\text{median}\left\{(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{\text{LMS}})^2\right\}_{i=1}^{n}$. Very high breakdown point, but very inefficient.
  - Least Trimmed Squares: choose $\hat{\beta}$ to minimize the sum of the p smallest elements of $\left\{(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{\text{LTS}})^2\right\}_{i=1}^{n}$. High breakdown point and more efficient, still not as efficient as some.
  - MM-estimator: what I recommend in practice (more in appendix)

M-estimation for Regression

- We can apply this to regression fairly straightforwardly. In robust M-estimators we choose $\rho()$ so that observations with large residuals get less weight.
- Can be very robust to outliers in the Y space (less so in the X space usually)
- Some options:
  - Least Median Squares: choose $\hat{\beta}$ to minimize $\text{median}\left\{(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{\text{LMS}})^2\right\}_{i=1}^{n}$. Very high breakdown point, but very inefficient.
  - Least Trimmed Squares: choose $\hat{\beta}$ to minimize the sum of the p smallest elements of $\left\{(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{\text{LTS}})^2\right\}_{i=1}^{n}$. High breakdown point and more efficient, still not as efficient as some.
  - MM-estimator: what I recommend in practice (more in appendix)
- You can find an asymptotic covariance matrix for M-estimators but I would bootstrap it if possible as the asymptotics kick in slowly.

```
library(MASS)
set.seed(588)
n <- 50
x <- rnorm(n)
y <- 10 - 2*x + rnorm(n)
x[1:5] <- rnorm(5, mean=5)
y[1:5] <- 10 +  rnorm(5)
ols.out <- lm(y~x)
m.out <- rlm(y~x, method="M")
lms.out <- lqs(y~x, method="lms")
lts.out <- lqs(y~x, method="lts")
s.out <- lqs(y~x, method="S")
mm.out <- rlm(y~x, method="MM")
```

Simulation Results



Thoughts on Robust Estimators

- Robust estimators aren't commonly seen in applied social science work but perhaps they should be.

Thoughts on Robust Estimators

- Robust estimators aren't commonly seen in applied social science work but perhaps they should be.
- Even though Gauss-Markov does not require normality, the L in BLUE is a fairly restrictive condition.

Thoughts on Robust Estimators

- Robust estimators aren't commonly seen in applied social science work but perhaps they should be.
- Even though Gauss-Markov does not require normality, the L in BLUE is a fairly restrictive condition.
- In most cases I personally would start with OLS, do diagnostics and then consider a robust alternative. If I don't have time for diagnostics, maybe robust is better from the outset.

Thoughts on Robust Estimators

- Robust estimators aren't commonly seen in applied social science work but perhaps they should be.
- Even though Gauss-Markov does not require normality, the L in BLUE is a fairly restrictive condition.
- In most cases I personally would start with OLS, do diagnostics and then consider a robust alternative. If I don't have time for diagnostics, maybe robust is better from the outset.
- See Baissa and Rainey (2018) "When BLUE is Not Best: Non-Normal Errors and the Linear Model" in *Political Science Research & Methods* for more on this topic.
- The Fox textbook Chapter 19 is also quite good on this and points out to the key references

Appendix: Characterizing Estimator Robustness (formally)

Definition (Breakdown Point)

The breakdown point of an estimator is the smallest fraction of the data that can be changed an arbitrary amount to produce an arbitrarily large change in the estimate (Seber and Lee 2003, pg 82)

Definition (Influence Function)

Let $F_p = (1-p)F + p\delta_{\mathbf{z}_0}$ where F is a probability measure, $\delta_{\mathbf{z}_0}$ is the point mass at $\mathbf{z}_0 \in \mathbb{R}^k$, and $p \in (0,1)$.
Let $T(\cdot)$ be a statistical functional. The influence function of T is

$$IF(\mathbf{z}_0; T, F) = \lim_{p \downarrow 0} \frac{T(F_p) - T(F)}{p}$$

The influence function is a function of $\mathbf{z}_0$ given T and F. It describes how T changes with small amounts of contamination at $\mathbf{z}_0$ (Hampel, Rousseeuw, Ronchetti, and Stahel, (1986), p. 84).

Appendix: S Estimators

To talk about $MM-$estimators we need a type of estimator called an S-estimator.

Appendix: S Estimators

To talk about $MM-$estimators we need a type of estimator called an S-estimator.

S-estimators work somewhat differently in that the goal is to minimize the scale estimate subject to a constraint.

Appendix: S Estimators

To talk about $MM-$estimators we need a type of estimator called an S-estimator.

S-estimators work somewhat differently in that the goal is to minimize the scale estimate subject to a constraint.

An S-estimator for the regression model is defined as the values of $\hat{\boldsymbol{\beta}}_S$ and s that minimize s subject to the constraint:

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_S}{s}\right) \geq K$$

where K is user-defined constant (typically set to 0.5) and $\rho : \mathbb{R} \to [0, 1]$ is a function with the following properties (Davies, 1990, p. 1653):

1. $\rho(0) = 1$
2. $\rho(u) = \rho(-u), u \in \mathbb{R}$
3. $\rho : \mathbb{R}_+ \to [0, 1]$ is nonincreasing, continuous at 0, and continuous on the left
4. for some $c > 0$, $\rho(u) > 0$ if $|u| < c$ and $\rho(u) = 0$ if $|u| > c$

Appendix: S Estimators

To talk about $MM-$estimators we need a type of estimator called an S-estimator.

S-estimators work somewhat differently in that the goal is to minimize the scale estimate subject to a constraint.

An S-estimator for the regression model is defined as the values of $\hat{\boldsymbol{\beta}}_S$ and s that minimize s subject to the constraint:

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_S}{s}\right) \geq K$$

where K is user-defined constant (typically set to 0.5) and $\rho : \mathbb{R} \to [0, 1]$ is a function with the following properties (Davies, 1990, p. 1653):

1. $\rho(0) = 1$
2. $\rho(u) = \rho(-u), u \in \mathbb{R}$
3. $\rho : \mathbb{R}_+ \to [0, 1]$ is nonincreasing, continuous at 0, and continuous on the left
4. for some $c > 0$, $\rho(u) > 0$ if $|u| < c$ and $\rho(u) = 0$ if $|u| > c$

Appendix: *MM*-Estimators

MM-estimators are, in some sense, the best of both worlds– very high breakdown point and good efficiency.

Appendix: *MM*-Estimators

MM-estimators are, in some sense, the best of both worlds– very high breakdown point and good efficiency.

The work by first calculating S-estimates of the scale and coefficients and then using these as starting values for a particular M-estimator.

Appendix: *MM*-Estimators

MM-estimators are, in some sense, the best of both worlds– very high breakdown point and good efficiency.

The work by first calculating S-estimates of the scale and coefficients and then using these as starting values for a particular M-estimator.

Good properties, but costly to compute (usually impossible to compute exactly).

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

- Last Week
  - multiple regression
- This Week
  - diagnosing problems and troubleshooting the linear model
  - unusual and influential data $\rightarrow$ robust estimation
  - non-linearity $\rightarrow$ generalized additive models
  - unusual errors $\rightarrow$ sandwich SEs
- Next Week
  - frameworks for causal inference
- Long Run
  - probability $\rightarrow$ inference $\rightarrow$ regression $\rightarrow$ causal inference

1. Opening Themes
2. Preliminaries: The Hat Matrix
3. Fun With (Misleading) Visualization
4. Extreme Values
   - Outliers
   - Leverage Points
   - Influence Points
   - Fun with Outliers
5. Robust Regression Methods
   - Appendix: Robustness
6. Nonlinearity
   - Linear Basis Function Models
   - Generalized Additive Models

Nonlinearity

Nonlinearity

- We know that linear regression asymptotically gets us the best linear approximation to the conditional expectation function.

Nonlinearity

- We know that linear regression asymptotically gets us the best linear approximation to the conditional expectation function.
- We can always add transformations of variables (like polynomials) to expand $\mathbf{X}$ in a way that makes non-linear shapes in the original variable.

Nonlinearity

- We know that linear regression asymptotically gets us the best linear approximation to the conditional expectation function.
- We can always add transformations of variables (like polynomials) to expand $\mathbf{X}$ in a way that makes non-linear shapes in the original variable.
- What happens when we don't know the shape of the non-linearity?

Nonlinearity

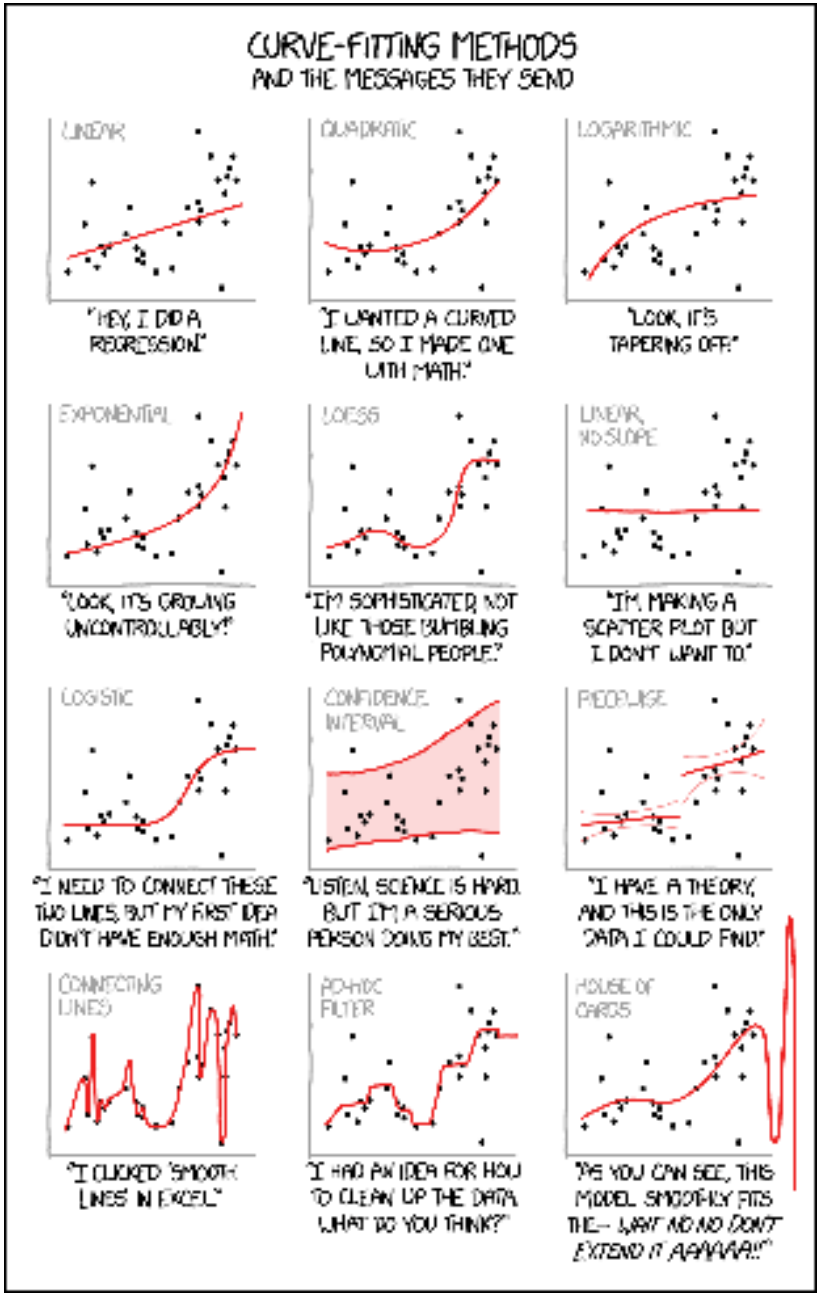
- We know that linear regression asymptotically gets us the best linear approximation to the conditional expectation function.
- We can always add transformations of variables (like polynomials) to expand $\mathbf{X}$ in a way that makes non-linear shapes in the original variable.
- What happens when we don't know the shape of the non-linearity?
- In Week 5 we talked about nonparametric regressions for settings with one independent variable.

Nonlinearity

- We know that linear regression asymptotically gets us the best linear approximation to the conditional expectation function.
- We can always add transformations of variables (like polynomials) to expand $\mathbf{X}$ in a way that makes non-linear shapes in the original variable.
- What happens when we don't know the shape of the non-linearity?
- In Week 5 we talked about nonparametric regressions for settings with one independent variable.
- Many forms of machine learning are best thought of as nonparametric regressions in higher dimensions.

Nonlinearity

- We know that linear regression asymptotically gets us the best linear approximation to the conditional expectation function.
- We can always add transformations of variables (like polynomials) to expand $\mathbf{X}$ in a way that makes non-linear shapes in the original variable.
- What happens when we don't know the shape of the non-linearity?
- In Week 5 we talked about nonparametric regressions for settings with one independent variable.
- Many forms of machine learning are best thought of as nonparametric regressions in higher dimensions.
- We can often see poor fits of the conditional expectation function in the residuals, but let's instead just do diagnosis by treatment and look at some of these other approaches to modeling.

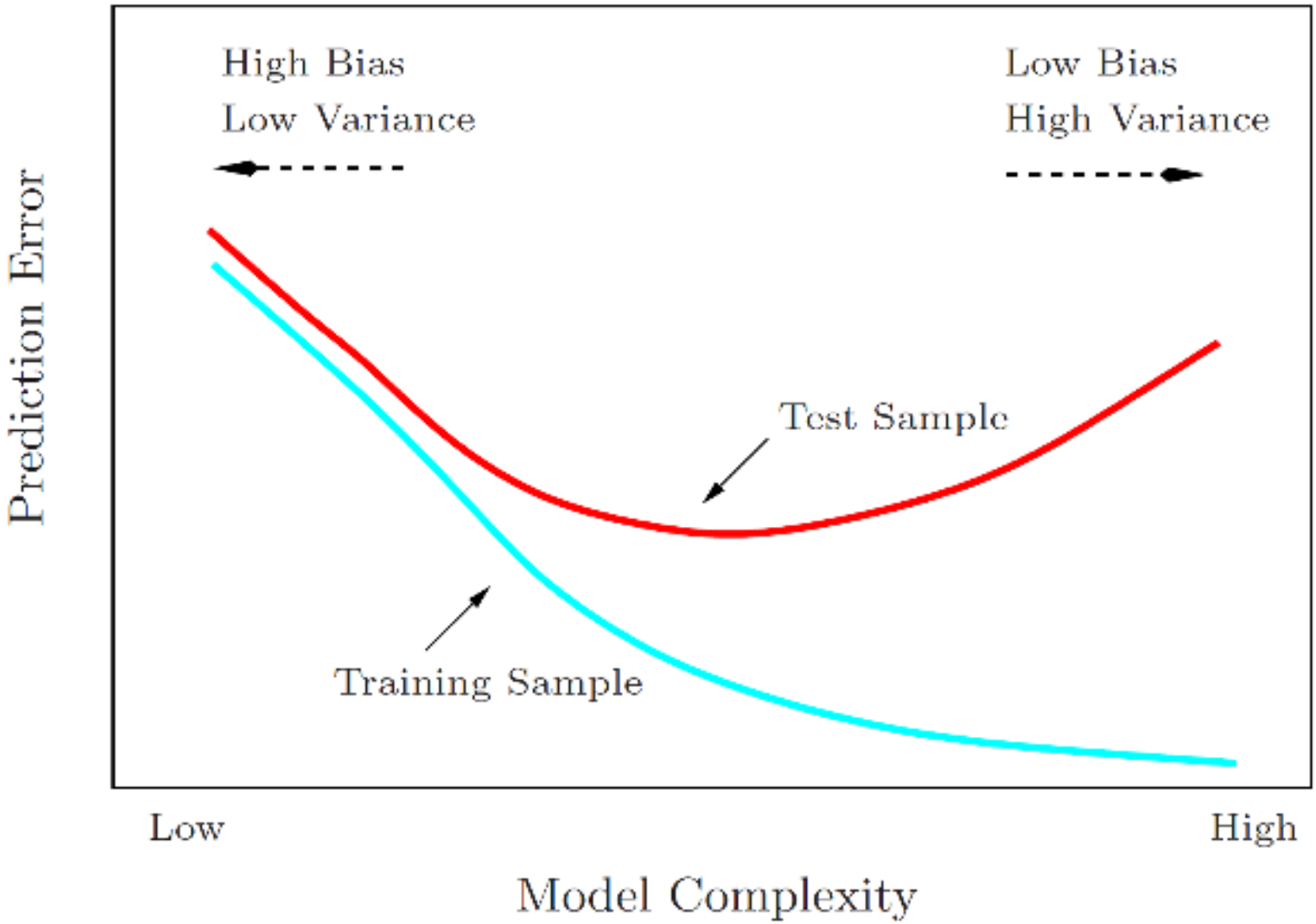


Thanks XKCD for having a comic for everything!

1. Opening Themes
2. Preliminaries: The Hat Matrix
3. Fun With (Misleading) Visualization
4. Extreme Values
   - Outliers
   - Leverage Points
   - Influence Points
   - Fun with Outliers
5. Robust Regression Methods
   - Appendix: Robustness
6. Nonlinearity
   - Linear Basis Function Models
   - Generalized Additive Models

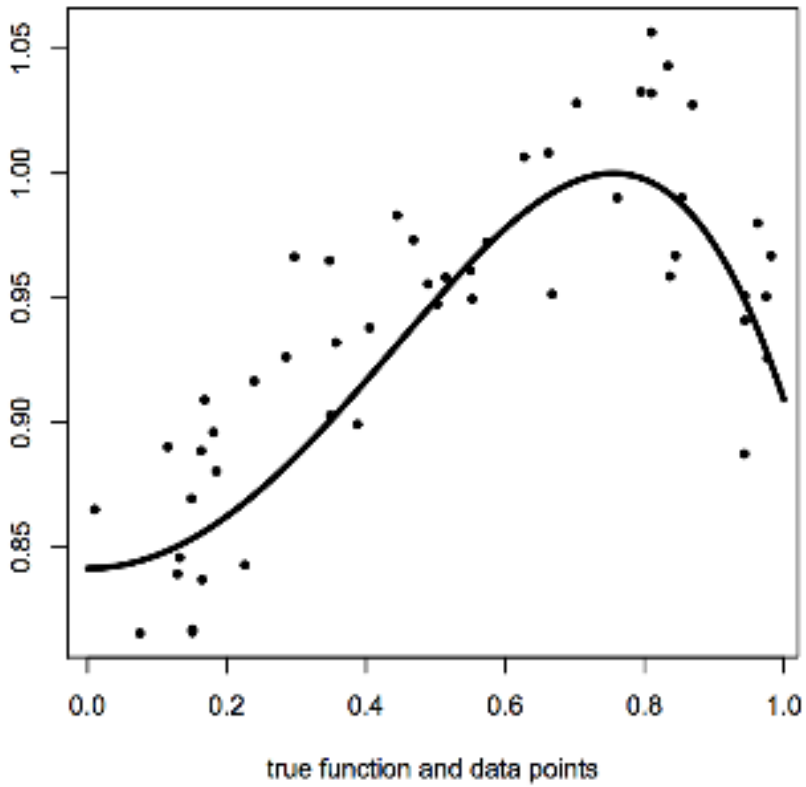
Bias-Variance Tradeoff

Bias-Variance Tradeoff



Example Synthetic Problem

$$y = \sin(1 + x^2) + \epsilon$$



This section adapted from slides by Radford Neal.

Linear Basis Function Models

Linear Basis Function Models

- We talked before about polynomials x^2, x^3, x^4 for modeling non-linearities, this is a linear basis function model.

Linear Basis Function Models

- We talked before about polynomials x^2, x^3, x^4 for modeling non-linearities, this is a linear basis function model.
- In general the idea is to do a linear regression of y on $\phi_1(x), \phi_2(x), \ldots, \phi_{m-1}(x)$ where ϕ_j are basis functions.

Linear Basis Function Models

- We talked before about polynomials x^2, x^3, x^4 for modeling non-linearities, this is a linear basis function model.
- In general the idea is to do a linear regression of y on $\phi_1(x), \phi_2(x), \ldots, \phi_{m-1}(x)$ where ϕ_j are basis functions.
- The model is now:

$$y = f(x, \beta) + \epsilon$$

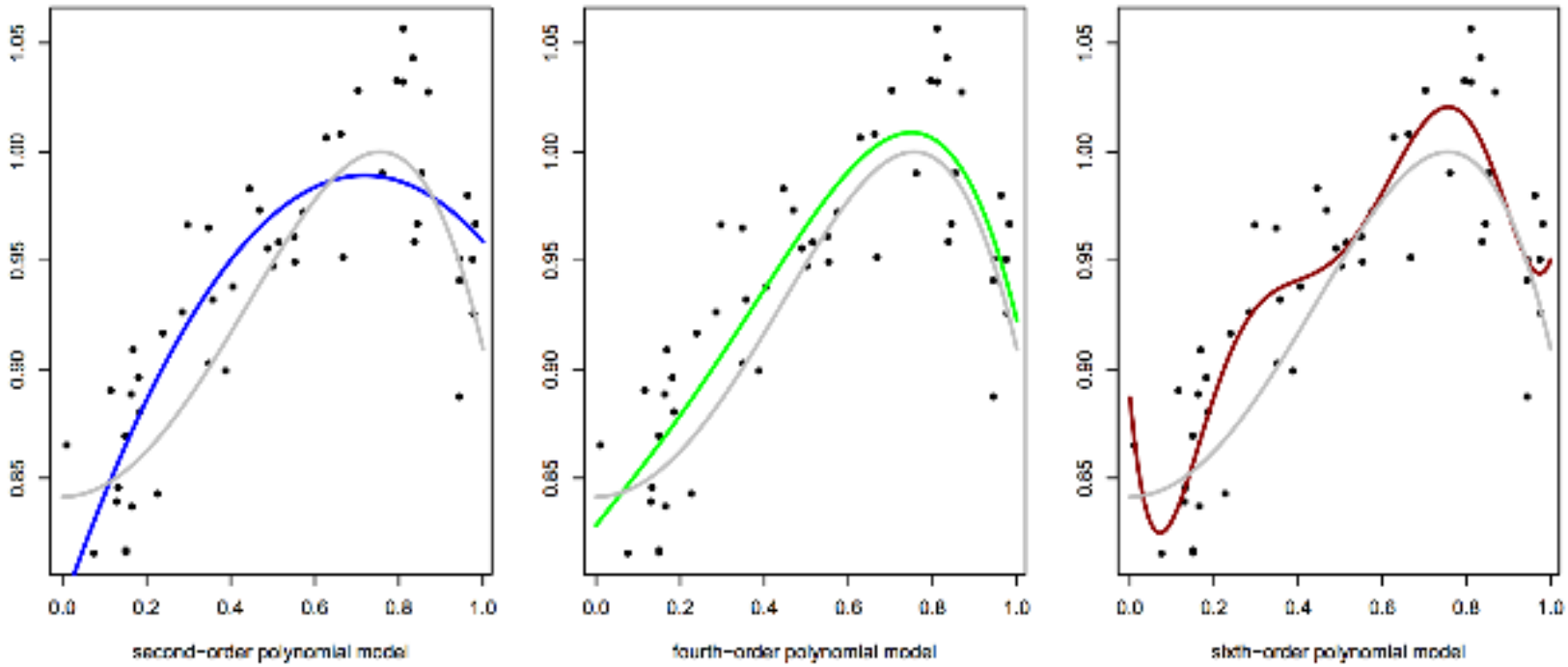
$$f(x, \beta) = \beta_0 + \sum_{j=1}^{m-1} \beta_j \phi_j(x) = \beta^T \phi(x)$$

Polynomial Basis Functions

We have already seen some basis functions. Here are OLS fits with polynomial basis functions of increasing order.

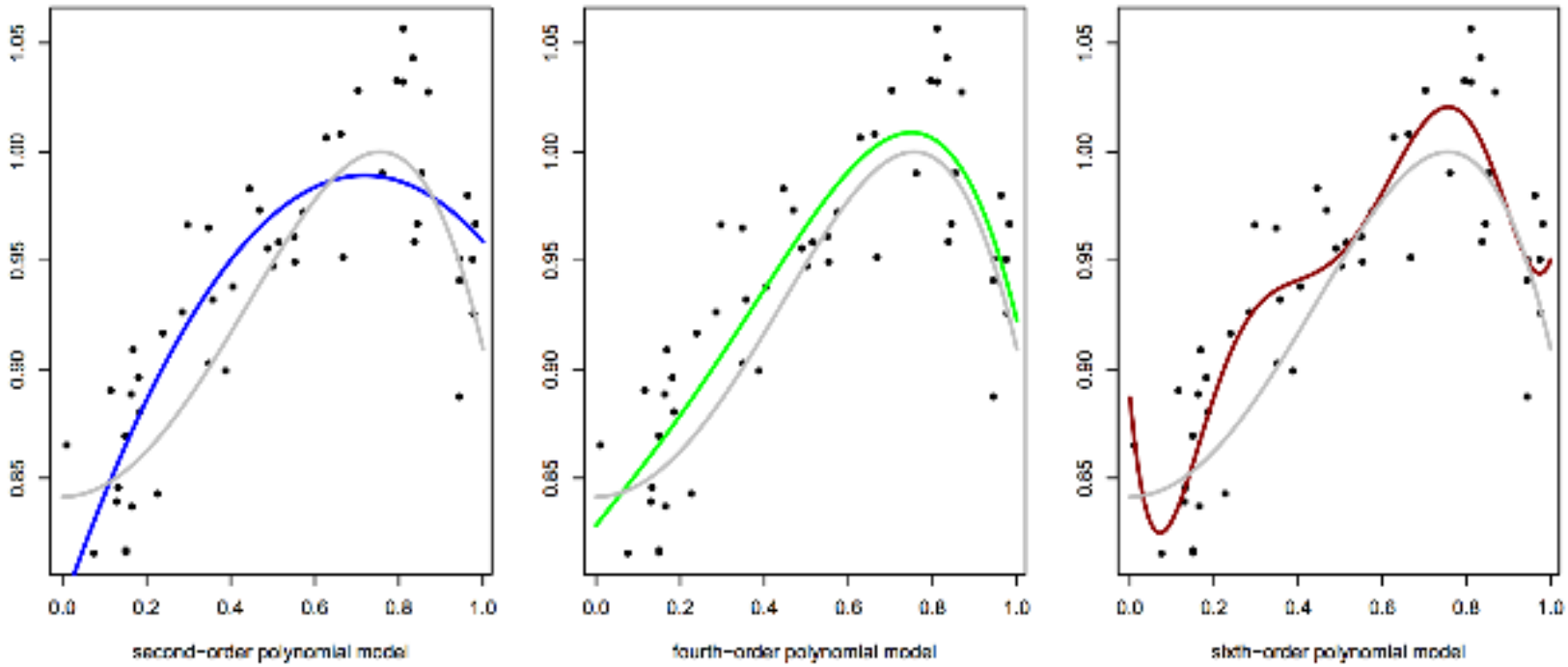
Polynomial Basis Functions

We have already seen some basis functions. Here are OLS fits with polynomial basis functions of increasing order.



Polynomial Basis Functions

We have already seen some basis functions. Here are OLS fits with polynomial basis functions of increasing order.



It appears that the last model is too complex and is overfitting a bit.

Local Basis Functions

Local Basis Functions

Polynomials are global basis functions, each affecting the prediction over the whole input space. Often local basis functions are more appropriate.

Local Basis Functions

Polynomials are global basis functions, each affecting the prediction over the whole input space. Often local basis functions are more appropriate.

One choice is a Gaussian basis function

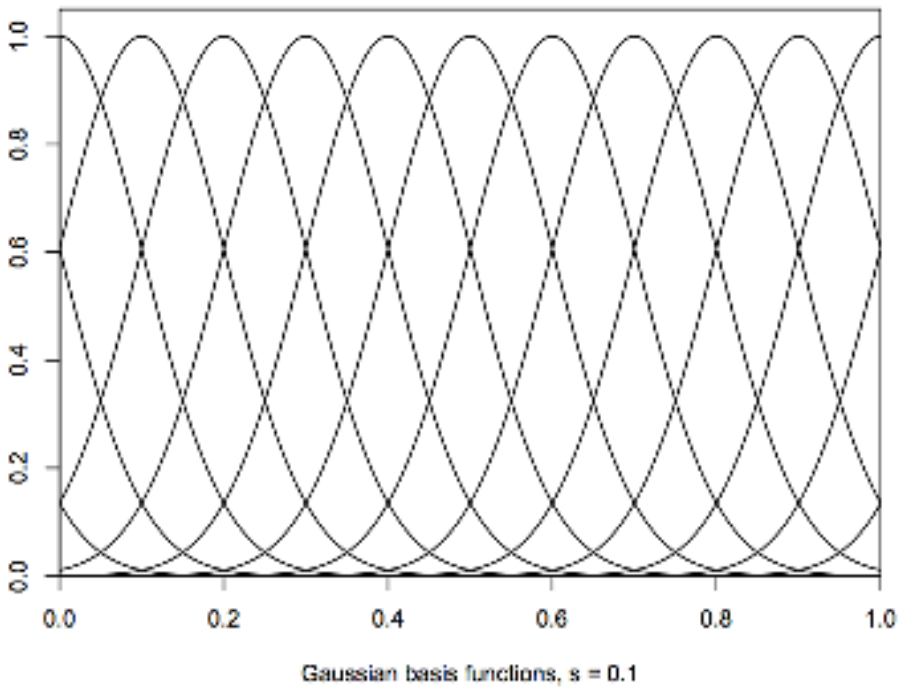
$$\phi_j(x) = \exp(-(x - \mu_j)^2)/2s^2)$$

Local Basis Functions

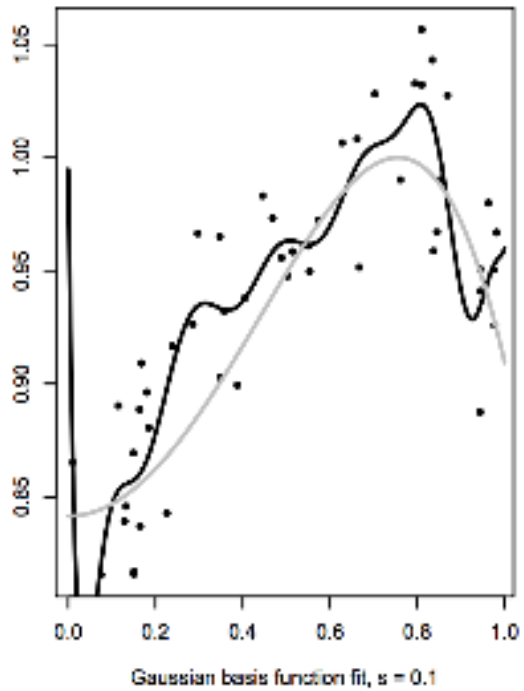
Polynomials are global basis functions, each affecting the prediction over the whole input space. Often local basis functions are more appropriate.

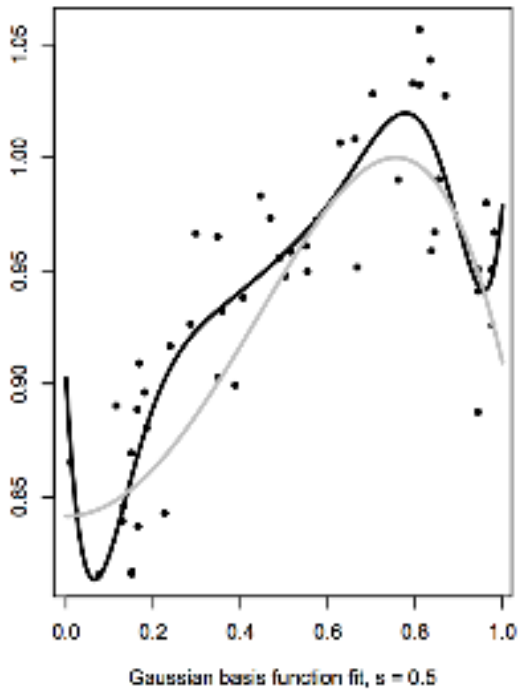
One choice is a Gaussian basis function

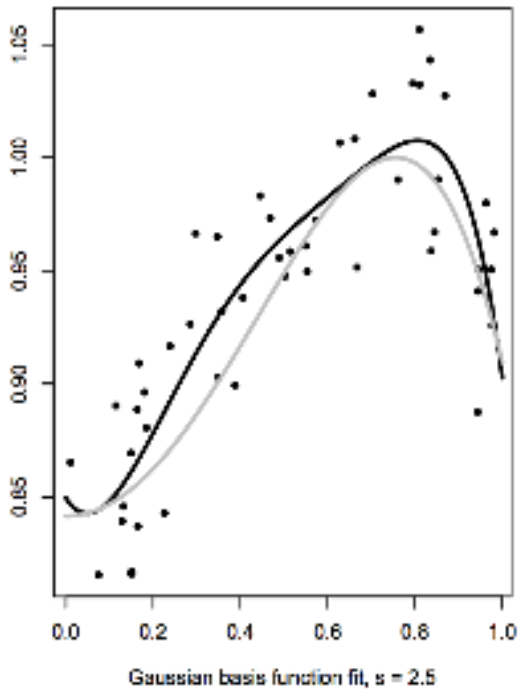
$$\phi_j(x) = \exp(-(x - \mu_j)^2)/2s^2)$$



Gaussian Basis Fits







Regularization

Regularization

- We've seen that flexible models can lead to overfitting

Regularization

- We've seen that flexible models can lead to overfitting
- Two ways to address: limit model flexibility or use a flexible model and regularize

Regularization

- We've seen that flexible models can lead to overfitting
- Two ways to address: limit model flexibility or use a flexible model and regularize
- Regularization is a way of expressing a preference for smoothness in our function by adding a penalty term to our optimization function.

Regularization

- We've seen that flexible models can lead to overfitting
- Two ways to address: limit model flexibility or use a flexible model and regularize
- Regularization is a way of expressing a preference for smoothness in our function by adding a penalty term to our optimization function.
- Here we will consider a penalty of the form $\lambda \sum_{j=1}^{m-1} \beta_j^2$ where λ controls the strength of the penalty.

Regularization

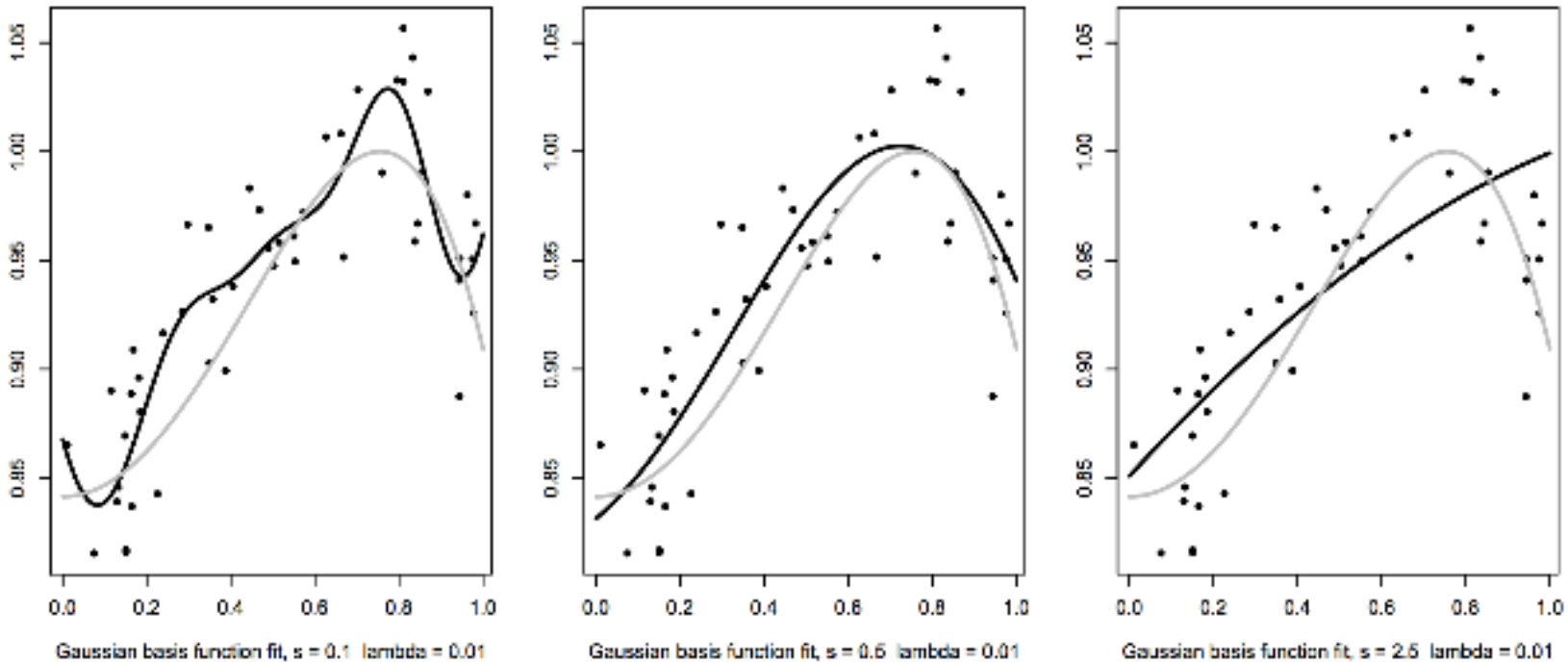
- We've seen that flexible models can lead to overfitting
- Two ways to address: limit model flexibility or use a flexible model and regularize
- Regularization is a way of expressing a preference for smoothness in our function by adding a penalty term to our optimization function.
- Here we will consider a penalty of the form $\lambda \sum_{j=1}^{m-1} \beta_j^2$ where λ controls the strength of the penalty.
- The penalty trades off some bias for an improvement in variance

Regularization

- We've seen that flexible models can lead to overfitting
- Two ways to address: limit model flexibility or use a flexible model and regularize
- Regularization is a way of expressing a preference for smoothness in our function by adding a penalty term to our optimization function.
- Here we will consider a penalty of the form $\lambda \sum_{j=1}^{m-1} \beta_j^2$ where λ controls the strength of the penalty.
- The penalty trades off some bias for an improvement in variance
- The trick in general is how to set λ

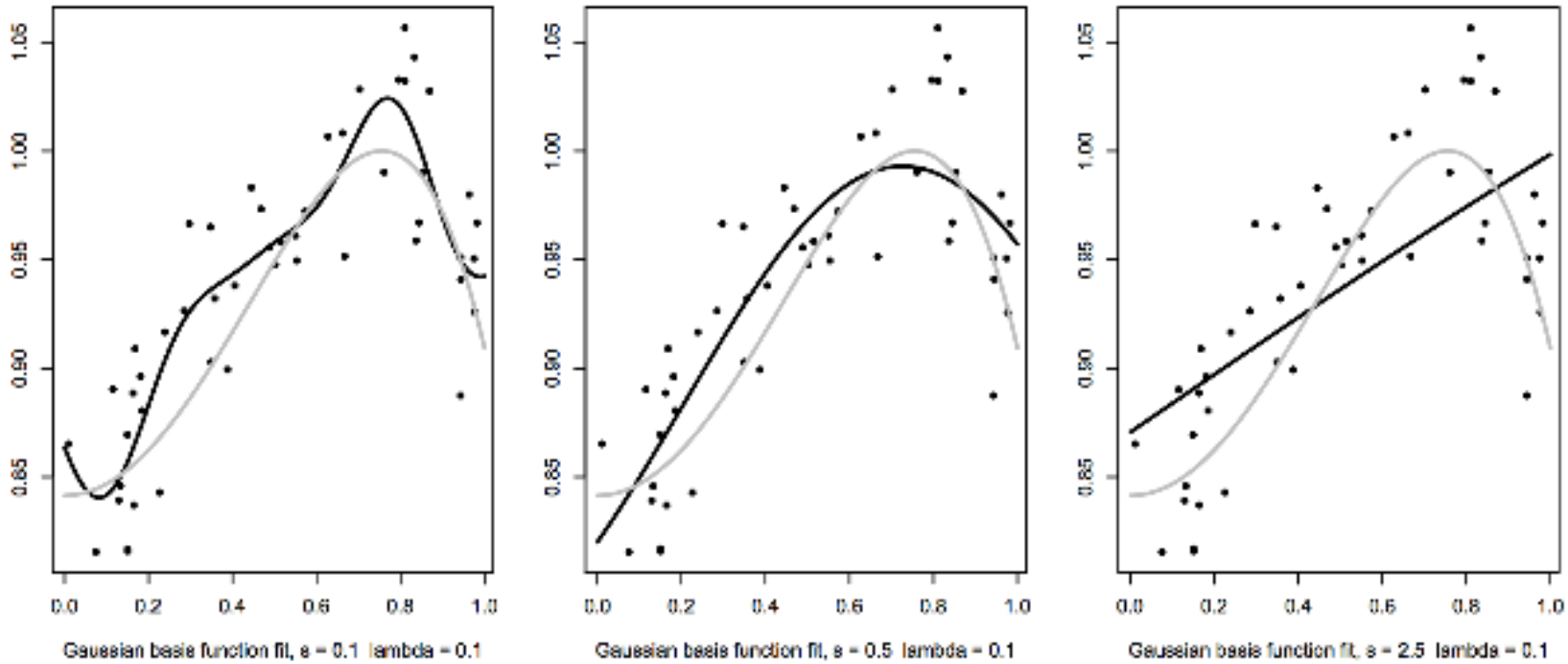
Results

Here are the results with $\lambda = 0.01$:



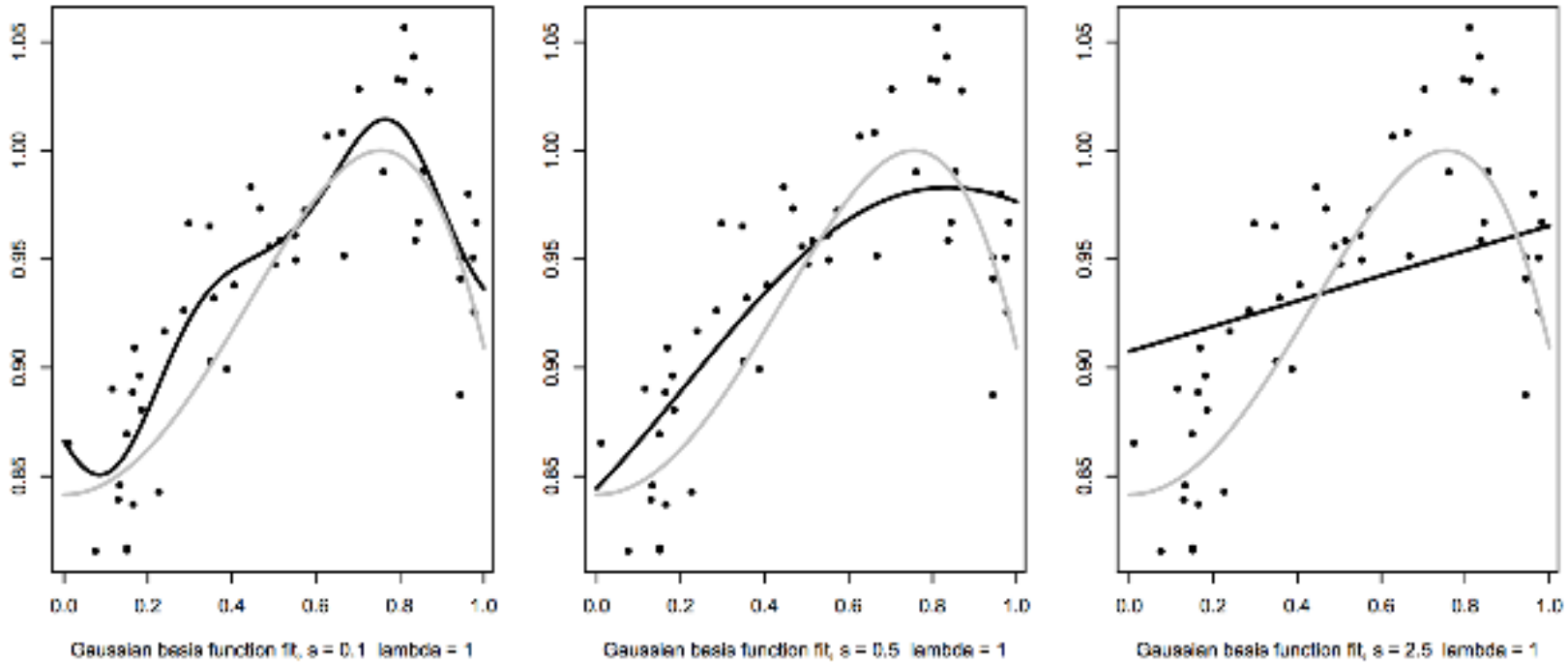
Results

Here are the results with $\lambda = 0.1$:



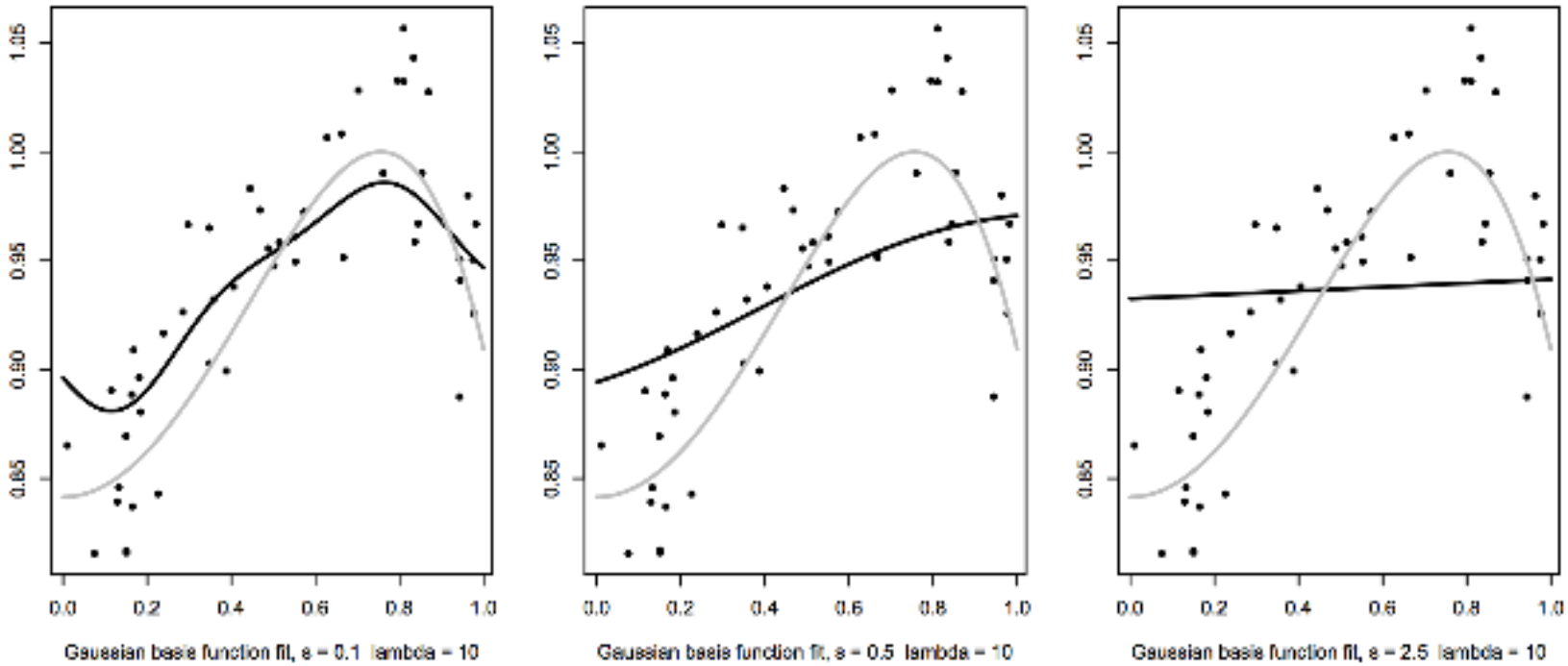
Results

Here are the results with $\lambda = 1$:



Results

Here are the results with $\lambda = 10$:



Conclusions from This Example

Conclusions from This Example

- we can control overfitting by modifying the width of the basis function s or with penalty

Conclusions from This Example

- we can control overfitting by modifying the width of the basis function s or with penalty
- we will need some way in general to tune these

Conclusions from This Example

- we can control overfitting by modifying the width of the basis function s or with penalty
- we will need some way in general to tune these
- we will also need some way to handle multivariate functions.

Conclusions from This Example

- we can control overfitting by modifying the width of the basis function s or with penalty
- we will need some way in general to tune these
- we will also need some way to handle multivariate functions.

Generalized Additive Models (GAM)

Recall the linear model,

$$y_i = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3 + u_i$$

Generalized Additive Models (GAM)

Recall the linear model,

$$y_i = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3 + u_i$$

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,

$$y_i = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3 + u_i$$

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:

$$y_i = \beta_0 + s_1(x_{1i}) + s_2(x_{2i}) + s_3(x_{3i}) + u_i$$

Generalized Additive Models (GAM)

$$y_i = \beta_0 + s_1(x_{1i}) + s_2(x_{2i}) + s_3(x_{3i}) + u_i$$

- GAMS are semi-parametric, they strike a compromise between nonparametric methods and parametric regression

Generalized Additive Models (GAM)

$$y_i = \beta_0 + s_1(x_{1i}) + s_2(x_{2i}) + s_3(x_{3i}) + u_i$$

- GAMS are semi-parametric, they strike a compromise between nonparametric methods and parametric regression
- $s_j(\cdot)$ are usually estimated with locally weighted regression smoothers or cubic smoothing splines (but many approaches are possible)

Generalized Additive Models (GAM)

$$y_i = \beta_0 + s_1(x_{1i}) + s_2(x_{2i}) + s_3(x_{3i}) + u_i$$

- GAMS are semi-parametric, they strike a compromise between nonparametric methods and parametric regression
- $s_j(\cdot)$ are usually estimated with locally weighted regression smoothers or cubic smoothing splines (but many approaches are possible)
- They do NOT give you a set of regression parameters $\hat{\beta}$. Instead one obtains a graphical summary of how $E[Y|X, X_2, ..., X_k]$ varies with X_1 (estimates of $s_j(\cdot)$ at every value of $X_{i,j}$)

Generalized Additive Models (GAM)

$$y_i = \beta_0 + s_1(x_{1i}) + s_2(x_{2i}) + s_3(x_{3i}) + u_i$$

- GAMS are semi-parametric, they strike a compromise between nonparametric methods and parametric regression
- $s_j(\cdot)$ are usually estimated with locally weighted regression smoothers or cubic smoothing splines (but many approaches are possible)
- They do NOT give you a set of regression parameters $\hat{\beta}$. Instead one obtains a graphical summary of how $E[Y|X, X_2, ..., X_k]$ varies with X_1 (estimates of $s_j(\cdot)$ at every value of $X_{i,j}$)
- 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.

$$y_i = \beta_0 + s_{12}(x_{1i}, x_{2i}) + s_3(x_{3i}) + u_i$$

Generalized Additive Models (GAM)

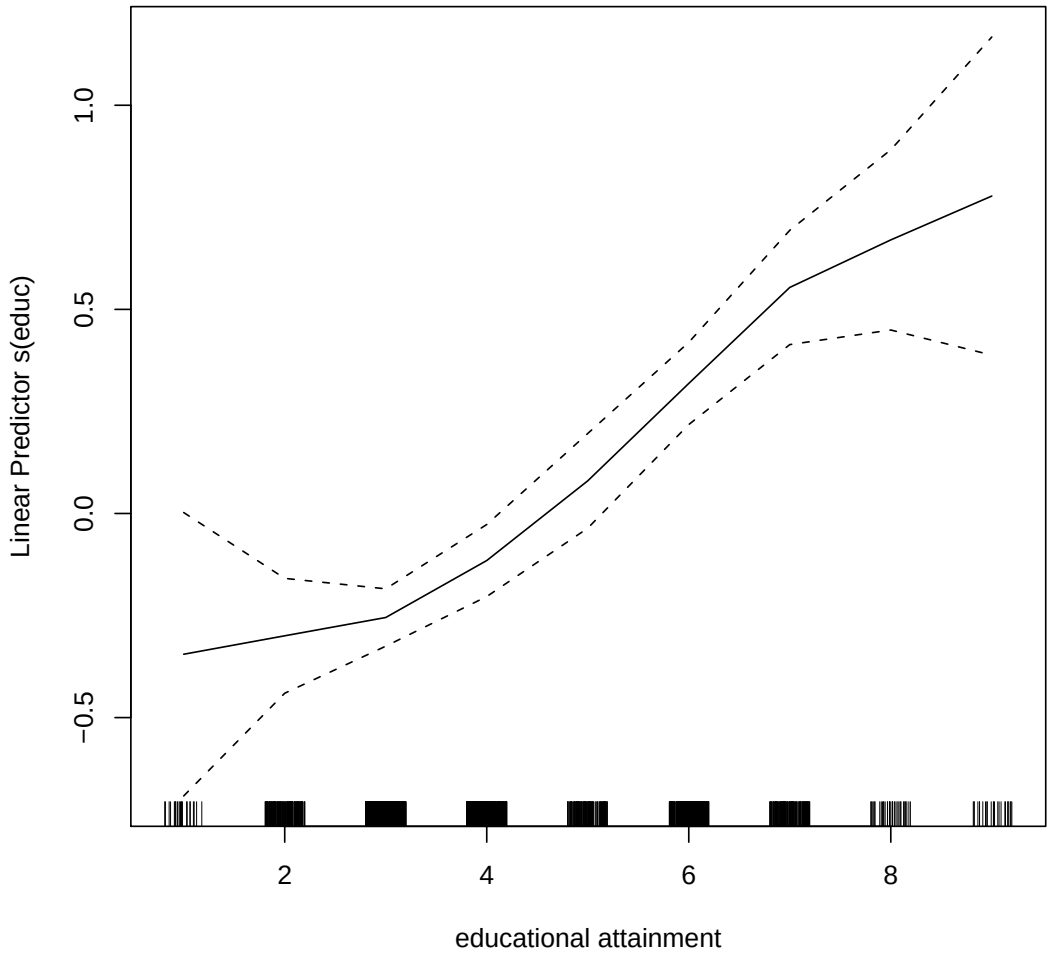
The GAM approach can be extended to allow interactions ($s_{12}(\cdot)$) between explanatory variables, but this eats up degrees of freedom so you need a lot of data.

$$y_i = \beta_0 + s_{12}(x_{1i}, x_{2i}) + s_3(x_{3i}) + u_i$$

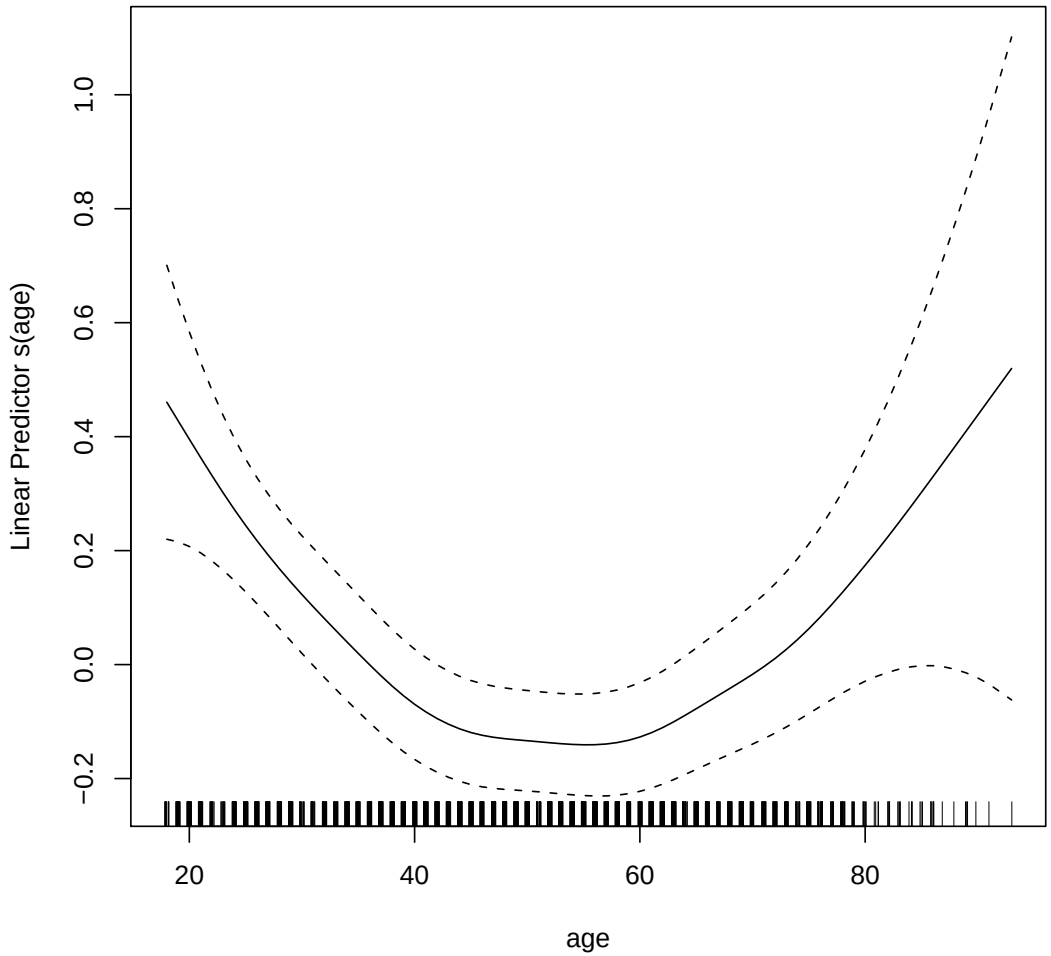
It can also be used for hybrid models where we model some variables as parametrically and other with a flexible function:

$$y_i = \beta_0 + \beta_1 x_{1i} + s_2(x_{2i}) + s_3(x_{3i}) + u_i$$

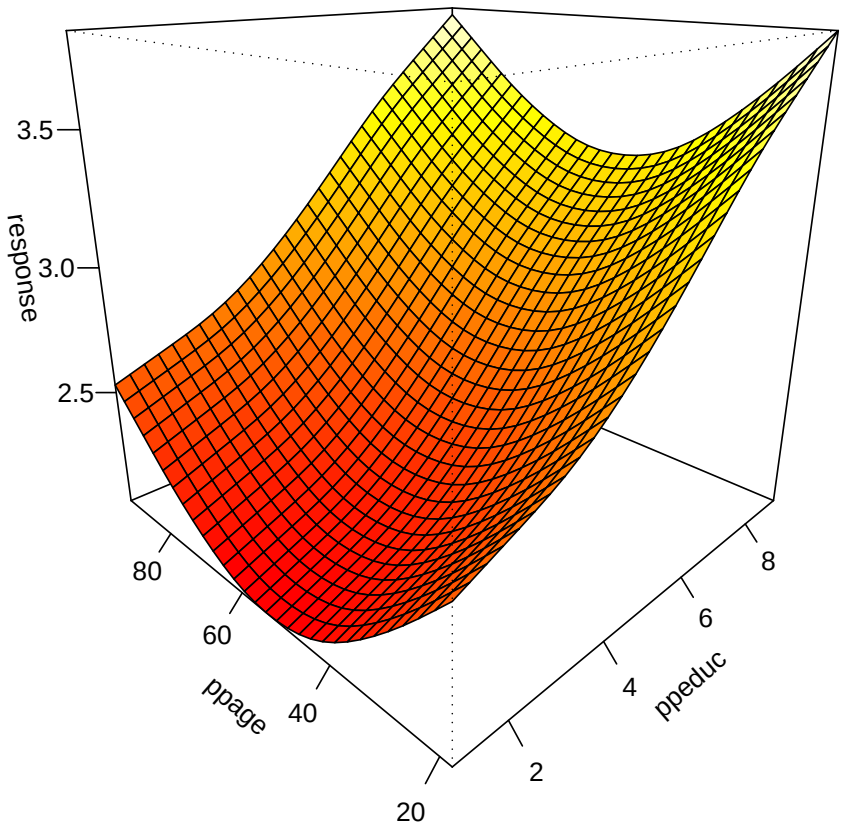
GAM Fit to Attitudes Toward Immigration



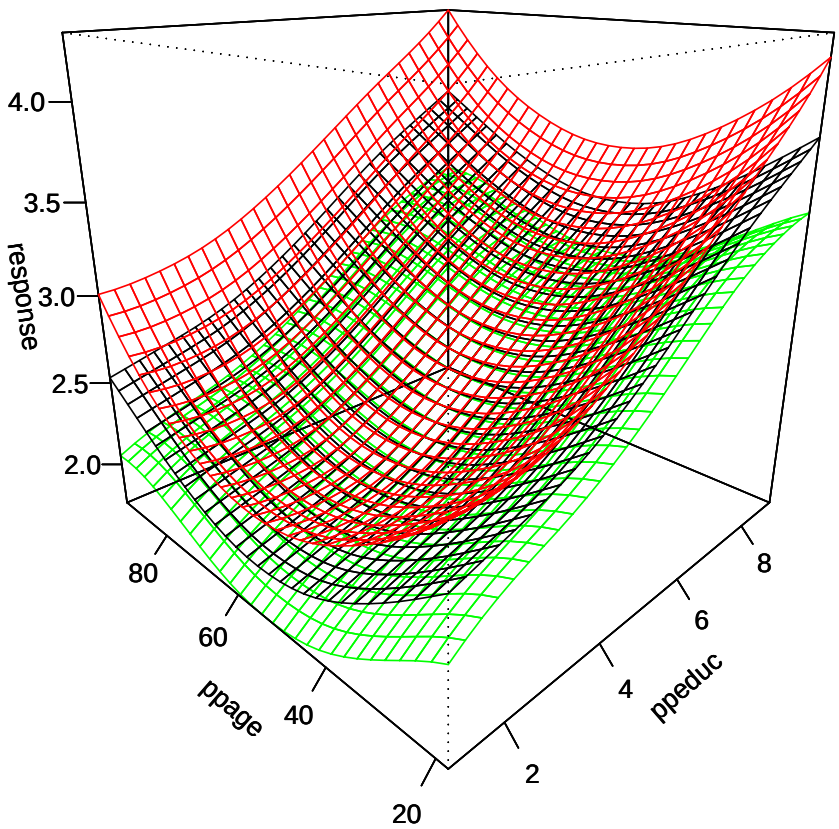
GAM Fit to Attitudes Toward Immigration



GAM Fit to Attitudes Toward Immigration

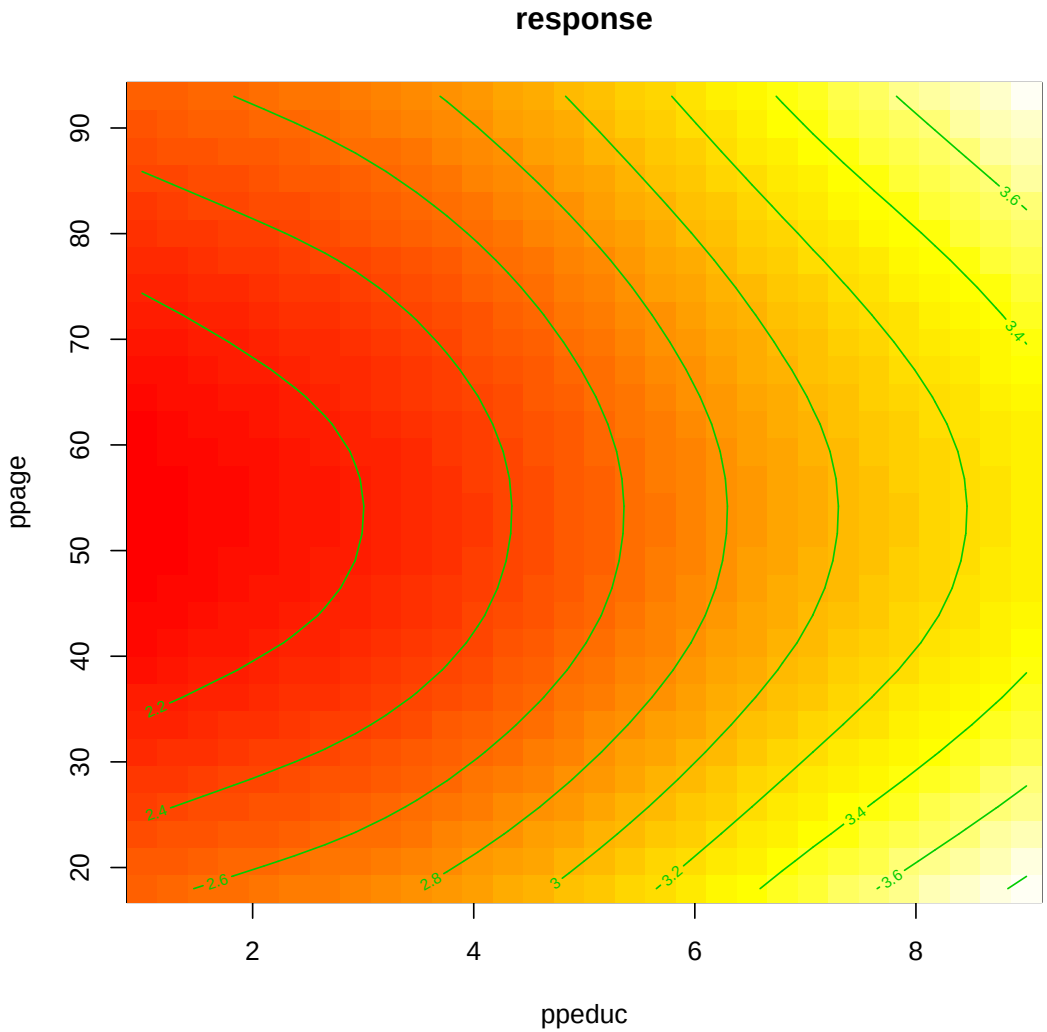


GAM Fit to Attitudes Toward Immigration



red/green are +/− 2 s.e.

GAM Fit to Attitudes Toward Immigration



Concluding Thoughts

Concluding Thoughts

- Non-linearity is pretty easy to detect and can substantially change our inferences

Concluding Thoughts

- Non-linearity is pretty easy to detect and can substantially change our inferences
- GAMs are a great way to model/detect non-linearity but transformations are often simpler

Concluding Thoughts

- Non-linearity is pretty easy to detect and can substantially change our inferences
- GAMs are a great way to model/detect non-linearity but transformations are often simpler
- However, be wary of the global properties of transformations and polynomials

Concluding Thoughts

- Non-linearity is pretty easy to detect and can substantially change our inferences
- GAMs are a great way to model/detect non-linearity but transformations are often simpler
- However, be wary of the global properties of transformations and polynomials
- Non-linearity concerns are most relevant for continuous covariates with a large range (age)

Concluding Thoughts

- Non-linearity is pretty easy to detect and can substantially change our inferences
- GAMs are a great way to model/detect non-linearity but transformations are often simpler
- However, be wary of the global properties of transformations and polynomials
- Non-linearity concerns are most relevant for continuous covariates with a large range (age)
- NB: it is okay if you didn't follow all of this today! GAMs are tricky.

This Week in Review

This Week in Review

- Analysis of extreme values: outliers, leverage points, and influence points.

This Week in Review

- Analysis of extreme values: outliers, leverage points, and influence points.
- Approaches to modeling non-linearity in the CEF.

This Week in Review

- Analysis of extreme values: outliers, leverage points, and influence points.
- Approaches to modeling non-linearity in the CEF.

Next week: Adjusting for Observed Confounding