Week 4: Testing/Regression

Brandon Stewart[1]

Princeton

September 26–28, 2022

[1] Many of the slides are adapted from Matt Blackwell, Adam Glynn, Jens Hainmueller, and Erin Hartman

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

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

- Last Week
  - inference and estimator properties
  - point estimates, confidence intervals

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

- Last Week
  - inference and estimator properties
  - point estimates, confidence intervals
- This Week

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

- Last Week
  - inference and estimator properties
  - point estimates, confidence intervals
- This Week
  - hypothesis testing

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

- Last Week
  - inference and estimator properties
  - point estimates, confidence intervals
- This Week
  - hypothesis testing
  - what is regression?

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

- Last Week
  - inference and estimator properties
  - point estimates, confidence intervals
- This Week
  - hypothesis testing
  - what is regression?
  - nonparametric and linear regression

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

- Last Week
  - inference and estimator properties
  - point estimates, confidence intervals
- This Week
  - hypothesis testing
  - what is regression?
  - nonparametric and linear regression
- Next Week
  - inference for simple regression
  - properties of ordinary least squares

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

- Last Week
  - inference and estimator properties
  - point estimates, confidence intervals
- This Week
  - hypothesis testing
  - what is regression?
  - nonparametric and linear regression
- Next Week
  - inference for simple regression
  - properties of ordinary least squares
- Long Run
  - probability $\rightarrow$ inference $\rightarrow$ regression $\rightarrow$ causal inference

We Secretly Already Covered This!

American Political Science Review Vol. 102, No. 1 February 2008
DOI: 10.1017/S000305540808009X

Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment

ALAN S. GERBER *Yale University*
DONALD P. GREEN *Yale University*
CHRISTOPHER W. LARIMER *University of Northern Iowa*

We Secretly Already Covered This!

Dear Registered Voter:

WHAT IF YOUR NEIGHBORS KNEW WHETHER YOU VOTED?

Why do so many people fail to vote? We've been talking about the problem for years, but it only seems to get worse. This year, we're taking a new approach. We're sending this mailing to you and your neighbors to publicize who does and does not vote.

The chart shows the names of some of your neighbors, showing which have voted in the past. After the August 8 election, we intend to mail an updated chart. You and your neighbors will all know who voted and who did not.

DO YOUR CIVIC DUTY — VOTE!

MAPLE DR	Aug 04	Nov 04	Aug 06
9995 JOSEPH JAMES SMITH	Voted	Voted	______
9995 JENNIFER KAY SMITH		Voted	______
9997 RICHARD B JACKSON		Voted	______
9999 KATHY MARIE JACKSON		Voted	______

Review of the Gerber, Green and Larimer Result

- Our estimand is the difference in population means: $\theta = \mu_y - \mu_x$. Where μ_y is those receiving the social pressure mailer and μ_x is those receiving the civic duty mailer.

Review of the Gerber, Green and Larimer Result

- Our estimand is the difference in population means: $\theta = \mu_y - \mu_x$. Where μ_y is those receiving the social pressure mailer and μ_x is those receiving the civic duty mailer.
- We used difference in means to get an estimate of 0.063

Review of the Gerber, Green and Larimer Result

- Our estimand is the difference in population means: $\theta = \mu_y - \mu_x$. Where μ_y is those receiving the social pressure mailer and μ_x is those receiving the civic duty mailer.
- We used difference in means to get an estimate of 0.063
- We estimated the estimator's standard error $\widehat{\text{SE}}(\hat{\theta})$.

Review of the Gerber, Green and Larimer Result

- Our estimand is the difference in population means: $\theta = \mu_y - \mu_x$. Where μ_y is those receiving the social pressure mailer and μ_x is those receiving the civic duty mailer.
- We used difference in means to get an estimate of 0.063
- We estimated the estimator's standard error $\widehat{\text{SE}}(\hat{\theta})$.
- What if there was no difference in means in the population ($\mu_y - \mu_x = 0$)?

Review of the Gerber, Green and Larimer Result

- Our estimand is the difference in population means: $\theta = \mu_y - \mu_x$. Where μ_y is those receiving the social pressure mailer and μ_x is those receiving the civic duty mailer.
- We used difference in means to get an estimate of 0.063
- We estimated the estimator's standard error $\widehat{\text{SE}}(\hat{\theta})$.
- What if there was no difference in means in the population ($\mu_y - \mu_x = 0$)?
- By asymptotic Normality $(\hat{\theta} - 0)/\text{SE}(\hat{\theta}) \sim N(0, 1)$

Review of the Gerber, Green and Larimer Result

- Our estimand is the difference in population means: $\theta = \mu_y - \mu_x$. Where μ_y is those receiving the social pressure mailer and μ_x is those receiving the civic duty mailer.
- We used difference in means to get an estimate of 0.063
- We estimated the estimator's standard error $\widehat{\text{SE}}(\hat{\theta})$.
- What if there was no difference in means in the population ($\mu_y - \mu_x = 0$)?
- By asymptotic Normality $(\hat{\theta} - 0)/\text{SE}(\hat{\theta}) \sim N(0, 1)$
- This implies that $\hat{\theta} \sim \mathcal{N}(0, \text{SE}(\hat{\theta})^2)$

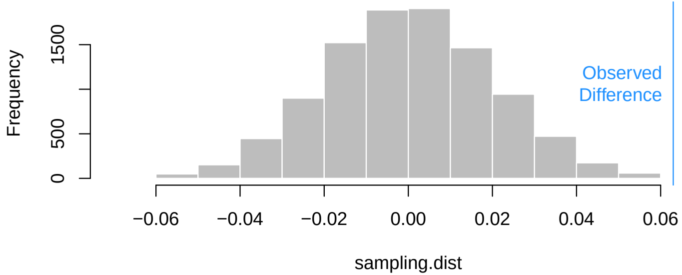
Review of the Gerber, Green and Larimer Result

- Our estimand is the difference in population means: $\theta = \mu_y - \mu_x$. Where μ_y is those receiving the social pressure mailer and μ_x is those receiving the civic duty mailer.
- We used difference in means to get an estimate of 0.063
- We estimated the estimator's standard error $\widehat{\text{SE}}(\hat{\theta})$.
- What if there was no difference in means in the population $(\mu_y - \mu_x = 0)$?
- By asymptotic Normality $(\hat{\theta} - 0)/\text{SE}(\hat{\theta}) \sim N(0, 1)$
- This implies that $\hat{\theta} \sim \mathcal{N}(0, \text{SE}(\hat{\theta})^2)$
- Which we then approximate with $\hat{\theta} \sim \mathcal{N}(0, \widehat{\text{SE}}(\hat{\theta})^2)$

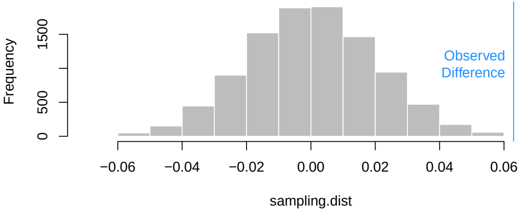
We Secretly Already Covered This!

We Secretly Already Covered This!

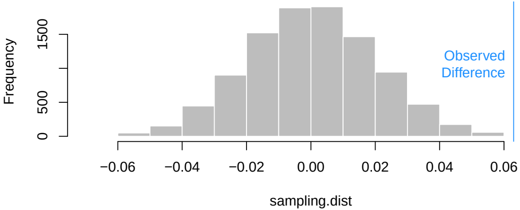
Example from Gerber, Green and Larimer (2008).



Overall Idea

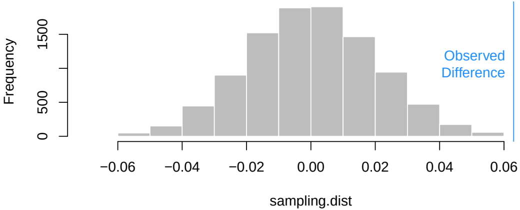


Overall Idea



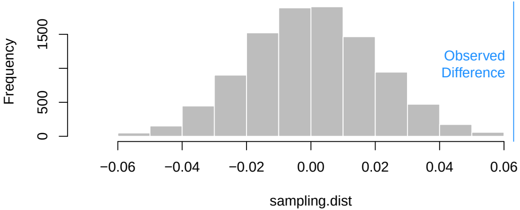
- We assumed we knew the difference in means was zero.

Overall Idea



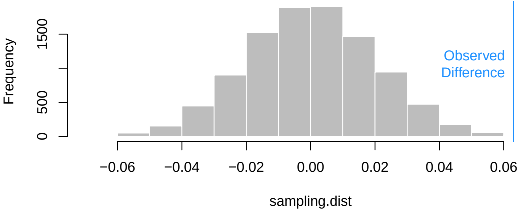
- We assumed we knew the difference in means was zero.
- We derived the sampling distribution under that value.

Overall Idea



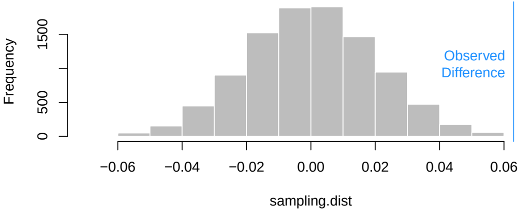
- We assumed we knew the difference in means was zero.
- We derived the sampling distribution under that value.
- We asked 'if the population difference was zero (i.e. this was the sampling distribution), how likely would it be to see an estimate as extreme (or more extreme) than our observed estimate?'

Overall Idea



- We assumed we knew the difference in means was zero.
- We derived the sampling distribution under that value.
- We asked 'if the population difference was zero (i.e. this was the sampling distribution), how likely would it be to see an estimate as extreme (or more extreme) than our observed estimate?'
- Our observed difference was so implausible we concluded it was unlikely the population difference was really zero.

Overall Idea



- We assumed we knew the difference in means was zero.
- We derived the sampling distribution under that value.
- We asked 'if the population difference was zero (i.e. this was the sampling distribution), how likely would it be to see an estimate as extreme (or more extreme) than our observed estimate?'
- Our observed difference was so implausible we concluded it was unlikely the population difference was really zero.

Basically, we did a hypothesis test.

What is a Hypothesis Test?

What is a Hypothesis Test?

- A hypothesis test is a way of assessing evidence against a particular hypothesis about the world—it is a statistical thought experiment.

What is a Hypothesis Test?

- A hypothesis test is a way of assessing evidence against a particular hypothesis about the world—it is a statistical thought experiment.
- By using probability, we define what the data would look like under a given state of the world, and then assess whether or not that's consistent with our observed data.

What is a Hypothesis Test?

- A hypothesis test is a way of assessing evidence against a particular hypothesis about the world—it is a statistical thought experiment.
- By using probability, we define what the data would look like under a given state of the world, and then assess whether or not that's consistent with our observed data.
- Hypothesis testing is an alternative way to think about inference than confidence intervals, but using much of the same infrastructure.

What is a Hypothesis Test?

- A hypothesis test is a way of assessing evidence against a particular hypothesis about the world—it is a statistical thought experiment.
- By using probability, we define what the data would look like under a given state of the world, and then assess whether or not that's consistent with our observed data.
- Hypothesis testing is an alternative way to think about inference than confidence intervals, but using much of the same infrastructure.
- Hypothesis tests lead to discrete decisions.

Background on Drug Testing

Background on Drug Testing

Statistics play an important role in determining which drugs are approved for sale by the FDA.

Background on Drug Testing

Statistics play an important role in determining which drugs are approved for sale by the FDA.

There are typically three phases of clinical trials before a drug is approved:

- Phase I: Toxicity (Will it kill you?)
- Phase II: Efficacy (Is there any evidence that it helps?)
- Phase III: Effectiveness (Is it better than existing treatments?)

Phase I trials are conducted on a small number of healthy volunteers, Phase II trial are either randomized experiments or within-patient comparisons, and Phase III trials are almost always randomized experiments with control groups.

Example

Example

Consider a Phase II efficacy trial reported in Sowers et al. (2006), for a drug combination designed to treat high blood pressure in patients with metabolic syndrome.

Example

Consider a Phase II efficacy trial reported in Sowers et al. (2006), for a drug combination designed to treat high blood pressure in patients with metabolic syndrome.

- The trial included 345 patients with initial systolic blood pressure between 140-159.

Example

Consider a Phase II efficacy trial reported in Sowers et al. (2006), for a drug combination designed to treat high blood pressure in patients with metabolic syndrome.

- The trial included 345 patients with initial systolic blood pressure between 140-159.
- Each subject was assigned to take the drug combination for 16 weeks.

Example

Consider a Phase II efficacy trial reported in Sowers et al. (2006), for a drug combination designed to treat high blood pressure in patients with metabolic syndrome.

- The trial included 345 patients with initial systolic blood pressure between 140-159.
- Each subject was assigned to take the drug combination for 16 weeks.
- Systolic blood pressure was measured on each subject before and after the treatment period.

Example

Subject	SBP_{before}	SBP_{after}	Decrease
1			
2			
3			
4			
$\vdots$			
345			

Example

Subject	SBP_{before}	SBP_{after}	Decrease
1	147		
2	153		
3	142		
4	141		
$\vdots$	$\vdots$		
345	155		

Example

Subject	SBP_{before}	SBP_{after}	Decrease
1	147	135	
2	153	122	
3	142	119	
4	141	134	
⋮	⋮	⋮	
345	155	115	

Example

Subject	SBP_{before}	SBP_{after}	Decrease
1	147	135	12
2	153	122	31
3	142	119	23
4	141	134	7
$\vdots$	$\vdots$	$\vdots$	$\vdots$
345	155	115	40

Example

Example

- The drug was administered to 345 patients.

Example

- The drug was administered to 345 patients.
- On average, blood pressure was 21 points lower after treatment.

Example

- The drug was administered to 345 patients.
- On average, blood pressure was 21 points lower after treatment.
- The standard deviation of changes in blood pressure was 14.3.

Example

- The drug was administered to 345 patients.
- On average, blood pressure was 21 points lower after treatment.
- The standard deviation of changes in blood pressure was 14.3.

Question: Should the FDA allow the drug to proceed to the next stage of testing?

The FDA's Decision

We can think of the FDA's problem in terms of two dimensions:

- The true state of the world
- The decision made by the FDA

	Drug works	Drug doesn't work
FDA approves		
FDA doesn't approve		

Two kinds of bad decisions:

- False positive
- False negative

The FDA's Decision

We can think of the FDA's problem in terms of two dimensions:

- The true state of the world
- The decision made by the FDA

	Drug works	Drug doesn't work
FDA approves	Good!	
FDA doesn't approve		Good!

Two kinds of bad decisions:

- False positive
- False negative

The FDA's Decision

We can think of the FDA's problem in terms of two dimensions:

- The true state of the world
- The decision made by the FDA

	Drug works	Drug doesn't work
FDA approves	Good!	
FDA doesn't approve	Bad!	Good!

Two kinds of bad decisions:

- False positive
- False negative

The FDA's Decision

We can think of the FDA's problem in terms of two dimensions:

- The true state of the world
- The decision made by the FDA

	Drug works	Drug doesn't work
FDA approves	Good!	Bad!
FDA doesn't approve	Bad!	Good!

Two kinds of bad decisions:

- False positive
- False negative

Elements of a Hypothesis Test

Important terms we are about to define:

Elements of a Hypothesis Test

Important terms we are about to define:

- Null Hypothesis (assumed state of world for test)
- Alternative Hypothesis (all other states of the world)
- Type I and Type II Errors (two types of errors)
- Test Statistic (what we will observe from the sample)
- Test Level (the probability of a type I error)
- Rejection Region (the basis of our decision)

Hypotheses

- Null Hypothesis (H_0): The conservatively assumed state of the world we are accumulating evidence against (often "no effect").

  Example: There is not a difference in voting rates between those who received the social pressure mailer and those that received the civic duty mailer.

Hypotheses

- Null Hypothesis (H_0): The conservatively assumed state of the world we are accumulating evidence against (often "no effect").

  Example: There is not a difference in voting rates between those who received the social pressure mailer and those that received the civic duty mailer.

- Alternative Hypothesis (H_a): The state of the world where the null hypothesis is not true and thus the claim to be indirectly tested.

  Example: There is a difference in voting rates between those who received the social pressure mailer and those that received the civic duty mailer.

Error Types

	(H_0 False)	(H_0 True)
Reject H_0		
Don't Reject H_0		

Error Types

	(H_0 False)	(H_0 True)
Reject H_0	Correct	
Don't Reject H_0		Correct

Error Types

	(H_0 False)	(H_0 True)
Reject H_0	Correct	Type I error
Don't Reject H_0		Correct

Error Types

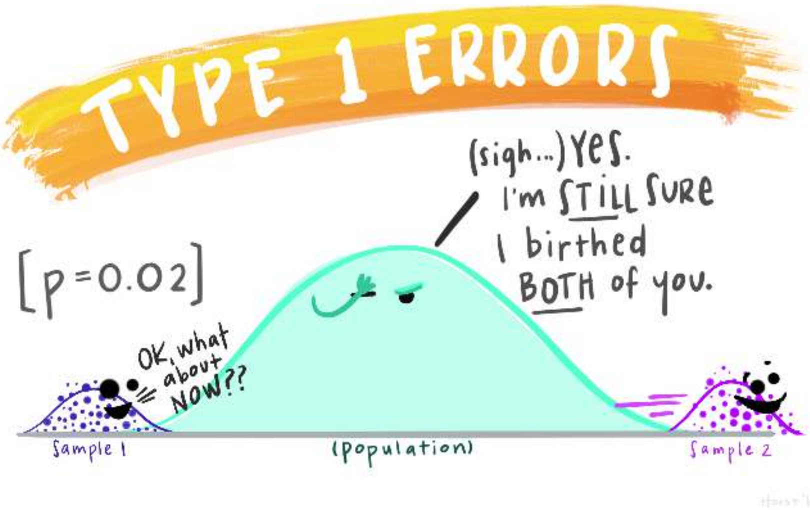
	(H_0 False)	(H_0 True)
Reject H_0	Correct	Type I error
Don't Reject H_0	Type II error	Correct

Error Types

	(H_0 False)	(H_0 True)
Reject H_0	Correct	Type I error
Don't Reject H_0	Type II error	Correct

We generally make the normative judgment that we prefer an undetected finding (Type II error) to a false discovery (Type I error).

A Visual Reminder from Allison Horst



Artwork by @allison_horst

A Visual Reminder from Allison Horst

Artwork by @allison_horst

Test Statistics and Null Distributions

Test Statistics and Null Distributions

Test Statistic: which we denote T_n is a function of the sample, the estimator and the null hypothesis.

$$T_n = \frac{\hat{\theta} - \theta_0}{\widehat{\mathsf{SE}}[\hat{\theta}]}$$

This is a random variable because it is a function of the sample.

Test Statistics and Null Distributions

Test Statistic: which we denote T_n is a function of the sample, the estimator and the null hypothesis.

$$T_n = \frac{\hat{\theta} - \theta_0}{\widehat{\text{SE}}[\hat{\theta}]}$$

This is a random variable because it is a function of the sample.

Null Distribution: the sampling distribution of the statistic/test statistic <u>assuming that the null is true</u>.

Test Statistics and Null Distribution

Returning to the voting experiment. We know from the Central Limit Theorem that the standardized difference in means has a standard normal distribution asymptotically,

$$T_n = \frac{\hat{\theta} - (\mu_y - \mu_x)}{\widehat{\mathsf{SE}}[\hat{\theta}]} \xrightarrow{d} \mathcal{N}(0, 1)$$

Test Statistics and Null Distribution

Returning to the voting experiment. We know from the Central Limit Theorem that the standardized difference in means has a standard normal distribution asymptotically,

$$T_n = \frac{\hat{\theta} - (\mu_y - \mu_x)}{\widehat{\mathsf{SE}}[\hat{\theta}]} \xrightarrow{d} \mathcal{N}(0,1)$$

Under the null hypothesis of $\mu_y - \mu_x = 0$, we have

$$T_n = \frac{\hat{\theta}}{\widehat{\mathsf{SE}}[\hat{\theta}]} \xrightarrow{d} \mathcal{N}(0,1)$$

If T_n is very far from zero—in the sense that it has low probability under $\mathcal{N}(0,1)$—then we reject the null hypothesis as not plausible.

Rejection Region

Rejection Region

- The rejection region R contains the values of the test statistics, T_n for which we reject the null. This is defined by the null distribution and our chosen tolerance for Type I error.

Rejection Region

- The rejection region R contains the values of the test statistics, T_n for which we reject the null. This is defined by the null distribution and our chosen tolerance for Type I error.
- Denote the probability of Type I error as α.

Rejection Region

- The rejection region R contains the values of the test statistics, T_n for which we reject the null. This is defined by the null distribution and our chosen tolerance for Type I error.
- Denote the probability of Type I error as α.
- If we have a two-sided test (i.e. our null hypothesis equals a given value and thus extreme values on either side could reject the null), we reject when $|T_n| > c$ where $P_{H_0}(|T_n| > c) = \alpha$.

Rejection Region

- The rejection region R contains the values of the test statistics, T_n for which we reject the null. This is defined by the null distribution and our chosen tolerance for Type I error.
- Denote the probability of Type I error as α.
- If we have a two-sided test (i.e. our null hypothesis equals a given value and thus extreme values on either side could reject the null), we reject when $|T_n| > c$ where $P_{H_0}(|T_n| > c) = \alpha$.
- We call c the critical value.

Rejection Region

- The rejection region R contains the values of the test statistics, T_n for which we reject the null. This is defined by the null distribution and our chosen tolerance for Type I error.
- Denote the probability of Type I error as α.
- If we have a two-sided test (i.e. our null hypothesis equals a given value and thus extreme values on either side could reject the null), we reject when $|T_n| > c$ where $P_{H_0}(|T_n| > c) = \alpha$.
- We call c the critical value.
- Much like the 95% confidence interval, we pick $\alpha = .05$ by convention.

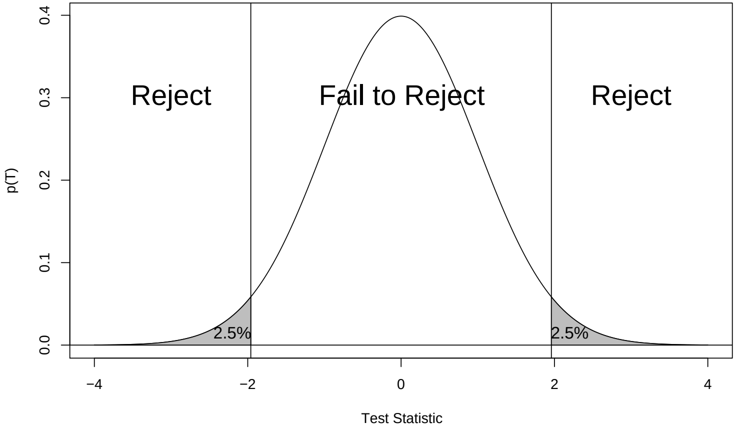
The value for which P=0.05, or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation ought to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant. Using this criterion we should be led to follow up a false indication only once in 22 trials, even if the statistics were the only guide available. Small effects will still escape notice if the data are insufficiently numerous to bring them out, but no lowering of the standard of significance would meet this difficulty.

- Ronald Fisher, *Design of Experiments* (1922)

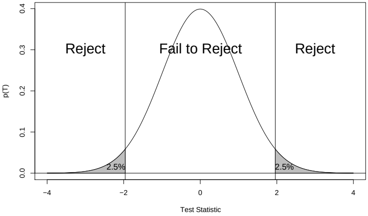
Two-sided Rejection Region

Two-sided Rejection Region

Rejection region with $\alpha = .05$, $H_0 : \theta = 0$, $H_A : \theta \neq 0$:

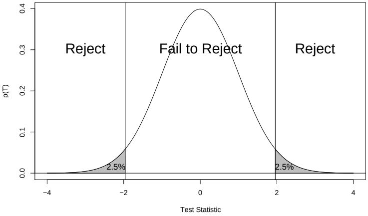


Defining the Rejection Region



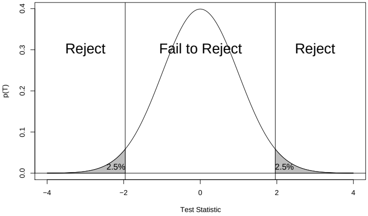
- We define the rejection region such that $|T_n| > c$ where $P_{H_0}(|T_n| > c) = \alpha$.

Defining the Rejection Region



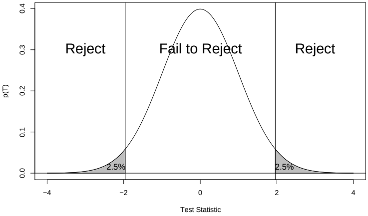
- We define the rejection region such that $|T_n| > c$ where $P_{H_0}(|T_n| > c) = \alpha$.
- This means that we will get a false rejection only 5% of the time.

Defining the Rejection Region



- We define the rejection region such that $|T_n| > c$ where $P_{H_0}(|T_n| > c) = \alpha$.
- This means that we will get a false rejection only 5% of the time.
- We want to find the point c such that $P_{H_0}(T_n < c) + P_{H_0}(T_n > c) = \alpha$ where we typically use equal probability on each side by convention.

Defining the Rejection Region



- We define the rejection region such that $|T_n| > c$ where $P_{H_0}(|T_n| > c) = \alpha$.
- This means that we will get a false rejection only 5% of the time.
- We want to find the point c such that $P_{H_0}(T_n < c) + P_{H_0}(T_n > c) = \alpha$ where we typically use equal probability on each side by convention.
- This is just the task of finding the quantile for $\alpha/2$. In the case of $\alpha = .05$, $\texttt{qnorm}(.05/2) = 1.96$

The Complete Recipe

The Complete Recipe

1. Define a null hypothesis and the alternative hypothesis.
2. Choose a test statistic T_n.
3. Determine the Type I error you will tolerate (α).
4. Determine your critical value and thus your rejection region.
5. Calculate your test statistic in your observed data.
6. If your observed data is sufficiently unlikely under your null hypothesis, reject your null.

The Gerber, Green and Larimer Example

```
diff <- mean(treated) - mean(control)
se_diff <- sqrt(var(treated)/length(treated) +
                var(control)/length(control))
test_statistic <- diff/se_diff
```

This yields 18.3 which is much larger in absolute value than our .05 critical value of 1.96.

We reject the null.

Back to the FDA

Let's go back to the FDA making a decision about drugs.

Back to the FDA

Let's go back to the FDA making a decision about drugs.

	Drug works	Drug doesn't work
FDA approves	Good!	Bad!
FDA doesn't approve	Bad!	Good!

Back to the FDA

Let's go back to the FDA making a decision about drugs.

	Drug works	Drug doesn't work
FDA approves	Good!	Bad!
FDA doesn't approve	Bad!	Good!

Drug trials are expensive and *ex ante* we can specify that we only care about one direction in particular. Consider the Sowers et al (2006) case which claimed to decrease blood pressure.

Sowers et al. Example

We can define our hypotheses for a one-sided test.

$$H_0 : \mu_{decrease} \leq 0 \quad (1)$$
$$H_a : \mu_{decrease} > 0 \quad (2)$$

Sowers et al. Example

We can define our hypotheses for a one-sided test.

$$H_0 : \mu_{decrease} \leq 0 \tag{1}$$
$$H_a : \mu_{decrease} > 0 \tag{2}$$

We can calculate the test statistic:

Sowers et al. Example

We can define our hypotheses for a one-sided test.

$$H_0 : \mu_{decrease} \leq 0 \quad (1)$$

$$H_a : \mu_{decrease} > 0 \quad (2)$$

We can calculate the test statistic:

- $\overline{x} = 21.0$
- $\hat{\sigma} = 14.3$
- $n = 345$

Sowers et al. Example

We can define our hypotheses for a one-sided test.

$$H_0 : \mu_{decrease} \leq 0 \tag{1}$$
$$H_a : \mu_{decrease} > 0 \tag{2}$$

We can calculate the test statistic:

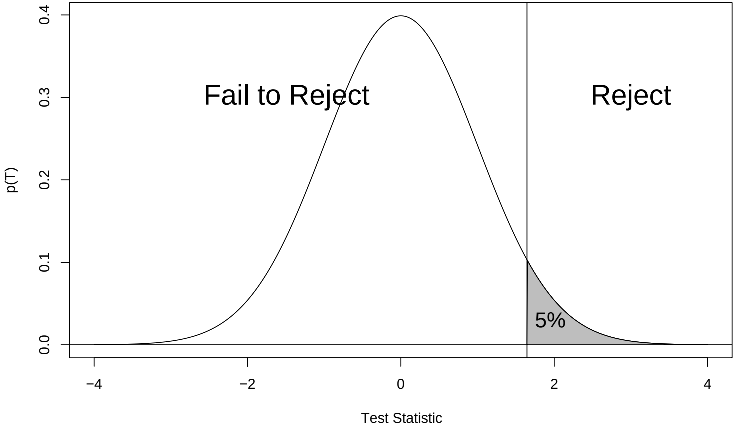
- $\overline{x} = 21.0$
- $\hat{\sigma} = 14.3$
- $n = 345$

Therefore,

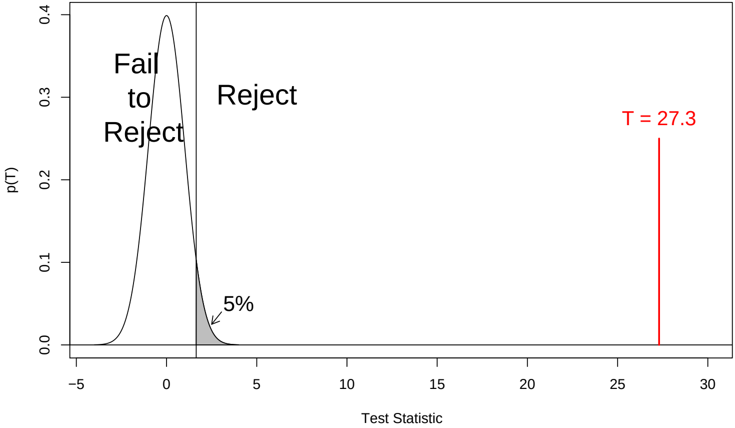
$$t_n = \frac{21.0 - 0}{\frac{14.3}{\sqrt{345}}} = 27.3$$

We construct our rejection region with $c = \texttt{qnorm(.95)} = 1.644$.

Rejection Region with $\alpha = .05$



Rejection Region with $\alpha = .05$



t-test

t-test

- This strategy works for any asymptotically normal estimator.

t-test

- This strategy works for any asymptotically normal estimator.
- A size-α t-test (also called a Wald test) rejects H_0 when $|T_n| > c$ where

$$T_n = \frac{\hat{\theta} - \theta_0}{\widehat{\text{SE}}[\hat{\theta}]}$$

t-test

- This strategy works for any asymptotically normal estimator.
- A size-α t-test (also called a Wald test) rejects H_0 when $|T_n| > c$ where

$$T_n = \frac{\hat{\theta} - \theta_0}{\widehat{\mathsf{SE}}[\hat{\theta}]}$$

- We get the critical value c by using the standard normal to get the probability such that $P_{H_0}(T_n \leq_c) = 1 - \alpha/2$

t-test

- This strategy works for any asymptotically normal estimator.
- A size-α t-test (also called a Wald test) rejects H_0 when $|T_n| > c$ where
$$T_n = \frac{\hat{\theta} - \theta_0}{\widehat{\text{SE}}[\hat{\theta}]}$$
- We get the critical value c by using the standard normal to get the probability such that $P_{H_0}(T_n \leq_c) = 1 - \alpha/2$
- This will guarantee the nominal probability of Type I error as n gets large.

Connections to Confidence Intervals

Connections to Confidence Intervals

You may have noticed that there is a 1:1 mapping between CIs and hypothesis tests (same math, different story).

Connections to Confidence Intervals

You may have noticed that there is a 1:1 mapping between CIs and hypothesis tests (same math, different story).

- If the confidence interval includes the null hypothesis, fail to reject your null.

Connections to Confidence Intervals

You may have noticed that there is a 1:1 mapping between CIs and hypothesis tests (same math, different story).

- If the confidence interval includes the null hypothesis, fail to reject your null.
- If the confidence interval does not include the null hypothesis, reject your null.

Connections to Confidence Intervals

You may have noticed that there is a 1:1 mapping between CIs and hypothesis tests (same math, different story).

- If the confidence interval includes the null hypothesis, fail to reject your null.
- If the confidence interval does not include the null hypothesis, reject your null.

A $100(1-\alpha)\%$ confidence interval represents all null hypotheses that we would not reject with a α level test.

Connections to Confidence Intervals

You may have noticed that there is a 1:1 mapping between CIs and hypothesis tests (same math, different story).

- If the confidence interval includes the null hypothesis, fail to reject your null.
- If the confidence interval does not include the null hypothesis, reject your null.

A $100(1-\alpha)\%$ confidence interval represents all null hypotheses that we would not reject with a α level test.

We can think of confidence intervals as a range of plausible values in the sense that we would not have rejected them had they been our null hypotheses.

Power

Power

We designed our tests to minimize Type I error (the false positive).

However we might also worry we are failing to detect important findings.

Power

We designed our tests to minimize Type I error (the false positive). However we might also worry we are failing to detect important findings.

Experiments and surveys are expensive and so we want to design our work to minimize the chance that we fail to reject the null at the end of the day.

Power

We designed our tests to minimize Type I error (the false positive). However we might also worry we are failing to detect important findings.

Experiments and surveys are expensive and so we want to design our work to minimize the chance that we fail to reject the null at the end of the day.

Definition (Power)

The power of a test is the probability that a tests rejects the null given some assumed population distribution $P_\theta(|T_n| > c)$.

$$\text{Power} = 1 - P(\text{Type II error})$$

Power

We designed our tests to minimize Type I error (the false positive). However we might also worry we are failing to detect important findings.

Experiments and surveys are expensive and so we want to design our work to minimize the chance that we fail to reject the null at the end of the day.

Definition (Power)

The power of a test is the probability that a tests rejects the null given some assumed population distribution $P_\theta(|T_n| > c)$.

$$\text{Power} = 1 - P(\text{Type II error})$$

If we fail to reject the null hypothesis, there are basically three possible states of the world:

Power

We designed our tests to minimize Type I error (the false positive).
However we might also worry we are failing to detect important findings.

Experiments and surveys are expensive and so we want to design our work to minimize the chance that we fail to reject the null at the end of the day.

Definition (Power)

The power of a test is the probability that a tests rejects the null given some assumed population distribution $P_\theta(|T_n| > c)$.

$$\text{Power} = 1 - P(\text{Type II error})$$

If we fail to reject the null hypothesis, there are basically three possible states of the world:

1. Null is true (no difference between the mailer populations).

Power

We designed our tests to minimize Type I error (the false positive). However we might also worry we are failing to detect important findings.

Experiments and surveys are expensive and so we want to design our work to minimize the chance that we fail to reject the null at the end of the day.

Definition (Power)

The power of a test is the probability that a tests rejects the null given some assumed population distribution $P_\theta(|T_n| > c)$.

$$\text{Power} = 1 - P(\text{Type II error})$$

If we fail to reject the null hypothesis, there are basically three possible states of the world:

1. Null is true (no difference between the mailer populations).
2. Null is false (there is a difference between the mailer populations), but test had low power.

Power

We designed our tests to minimize Type I error (the false positive). However we might also worry we are failing to detect important findings.

Experiments and surveys are expensive and so we want to design our work to minimize the chance that we fail to reject the null at the end of the day.

Definition (Power)

The power of a test is the probability that a tests rejects the null given some assumed population distribution $P_\theta(|T_n| > c)$.

$$\text{Power} = 1 - P(\text{Type II error})$$

If we fail to reject the null hypothesis, there are basically three possible states of the world:

1. Null is true (no difference between the mailer populations).
2. Null is false (there is a difference between the mailer populations), but test had low power.
3. Null is false, the test is well-powered and we got incredibly unlucky.

Power is Important

Power is Important

- Imagine that you are trying to prove that your approximate measurement strategy is good.

Power is Important

- Imagine that you are trying to prove that your approximate measurement strategy is good.
- You take a sample of 20 measurements where you can obtain the truth and do a hypothesis test.

Power is Important

- Imagine that you are trying to prove that your approximate measurement strategy is good.
- You take a sample of 20 measurements where you can obtain the truth and do a hypothesis test.
- You fail to reject the null. Is this good evidence that there are no differences and your measurement is perfect?

Power is Important

- Imagine that you are trying to prove that your approximate measurement strategy is good.
- You take a sample of 20 measurements where you can obtain the truth and do a hypothesis test.
- You fail to reject the null. Is this good evidence that there are no differences and your measurement is perfect?
- No! You haven't been able to confidently reject the possibility that there is no difference, but that's different than rejecting the possibility that there is a difference.

Power is Important

- Imagine that you are trying to prove that your approximate measurement strategy is good.
- You take a sample of 20 measurements where you can obtain the truth and do a hypothesis test.
- You fail to reject the null. Is this good evidence that there are no differences and your measurement is perfect?
- No! You haven't been able to confidently reject the possibility that there is no difference, but that's different than rejecting the possibility that there is a difference.
- This might seem obvious but conflating a lack of evidence with evidence for a zero effect is a problem that crops up all the time.

Power is Important

- Imagine that you are trying to prove that your approximate measurement strategy is good.
- You take a sample of 20 measurements where you can obtain the truth and do a hypothesis test.
- You fail to reject the null. Is this good evidence that there are no differences and your measurement is perfect?
- No! You haven't been able to confidently reject the possibility that there is no difference, but that's different than rejecting the possibility that there is a difference.
- This might seem obvious but conflating a lack of evidence with evidence for a zero effect is a problem that crops up all the time.
- Power analysis is a way of guiding the choice of sample size prior to an experiment to avoid this kind of mistake.

How Well Powered Are Social Science Studies?

How Well Powered Are Social Science Studies?

Quantitative Political Science Research is Greatly Underpowered[*]

Vincent Arel-Bundock[†] Ryan C. Briggs[‡]
Hristos Doucouliagos[§] Marco Mendoza Aviña[¶] T.D. Stanley[⊠]

July 05, 2022

Abstract

We analyze the statistical power of political science research by collating over 16,000 hypothesis tests from about 2,000 articles. Even with generous assumptions, the median analysis has about 10% power, and only about 1 in 10 tests have at least 80% power to detect the consensus effects reported in the literature. There is also substantial heterogeneity in tests across research areas, with some being characterized by high-power but most having very low power. To contextualize our findings, we survey political methodologists to assess their expectations about power levels. Most methodologists greatly overestimate the statistical power of political science research.

The median analysis has about 10% power ($\alpha = 0.05$), and only about 1 in 10 statistical tests have at least 80% power.

Steps for Power Analysis

Steps for Power Analysis

1) Specify the null hypothesis to be tested at significance level α

Steps for Power Analysis

1) Specify the null hypothesis to be tested at significance level α
2) Choose a true value for population parameters and derive the sampling distribution of test statistic

Steps for Power Analysis

1) Specify the null hypothesis to be tested at significance level α
2) Choose a true value for population parameters and derive the sampling distribution of test statistic
3) Calculate the probability of rejecting the null hypothesis under this sampling distribution.

Steps for Power Analysis

1) Specify the null hypothesis to be tested at significance level α
2) Choose a true value for population parameters and derive the sampling distribution of test statistic
3) Calculate the probability of rejecting the null hypothesis under this sampling distribution.
4) Find the smallest sample size such that this rejection probability equals a pre-specified power level.

Steps for Power Analysis

1) Specify the null hypothesis to be tested at significance level α
2) Choose a true value for population parameters and derive the sampling distribution of test statistic
3) Calculate the probability of rejecting the null hypothesis under this sampling distribution.
4) Find the smallest sample size such that this rejection probability equals a pre-specified power level.
5) Possibly repeat under different assumptions about the population.

Example: Power Analysis

Example: Power Analysis

TABLE 2. Effects of Four Mail Treatments on Voter Turnout in the August 2006 Primary Election

	Experimental Group				
	Control	Civic Duty	Hawthorne	Self	Neighbors
Percentage Voting	29.7%	31.5%	32.2%	34.5%	37.8%
N of Individuals	191,243	38,218	38,204	38,218	38,201

Example: Power Analysis

TABLE 2. Effects of Four Mail Treatments on Voter Turnout in the August 2006 Primary Election

	Experimental Group				
	Control	Civic Duty	Hawthorne	Self	Neighbors
Percentage Voting	29.7%	31.5%	32.2%	34.5%	37.8%
N of Individuals	191,243	38,218	38,204	38,218	38,201

- Why did they use 38 thousand people per group? Power analysis!

Example: Power Analysis

TABLE 2. Effects of Four Mail Treatments on Voter Turnout in the August 2006 Primary Election

	Experimental Group				
	Control	Civic Duty	Hawthorne	Self	Neighbors
Percentage Voting	29.7%	31.5%	32.2%	34.5%	37.8%
N of Individuals	191,243	38,218	38,204	38,218	38,201

- Why did they use 38 thousand people per group? Power analysis!
- If their results were exactly true, would we be expect to be able to confirm that the effect is non-zero with a replication using only 500 mailers?

Example: Power Analysis

TABLE 2. Effects of Four Mail Treatments on Voter Turnout in the August 2006 Primary Election

	Experimental Group				
	Control	Civic Duty	Hawthorne	Self	Neighbors
Percentage Voting	29.7%	31.5%	32.2%	34.5%	37.8%
N of Individuals	191,243	38,218	38,204	38,218	38,201

- Why did they use 38 thousand people per group? Power analysis!
- If their results were exactly true, would we be expect to be able to confirm that the effect is non-zero with a replication using only 500 mailers?
- For this example, let's ignore the household sampling (which is really important in practice, but we want to make this simple!).

What Do We Know?

TABLE 2. Effects of Four Mail Treatments on Voter Turnout in the August 2006 Primary Election

	Experimental Group				
	Control	Civic Duty	Hawthorne	Self	Neighbors
Percentage Voting	29.7%	31.5%	32.2%	34.5%	37.8%
N of Individuals	191,243	38,218	38,204	38,218	38,201

$$\mu_y = .378$$

What Do We Know?

TABLE 2. Effects of Four Mail Treatments on Voter Turnout in the August 2006 Primary Election

	Experimental Group				
	Control	Civic Duty	Hawthorne	Self	Neighbors
Percentage Voting	29.7%	31.5%	32.2%	34.5%	37.8%
N of Individuals	191,243	38,218	38,204	38,218	38,201

$$\begin{aligned}\mu_y &= .378\\ \sigma_y^2 &= (.378)(1-.378) = 0.235\end{aligned}$$

What Do We Know?

TABLE 2. Effects of Four Mail Treatments on Voter Turnout in the August 2006 Primary Election

	Experimental Group				
	Control	Civic Duty	Hawthorne	Self	Neighbors
Percentage Voting	29.7%	31.5%	32.2%	34.5%	37.8%
N of Individuals	191,243	38,218	38,204	38,218	38,201

$$\begin{aligned}
\mu_y &= .378 \\
\sigma_y^2 &= (.378)(1-.378) = 0.235 \\
\mu_x &= .315 \\
\sigma_x^2 &= (.315)(1-.315) = 0.216
\end{aligned}$$

What Do We Know?

TABLE 2. Effects of Four Mail Treatments on Voter Turnout in the August 2006 Primary Election

	Experimental Group				
	Control	Civic Duty	Hawthorne	Self	Neighbors
Percentage Voting	29.7%	31.5%	32.2%	34.5%	37.8%
N of Individuals	191,243	38,218	38,204	38,218	38,201

$$
\begin{aligned}
\mu_y &= .378 \\
\sigma_y^2 &= (.378)(1 - .378) = 0.235 \\
\mu_x &= .315 \\
\sigma_x^2 &= (.315)(1 - .315) = 0.216 \\
\theta &= \mu_y - \mu_x = .063 \\
\theta_0 &= 0
\end{aligned}
$$

What Do We Know?

TABLE 2. Effects of Four Mail Treatments on Voter Turnout in the August 2006 Primary Election

	Experimental Group				
	Control	Civic Duty	Hawthorne	Self	Neighbors
Percentage Voting	29.7%	31.5%	32.2%	34.5%	37.8%
N of Individuals	191,243	38,218	38,204	38,218	38,201

$$
\begin{aligned}
\mu_y &= .378 \\
\sigma_y^2 &= (.378)(1-.378) = 0.235 \\
\mu_x &= .315 \\
\sigma_x^2 &= (.315)(1-.315) = 0.216 \\
\theta &= \mu_y - \mu_x = .063 \\
\theta_0 &= 0 \\
V[\hat{\theta}_n] &= .235/250 + 0.216/250 = .001804
\end{aligned}
$$

What Do We Know?

TABLE 2. Effects of Four Mail Treatments on Voter Turnout in the August 2006 Primary Election

	Experimental Group				
	Control	Civic Duty	Hawthorne	Self	Neighbors
Percentage Voting	29.7%	31.5%	32.2%	34.5%	37.8%
N of Individuals	191,243	38,218	38,204	38,218	38,201

$$
\begin{aligned}
\mu_y &= .378 \\
\sigma_y^2 &= (.378)(1-.378) = 0.235 \\
\mu_x &= .315 \\
\sigma_x^2 &= (.315)(1-.315) = 0.216 \\
\theta &= \mu_y - \mu_x = .063 \\
\theta_0 &= 0 \\
V[\hat{\theta}_n] &= .235/250 + 0.216/250 = .001804 \\
\hat{\theta}_n &\xrightarrow{d} \mathcal{N}(0.63, 0.001804)
\end{aligned}
$$

Calculating Probability of Rejecting the Null

Calculating Probability of Rejecting the Null

- We reject when

$$|T| = \frac{|\hat{\theta}_n - 0|}{\hat{\mathrm{SE}}[\hat{\theta}_n]} > 1.96$$

Calculating Probability of Rejecting the Null

- We reject when

$$|T| = \frac{|\hat{\theta}_n - 0|}{\hat{\text{SE}}[\hat{\theta}_n]} > 1.96$$

- Rearranging we get rejection when

$$|\hat{\theta}_n| > 1.96\hat{\text{SE}}[\hat{\theta}_n]$$

Calculating Probability of Rejecting the Null

- We reject when

$$|T| = \frac{|\hat{\theta}_n - 0|}{\hat{\text{SE}}[\hat{\theta}_n]} > 1.96$$

- Rearranging we get rejection when

$$|\hat{\theta}_n| > 1.96\hat{\text{SE}}[\hat{\theta}_n]$$

- Under our assumption of the truth we know $V[\hat{\theta}_n] = .001804$ and thus:

$$\left\{\hat{\theta}_n < -1.96\sqrt{.001804}\right\} \cup \left\{\hat{\theta}_n > 1.96\sqrt{.001804}\right\}$$

Calculating Probability of Rejecting the Null

- We reject when
$$|T| = \frac{|\hat{\theta}_n - 0|}{\hat{\text{SE}}[\hat{\theta}_n]} > 1.96$$
- Rearranging we get rejection when
$$|\hat{\theta}_n| > 1.96\hat{\text{SE}}[\hat{\theta}_n]$$
- Under our assumption of the truth we know $V[\hat{\theta}_n] = .001804$ and thus:
$$\left\{\hat{\theta}_n < -1.96\sqrt{.001804}\right\} \cup \left\{\hat{\theta}_n > 1.96\sqrt{.001804}\right\}$$
- Using the sampling distribution we derived can calculate:
$$P\left(\hat{\theta}_n < -1.96\sqrt{.001804}\right) + P\left(\hat{\theta}_n > 1.96\sqrt{.001804}\right)$$

Getting an Answer in R

Getting an Answer in R

Let's calculate this using the fact that $\hat{\theta}_n \xrightarrow{d} \mathcal{N}(0.63, 0.001804)$

Getting an Answer in R

Let's calculate this using the fact that $\hat{\theta}_n \xrightarrow{d} \mathcal{N}(0.63, 0.001804)$

```
se <- sqrt(.001804)
pnorm(-1.96*se, mean=.05, sd=se) +
pnorm(1.96*se, mean=.05, sd=se, lower.tail=FALSE)
```

Getting an Answer in R

Let's calculate this using the fact that $\hat{\theta}_n \xrightarrow{d} \mathcal{N}(0.63, 0.001804)$

```
se <- sqrt(.001804)
pnorm(-1.96*se, mean=.05, sd=se) +
pnorm(1.96*se, mean=.05, sd=se, lower.tail=FALSE)
```

We get 0.2177 which tells us that if the true population means were those calculated in the experiment, we would be able to reject the null of no effect about 22% of the time using 500 mailers.

Getting an Answer in R

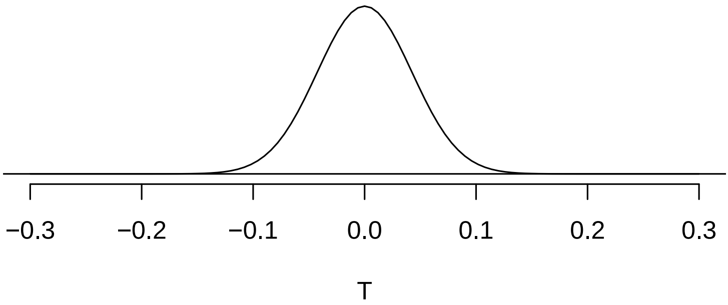
Let's calculate this using the fact that $\hat{\theta}_n \xrightarrow{d} \mathcal{N}(0.63, 0.001804)$

```
se <- sqrt(.001804)
pnorm(-1.96*se, mean=.05, sd=se) +
pnorm(1.96*se, mean=.05, sd=se, lower.tail=FALSE)
```

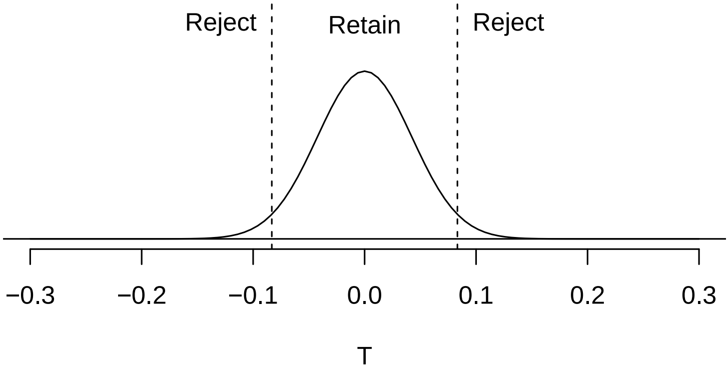
We get 0.2177 which tells us that if the true population means were those calculated in the experiment, we would be able to reject the null of no effect about 22% of the time using 500 mailers.

Yikes! That is not well powered.

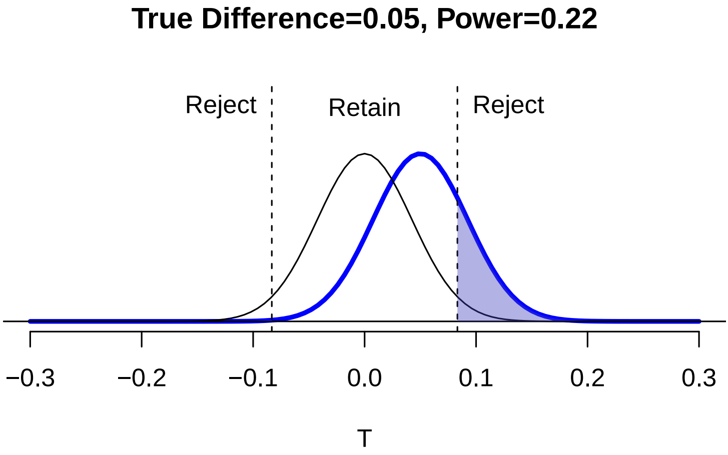
Power Graph



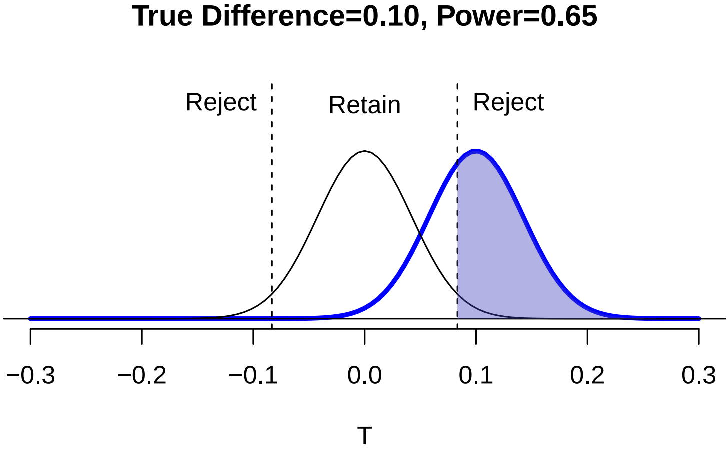
Power Graph



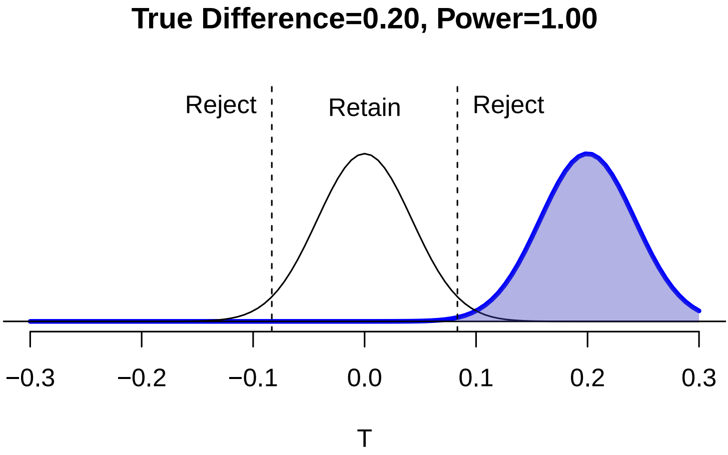
Power Graph


True Difference=0.05, Power=0.22
Reject
Retain
Reject
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
T

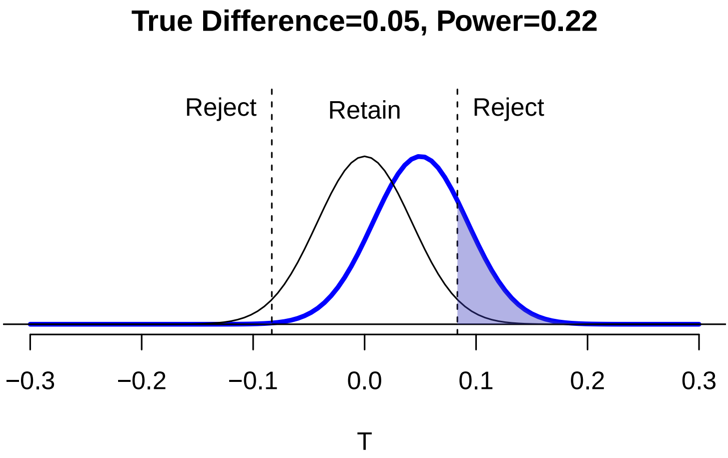
Power Graph


True Difference=0.10, Power=0.65
Reject
Retain
Reject
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
T

Power Graph

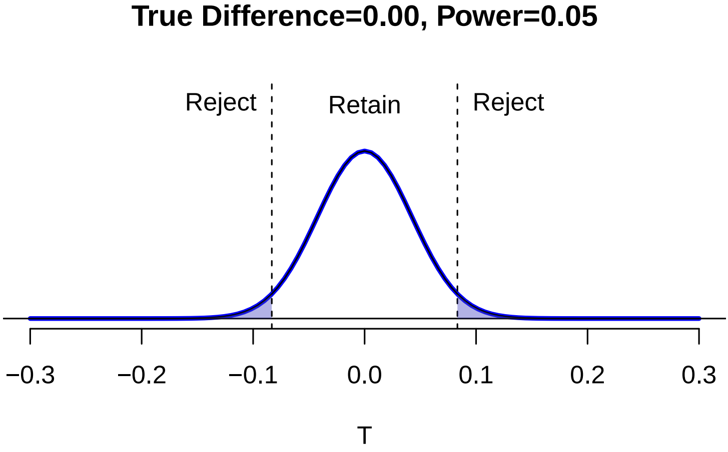

True Difference=0.20, Power=1.00
Reject
Retain
Reject
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
T

Power Graph

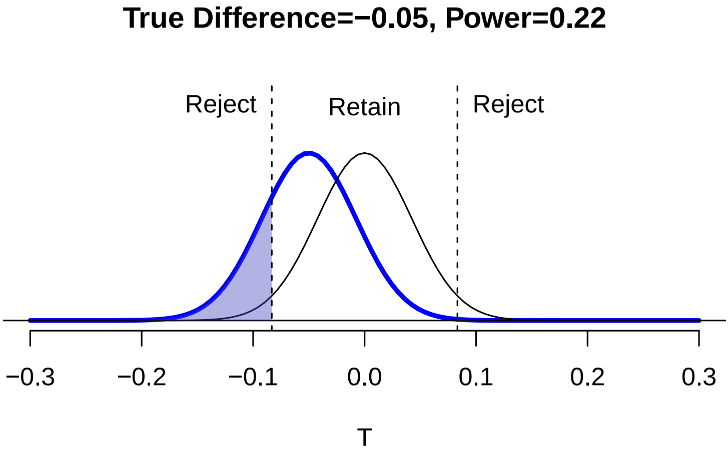


True Difference=0.05, Power=0.22
Reject
Retain
Reject
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
T

Power Graph

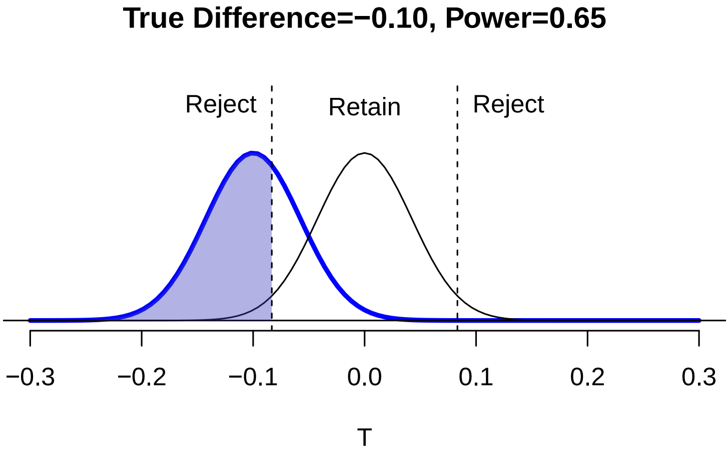

True Difference=0.00, Power=0.05
Reject
Retain
Reject
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
T

Power Graph



True Difference=−0.05, Power=0.22
Reject
Retain
Reject
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
T

Power Graph



True Difference=−0.10, Power=0.65
Reject
Retain
Reject
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
T

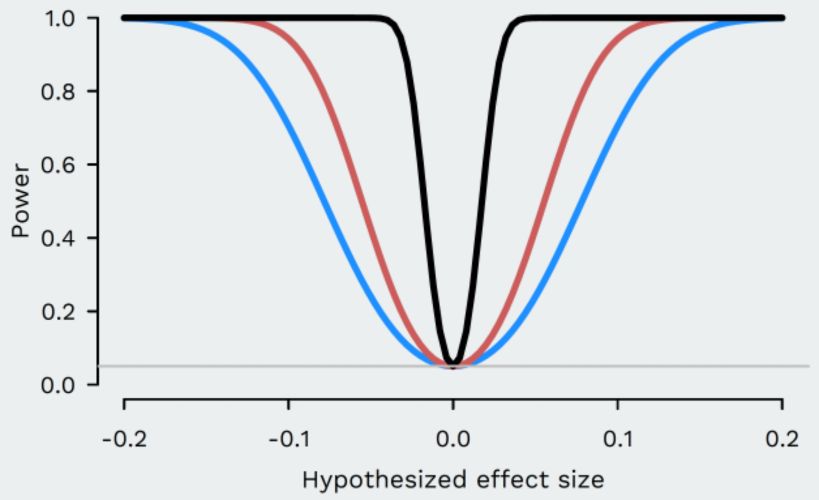
Wait! Does that mean I need to know the truth before I start to do power analysis?

Wait! Does that mean I need to know the truth before I start to do power analysis?

Yeah. . . kind of. In practice, this is why we calculate under many possible configurations.

Power Curve

You can graph power for various possible effect sizes (here for 500, 1000, and 10000 samples).



Power calculations

Power calculations

- Power calculations depend on your assumed difference and the population variance, neither of which is typically known.

Power calculations

- Power calculations depend on your assumed difference and the population variance, neither of which is typically known.
- Instead you may want to power depending on the minimal difference of interest with a conservative (high) estimate of population variance.

Power calculations

- Power calculations depend on your assumed difference and the population variance, neither of which is typically known.
- Instead you may want to power depending on the minimal difference of interest with a conservative (high) estimate of population variance.
- In general power is favorable when:
  - large n
  - bigger difference (pushes the alternative distribution away)
  - smaller variance (it squeezes the distribution in)

Power calculations

- Power calculations depend on your assumed difference and the population variance, neither of which is typically known.
- Instead you may want to power depending on the minimal difference of interest with a conservative (high) estimate of population variance.
- In general power is favorable when:
    - large n
    - bigger difference (pushes the alternative distribution away)
    - smaller variance (it squeezes the distribution in)
- Power analysis is really important if you are planning experiments, but we will touch on it only cursorily in this class. The Gerber and Green Field Experiments book is an amazing resource for more on experiments in general.

We Covered

- Hypothesis testing provides a principled framework for making decisions between alternatives.

We Covered

- Hypothesis testing provides a principled framework for making decisions between alternatives.
- The level of a test determines how often the researcher is willing to reject a correct null hypothesis.

We Covered

- Hypothesis testing provides a principled framework for making decisions between alternatives.
- The level of a test determines how often the researcher is willing to reject a correct null hypothesis.
- There is a close relationship between the results of an α level hypothesis test and the coverage of a $(1 - \alpha)\%$ confidence interval.

We Covered

- Hypothesis testing provides a principled framework for making decisions between alternatives.
- The level of a test determines how often the researcher is willing to reject a correct null hypothesis.
- There is a close relationship between the results of an α level hypothesis test and the coverage of a $(1 - \alpha)\%$ confidence interval.
- Power analysis is a way to assess your probability of missing a finding of interest.

We Covered

- Hypothesis testing provides a principled framework for making decisions between alternatives.
- The level of a test determines how often the researcher is willing to reject a correct null hypothesis.
- There is a close relationship between the results of an α level hypothesis test and the coverage of a $(1-\alpha)\%$ confidence interval.
- Power analysis is a way to assess your probability of missing a finding of interest.
- We will cover more on subtleties of interpretation next.

p-values

p-values

The appropriate level (α) for a hypothesis test depends on the relative costs of Type I and Type II errors.

p-values

The appropriate level (α) for a hypothesis test depends on the relative costs of Type I and Type II errors.

What if there is disagreement about these costs?

p-values

The appropriate level (α) for a hypothesis test depends on the relative costs of Type I and Type II errors.

What if there is disagreement about these costs?

We might like a quantity that summarizes the strength of evidence against the null hypothesis without making a yes or no decision.

p-values

The appropriate level (α) for a hypothesis test depends on the relative costs of Type I and Type II errors.

What if there is disagreement about these costs?

We might like a quantity that summarizes the strength of evidence against the null hypothesis without making a yes or no decision.

Definition (p-value)

The p-value is the smallest value α such that an α-level test would reject the null hypothesis.

p-values

The appropriate level (α) for a hypothesis test depends on the relative costs of Type I and Type II errors.

What if there is disagreement about these costs?

We might like a quantity that summarizes the strength of evidence against the null hypothesis without making a yes or no decision.

Definition (p-value)

The p-value is the smallest value α such that an α-level test would reject the null hypothesis.

Under the null hypothesis, this corresponds to the probability of observing a test statistic as extreme or more extreme than the one in the observed data (where extreme is defined in terms of the alternative hypothesis).

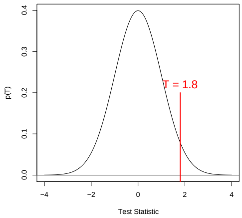
p-values

The p-value depends on both the realized value of the test statistic and the alternative hypothesis.

p-values

The p-value depends on both the realized value of the test statistic and the alternative hypothesis.

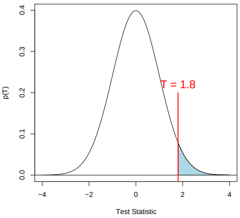
$$H_0 : \theta \leq 0$$
$$H_a : \theta > 0$$



p-values

The p-value depends on both the realized value of the test statistic and the alternative hypothesis.

$$H_0 : \theta \leq 0$$
$$H_a : \theta > 0$$

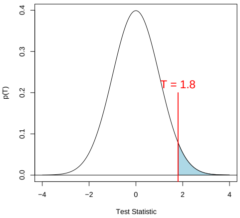


$$p = 0.036$$

p-values

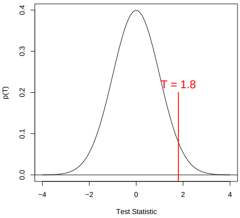
The p-value depends on both the realized value of the test statistic and the alternative hypothesis.

$$H_0 : \theta \leq 0$$
$$H_a : \theta > 0$$



$$p = 0.036$$

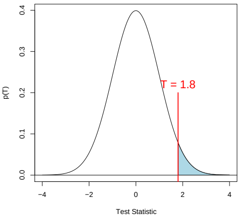
$$H_0 : \theta = 0$$
$$H_a : \theta \neq 0$$



p-values

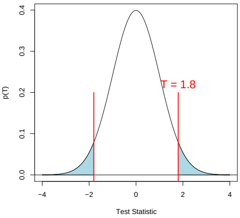
The p-value depends on both the realized value of the test statistic and the alternative hypothesis.

$$H_0 : \theta \leq 0$$
$$H_a : \theta > 0$$



$$p = 0.036$$

$$H_0 : \theta = 0$$
$$H_a : \theta \neq 0$$

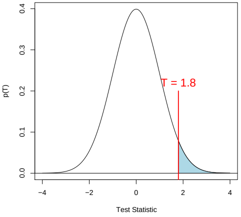


$$p = .072$$

p-values

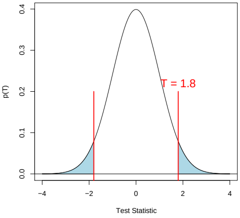
The p-value depends on both the realized value of the test statistic and the alternative hypothesis.

$$H_0 : \theta \leq 0$$
$$H_a : \theta > 0$$



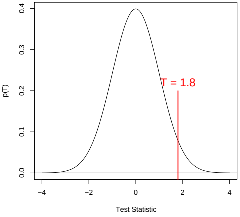
$$p = 0.036$$

$$H_0 : \theta = 0$$
$$H_a : \theta \neq 0$$



$$p = .072$$

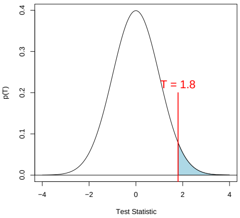
$$H_0 : \theta \geq 0$$
$$H_a : \theta < 0$$



p-values

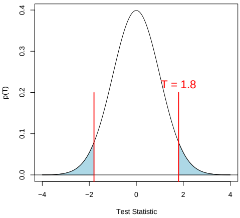
The p-value depends on both the realized value of the test statistic and the alternative hypothesis.

$$H_0 : \theta \leq 0$$
$$H_a : \theta > 0$$



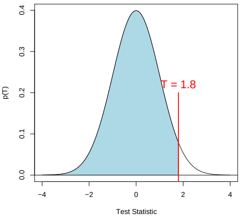
$p = 0.036$

$$H_0 : \theta = 0$$
$$H_a : \theta \neq 0$$



$p = .072$

$$H_0 : \theta \geq 0$$
$$H_a : \theta < 0$$



$p = 0.964$

Rejection Regions and p-values

Rejection Regions and p-values

- Recall for the social pressure experiment we got a test statistics of $t_{obs} = 18.5$.

Rejection Regions and p-values

- Recall for the social pressure experiment we got a test statistics of $t_{obs} = 18.5$.
- How likely was it we would get a test statistics this extreme or more extreme under our null hypothesis?

Rejection Regions and p-values

- Recall for the social pressure experiment we got a test statistics of $t_{obs} = 18.5$.
- How likely was it we would get a test statistics this extreme or more extreme under our null hypothesis?

$$\begin{aligned} P_0(|T_n| > 18.5) &= P_0(T_n > 18.5) + P_0(T_n < -18.5) \\ &= 2P_0(T_n < -18.5) \end{aligned}$$

Rejection Regions and p-values

- Recall for the social pressure experiment we got a test statistics of $t_{obs} = 18.5$.
- How likely was it we would get a test statistics this extreme or more extreme under our null hypothesis?

$$\begin{aligned} P_0(|T_n| > 18.5) &= P_0(T_n > 18.5) + P_0(T_n < -18.5) \\ &= 2P_0(T_n < -18.5) \end{aligned}$$

- We can get this in R with $2 * \texttt{pnorm}(-18.5)$

Rejection Regions and p-values

- Recall for the social pressure experiment we got a test statistics of $t_{obs} = 18.5$.
- How likely was it we would get a test statistics this extreme or more extreme under our null hypothesis?

$$\begin{aligned} P_0(|T_n| > 18.5) &= P_0(T_n > 18.5) + P_0(T_n < -18.5) \\ &= 2P_0(T_n < -18.5) \end{aligned}$$

- We can get this in R with $2 * \texttt{pnorm}(-18.5)$
- That yields a p-value of 2.06×10^{-76}.

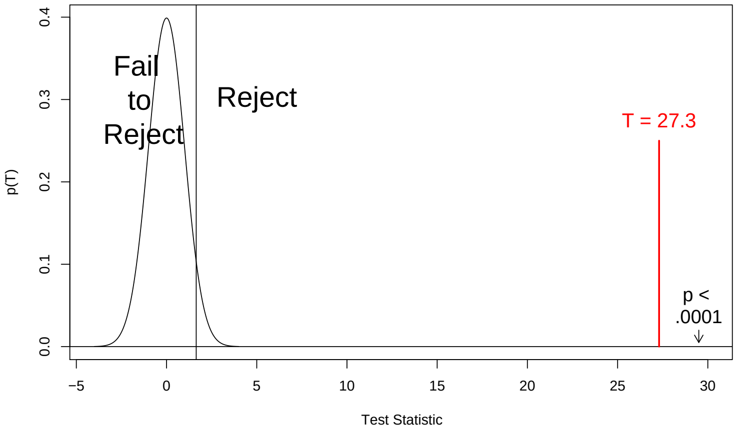
Rejection Regions and p-values

- Recall for the social pressure experiment we got a test statistics of $t_{obs} = 18.5$.
- How likely was it we would get a test statistics this extreme or more extreme under our null hypothesis?

$$\begin{aligned} P_0(|T_n| > 18.5) &= P_0(T_n > 18.5) + P_0(T_n < -18.5) \\ &= 2P_0(T_n < -18.5) \end{aligned}$$

- We can get this in R with $2 * \texttt{pnorm}(-18.5)$
- That yields a p-value of 2.06×10^{-76}.
- By convention we would say it is statistically significant at level α for some α that the p-value is below.

The Sowers et. al. Example

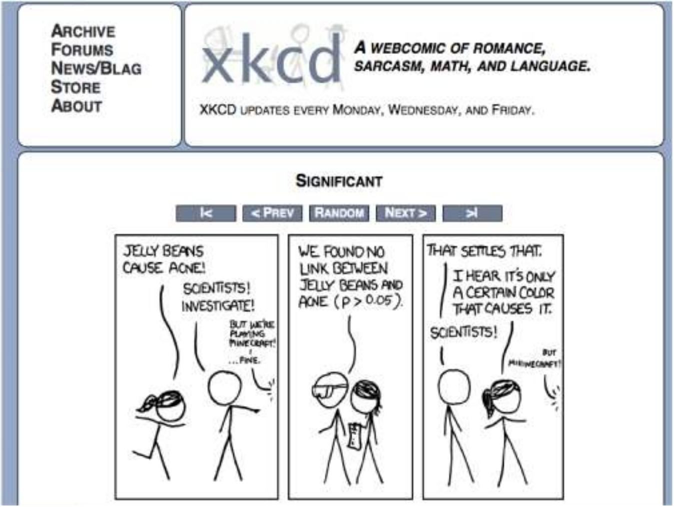


All those guarantees on Type I error?

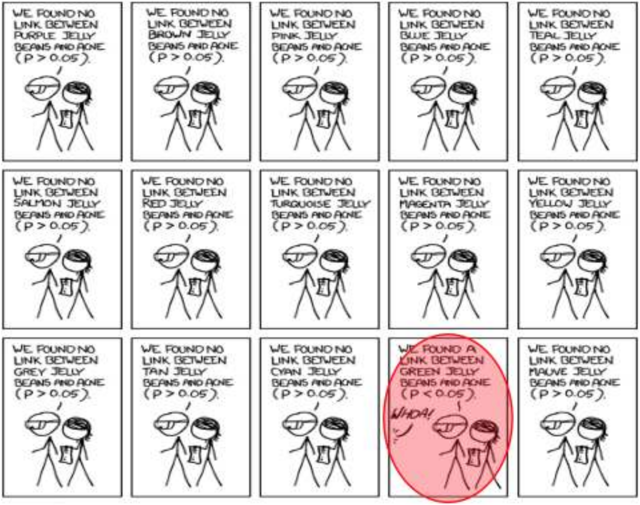
All those guarantees on Type I error?

It only works for one test.

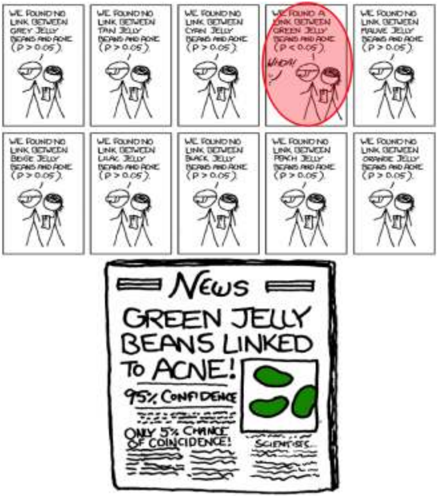
Star Chasing (aka there is an XKCD for everything)



Star Chasing (aka there is an XKCD for everything)



Star Chasing (aka there is an XKCD for everything)



Multiple Testing

- If we test all of the coefficients separately with a t-test, then we should expect that 5% of them will be significant just due to random chance.

Multiple Testing

- If we test all of the coefficients separately with a t-test, then we should expect that 5% of them will be significant just due to random chance.
- Illustration: randomly draw 21 variables, and run a regression of the first variable on the rest.

Multiple Testing

- If we test all of the coefficients separately with a t-test, then we should expect that 5% of them will be significant just due to random chance.
- Illustration: randomly draw 21 variables, and run a regression of the first variable on the rest.
- By design, no effect of any variable on any other, but when we run the regression:

Multiple Test Example

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.0280393  0.1138198  -0.246  0.80605
## X2          -0.1503904  0.1121808  -1.341  0.18389
## X3           0.0791578  0.0950278   0.833  0.40736
## X4          -0.0717419  0.1045788  -0.686  0.49472
## X5           0.1720783  0.1140017   1.509  0.13518
## X6           0.0808522  0.1083414   0.746  0.45772
## X7           0.1029129  0.1141562   0.902  0.37006
## X8          -0.3210531  0.1206727  -2.661  0.00945 **
## X9          -0.0531223  0.1079834  -0.492  0.62412
## X10          0.1801045  0.1264427   1.424  0.15827
## X11          0.1663864  0.1109471   1.500  0.13768
## X12          0.0080111  0.1037663   0.077  0.93866
## X13          0.0002117  0.1037845   0.002  0.99838
## X14         -0.0659690  0.1122145  -0.588  0.55829
## X15         -0.1296539  0.1115753  -1.162  0.24872
## X16         -0.0544456  0.1251395  -0.435  0.66469
## X17          0.0043351  0.1120122   0.039  0.96923
## X18         -0.0807963  0.1098525  -0.735  0.46421
## X19         -0.0858057  0.1185529  -0.724  0.47134
## X20         -0.1860057  0.1045602  -1.779  0.07910 .
## X21          0.0021111  0.1081179   0.020  0.98447
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9992 on 79 degrees of freedom
## Multiple R-squared:  0.2009, Adjusted R-squared:  -0.00142
## F-statistic: 0.993 on 20 and 79 DF,  p-value: 0.4797
```

Multiple Testing Gives False Positives

- Notice that out of 20 variables, one of the variables is significant at the 0.05 level (in fact, at the 0.01 level).

Multiple Testing Gives False Positives

- Notice that out of 20 variables, one of the variables is significant at the 0.05 level (in fact, at the 0.01 level).
- But this is exactly what we expect: $1/20 = 0.05$ of the tests are false positives at the 0.05 level

Multiple Testing Gives False Positives

- Notice that out of 20 variables, one of the variables is significant at the 0.05 level (in fact, at the 0.01 level).
- But this is exactly what we expect: $1/20 = 0.05$ of the tests are false positives at the 0.05 level
- Also note that $2/20 = 0.1$ are significant at the 0.1 level. Totally expected!

Multiple Testing Gives False Positives

- Notice that out of 20 variables, one of the variables is significant at the 0.05 level (in fact, at the 0.01 level).
- But this is exactly what we expect: $1/20 = 0.05$ of the tests are false positives at the 0.05 level
- Also note that $2/20 = 0.1$ are significant at the 0.1 level. Totally expected!
- The procedure by which tests/analyses are performed and shown to us matters a lot!

Problem of Multiple Testing

Problem of Multiple Testing

- The multiple testing (or "multiple comparison") problem occurs when one considers a set of statistical tests simultaneously.

Problem of Multiple Testing

- The multiple testing (or "multiple comparison") problem occurs when one considers a set of statistical tests simultaneously.
- Consider $k = 1, ..., m$ independent hypothesis tests (e.g. control versus various treatment groups). Even if each test is carried out at a low significance level (e.g., $\alpha = 0.05$) the overall type I error rate grows very fast: $\alpha_{overall} = 1 - (1 - \alpha_k)^m$.

Problem of Multiple Testing

- The multiple testing (or “multiple comparison”) problem occurs when one considers a set of statistical tests simultaneously.
- Consider $k = 1, ..., m$ independent hypothesis tests (e.g. control versus various treatment groups). Even if each test is carried out at a low significance level (e.g., $\alpha = 0.05$) the overall type I error rate grows very fast: $\alpha_{overall} = 1 - (1 - \alpha_k)^m$.
- That’s right - it grows exponentially. E.g., given test 7 tests at $\alpha = .1$ level the overall type I error is .52.

Problem of Multiple Testing

- The multiple testing (or "multiple comparison") problem occurs when one considers a set of statistical tests simultaneously.
- Consider $k = 1, ..., m$ independent hypothesis tests (e.g. control versus various treatment groups). Even if each test is carried out at a low significance level (e.g., $\alpha = 0.05$) the overall type I error rate grows very fast: $\alpha_{overall} = 1 - (1 - \alpha_k)^m$.
- That's right - it grows exponentially. E.g., given test 7 tests at $\alpha = .1$ level the overall type I error is .52.
- Even if all null hypotheses are true we will reject at least one of them with probability .52.

Problem of Multiple Testing

- The multiple testing (or "multiple comparison") problem occurs when one considers a set of statistical tests simultaneously.
- Consider $k = 1, ..., m$ independent hypothesis tests (e.g. control versus various treatment groups). Even if each test is carried out at a low significance level (e.g., $\alpha = 0.05$) the overall type I error rate grows very fast: $\alpha_{overall} = 1 - (1 - \alpha_k)^m$.
- That's right - it grows exponentially. E.g., given test 7 tests at $\alpha = .1$ level the overall type I error is .52.
- Even if all null hypotheses are true we will reject at least one of them with probability .52.
- Same for confidence intervals: probability that all 7 CI cover the true values simultaneously over repeated samples is .52.

Problem of Multiple Testing

- The multiple testing (or "multiple comparison") problem occurs when one considers a set of statistical tests simultaneously.
- Consider $k = 1, ..., m$ independent hypothesis tests (e.g. control versus various treatment groups). Even if each test is carried out at a low significance level (e.g., $\alpha = 0.05$) the overall type I error rate grows very fast: $\alpha_{overall} = 1 - (1 - \alpha_k)^m$.
- That's right - it grows exponentially. E.g., given test 7 tests at $\alpha = .1$ level the overall type I error is .52.
- Even if all null hypotheses are true we will reject at least one of them with probability .52.
- Same for confidence intervals: probability that all 7 CI cover the true values simultaneously over repeated samples is .52.
  So for each coefficient you have a .90 confidence interval, but overall a .52 percent confidence interval.

This all seems rather bad. Fixing this is an active research area for a lot of people.

This all seems rather bad. Fixing this is an active research area for a lot of people.

A Sketch of Two Styles of Solutions:
(1) statistical
and
(2) procedural.

One Step Deeper: Statistical Paths Forward

One Step Deeper: Statistical Paths Forward

- Control the Family-wise Error Rate:

$$P(\text{making at least one Type I error})$$

One Step Deeper: Statistical Paths Forward

- Control the Family-wise Error Rate:

$$P(\text{making at least one Type I error})$$

  - Bonferroni correction: reject the jth null hypothesis H_j if $p_j < \alpha/m$ where m is the total number of tests

One Step Deeper: Statistical Paths Forward

- Control the Family-wise Error Rate:

$$P(\text{making at least one Type I error})$$

  - Bonferroni correction: reject the jth null hypothesis H_j if $p_j < \alpha/m$ where m is the total number of tests
    - ★ NB: VERY conservative

One Step Deeper: Statistical Paths Forward

- Control the Family-wise Error Rate:

$$P(\text{making at least one Type I error})$$

  - Bonferroni correction: reject the jth null hypothesis H_j if $p_j < \alpha/m$ where m is the total number of tests
    - ★ NB: VERY conservative
- Control the False Discovery Rate

$$E\left[\frac{\#\text{ of false rejections}}{\max(\text{total }\#\text{ of rejections, } 1)}\right]$$

One Step Deeper: Statistical Paths Forward

- Control the Family-wise Error Rate:

$$P(\text{making at least one Type I error})$$

  - Bonferroni correction: reject the jth null hypothesis H_j if $p_j < \alpha/m$ where m is the total number of tests
    - ★ NB: VERY conservative

- Control the False Discovery Rate

$$E\left[\frac{\#\text{ of false rejections}}{\max(\text{total }\#\text{ of rejections, }1)}\right]$$

  - Benjamini-Hochberg Procedure: Order the p-values $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$

One Step Deeper: Statistical Paths Forward

- Control the Family-wise Error Rate:

$$P(\text{making at least one Type I error})$$

  - Bonferroni correction: reject the jth null hypothesis H_j if $p_j < \alpha/m$ where m is the total number of tests
    - ★ NB: VERY conservative

- Control the False Discovery Rate

$$E\left[\frac{\#\text{ of false rejections}}{\max(\text{total }\#\text{ of rejections, }1)}\right]$$

  - Benjamini-Hochberg Procedure: Order the p-values $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$
    - ★ Find the largest k such that $p_{(i)} \leq \frac{\alpha * k}{m}$ and call it k^*

One Step Deeper: Statistical Paths Forward

- Control the Family-wise Error Rate:

$$P(\text{making at least one Type I error})$$

  - Bonferroni correction: reject the jth null hypothesis H_j if $p_j < \alpha/m$ where m is the total number of tests
    - ★ NB: VERY conservative

- Control the False Discovery Rate

$$E\left[\frac{\#\text{ of false rejections}}{\max(\text{total }\#\text{ of rejections, }1)}\right]$$

  - Benjamini-Hochberg Procedure: Order the p-values $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$
    - ★ Find the largest k such that $p_{(i)} \leq \frac{\alpha * k}{m}$ and call it k^*
    - ★ Reject all H_k for $k \leq k^*$

One Step Deeper: Benjamini-Hochberg Example

Pvalues: 0.011, 0.13, 0.06, 0.54, 0.008, 0.024, 0.001, 0.201, 0.78, 0.023

Step 1: Sort	Step 2: Find New Threshold	Step 3: Find k	Step 4: Reject
0.001			
0.008			
0.011			
0.023			
0.024			
0.06			
0.13			
0.201			
0.54			
0.78			

One Step Deeper: Benjamini-Hochberg Example

Pvalues: 0.011, 0.13, 0.06, 0.54, 0.008, 0.024, 0.001, 0.201, 0.78, 0.023

Step 1: Sort	Step 2: Find New Threshold	Step 3: Find k	Step 4: Reject
0.001	0.05*1/10 = 0.005		
0.008	0.05*2/10 = 0.01		
0.011	0.05*3/10 = 0.015		
0.023	0.05*4/10 = 0.02		
0.024	0.05*5/10 = 0.025		
0.06	0.05*6/10 = 0.03		
0.13	0.05*7/10 = 0.035		
0.201	0.05*8/10 = 0.04		
0.54	0.05*9/10 = 0.045		
0.78	0.05*10/10 = 0.05		

One Step Deeper: Benjamini-Hochberg Example

Pvalues: 0.011, 0.13, 0.06, 0.54, 0.008, 0.024, 0.001, 0.201, 0.78, 0.023

Step 1: Sort	Step 2: Find New Threshold	Step 3: Find k	Step 4: Reject
0.001	0.05*1/10 = 0.005	*	
0.008	0.05*2/10 = 0.01	*	
0.011	0.05*3/10 = 0.015	*	
0.023	0.05*4/10 = 0.02		
0.024	0.05*5/10 = 0.025	*	
0.06	0.05*6/10 = 0.03		
0.13	0.05*7/10 = 0.035		
0.201	0.05*8/10 = 0.04		
0.54	0.05*9/10 = 0.045		
0.78	0.05*10/10 = 0.05		

One Step Deeper: Benjamini-Hochberg Example

Pvalues: 0.011, 0.13, 0.06, 0.54, 0.008, 0.024, 0.001, 0.201, 0.78, 0.023

Step 1: Sort	Step 2: Find New Threshold	Step 3: Find k	Step 4: Reject
0.001	0.05*1/10 = 0.005	*	Reject
0.008	0.05*2/10 = 0.01	*	Reject
0.011	0.05*3/10 = 0.015	*	Reject
0.023	0.05*4/10 = 0.02		Reject
0.024	0.05*5/10 = 0.025	*	Reject
0.06	0.05*6/10 = 0.03		
0.13	0.05*7/10 = 0.035		
0.201	0.05*8/10 = 0.04		
0.54	0.05*9/10 = 0.045		
0.78	0.05*10/10 = 0.05		

One Step Deeper: Procedural Paths Forward

One Step Deeper: Procedural Paths Forward

- Preregistration

One Step Deeper: Procedural Paths Forward

- Preregistration
  - in theory, forces people to pre-commit to analyses.

One Step Deeper: Procedural Paths Forward

- Preregistration
  - in theory, forces people to pre-commit to analyses.
  - doesn't directly address multiple comparisons, but may credibly limit them.

One Step Deeper: Procedural Paths Forward

- Preregistration
  - in theory, forces people to pre-commit to analyses.
  - doesn't directly address multiple comparisons, but may credibly limit them.
  - doesn't work if pre-analysis plans aren't abided by or if people register many, many hypotheses.

One Step Deeper: Procedural Paths Forward

- Preregistration
  - in theory, forces people to pre-commit to analyses.
  - doesn't directly address multiple comparisons, but may credibly limit them.
  - doesn't work if pre-analysis plans aren't abided by or if people register many, many hypotheses.
- Sample Splits

One Step Deeper: Procedural Paths Forward

- Preregistration
  - in theory, forces people to pre-commit to analyses.
  - doesn't directly address multiple comparisons, but may credibly limit them.
  - doesn't work if pre-analysis plans aren't abided by or if people register many, many hypotheses.
- Sample Splits
  - Set-aside one sample for discovery where you can search over lots of different options.

One Step Deeper: Procedural Paths Forward

- Preregistration
  - in theory, forces people to pre-commit to analyses.
  - doesn't directly address multiple comparisons, but may credibly limit them.
  - doesn't work if pre-analysis plans aren't abided by or if people register many, many hypotheses.
- Sample Splits
  - Set-aside one sample for discovery where you can search over lots of different options.
  - A second sample can be used to test a small set of hypotheses.

One Step Deeper: Procedural Paths Forward

- Preregistration
  - in theory, forces people to pre-commit to analyses.
  - doesn't directly address multiple comparisons, but may credibly limit them.
  - doesn't work if pre-analysis plans aren't abided by or if people register many, many hypotheses.
- Sample Splits
  - Set-aside one sample for discovery where you can search over lots of different options.
  - A second sample can be used to test a small set of hypotheses.
  - Similar to preregistration in that it doesn't directly address multiple comparisons, but limits them.

Fun With Salmon

Fun With Salmon

Bennett, Baird, Miller and Wolford. (2009). “Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: an argument for multiple comparisons correction.”

Methods

Methods

(a.k.a. the greatest methods section of all time)

Methods

(a.k.a. the greatest methods section of all time)

- <u>Subject</u>

Methods

(a.k.a. the greatest methods section of all time)

- Subject
  "One mature Atlantic Salmon (Salmo salar) participated in the fMRI study.

Methods

(a.k.a. the greatest methods section of all time)

- Subject
  "One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs,

Methods

(a.k.a. the greatest methods section of all time)

- Subject
  "One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning."

Methods

(a.k.a. the greatest methods section of all time)

- Subject
  "One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning."
- Task

Methods

(a.k.a. the greatest methods section of all time)

- Subject
  "One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning."
- Task
  "The task administered to the salmon involved completing an open-ended mentalizing task.

Methods

(a.k.a. the greatest methods section of all time)

- Subject
  "One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning."
- Task
  "The task administered to the salmon involved completing an open-ended mentalizing task. The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence.

Methods

(a.k.a. the greatest methods section of all time)

- Subject
  "One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning."
- Task
  "The task administered to the salmon involved completing an open-ended mentalizing task. The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence. The salmon was asked to determine what emotion the individual in the photo must have been experiencing."

Methods

(a.k.a. the greatest methods section of all time)

- Subject
  "One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning."
- Task
  "The task administered to the salmon involved completing an open-ended mentalizing task. The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence. The salmon was asked to determine what emotion the individual in the photo must have been experiencing."
- Design

Methods

(a.k.a. the greatest methods section of all time)

- Subject
  "One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning."
- Task
  "The task administered to the salmon involved completing an open-ended mentalizing task. The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence. The salmon was asked to determine what emotion the individual in the photo must have been experiencing."
- Design
  "Stimuli were presented in a block design with each photo presented for 10 seconds followed by 12 seconds of rest.

Methods

(a.k.a. the greatest methods section of all time)

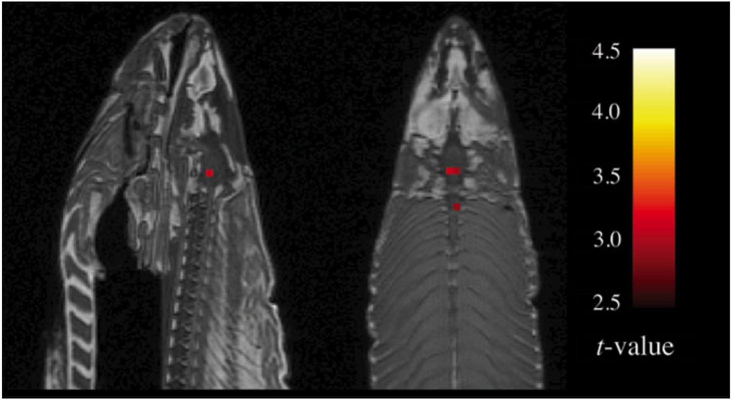
- Subject
  "One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning."
- Task
  "The task administered to the salmon involved completing an open-ended mentalizing task. The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence. The salmon was asked to determine what emotion the individual in the photo must have been experiencing."
- Design
  "Stimuli were presented in a block design with each photo presented for 10 seconds followed by 12 seconds of rest.A total of 15 photos were displayed.

Methods

(a.k.a. the greatest methods section of all time)

- Subject
  "One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning."
- Task
  "The task administered to the salmon involved completing an open-ended mentalizing task. The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence. The salmon was asked to determine what emotion the individual in the photo must have been experiencing."
- Design
  "Stimuli were presented in a block design with each photo presented for 10 seconds followed by 12 seconds of rest.A total of 15 photos were displayed. Total scan time was 5.5 minutes."

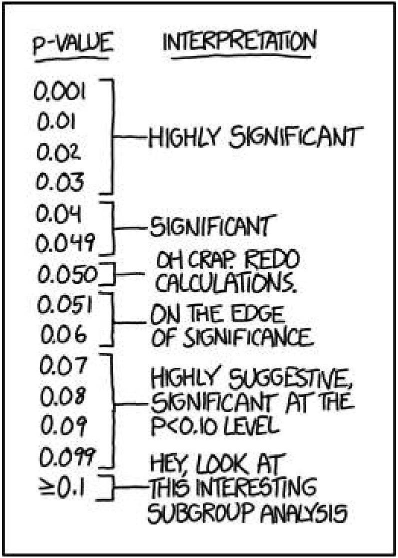
Results



"Several active voxels were discovered in a cluster located within the salmon's brain cavity. The size of this cluster was 81 mm^3 with a cluster-level significance of $p = .001$."

Okay, but what do they mean?

The Meaning of p-values (courtesy of XKCD)



The value of the p-value

Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis.

Ronald Fisher (1935)

The value of the p-value

Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis.

Ronald Fisher (1935)

In social science (and I think in psychology as well), the null hypothesis is almost certainly false, false, false, and you don't need a p-value to tell you this. The p-value tells you the extent to which a certain aspect of your data are consistent with the null hypothesis. A lack of rejection doesn't tell you that the null hypothesis is likely true; rather, it tells you that you don't have enough data to reject the null hypothesis.

Andrew Gelman (2010)

Practical versus Statistical Significance

$$T_n = \frac{\hat{\theta} - \theta_0}{\widehat{\text{SE}}[\hat{\theta}]} = \frac{\bar{X} - \mu_0}{\sqrt{\hat{\sigma}^2 / n}}$$

- What are the possible reasons for rejecting the null?

Practical versus Statistical Significance

$$T_n = \frac{\hat{\theta} - \theta_0}{\widehat{\text{SE}}[\hat{\theta}]} = \frac{\bar{X} - \mu_0}{\sqrt{\hat{\sigma}^2 / n}}$$

- What are the possible reasons for rejecting the null?
  1. $\overline{X} - \mu_0$ is large (big difference between sample mean and mean assumed by H_0)

Practical versus Statistical Significance

$$T_n = \frac{\hat{\theta} - \theta_0}{\widehat{\text{SE}}[\hat{\theta}]} = \frac{\bar{X} - \mu_0}{\sqrt{\hat{\sigma^2}/n}}$$

- What are the possible reasons for rejecting the null?
  1. $\overline{X} - \mu_0$ is large (big difference between sample mean and mean assumed by H_0)
  2. n is large (you have a lot of data so you have a lot of precision)

Practical versus Statistical Significance

$$T_n = \frac{\hat{\theta} - \theta_0}{\widehat{\mathrm{SE}}[\hat{\theta}]} = \frac{\bar{X} - \mu_0}{\sqrt{\hat{\sigma}^2/n}}$$

- What are the possible reasons for rejecting the null?
  1. $\overline{X} - \mu_0$ is large (big difference between sample mean and mean assumed by H_0)
  2. n is large (you have a lot of data so you have a lot of precision)
  3. $\hat{\sigma}^2$ is small (the outcome has low variability)

Practical versus Statistical Significance

$$T_n = \frac{\hat{\theta} - \theta_0}{\widehat{\text{SE}}[\hat{\theta}]} = \frac{\bar{X} - \mu_0}{\sqrt{\hat{\sigma}^2/n}}$$

- What are the possible reasons for rejecting the null?
  1. $\overline{X} - \mu_0$ is large (big difference between sample mean and mean assumed by H_0)
  2. n is large (you have a lot of data so you have a lot of precision)
  3. $\hat{\sigma}^2$ is small (the outcome has low variability)
- We need to be careful to distinguish:
  - practical significance (e.g. a big effect)
  - statistical significance (i.e. we reject the null)

Practical versus Statistical Significance

$$T_n = \frac{\hat{\theta} - \theta_0}{\widehat{\text{SE}}[\hat{\theta}]} = \frac{\bar{X} - \mu_0}{\sqrt{\hat{\sigma}^2 / n}}$$

- What are the possible reasons for rejecting the null?
  1. $\overline{X} - \mu_0$ is large (big difference between sample mean and mean assumed by H_0)
  2. n is large (you have a lot of data so you have a lot of precision)
  3. $\hat{\sigma}^2$ is small (the outcome has low variability)
- We need to be careful to distinguish:
  - practical significance (e.g. a big effect)
  - statistical significance (i.e. we reject the null)
- In large samples even tiny effects will be significant, but the results may not be very important substantively. Always discuss both!

Problems with p-values

Problems with p-values

- p-values are extremely common in the social sciences and are often the standard by which the value of the finding is judged.

Problems with p-values

- p-values are extremely common in the social sciences and are often the standard by which the value of the finding is judged.
- p-values do not:

Problems with p-values

- p-values are extremely common in the social sciences and are often the standard by which the value of the finding is judged.
- p-values do not:
  - measure the size or importance of a result.

Problems with p-values

- p-values are extremely common in the social sciences and are often the standard by which the value of the finding is judged.
- p-values do not:
  - measure the size or importance of a result.
  - measure the probability that the alternative hypothesis is false.

Problems with p-values

- p-values are extremely common in the social sciences and are often the standard by which the value of the finding is judged.
- p-values do not:
  - measure the size or importance of a result.
  - measure the probability that the alternative hypothesis is false.
  - measure the probability that the null hypothesis is true.

Problems with p-values

- p-values are extremely common in the social sciences and are often the standard by which the value of the finding is judged.
- p-values do not:
  - measure the size or importance of a result.
  - measure the probability that the alternative hypothesis is false.
  - measure the probability that the null hypothesis is true.
  - indicate how predictive a variable is of another.

Problems with p-values

- p-values are extremely common in the social sciences and are often the standard by which the value of the finding is judged.
- p-values do not:
  - measure the size or importance of a result.
  - measure the probability that the alternative hypothesis is false.
  - measure the probability that the null hypothesis is true.
  - indicate how predictive a variable is of another.
  - provide a good indication of whether a paper should be published.

Problems with p-values

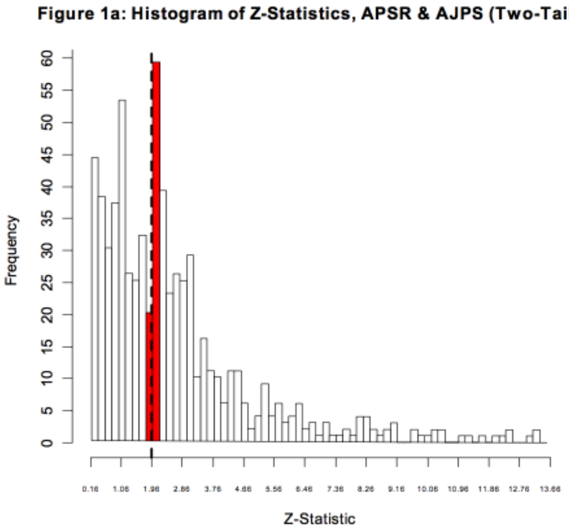
- p-values are extremely common in the social sciences and are often the standard by which the value of the finding is judged.
- p-values do not:
  - measure the size or importance of a result.
  - measure the probability that the alternative hypothesis is false.
  - measure the probability that the null hypothesis is true.
  - indicate how predictive a variable is of another.
  - provide a good indication of whether a paper should be published.
- a large p-value could mean either that we are in the null world OR that we had insufficient power.

Problems with p-values

- p-values are extremely common in the social sciences and are often the standard by which the value of the finding is judged.
- p-values do not:
  - measure the size or importance of a result.
  - measure the probability that the alternative hypothesis is false.
  - measure the probability that the null hypothesis is true.
  - indicate how predictive a variable is of another.
  - provide a good indication of whether a paper should be published.
- a large p-value could mean either that we are in the null world OR that we had insufficient power.

See also: The ASA's (American Statistical Association) Statement on p-Values: Context, Process, and Purpose (http://dx.doi.org/10.1080/00031305.2016.1154108)

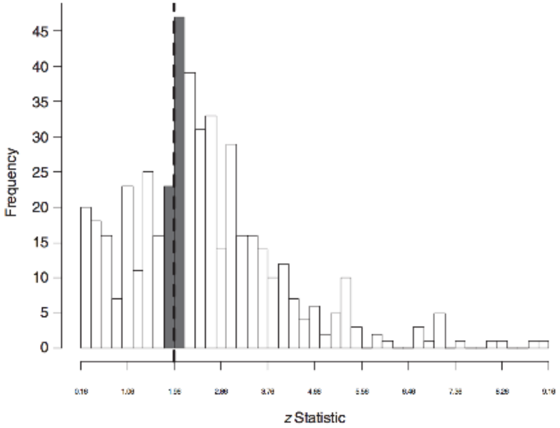
Arbitrary Publication Cutoffs



Gerber and Malhotra (2006) Top Political Science Journals

Arbitrary Publication Cutoffs

Figure 1
Histogram of *z* Statistics From the *American Sociological Review*, the *American Journal of Sociology*, and *The Sociological Quarterly* (Two-Tailed)



Gerber and Malhotra (2008) Top Sociology Journals

Arbitrary Publication Cutoffs

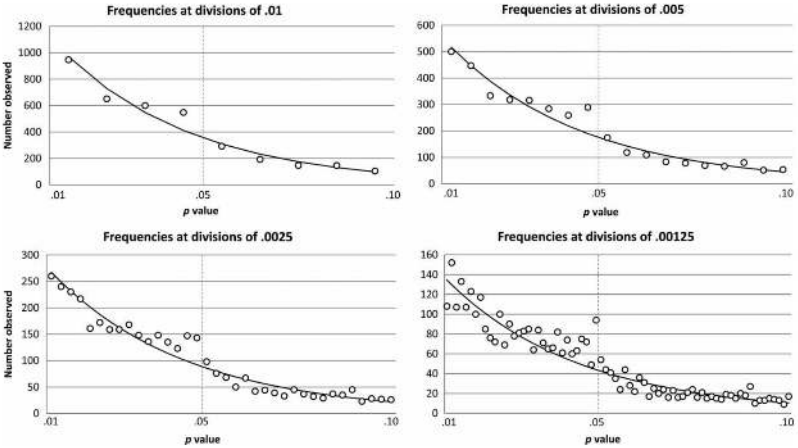


Figure 1.. *The graphs show the distribution of 3,627* p *values from three major psychology journals.*

Masicampo and Lalande (2012) Top Psychology Journals

Still Not Convinced?
The Real Harm of Misinterpreted p-values

ELSEVIER

Accident Analysis and Prevention 36 (2004) 495–500

ACCIDENT ANALYSIS & PREVENTION

www.elsevier.com/locate/aap

Viewpoint

The harm done by tests of significance

Ezra Hauer*

35 Merton Street, Apt. 1706, Toronto, Ont., Canada M4S 3G4

Abstract

Three historical episodes in which the application of null hypothesis significance testing (NHST) led to the mis-interpretation of data are described. It is argued that the pervasive use of this statistical ritual impedes the accumulation of knowledge and is unfit for use.

Keywords: Significance; Statistical hypothesis; Scientific method

Example from Hauer: Right-Turn-On-Red

Table 1
The Virginia RTOR study

	Before RTOR signing	After RTOR signing
Fatal crashes	0	0
Personal injury crashes	43	60
Persons injured	69	72
Property damage crashes	265	277
Property damage (US$)	161243	170807
Total crashes	308	337

The Point in Hauer

The Point in Hauer

- Two other interesting examples in Hauer (2004)

The Point in Hauer

- Two other interesting examples in Hauer (2004)
- Core issue is that lack of significance is not an indication of a zero effect, it could also be a lack of power (i.e. a small sample size relative to the difficulty of detecting the effect)

The Point in Hauer

- Two other interesting examples in Hauer (2004)
- Core issue is that lack of significance is not an indication of a zero effect, it could also be a lack of power (i.e. a small sample size relative to the difficulty of detecting the effect)
- On the opposite end, large tech companies rarely use significance testing because they have huge samples which essentially always find some non-zero effect. But that doesn't make the finding significant in a colloquial sense of important.

What if I need to show evidence of a zero effect?

AMERICAN JOURNAL of POLITICAL SCIENCE

An Equivalence Approach to Balance and Placebo Tests

Erin Hartman University of California Los Angeles
F. Daniel Hidalgo Massachusetts Institute of Technology

Abstract: *Recent emphasis on credible causal designs has led to the expectation that scholars justify their research designs by testing the plausibility of their causal identification assumptions, often through balance and placebo tests. Yet current practice is to use statistical tests with an inappropriate null hypothesis of no difference, which can result in equating nonsignificant differences with significant homogeneity. Instead, we argue that researchers should begin with the initial hypothesis that the data are inconsistent with a valid research design, and provide sufficient statistical evidence in favor of a valid design. When tests are correctly specified so that difference is the null and equivalence is the alternative, the problems afflicting traditional tests are alleviated. We argue that equivalence tests are better able to incorporate substantive considerations about what constitutes good balance on covariates and placebo outcomes than traditional tests. We demonstrate these advantages with applications to natural experiments.*

Replication Materials: The data, code, and any additional materials required to replicate all analyses in this article are available on the *American Journal of Political Science* Dataverse within the Harvard Dataverse Network, at: https://doi.org/10.7910/DVN/RYNSDG.

One Step Deeper: Equivalence Tests

Solution: Flip the hypotheses:

$$H_0 : \theta_T - \theta_C \leq \epsilon_L \text{ or } \theta_T - \theta_C \geq \epsilon_U$$
$$\text{versus}$$
$$H_A : \epsilon_L < \theta_T - \theta_C < \epsilon_U$$

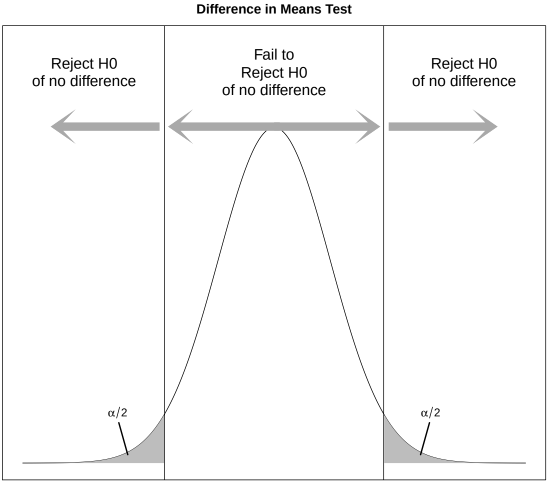
Tests of Difference vs. Equivalence Tests

$$H_0 : \frac{\mu_T - \mu_C}{\sigma} = 0 \qquad \text{versus} \qquad H_A : \frac{\mu_T - \mu_C}{\sigma} \neq 0$$

Type I Error

Test has α probability of declaring the two means different when they are, in fact, the same.

Problem: Controlling for the incorrect type of error if we're trying to provide evidence in favor of equivalence.



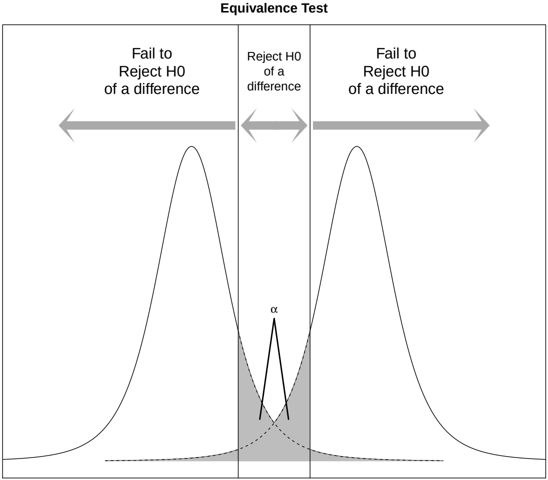
Tests of Difference vs. **Equivalence Tests**

$H_0 : \frac{\mu_T - \mu_C}{\sigma} \geq \epsilon_U$ or $\frac{\mu_T - \mu_C}{\sigma} \leq \epsilon_L$ versus $H_A : \epsilon_L < \frac{\mu_T - \mu_C}{\sigma} < \epsilon_U$

Type I Error

Test has α probability of declaring the two means equivalent when they are, in fact, the different.

Solution: Now control for the correct type of false positive.



p-values

p-values

General Message:

Don't misinterpret, or rely too heavily, on your p-values. They are evidence against your null, not evidence in favor of your alternative.

We Covered

We Covered

- hypothesis testing

We Covered

- hypothesis testing
- power

We Covered

- hypothesis testing
- power
- p-values

We Covered

- hypothesis testing
- power
- *p*-values
- multiple testing

We Covered

- hypothesis testing
- power
- p-values
- multiple testing
- the problems with p-values

We Covered

- hypothesis testing
- power
- p-values
- multiple testing
- the problems with p-values

Next Time: What is Regression?

We Covered

- hypothesis testing
- power
- p-values
- multiple testing
- the problems with p-values

Next Time: What is Regression?

Bonus reading for those interested to learn more:

- Hauer. 2004 "The harm done by tests of significance." *Accident Analysis & Prevention*.
- Gigerenzer. 2004. "Mindless statistics." *Journal of Socio-Economics*.
- Nuzzo. 2014. "Statistical Errors." *Nature*
- Ward et al. 2010. "The perils of policy by p-value: Predicting civil conflicts." *Journal of Peace Research*
- Cohen. 1994. "The Earth is Round ($p < 0.05$)." *American Psychologist*
- Schwab. 2011. "Researchers should make thoughtful assessments instead of null-hypothesis significance tests." *Organizational Science*.

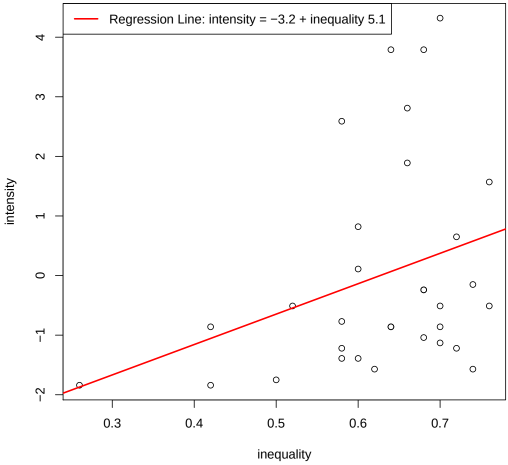
Where We've Been and Where We're Going...

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

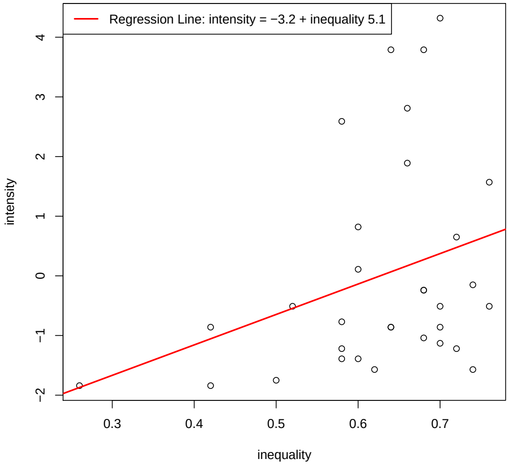
- Last Week
  - inference and estimator properties
  - point estimates, confidence intervals
- This Week
  - hypothesis testing
  - what is regression?
  - nonparametric and linear regression
- Next Week
  - inference for simple regression
  - properties of ordinary least squares
- Long Run
  - probability $\rightarrow$ inference $\rightarrow$ regression $\rightarrow$ causal inference

1. Hypothesis Testing
   - Terminology and Procedure
   - One-Sided Tests
   - Connections
   - Power
2. p-values
   - Mechanics
   - Multiple Testing
   - Fun With Salmon
   - The Significance of Significance
3. What is Regression?
   - Conditional Expectation Functions
   - Nonparametric Regression
   - Best Linear Predictor
   - Ordinary Least Squares
4. Interpreting Regression
   - Fun With Linearity

What You've Probably Seen This



What You've Probably Seen This



We are going to go about this a slightly different way.

What is a relationship and why do we care?

What is a relationship and why do we care?

- Most of what we want to do in the social science is learn about how two variables are related.

What is a relationship and why do we care?

- Most of what we want to do in the social science is learn about how two variables are related.
  - Does turnout vary by types of mailers received?
  - Does a the probability of getting a job vary by applicant race?
  - Is intergenerational mobility less likely for children with an incarcerated parent?

What is a relationship and why do we care?

- Most of what we want to do in the social science is learn about how two variables are related.
  - Does turnout vary by types of mailers received?
  - Does a the probability of getting a job vary by applicant race?
  - Is intergenerational mobility less likely for children with an incarcerated parent?
- By convention variables tend to be called X and Y.

What is a relationship and why do we care?

- Most of what we want to do in the social science is learn about how two variables are related.
  - Does turnout vary by types of mailers received?
  - Does a the probability of getting a job vary by applicant race?
  - Is intergenerational mobility less likely for children with an incarcerated parent?
- By convention variables tend to be called X and Y.
- Y - the dependent variable or outcome or regressand or left-hand-side variable or response
  - Voter turnout
  - Receiving a job
  - Income relative to parent

What is a relationship and why do we care?

- Most of what we want to do in the social science is learn about how two variables are related.
  - Does turnout vary by types of mailers received?
  - Does a the probability of getting a job vary by applicant race?
  - Is intergenerational mobility less likely for children with an incarcerated parent?
- By convention variables tend to be called X and Y.
- Y - the dependent variable or outcome or regressand or left-hand-side variable or response
  - Voter turnout
  - Receiving a job
  - Income relative to parent
- X - the independent variable or explanatory variable or regressor or right-hand-side variable or treatment or predictor
  - Social pressure mailer versus Civic Duty Mailer
  - Applicant race
  - Incarcerated parent

Characterizing the Joint Distribution

Characterizing the Joint Distribution

- A first step in this process is to characterize the joint distribution between the two random variables, $f_{X,Y}$, based on pairs of draws for the same unit (X_i, Y_i).

Characterizing the Joint Distribution

- A first step in this process is to characterize the joint distribution between the two random variables, $f_{X,Y}$, based on pairs of draws for the same unit (X_i, Y_i).
- Generally we are trying to characterize some properties of the conditional distribution $f_{Y|X}$ and often we will use the conditional expectation function, $\mu(x) = E[Y|X = x]$, as a summary.

Characterizing the Joint Distribution

- A first step in this process is to characterize the joint distribution between the two random variables, $f_{X,Y}$, based on pairs of draws for the same unit (X_i, Y_i).
- Generally we are trying to characterize some properties of the conditional distribution $f_{Y|X}$ and often we will use the conditional expectation function, $\mu(x) = E[Y|X = x]$, as a summary.
- We can use the conditional expectation function to help us perform important social science tasks:
    - Description: what is average value of Y among people with $X = x$ in the population.

Characterizing the Joint Distribution

- A first step in this process is to characterize the joint distribution between the two random variables, $f_{X,Y}$, based on pairs of draws for the same unit (X_i, Y_i).
- Generally we are trying to characterize some properties of the conditional distribution $f_{Y|X}$ and often we will use the conditional expectation function, $\mu(x) = E[Y|X = x]$, as a summary.
- We can use the conditional expectation function to help us perform important social science tasks:
  - Description: what is average value of Y among people with $X = x$ in the population.
  - Prediction: for a random sample of the population and given a value of x, $\mu(x)$ is the best predictor in terms of squared error.

Characterizing the Joint Distribution

- A first step in this process is to characterize the joint distribution between the two random variables, $f_{X,Y}$, based on pairs of draws for the same unit (X_i, Y_i).
- Generally we are trying to characterize some properties of the conditional distribution $f_{Y|X}$ and often we will use the conditional expectation function, $\mu(x) = E[Y|X = x]$, as a summary.
- We can use the conditional expectation function to help us perform important social science tasks:
  - Description: what is average value of Y among people with $X = x$ in the population.
  - Prediction: for a random sample of the population and given a value of x, $\mu(x)$ is the best predictor in terms of squared error.
  - Causal Inference: with additional assumptions (later in the semester) we can talk about how intervening to change the value of X will change Y.

Reminder of Definitions

Definition (Conditional Expectation (Discrete))

Let Y and X be discrete random variables. The conditional expectation of Y given $X = x$ is defined as:

$$E[Y|X = x] = \sum_y y\, P(Y = y|X = x) = \sum_y y\, p_{Y|X}(y|x)$$

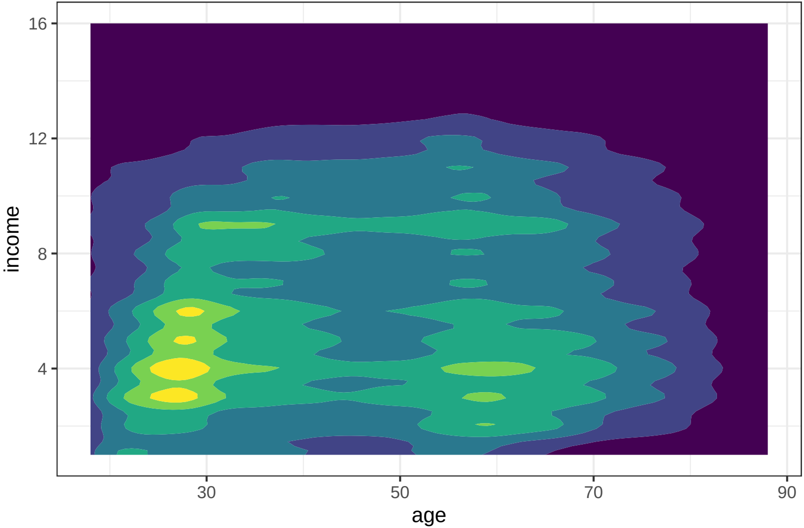
Definition (Conditional Expectation (Continuous))

Let Y and X be continuous random variables. The conditional expectation of Y given $X = x$ is given by:

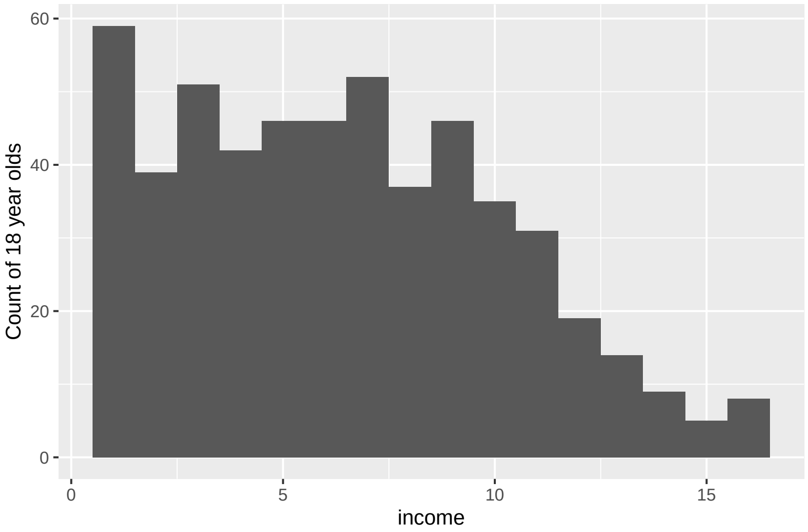
$$E[Y|X = x] = \int_{-\infty}^{\infty} y\, f_{Y|X}(y|x) dy$$

Review of CEF Concepts

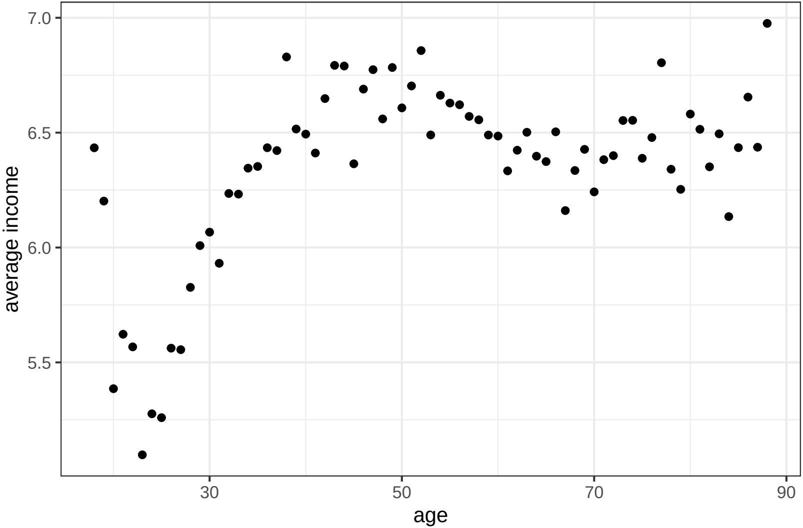
Review of CEF Concepts



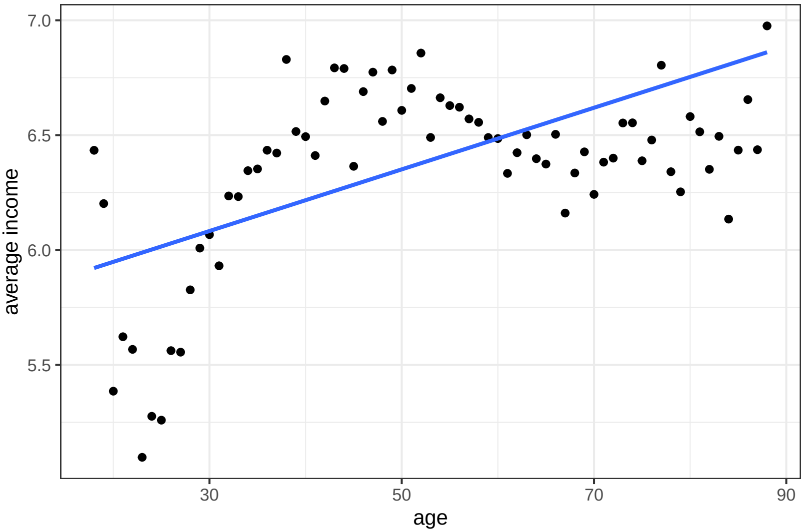
Review of CEF Concepts



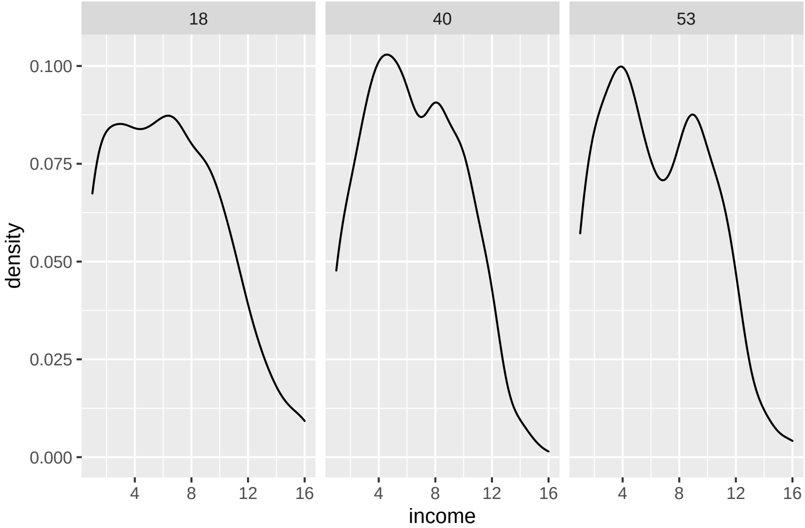
Review of CEF Concepts



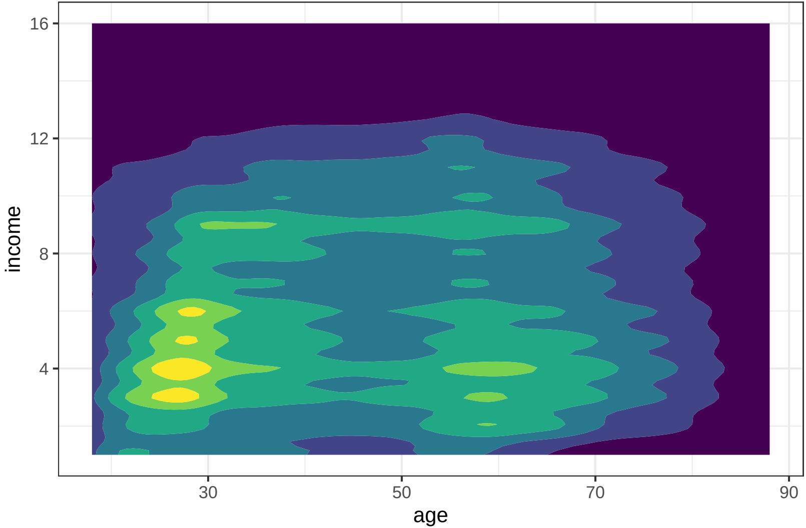
Review of CEF Concepts



Review of CEF Concepts



Review of CEF Concepts



Implications of the CEF Definition

Implications of the CEF Definition

Let X and Y be random variables and decompose Y_i into the conditional expectation function plus error such that,

$$Y_i = E[Y_i|X_i] + \epsilon_i$$

Implications of the CEF Definition

Let X and Y be random variables and decompose Y_i into the conditional expectation function plus error such that,

$$Y_i = E[Y_i|X_i] + \epsilon_i$$

- Then, the expectation of the error doesn't depend on X_i

$$E[\epsilon_i|X_i] = E[\epsilon_i] = 0$$

Implications of the CEF Definition

Let X and Y be random variables and decompose Y_i into the conditional expectation function plus error such that,

$$Y_i = E[Y_i|X_i] + \epsilon_i$$

- Then, the expectation of the error doesn't depend on X_i

$$E[\epsilon_i|X_i] = E[\epsilon_i] = 0$$

- the error is uncorrelated with any function of X_i

$$Cov[g(X_i), \epsilon_i] = 0$$

Implications of the CEF Definition

Let X and Y be random variables and decompose Y_i into the conditional expectation function plus error such that,

$$Y_i = E[Y_i|X_i] + \epsilon_i$$

- Then, the expectation of the error doesn't depend on X_i

$$E[\epsilon_i|X_i] = E[\epsilon_i] = 0$$

- the error is uncorrelated with any function of X_i

$$Cov[g(X_i), \epsilon_i] = 0$$

- the variance of the outcome given X is the variance of the error given X.

$$V[\epsilon_i|X_i] = V[Y_i|X_i]$$

Implications of the CEF Definition

Let X and Y be random variables and decompose Y_i into the conditional expectation function plus error such that,

$$Y_i = E[Y_i|X_i] + \epsilon_i$$

- Then, the expectation of the error doesn't depend on X_i

$$E[\epsilon_i|X_i] = E[\epsilon_i] = 0$$

- the error is uncorrelated with any function of X_i

$$Cov[g(X_i), \epsilon_i] = 0$$

- the variance of the outcome given X is the variance of the error given X.

$$V[\epsilon_i|X_i] = V[Y_i|X_i]$$

- and the CEF is the lowest mean squared error predictor of Y_i given X_i

CEF for binary covariates

CEF for binary covariates

- We've been writing μ_1 and μ_0 for the means in different groups.

CEF for binary covariates

- We've been writing μ_1 and μ_0 for the means in different groups.
- For example, on the problem set, we looked at the expected value of the loan amount given income group.

CEF for binary covariates

- We've been writing μ_1 and μ_0 for the means in different groups.
- For example, on the problem set, we looked at the expected value of the loan amount given income group.
- Note that these are just conditional expectations. Define Y to be the loan amount, $X = 1$ to indicate the high income group and $X = 0$ to indicate the low income group:

$$\mu_1 = E[Y|X = 1]$$
$$\mu_0 = E[Y|X = 0]$$

CEF for binary covariates

- We've been writing μ_1 and μ_0 for the means in different groups.
- For example, on the problem set, we looked at the expected value of the loan amount given income group.
- Note that these are just conditional expectations. Define Y to be the loan amount, $X = 1$ to indicate the high income group and $X = 0$ to indicate the low income group:

$$\mu_1 = E[Y|X = 1]$$
$$\mu_0 = E[Y|X = 0]$$

- Notice here that since X can only take on two values, 0 and 1, then these two conditional means completely summarize the CEF.

CEF for binary covariates

- We've been writing μ_1 and μ_0 for the means in different groups.
- For example, on the problem set, we looked at the expected value of the loan amount given income group.
- Note that these are just conditional expectations. Define Y to be the loan amount, $X = 1$ to indicate the high income group and $X = 0$ to indicate the low income group:

$$\mu_1 = E[Y|X = 1]$$
$$\mu_0 = E[Y|X = 0]$$

- Notice here that since X can only take on two values, 0 and 1, then these two conditional means completely summarize the CEF.
- Estimation just involves taking the means within the groups.

Non-binary CEFs

- If X is discrete with not too many categories, we can estimate the conditional expectation using the means within groups:
  - $\hat{\mu}(1) = \hat{E}[Y|X = 1] = \frac{1}{n_{x=1}} \sum_{i:X_i=1} Y_i$
  - $\hat{\mu}(2) = \hat{E}[Y|X = 2] = \frac{1}{n_{x=2}} \sum_{i:X_i=2} Y_i$
  - ...

Non-binary CEFs

- If X is discrete with not too many categories, we can estimate the conditional expectation using the means within groups:
  - $\hat{\mu}(1) = \hat{E}[Y|X = 1] = \frac{1}{n_{x=1}} \sum_{i:X_i=1} Y_i$
  - $\hat{\mu}(2) = \hat{E}[Y|X = 2] = \frac{1}{n_{x=2}} \sum_{i:X_i=2} Y_i$
  - ...
- This kind of estimation is nonparametric in the sense that it makes no assumptions about the specific functional form of $\mu(x)$.

Non-binary CEFs

- If X is discrete with not too many categories, we can estimate the conditional expectation using the means within groups:
  - $\hat{\mu}(1) = \hat{E}[Y|X = 1] = \frac{1}{n_{x=1}} \sum_{i:X_i=1} Y_i$
  - $\hat{\mu}(2) = \hat{E}[Y|X = 2] = \frac{1}{n_{x=2}} \sum_{i:X_i=2} Y_i$
  - ...
- This kind of estimation is nonparametric in the sense that it makes no assumptions about the specific functional form of $\mu(x)$.
- When X can take on many possible values (think income) or we have few observation for a given value of X, we have to write out a more general function.

Non-binary CEFs

- If X is discrete with not too many categories, we can estimate the conditional expectation using the means within groups:
  - $\hat{\mu}(1) = \hat{E}[Y|X=1] = \frac{1}{n_{x=1}} \sum_{i:X_i=1} Y_i$
  - $\hat{\mu}(2) = \hat{E}[Y|X=2] = \frac{1}{n_{x=2}} \sum_{i:X_i=2} Y_i$
  - $\ldots$
- This kind of estimation is nonparametric in the sense that it makes no assumptions about the specific functional form of $\mu(x)$.
- When X can take on many possible values (think income) or we have few observation for a given value of X, we have to write out a more general function.
- These functional forms are unknown which makes life hard.

Nonparametric Regression with Discrete X

Nonparametric Regression with Discrete X

- Let's take a look at some data on education and income from the American National Election Study

Nonparametric Regression with Discrete X

- Let's take a look at some data on education and income from the American National Election Study
- We use two variables:
  - Y: income
  - X: educational attainment

Nonparametric Regression with Discrete X

- Let's take a look at some data on education and income from the American National Election Study
- We use two variables:
  - Y: income
  - X: educational attainment
- Goal is to characterize the conditional expectation $E[Y|X = x]$, i.e. how average income varies with education level

Nonparametric Regression with Discrete X

Nonparametric Regression with Discrete X

`educ`: Respondent's education:

- 1. 8 grades or less and no diploma or
- 2. 9-11 grades
- 3. High school diploma or equivalency test
- 4. More than 12 years of schooling, no higher degree
- 5. Junior or community college level degree (AA degrees)
- 6. BA level degrees; 17+ years, no postgraduate degree
- 7. Advanced degree

Nonparametric Regression with Discrete X

`income`: Respondent's family income:

- 1. None or less than \$2,999
- 2. \$3,000-\$4,999
- 3. \$5,000-\$6,999
- 4. \$7,000-\$8,999
- 5. \$9,000-\$9,999
- 6. \$10,000-\$10,999

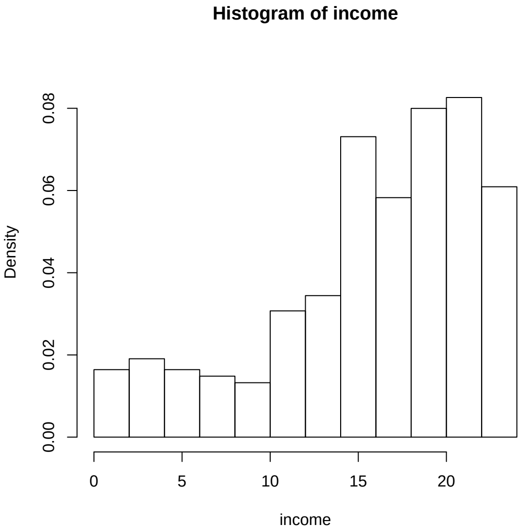
$\vdots$

- 17. \$35,000-\$39,999
- 18. \$40,000-\$44,999

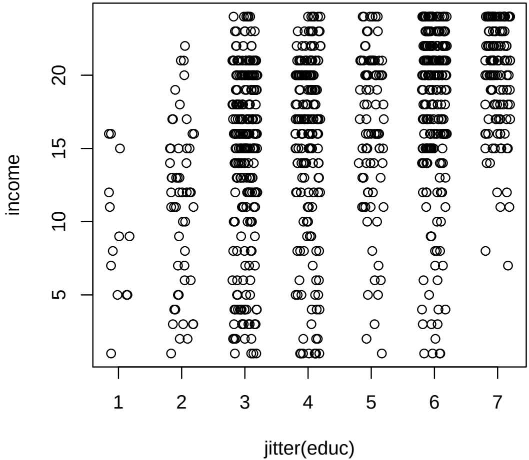
$\vdots$

- 23. \$90,000-\$104,999
- 24. \$105,000 and over

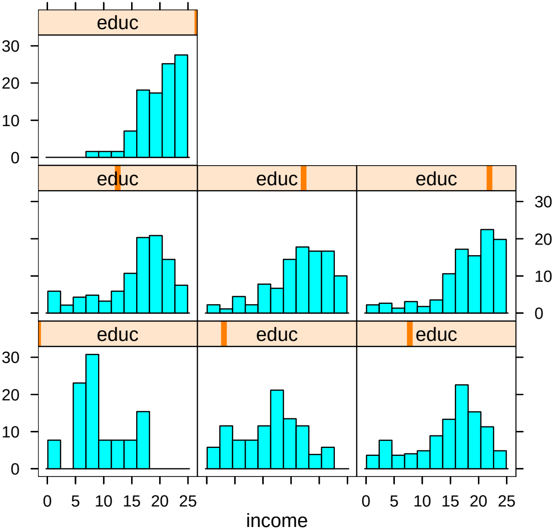
Marginal Distribution of Y (income)



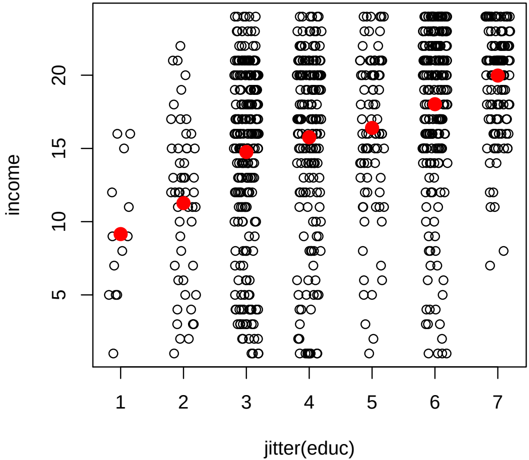
Joint Distribution of X and Y (Income and Education)



Distribution of income given education $p(y|x)$



Nonparametric Regression with Discrete X



Nonparametric Regression

Nonparametric Regression

- This approach works well as long as

Nonparametric Regression

- This approach works well as long as
  - X is discrete

Nonparametric Regression

- This approach works well as long as
  - X is discrete
  - there are a small number values of X

Nonparametric Regression

- This approach works well as long as
  - X is discrete
  - there are a small number values of X
  - a small number of X variables

Nonparametric Regression

- This approach works well as long as
  - X is discrete
  - there are a small number values of X
  - a small number of X variables
  - a lot of observations at each X value

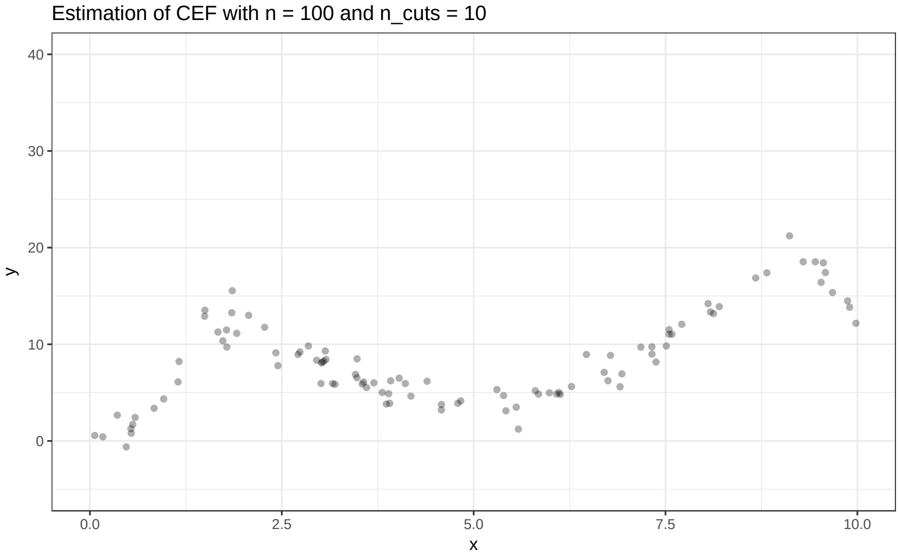
Nonparametric Regression

- This approach works well as long as
  - X is discrete
  - there are a small number values of X
  - a small number of X variables
  - a lot of observations at each X value
- But what do we do when X is continuous and has many values?

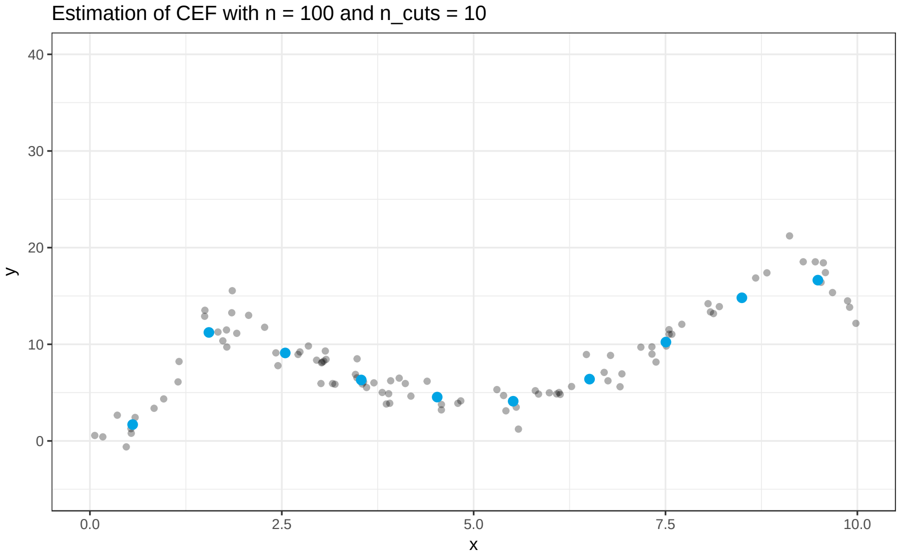
Nonparametric Regression

- This approach works well as long as
  - X is discrete
  - there are a small number values of X
  - a small number of X variables
  - a lot of observations at each X value
- But what do we do when X is continuous and has many values?
- Let's talk through a few options.

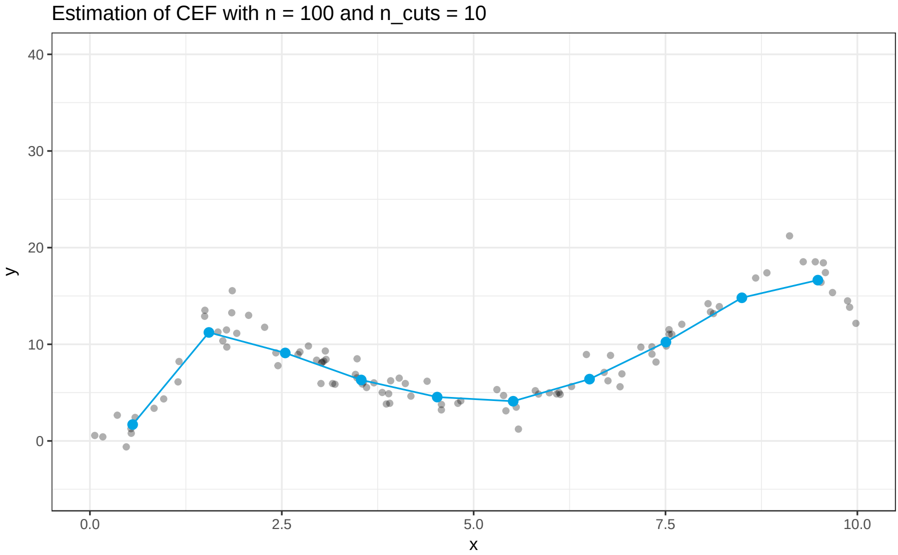
A Binning Approach to CEF for continuous random variables



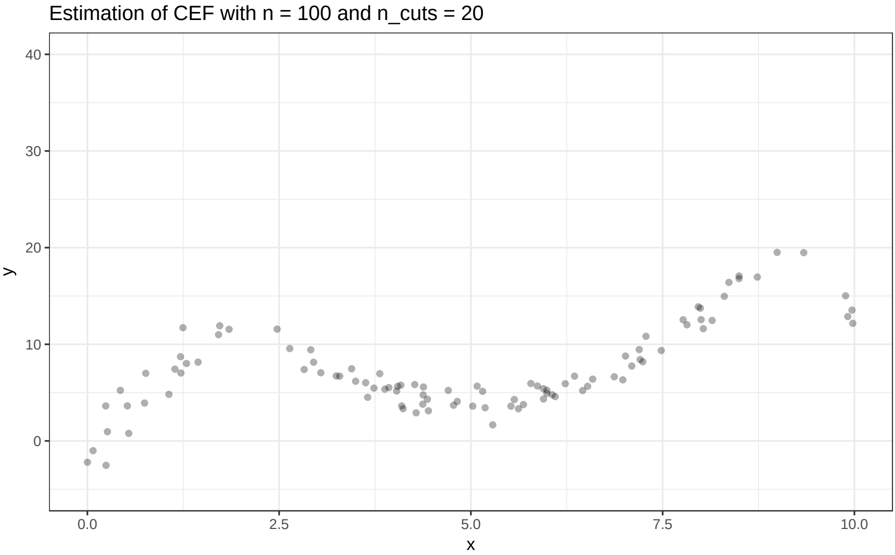
A Binning Approach to CEF for continuous random variables



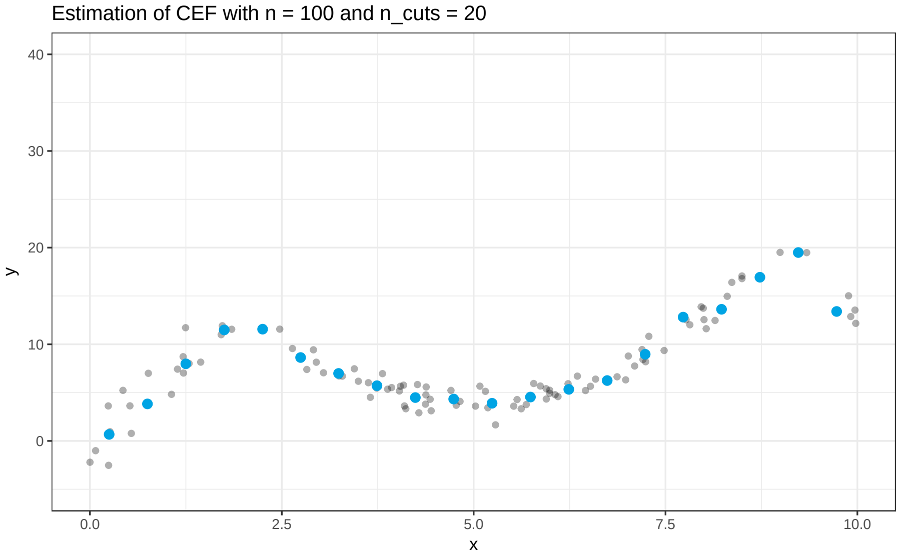
A Binning Approach to CEF for continuous random variables



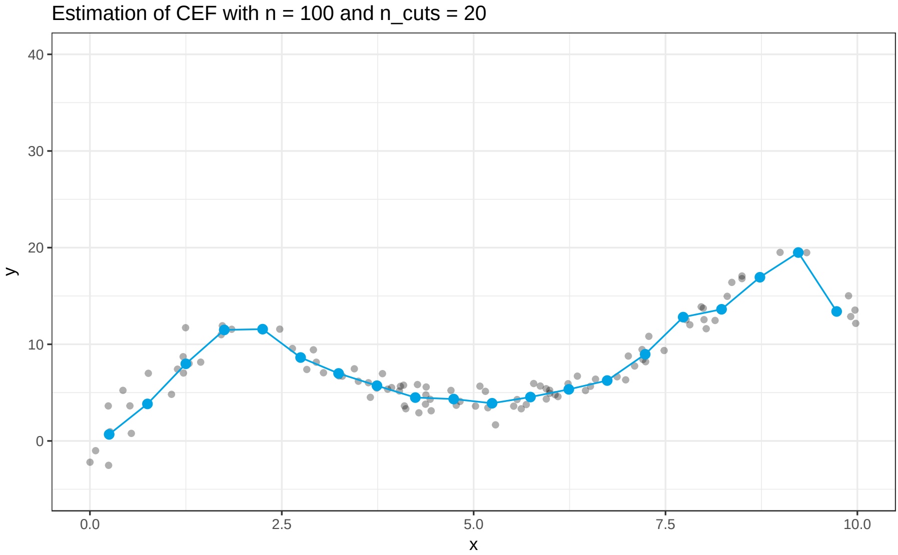
A Binning Approach to CEF for continuous random variables



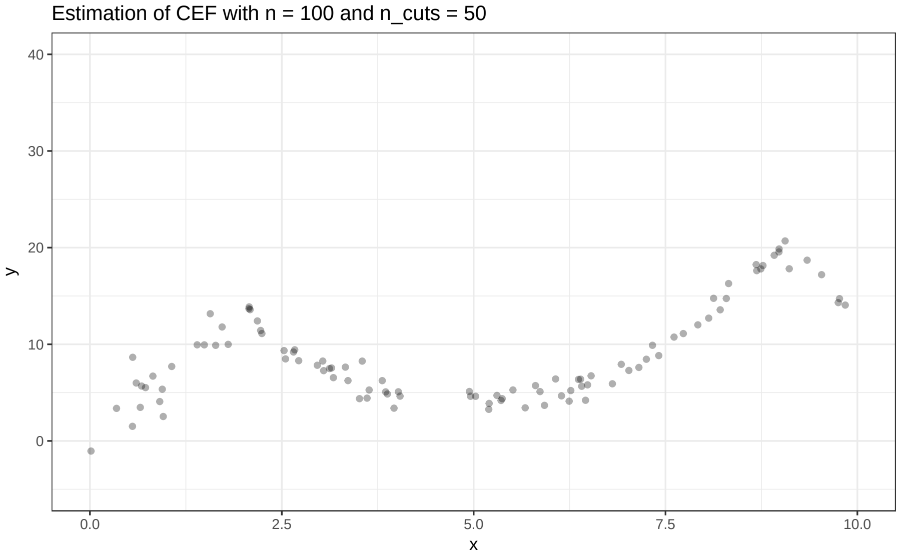
A Binning Approach to CEF for continuous random variables



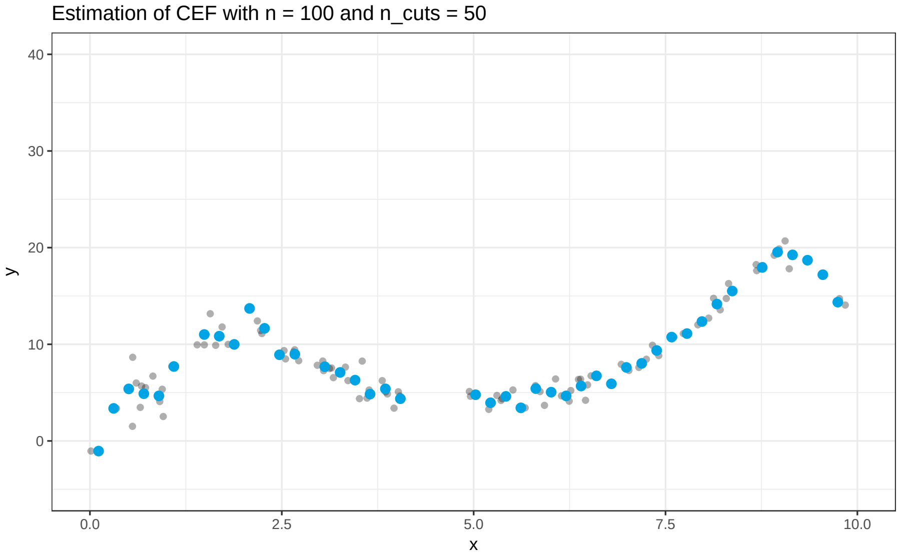
A Binning Approach to CEF for continuous random variables



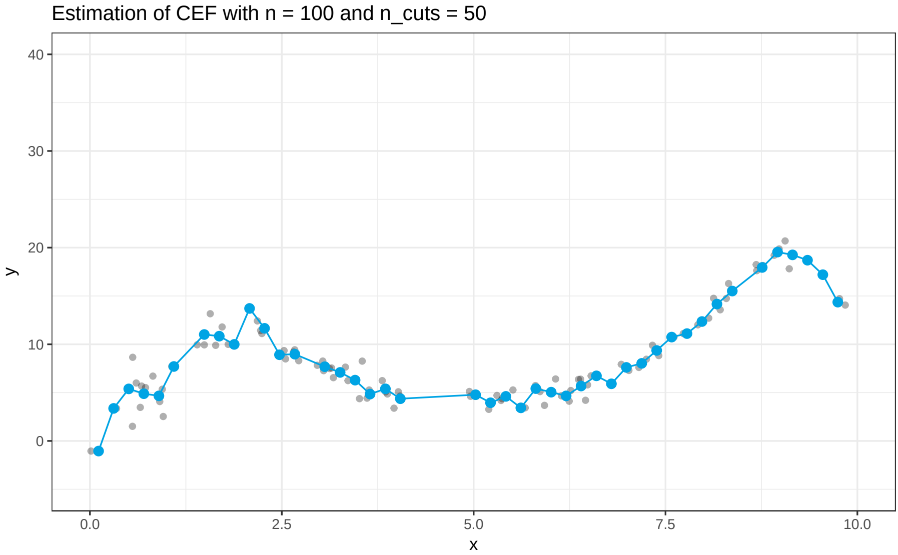
A Binning Approach to CEF for continuous random variables



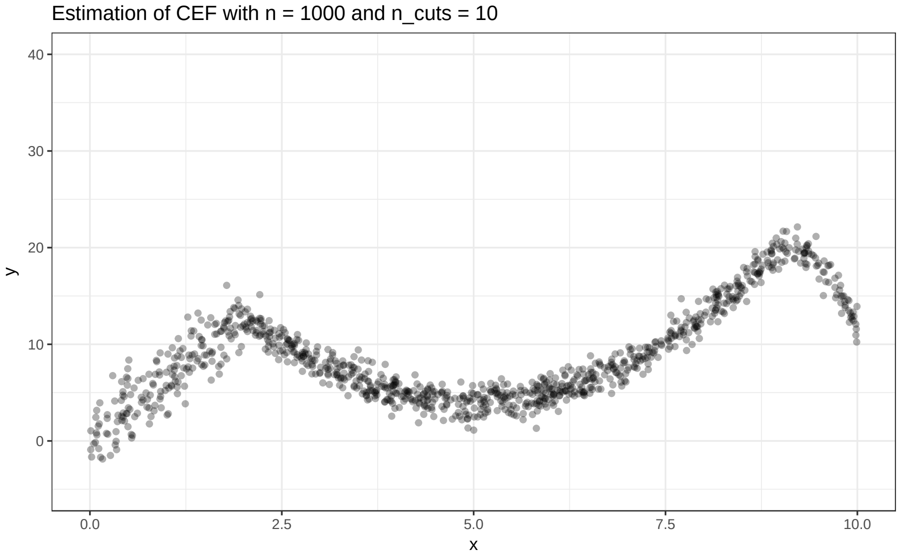
A Binning Approach to CEF for continuous random variables



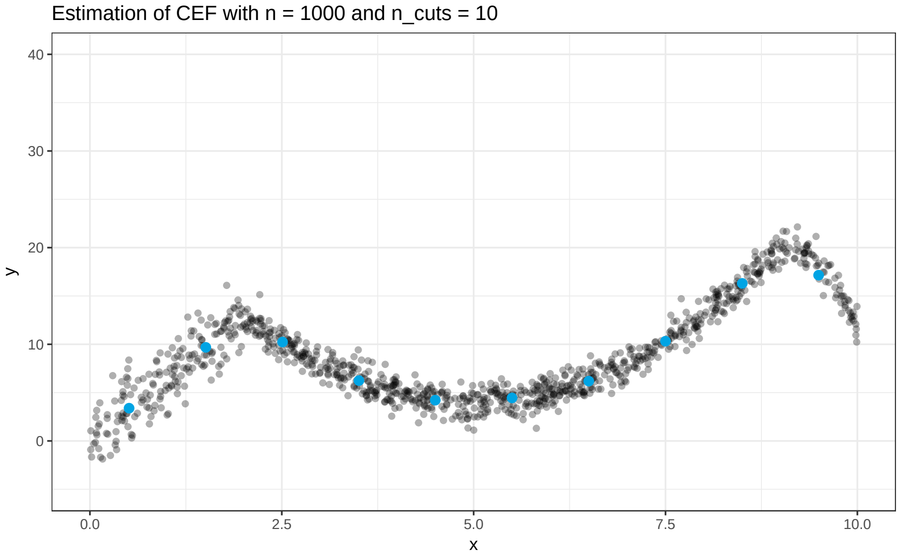
A Binning Approach to CEF for continuous random variables



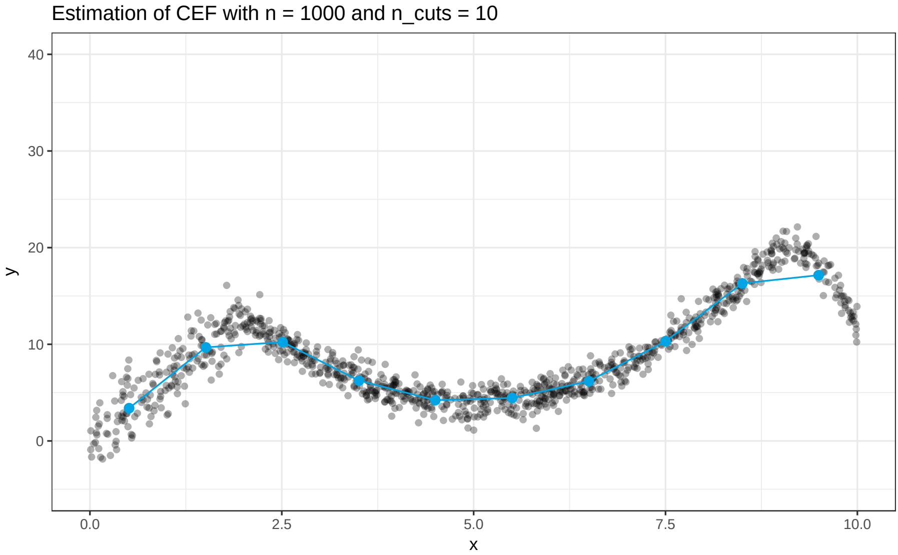
A Binning Approach to CEF for continuous random variables



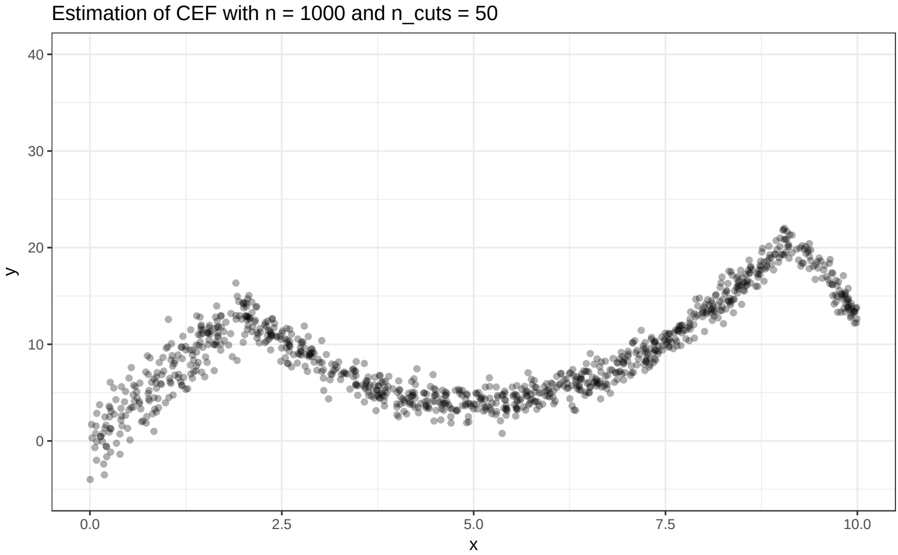
A Binning Approach to CEF for continuous random variables



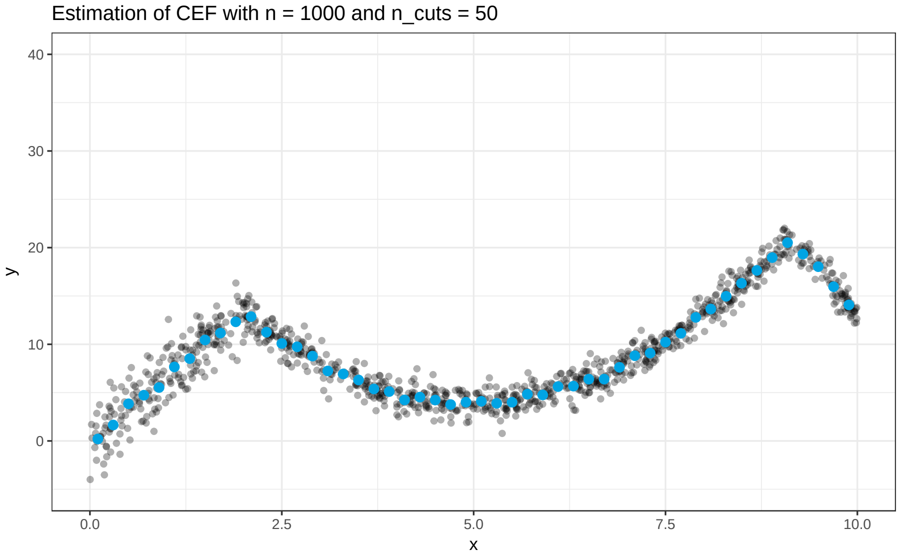
A Binning Approach to CEF for continuous random variables



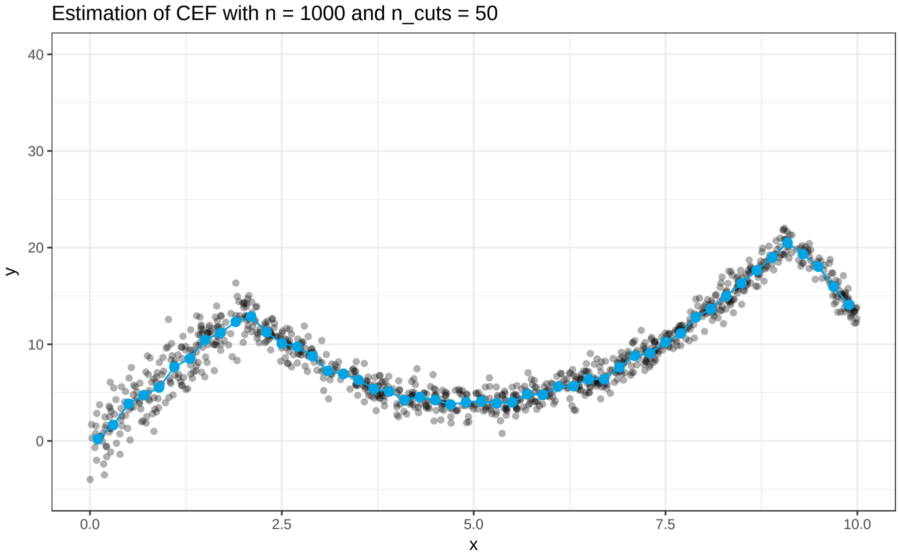
A Binning Approach to CEF for continuous random variables



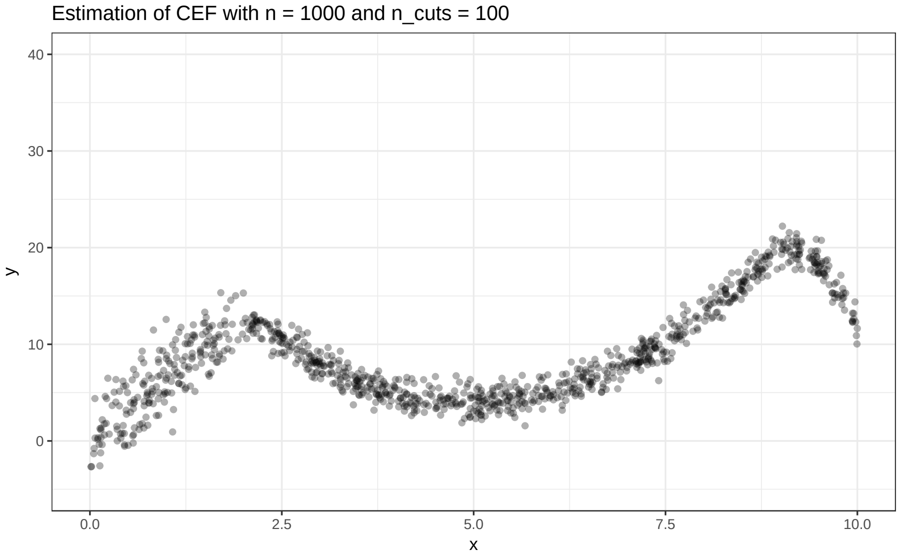
A Binning Approach to CEF for continuous random variables



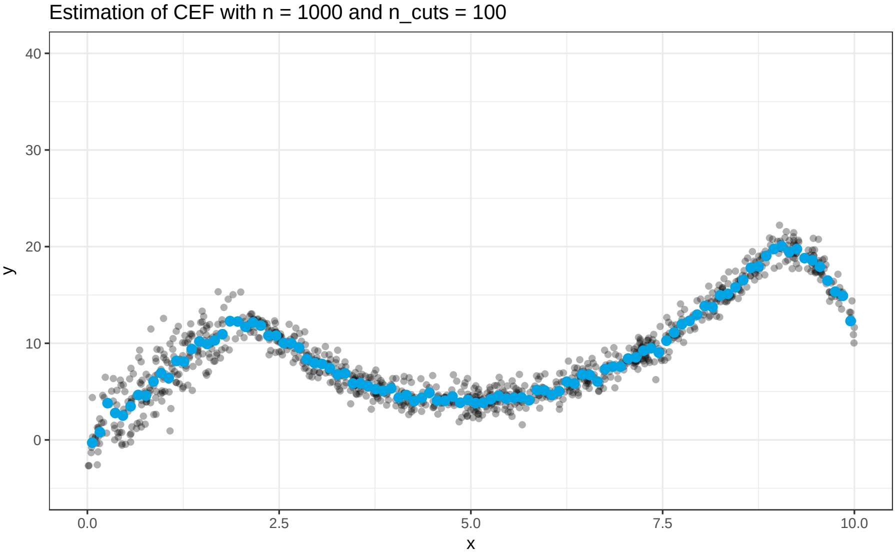
A Binning Approach to CEF for continuous random variables



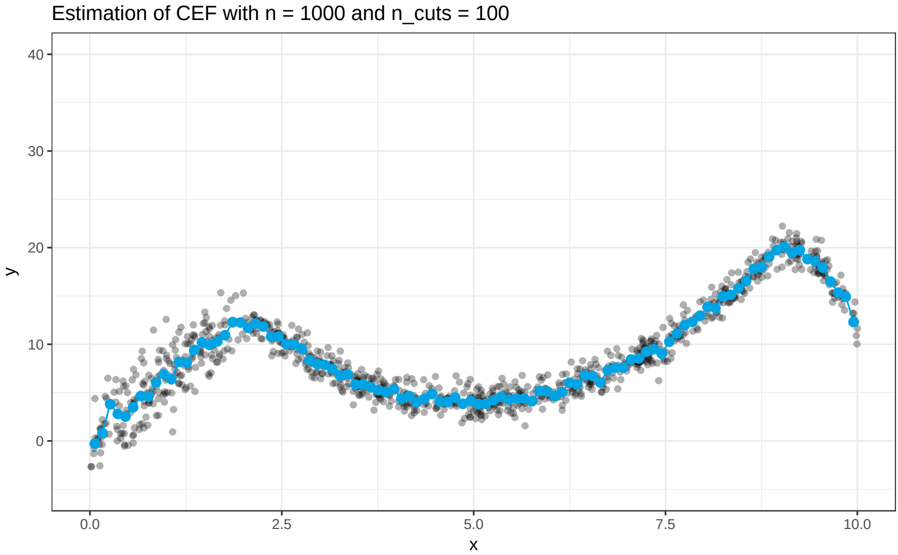
A Binning Approach to CEF for continuous random variables



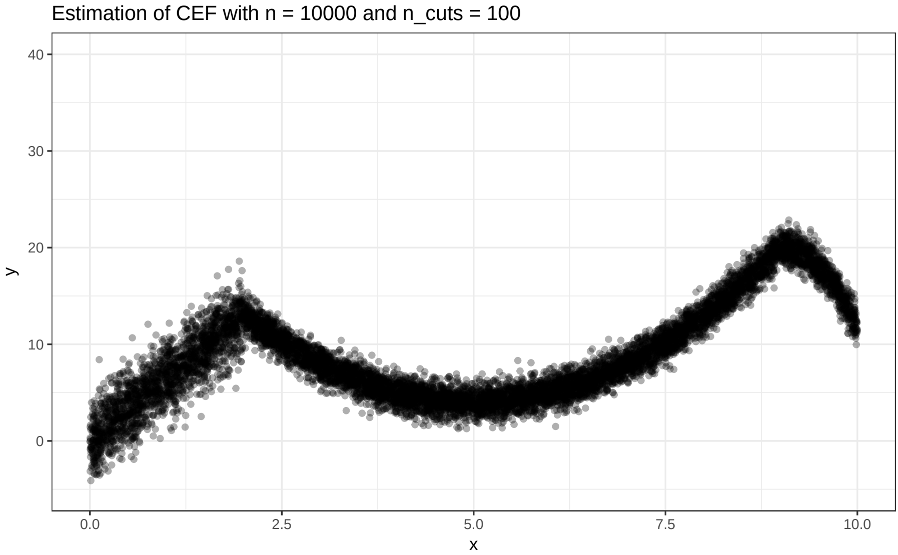
A Binning Approach to CEF for continuous random variables



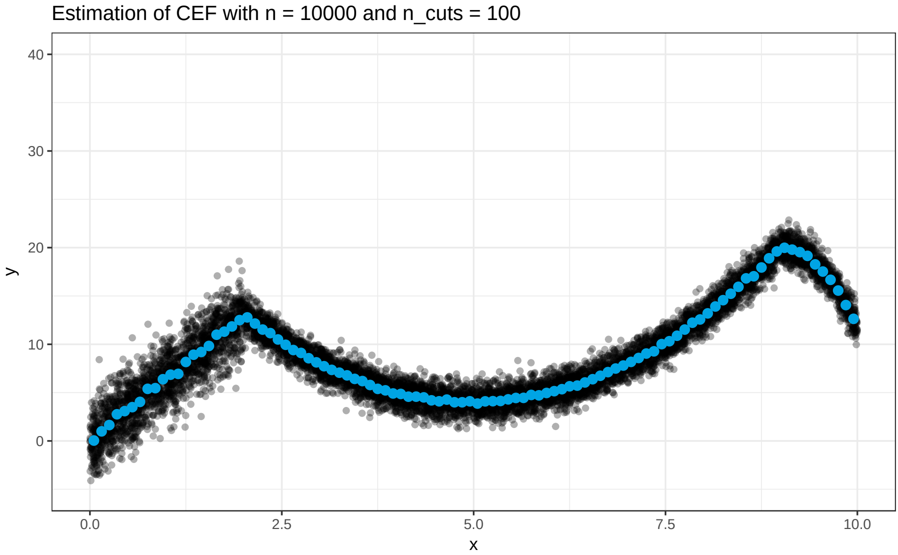
A Binning Approach to CEF for continuous random variables



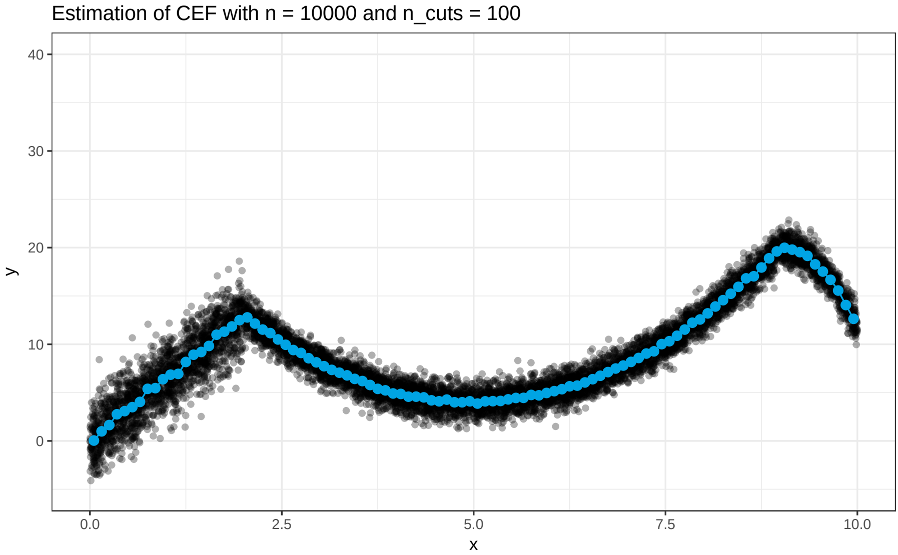
A Binning Approach to CEF for continuous random variables



A Binning Approach to CEF for continuous random variables



A Binning Approach to CEF for continuous random variables



Uniform Kernel Regression: Simple Local Averages

Uniform Kernel Regression: Simple Local Averages

- Dividing into discrete bins can get pretty noisy.

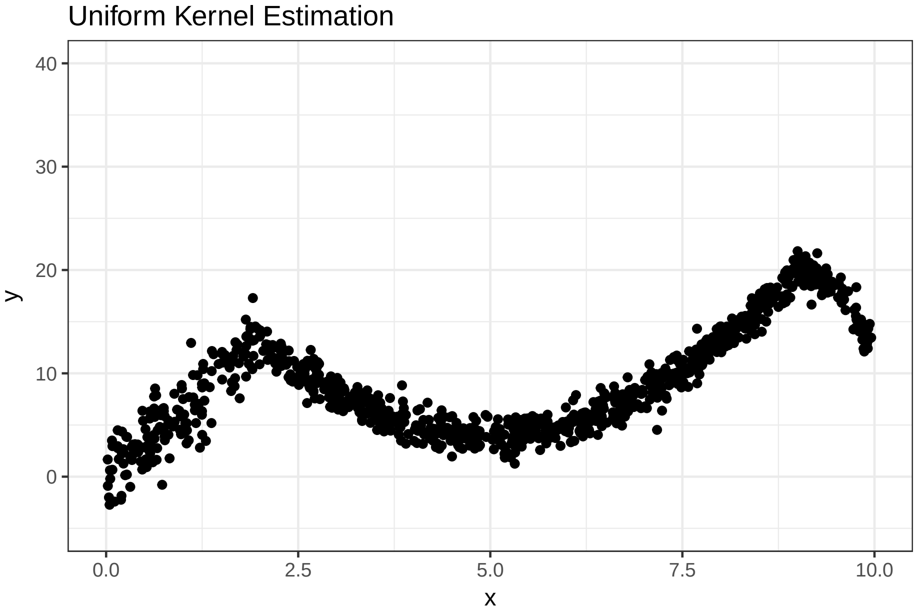
Uniform Kernel Regression: Simple Local Averages

- Dividing into discrete bins can get pretty noisy.
- Another approach is to use a moving local average to estimate $E[Y|X]$.

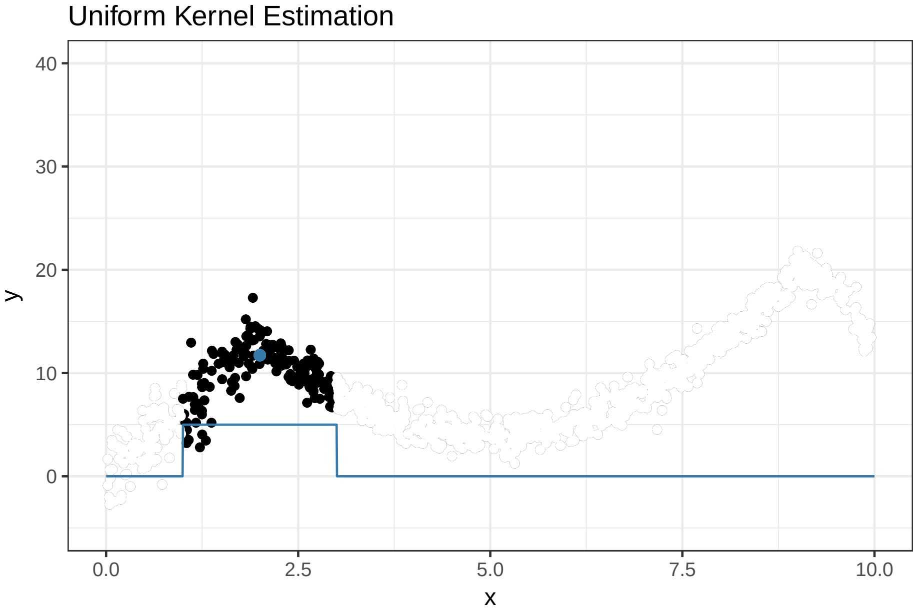
Uniform Kernel Regression: Simple Local Averages

- Dividing into discrete bins can get pretty noisy.
- Another approach is to use a moving local average to estimate $E[Y|X]$.
- We will call this approach uniform kernel regression for a reason that will become clear shortly.

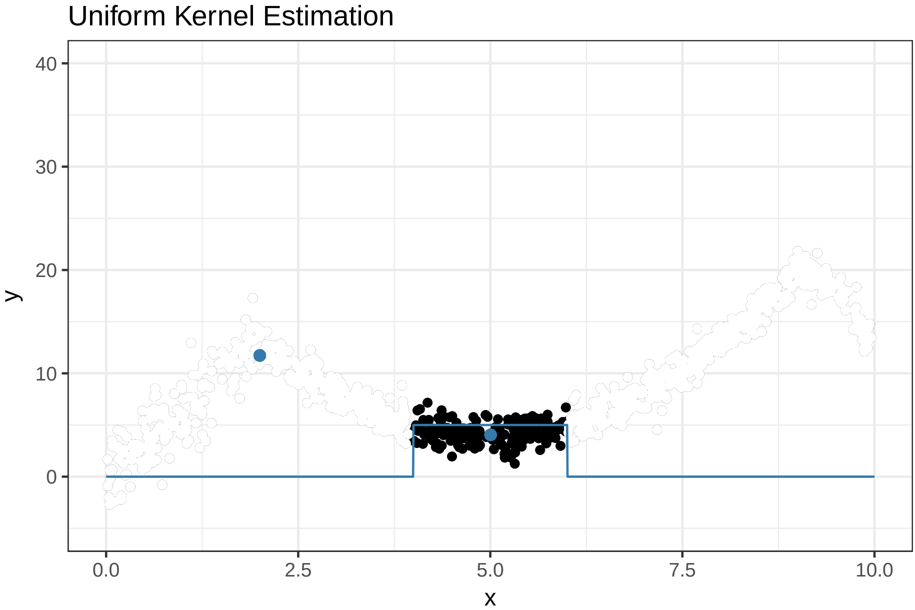
Example of Kernel CEF Estimation



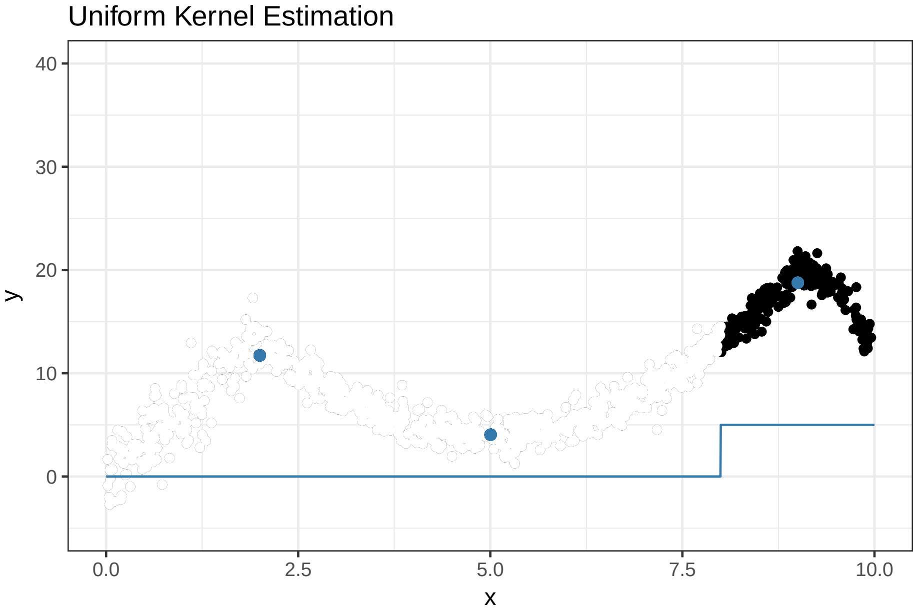
Example of Kernel CEF Estimation



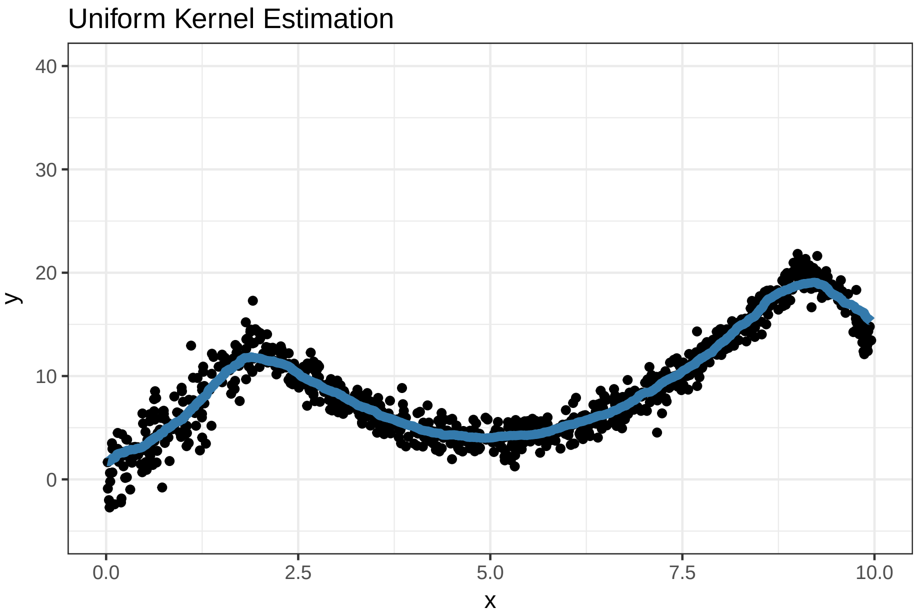
Example of Kernel CEF Estimation



Example of Kernel CEF Estimation



Example of Kernel CEF Estimation



Uniform Kernel Regression: Simple Local Averages

Uniform Kernel Regression: Simple Local Averages

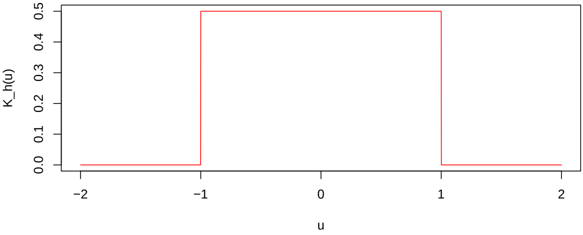
- Calculate the average of the observed y points that have x values in the interval $[x_0 - h\,,\, x_0 + h]$

Uniform Kernel Regression: Simple Local Averages

- Calculate the average of the observed y points that have x values in the interval $[x_0 - h\,,\, x_0 + h]$
- $h =$ some positive number (called the bandwidth)

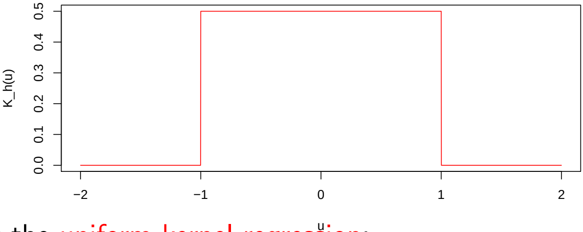
Uniform Kernel Regression: Simple Local Averages

- Calculate the average of the observed y points that have x values in the interval $[x_0 - h\,,\, x_0 + h]$
- $h =$ some positive number (called the bandwidth)
- Uniform kernel: every observation in the interval is equally weighted



Uniform Kernel Regression: Simple Local Averages

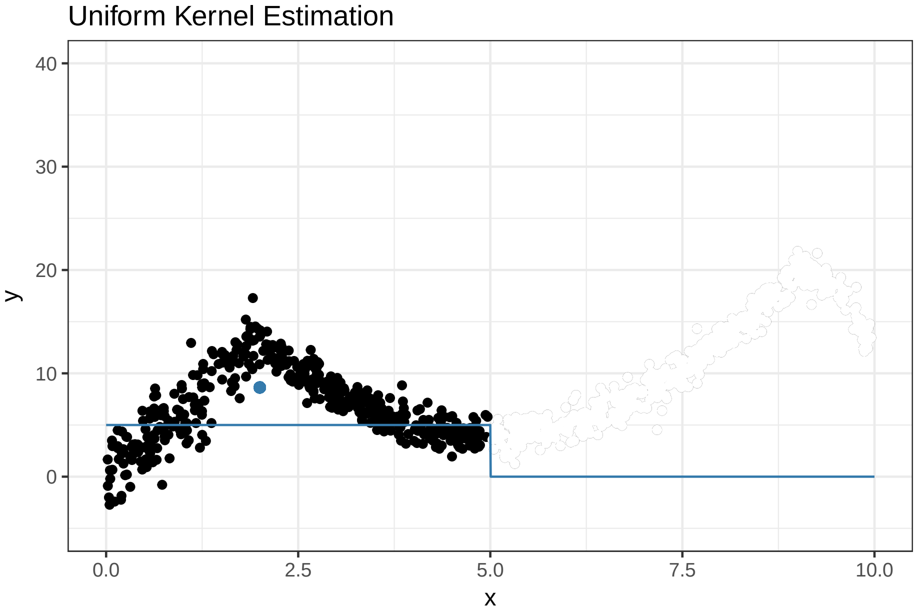
- Calculate the average of the observed y points that have x values in the interval $[x_0 - h\,,\, x_0 + h]$
- $h =$ some positive number (called the bandwidth)
- Uniform kernel: every observation in the interval is equally weighted



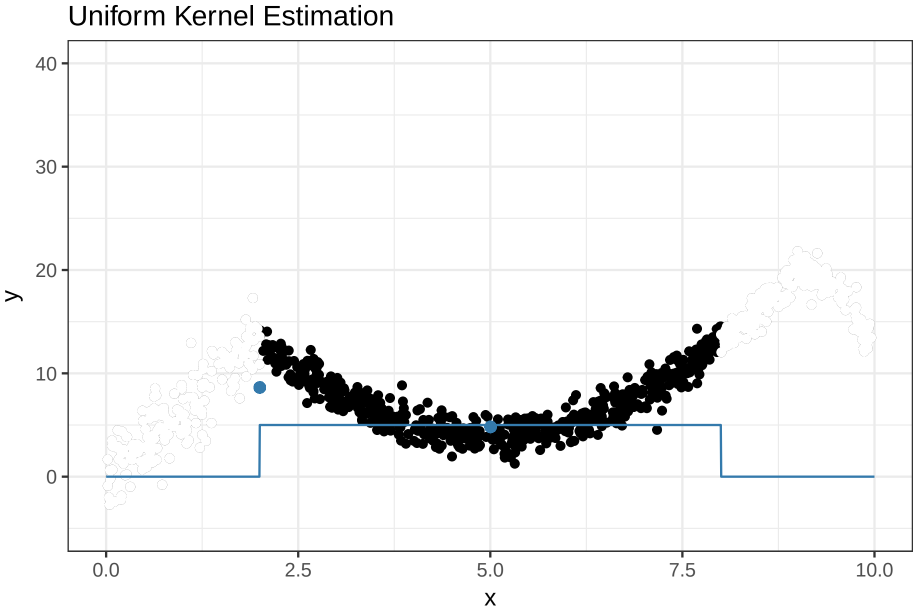
- This gives the uniform kernel regression:

$$\widehat{E}[Y|X = x_0] = \frac{\sum_{i=1}^{N} K_h((X_i - x_0)/h)Y_i}{\sum_{i=1}^{N} K_h((X_i - x_0)/h)} \text{ where } K_h(u) = \frac{1}{2}\,\mathbf{1}_{\{|u|\leq 1\}}$$

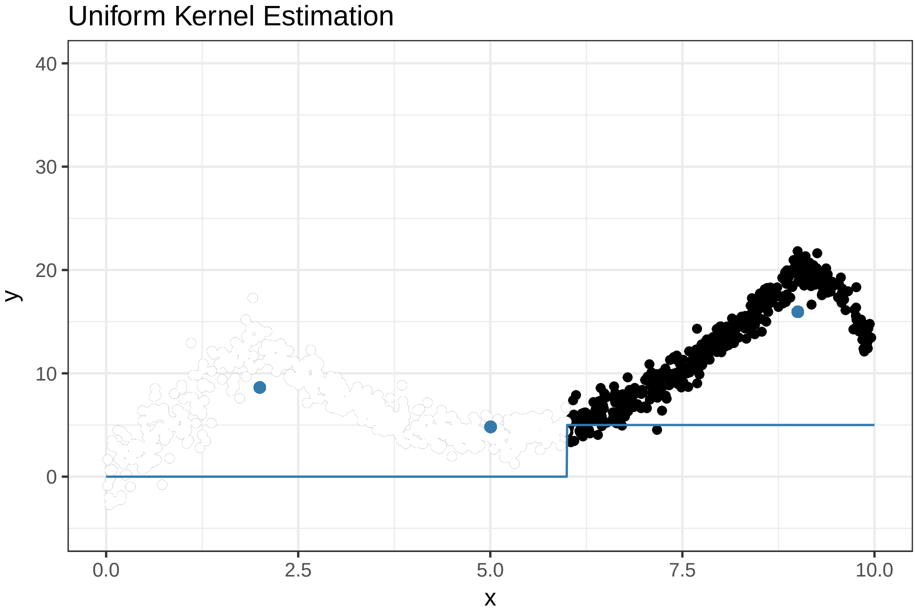
Changing the Bandwidth



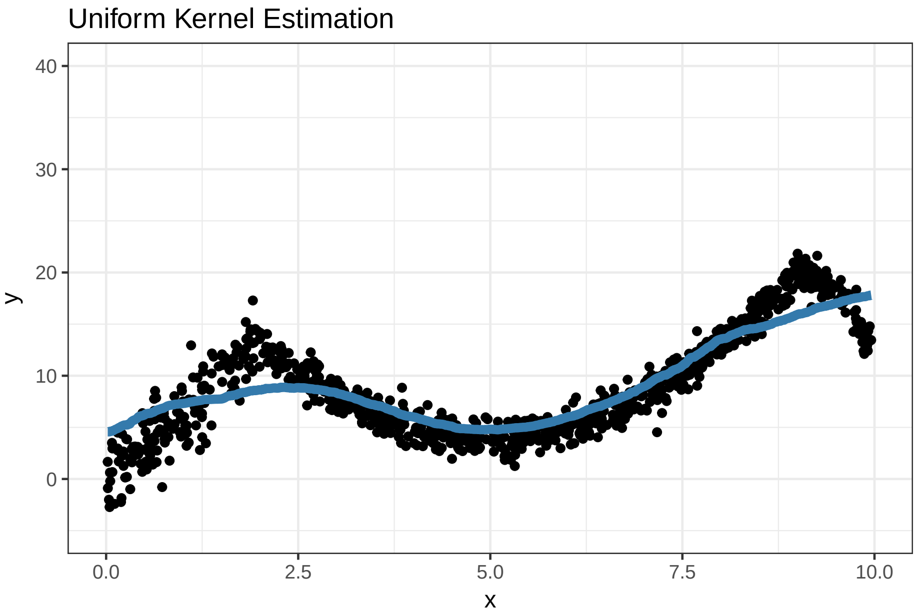
Changing the Bandwidth



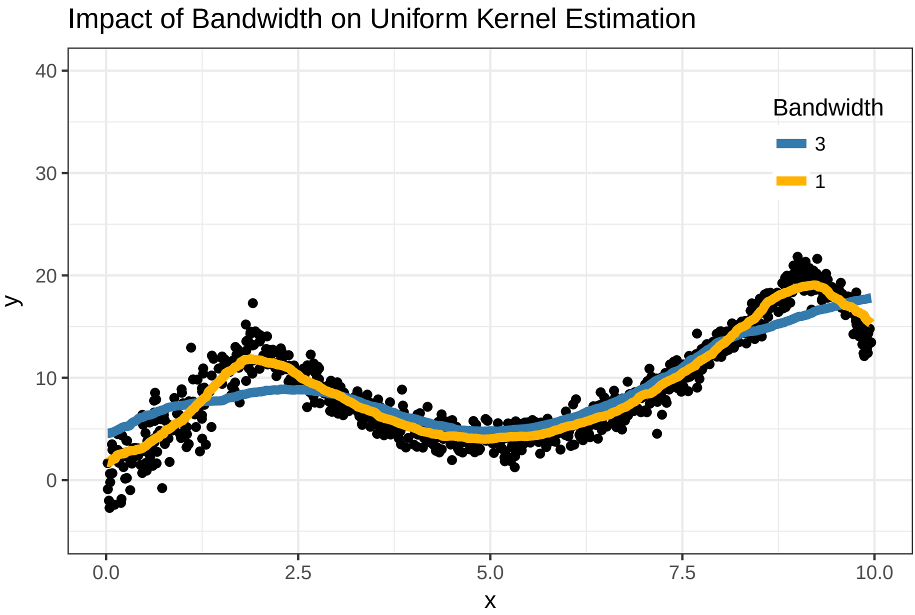
Changing the Bandwidth



Changing the Bandwidth



Changing the Bandwidth



Uniform Kernel Regression: Properties

Theorem (Consistency of the Uniform Kernel Density Estimator)

For iid continuous random variables $X_1, X_2, \ldots X_n$, $\forall x \in \mathbb{R}$,

- *if the kernel is uniform, and*
- *if $h \to 0$ and*
- *$nh \to \infty$ as*
- *$n \to \infty$, then*

$$\hat{f}_K(x) \xrightarrow{p} f(x).$$

Aronow and Miller Theorem 3.3.8. Proof by weak law of large numbers and the plug-in principle.

The More General Form of the Estimator

The More General Form of the Estimator

Definition (Kernel Density Estimator)

Let $X_1, X_2, \ldots X_n$ be iid continuous random variables with common PDF f.

The More General Form of the Estimator

Definition (Kernel Density Estimator)

Let $X_1, X_2, \ldots X_n$ be iid continuous random variables with common PDF f.

Let $K : \mathbb{R} \to \mathbb{R}$ be a function which is symmetric about the y-axis satisfying $\int_{-\infty}^{\infty} K(x)dx = 1$, and let $K_h(x) = \frac{1}{h}K(\frac{x}{h}) \forall x \in \mathbb{R}$ and $h > 0$.

The More General Form of the Estimator

Definition (Kernel Density Estimator)

Let $X_1, X_2, \dots X_n$ be iid continuous random variables with common PDF f.

Let $K : \mathbb{R} \to \mathbb{R}$ be a function which is symmetric about the y-axis satisfying $\int_{-\infty}^{\infty} K(x)dx = 1$, and let $K_h(x) = \frac{1}{h}K(\frac{x}{h}) \forall x \in \mathbb{R}$ and $h > 0$.

Then a kernel density estimator of $f(x)$ is

$$\hat{f}_K(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i), \forall x \in \mathbb{R}$$

The function K is called the kernel and the scaling parameter h is called the bandwidth.
(Aronow and Miller Definition 3.3.7)

Kernel Estimation

Kernel Estimation

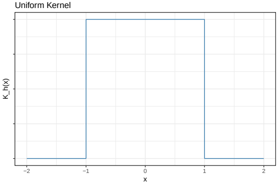
- We can create other estimators by filling in other kernels.

Kernel Estimation

- We can create other estimators by filling in other kernels.
- Generally the idea is to weight things further from the focal point by less than those closer to the focal point.

Kernel Estimation

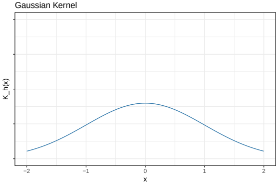
- We can create other estimators by filling in other kernels.
- Generally the idea is to weight things further from the focal point by less than those closer to the focal point.
- Uniform Kernel
- Calculate the average of the observed y points that have x values in the interval $[x_0 - h, x_0 + h]$
- Each observation within the interval is given equal weight, each observation outside the interval is given 0 weight



Kernel Estimation

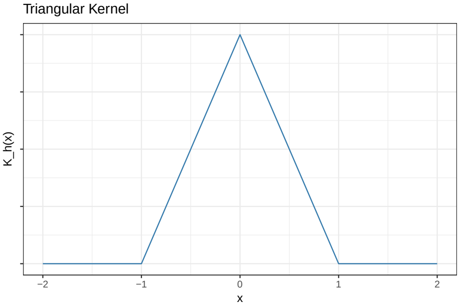
- We can create other estimators by filling in other kernels.
- Generally the idea is to weight things further from the focal point by less than those closer to the focal point.
- Gaussian Kernel
- Distance weighted by how far from x_0 following the normal density

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$



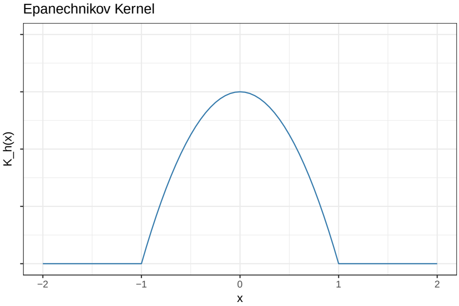
Kernel Estimation

- We can create other estimators by filling in other kernels.
- Generally the idea is to weight things further from the focal point by less than those closer to the focal point.
- Triangular
- Distance weighted by how far from x_0 using linear distance

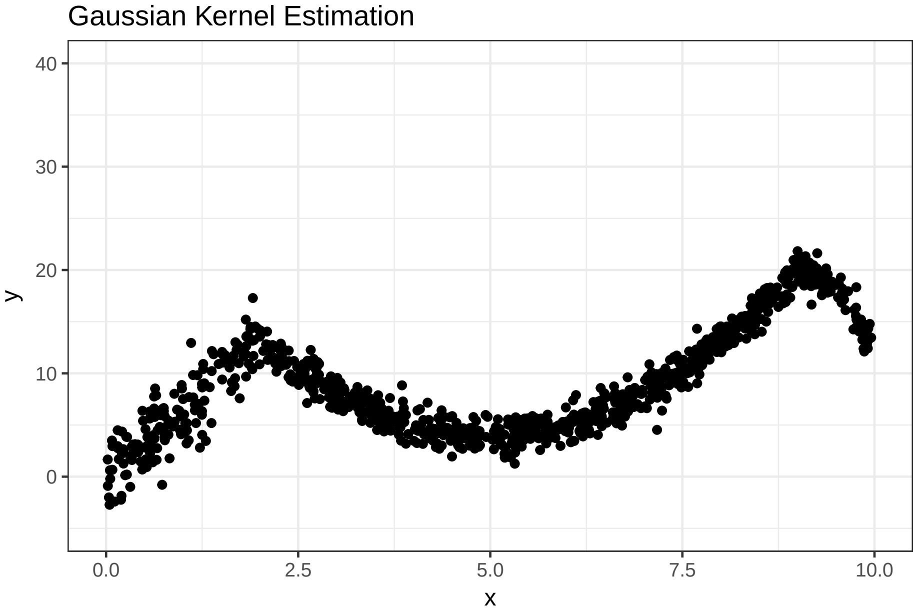


Kernel Estimation

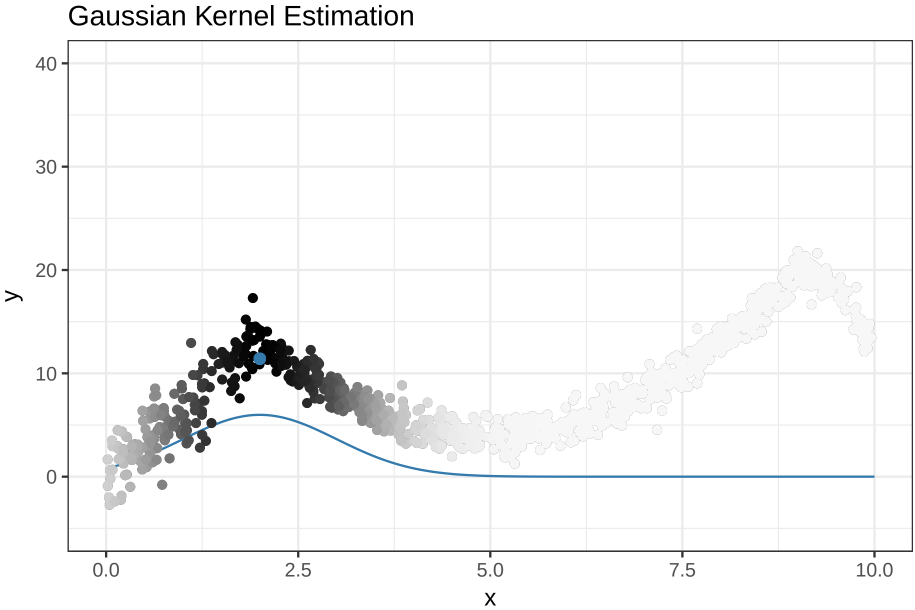
- We can create other estimators by filling in other kernels.
- Generally the idea is to weight things further from the focal point by less than those closer to the focal point.
- Epanechnikov
- Distance weighted by how far from x_0 using a parabolic function



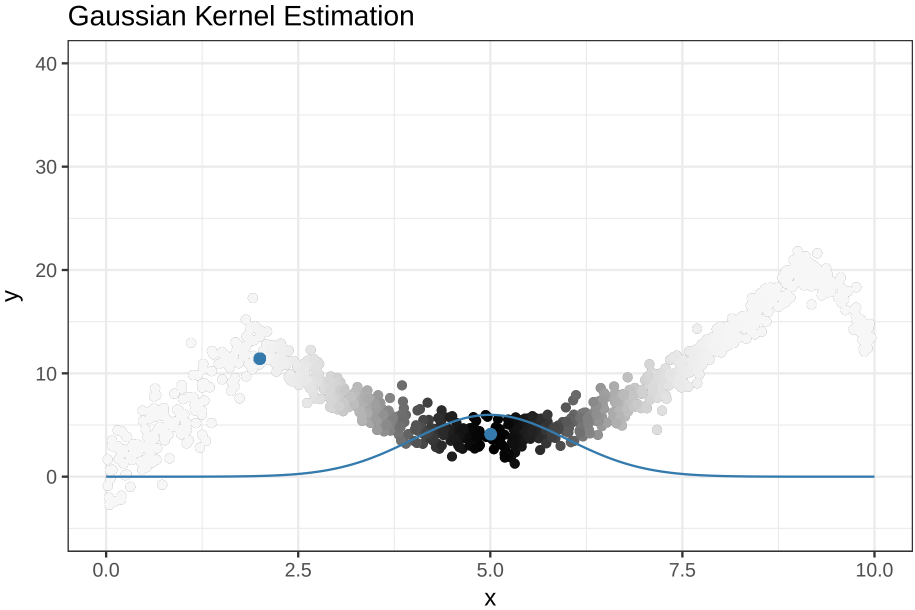
Example of Gaussian Kernel



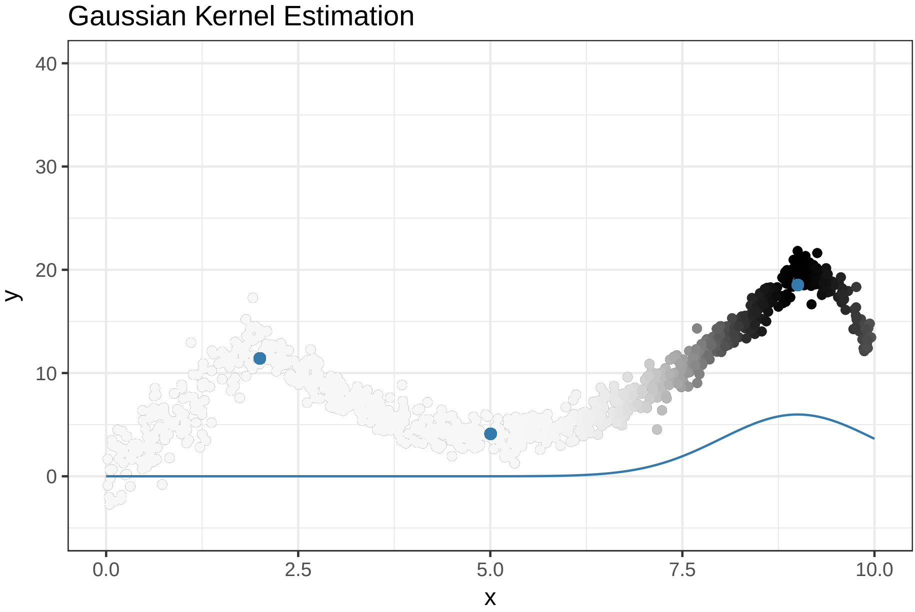
Example of Gaussian Kernel



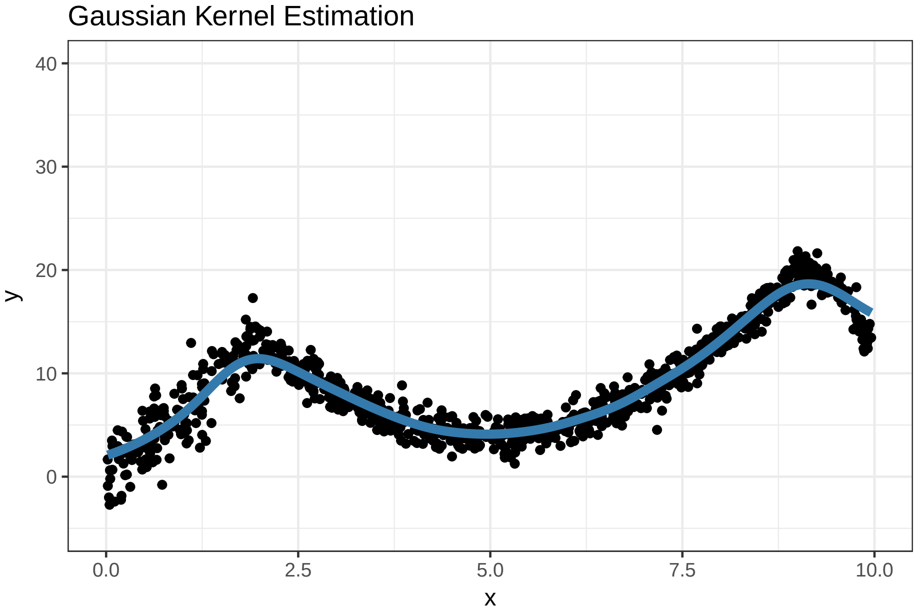
Example of Gaussian Kernel



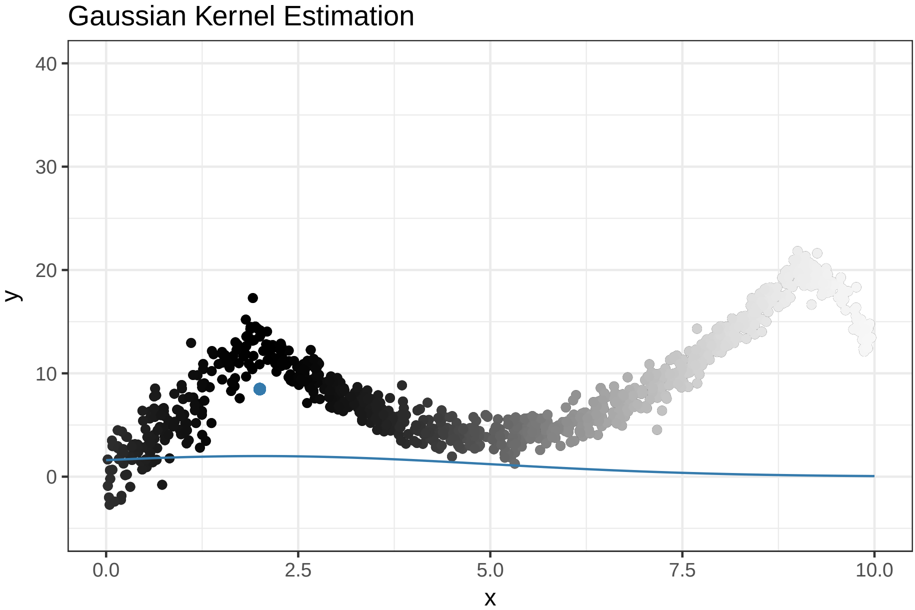
Example of Gaussian Kernel



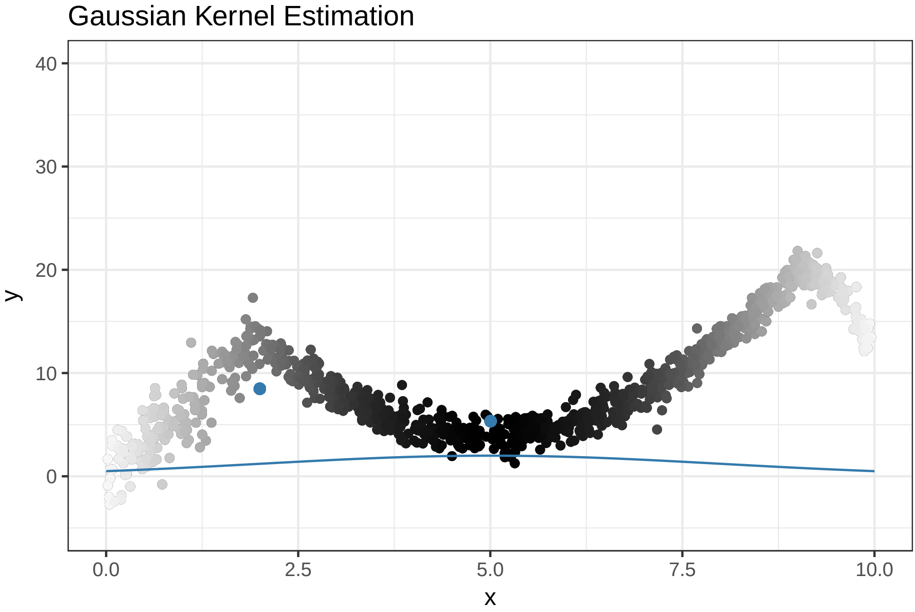
Example of Gaussian Kernel



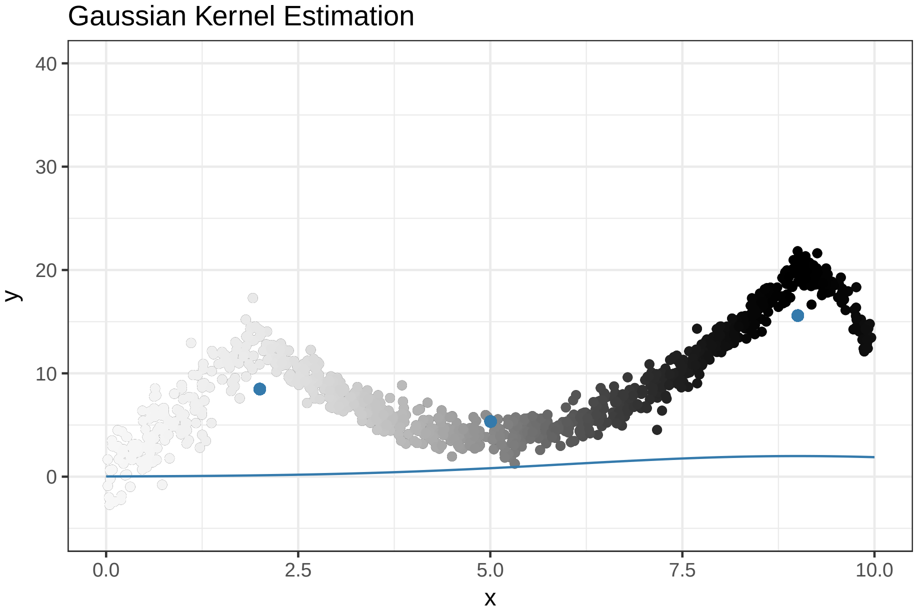
Example of Gaussian Kernel



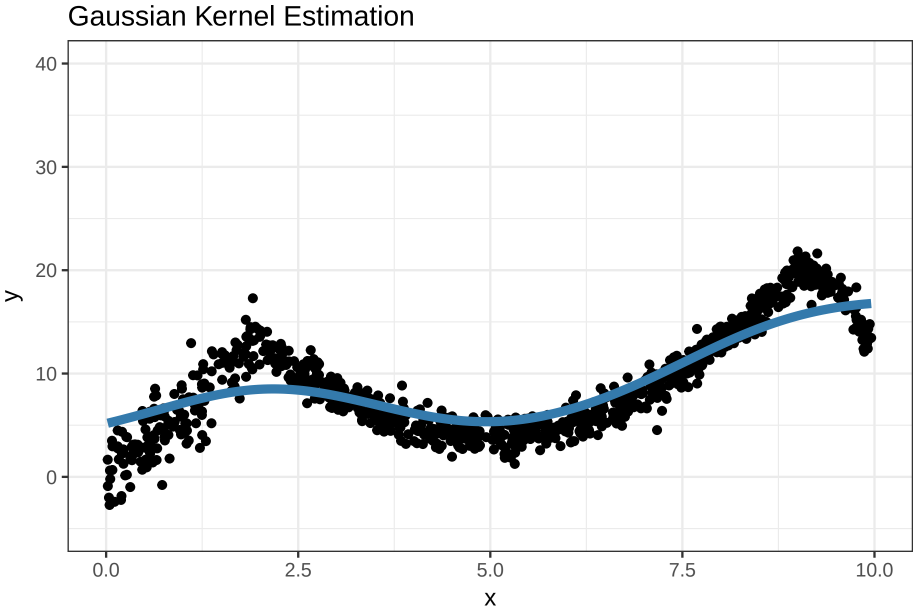
Example of Gaussian Kernel



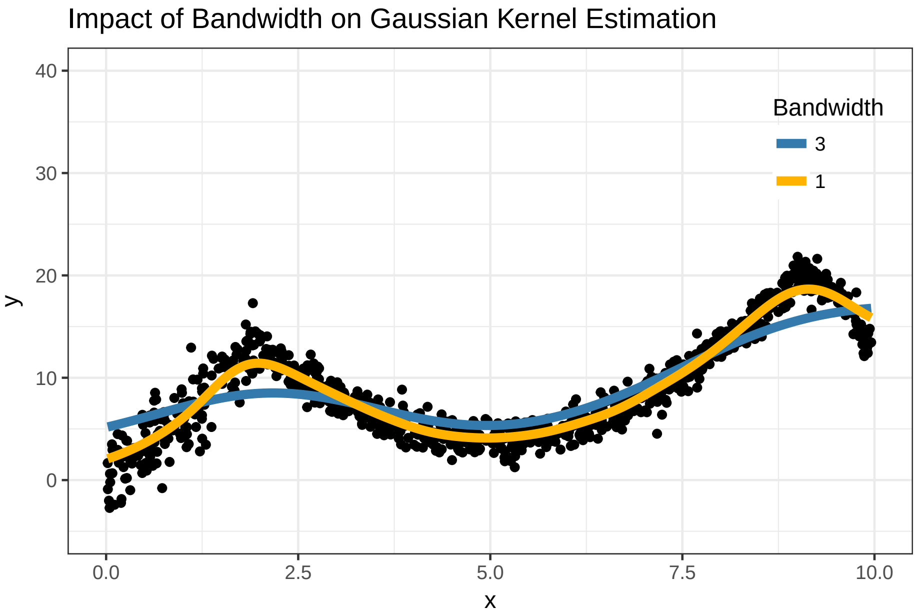
Example of Gaussian Kernel



Example of Gaussian Kernel



Example of Gaussian Kernel



Bias-Variance Tradeoff

Bias-Variance Tradeoff

- When choosing an estimator $\widehat{E}[Y|X]$ for $E[Y|X]$, we face a bias-variance tradeoff

Bias-Variance Tradeoff

- When choosing an estimator $\widehat{E}[Y|X]$ for $E[Y|X]$, we face a bias-variance tradeoff
- Notice that we can chose models with various levels of flexibility:

Bias-Variance Tradeoff

- When choosing an estimator $\widehat{E}[Y|X]$ for $E[Y|X]$, we face a bias-variance tradeoff
- Notice that we can chose models with various levels of flexibility:
  - A very flexible estimator allows the shape of the function to vary (e.g. a kernel regression with a small bandwidth)

Bias-Variance Tradeoff

- When choosing an estimator $\widehat{E}[Y|X]$ for $E[Y|X]$, we face a bias-variance tradeoff
- Notice that we can chose models with various levels of flexibility:
  - A very flexible estimator allows the shape of the function to vary (e.g. a kernel regression with a small bandwidth)
  - A very inflexible estimator restricts the shape of the function to a particular form
    (e.g. a kernel regression with a very wide bandwidth)

Best Linear Predictor

Best Linear Predictor

- The CEF can be infinitely complex.

Best Linear Predictor

- The CEF can be infinitely complex.
- Estimating the CEF can, therefore, be difficult.

Best Linear Predictor

- The CEF can be infinitely complex.
- Estimating the CEF can, therefore, be difficult.
- So, instead we can limit ourselves to classes of functions that approximate the CEF.

Best Linear Predictor

- The CEF can be infinitely complex.
- Estimating the CEF can, therefore, be difficult.
- So, instead we can limit ourselves to classes of functions that approximate the CEF.
  - I.e. we are using a more inflexible estimator.

Best Linear Predictor

- The CEF can be infinitely complex.
- Estimating the CEF can, therefore, be difficult.
- So, instead we can limit ourselves to classes of functions that approximate the CEF.
  - I.e. we are using a more inflexible estimator.
- In particular, what if we limit ourselves to the class of linear predictors:

$$g(X) = a + bX$$

Best Linear Predictor

- The CEF can be infinitely complex.
- Estimating the CEF can, therefore, be difficult.
- So, instead we can limit ourselves to classes of functions that approximate the CEF.
  - I.e. we are using a more inflexible estimator.
- In particular, what if we limit ourselves to the class of linear predictors:

$$g(X) = a + bX$$

- We call this the linear regression line

Best Linear Predictor

- The CEF can be infinitely complex.
- Estimating the CEF can, therefore, be difficult.
- So, instead we can limit ourselves to classes of functions that approximate the CEF.
  - I.e. we are using a more inflexible estimator.
- In particular, what if we limit ourselves to the class of linear predictors:

$$g(X) = a + bX$$

- We call this the linear regression line
- What function minimizes MSE, among this class of functional forms?

Best Linear Predictor

Theorem

For a random variable X and Y, if $V[X] > 0$, then the best linear predictor (BLP) of Y given X is $g(X) = \alpha + \beta X$ where,

$$\alpha = E[Y] - \frac{Cov[X, Y]}{V[X]} E[X]$$

$$\beta = \frac{Cov(X, Y)}{V(X)}$$

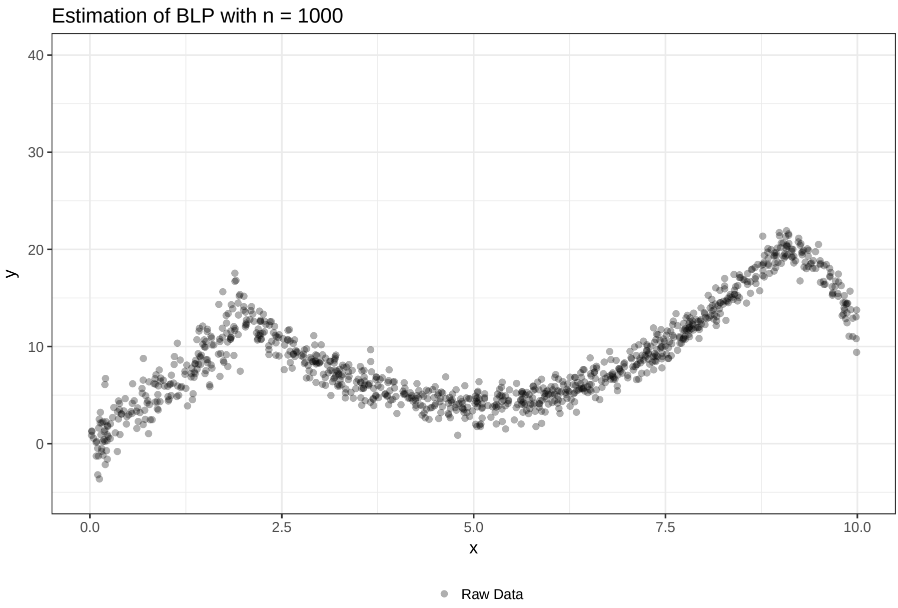
Corollary

- *The BLP is the best linear predictor of the CEF. I.e. setting $a = \alpha$ and $b = \beta$ minimizes*

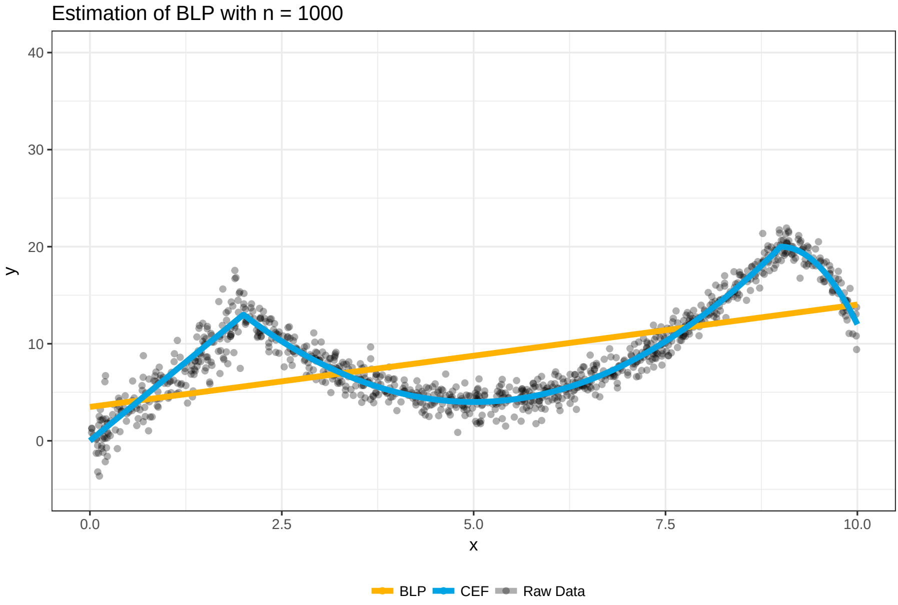
$$E\left[(E[Y \mid X] - (a + bX))^2\right]$$

- *If the CEF is linear, the CEF is the BLP*

Best Linear Predictor



Best Linear Predictor



Linear Approximations

- But what if the CEF isn't linear?

Linear Approximations

- But what if the CEF isn't linear?
- Well it probably isn't, but the best linear predictor is still well-defined.

Linear Approximations

- But what if the CEF isn't linear?
- Well it probably isn't, but the best linear predictor is still well-defined.
- It is the linear projection of Y_i onto X_i.

Linear Approximations

- But what if the CEF isn't linear?
- Well it probably isn't, but the best linear predictor is still well-defined.
- It is the linear projection of Y_i onto X_i.
- In general, this is distinct from the CEF:

Linear Approximations

- But what if the CEF isn't linear?
- Well it probably isn't, but the best linear predictor is still well-defined.
- It is the linear projection of Y_i onto X_i.
- In general, this is distinct from the CEF:
  - CEF is the best predictor of Y_i among all functions.

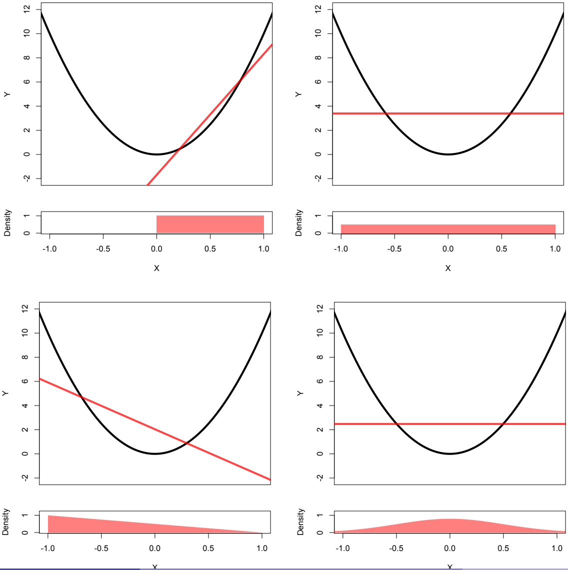
Linear Approximations

- But what if the CEF isn't linear?
- Well it probably isn't, but the best linear predictor is still well-defined.
- It is the linear projection of Y_i onto X_i.
- In general, this is distinct from the CEF:
  - CEF is the best predictor of Y_i among all functions.
  - Linear projection is the best predictor among linear functions.

Linear Approximations

- But what if the CEF isn't linear?
- Well it probably isn't, but the best linear predictor is still well-defined.
- It is the linear projection of Y_i onto X_i.
- In general, this is distinct from the CEF:
  - CEF is the best predictor of Y_i among all functions.
  - Linear projection is the best predictor among linear functions.
- The nice thing about the linear projection is that it exists and is well-defined even if the CEF is non-linear.

BLP Approximations Depend on the Marginal Distribution of X



Best Linear Predictor

Best Linear Predictor

Warning: the BLP won't always be a good fit for the data
(even though it really wants to be)

Best Linear Predictor

Warning: the BLP won't always be a good fit for the data
(even though it really wants to be)

Figure: 'If I fits, I sits'

Best Linear Predictor

Warning: the BLP won't always be a good fit for the data
(even though it really wants to be)

Figure: 'If I fits, I sits'

The BLP is always a line regardless of the data.

Ordinary Least Squares

Ordinary Least Squares

- Okay, we've finally made it to the OLS model you've heard so much about!

Ordinary Least Squares

- Okay, we've finally made it to the OLS model you've heard so much about!
- We've defined a population line of best fit $\beta_0 + \beta_1 X_i$ that approximates the CEF.

Ordinary Least Squares

- Okay, we've finally made it to the OLS model you've heard so much about!
- We've defined a population line of best fit $\beta_0 + \beta_1 X_i$ that approximates the CEF.
- We can now estimate β_0 and β_1 like any other population parameters using samples from the joint distribution $f_{(Y,X)}(y, x)$

Ordinary Least Squares

- Okay, we've finally made it to the OLS model you've heard so much about!
- We've defined a population line of best fit $\beta_0 + \beta_1 X_i$ that approximates the CEF.
- We can now estimate β_0 and β_1 like any other population parameters using samples from the joint distribution $f_{(Y,X)}(y, x)$
- The core idea will be to use the plug-in principle. We want the line that minimizes:

$$(\beta_0, \beta_1) = \underset{\beta_0, \beta_1}{\arg\min}\, E[(Y_i - \beta_0 - \beta_1 X_i)^2]$$

Ordinary Least Squares

- Okay, we've finally made it to the OLS model you've heard so much about!
- We've defined a population line of best fit $\beta_0 + \beta_1 X_i$ that approximates the CEF.
- We can now estimate β_0 and β_1 like any other population parameters using samples from the joint distribution $f_{(Y,X)}(y,x)$
- The core idea will be to use the plug-in principle. We want the line that minimizes:

$$(\beta_0, \beta_1) = \arg\min_{\beta_0, \beta_1} E[(Y_i - \beta_0 - \beta_1 X_i)^2]$$

- So we will plug in the sample analogs to the population:

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{\widetilde{\beta}_0, \widetilde{\beta}_1} \sum_{i=1}^{n} (Y_i - \widetilde{\beta}_0 - \widetilde{\beta}_1 X_i)^2$$

Ordinary Least Squares

- Okay, we've finally made it to the OLS model you've heard so much about!
- We've defined a population line of best fit $\beta_0 + \beta_1 X_i$ that approximates the CEF.
- We can now estimate β_0 and β_1 like any other population parameters using samples from the joint distribution $f_{(Y,X)}(y,x)$
- The core idea will be to use the plug-in principle. We want the line that minimizes:

$$(\beta_0, \beta_1) = \underset{\beta_0, \beta_1}{\arg\min}\, E[(Y_i - \beta_0 - \beta_1 X_i)^2]$$

- So we will plug in the sample analogs to the population:

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{\widetilde{\beta}_0, \widetilde{\beta}_1}{\arg\min} \sum_{i=1}^{n} (Y_i - \widetilde{\beta}_0 - \widetilde{\beta}_1 X_i)^2$$

- This is called the ordinary least squares (OLS) estimator.

Plug-in Estimation of the BLP

Noting we can rewrite $\frac{Cov[X,Y]}{V[X]} = \frac{E[XY]-E[X]E[Y]}{E[X^2]-E[X]^2}$, and using the plug-in principle, we can estimate the parameters of the BLP as:

Plug-in Estimation of the BLP

Noting we can rewrite $\frac{Cov[X,Y]}{V[X]} = \frac{E[XY]-E[X]E[Y]}{E[X^2]-E[X]^2}$, and using the plug-in principle, we can estimate the parameters of the BLP as:

$$\hat{\alpha} = \overline{Y} - \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\overline{X^2} - \overline{X}^2}\overline{X} = \overline{Y} - \hat{\beta}\overline{X}$$

$$\hat{\beta} = \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\overline{X^2} - \overline{X}^2} = \frac{\widehat{Cov}(X, Y)}{\widehat{V}(X)} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Plug-in Estimation of the BLP

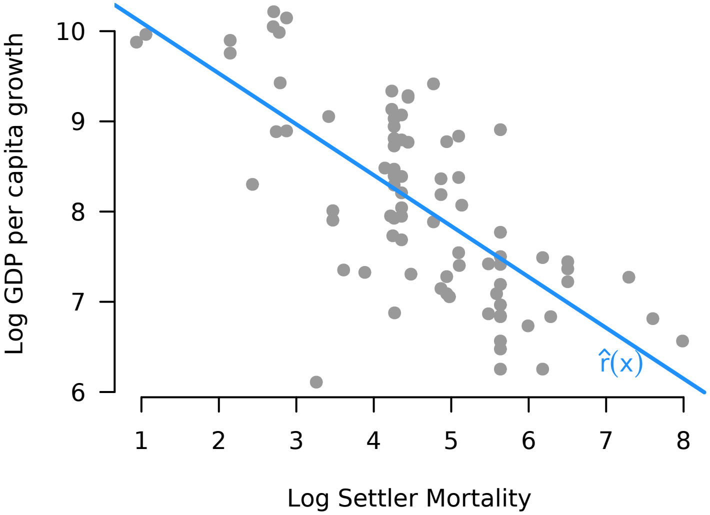
Noting we can rewrite $\frac{Cov[X,Y]}{V[X]} = \frac{E[XY]-E[X]E[Y]}{E[X^2]-E[X]^2}$, and using the plug-in principle, we can estimate the parameters of the BLP as:

$$\hat{\alpha} = \overline{Y} - \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\overline{X^2} - \overline{X}^2}\overline{X} = \overline{Y} - \hat{\beta}\overline{X}$$

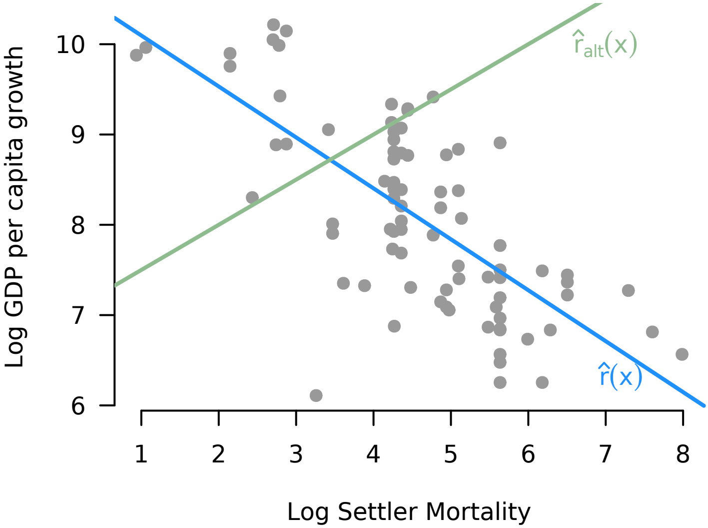
$$\hat{\beta} = \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\overline{X^2} - \overline{X}^2} = \frac{\widehat{Cov}(X,Y)}{\widehat{V}(X)} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

This corresponds to the linear projection which minimizes the sum of squared errors.

Fitted linear CEF/regression function



Fitted linear CEF/regression function



Fitted values and residuals

Fitted values and residuals

Definition (Fitted Value)

A **fitted value** or **predicted value** is the estimated conditional mean of Y_i for a particular observation with independent variable X_i:

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$$

Fitted values and residuals

Definition (Fitted Value)

A **fitted value** or **predicted value** is the estimated conditional mean of Y_i for a particular observation with independent variable X_i:

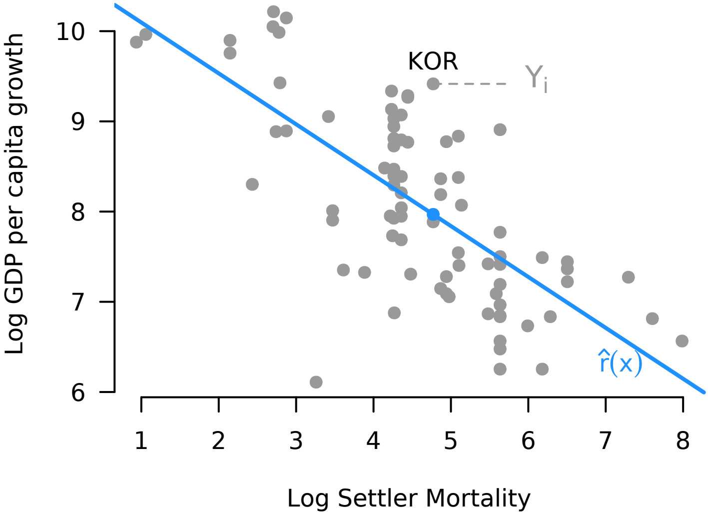
$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$$

Definition (Residual)

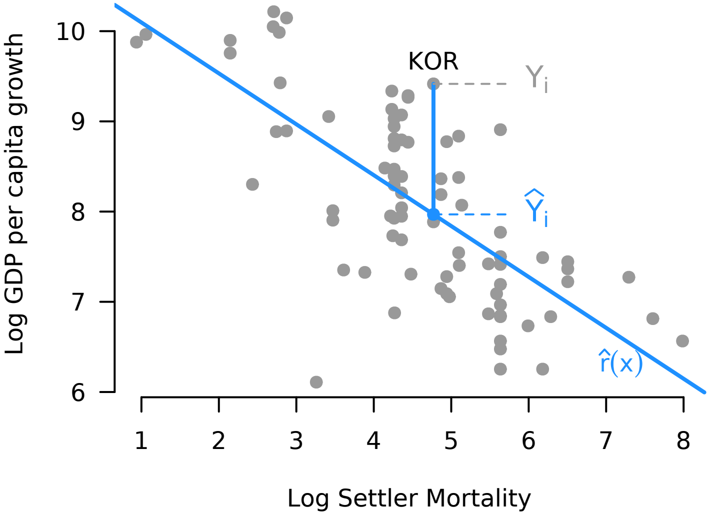
The **residual** is the difference between the actual value of Y_i and the predicted value, $\widehat{Y}_i$:

$$\widehat{u}_i = Y_i - \widehat{Y}_i = Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i$$

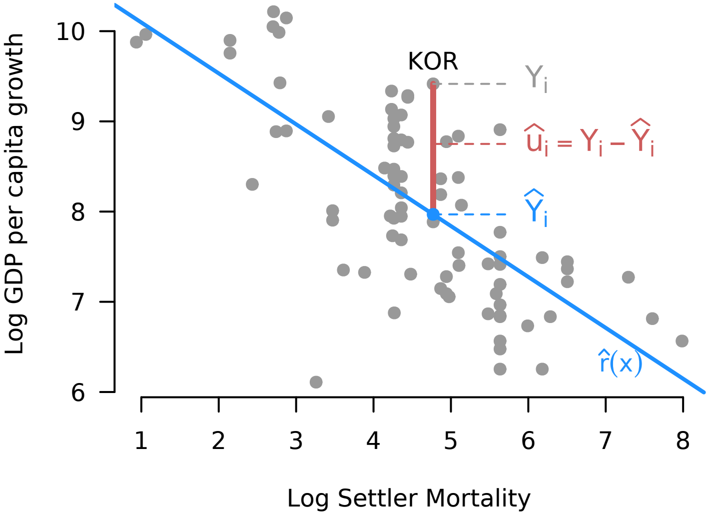
Fitted linear CEF/regression function



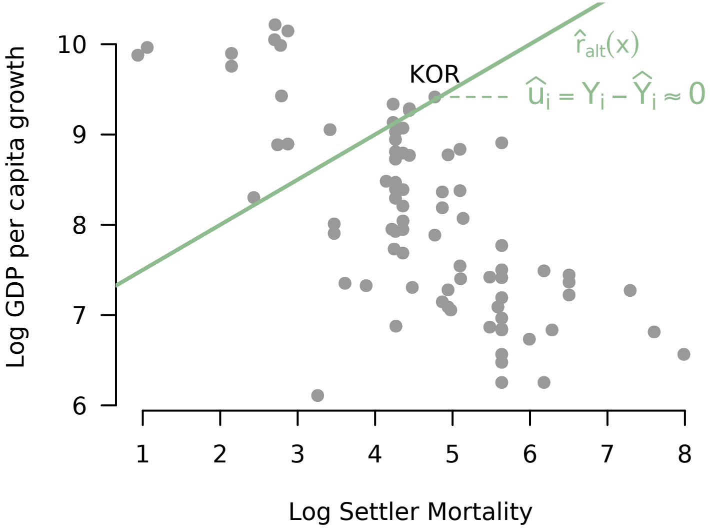
Fitted linear CEF/regression function



Fitted linear CEF/regression function



Why not this line?



Minimize the residuals

Minimize the residuals

- The residuals, $\widehat{u}_i = Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i$, tell us how well the line fits the data.

Minimize the residuals

- The residuals, $\widehat{u}_i = Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i$, tell us how well the line fits the data.
  - Larger magnitude residuals means that points are very far from the line

Minimize the residuals

- The residuals, $\widehat{u}_i = Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i$, tell us how well the line fits the data.
  - Larger magnitude residuals means that points are very far from the line
  - Residuals close to 0 mean points very close to the line

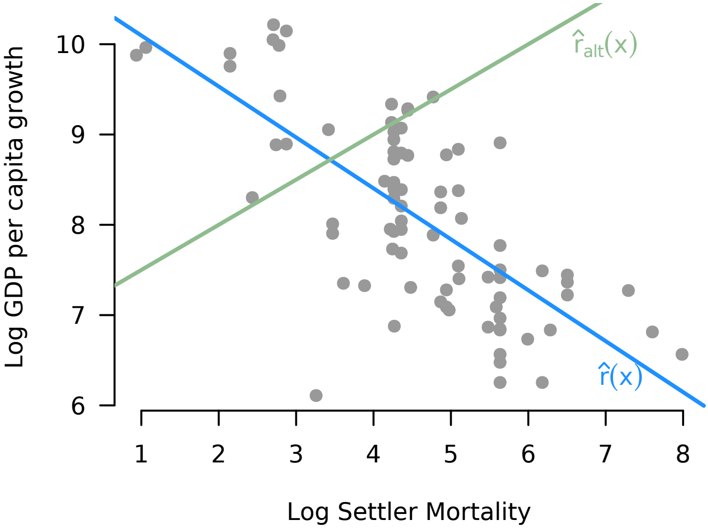
Minimize the residuals

- The residuals, $\widehat{u}_i = Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i$, tell us how well the line fits the data.
  - Larger magnitude residuals means that points are very far from the line
  - Residuals close to 0 mean points very close to the line
- The smaller the magnitude of the residuals, the better we are doing at predicting Y

Minimize the residuals

- The residuals, $\widehat{u}_i = Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i$, tell us how well the line fits the data.
  - Larger magnitude residuals means that points are very far from the line
  - Residuals close to 0 mean points very close to the line
- The smaller the magnitude of the residuals, the better we are doing at predicting Y
- Choose the line that minimizes the residuals

Which is better at minimizing residuals?



Linear Regression: Justification

Linear Regression: Justification

- The BLP may be a very loose approximation to $E[Y|X]$

Linear Regression: Justification

- The BLP may be a very loose approximation to $E[Y|X]$
- Why would we ever want to do this?

Linear Regression: Justification

- The BLP may be a very loose approximation to $E[Y|X]$
- Why would we ever want to do this?
  - Theoretical reason to assume linearity is close enough

Linear Regression: Justification

- The BLP may be a very loose approximation to $E[Y|X]$
- Why would we ever want to do this?
    - Theoretical reason to assume linearity is close enough
    - Ease of interpretation

Linear Regression: Justification

- The BLP may be a very loose approximation to $E[Y|X]$
- Why would we ever want to do this?
  - Theoretical reason to assume linearity is close enough
  - Ease of interpretation
  - Bias-variance tradeoff

Linear Regression: Justification

- The BLP may be a very loose approximation to $E[Y|X]$
- Why would we ever want to do this?
  - Theoretical reason to assume linearity is close enough
  - Ease of interpretation
  - Bias-variance tradeoff
  - Analytical derivation of sampling distributions (next few weeks)

Linear Regression: Justification

- The BLP may be a very loose approximation to $E[Y|X]$
- Why would we ever want to do this?
  - Theoretical reason to assume linearity is close enough
  - Ease of interpretation
  - Bias-variance tradeoff
  - Analytical derivation of sampling distributions (next few weeks)
  - We can make the model more flexible, even in a linear framework (e.g. we can add polynomials, use log transformations, etc.)

Linear Regression: Justification

- The BLP may be a very loose approximation to $E[Y|X]$
- Why would we ever want to do this?
  - Theoretical reason to assume linearity is close enough
  - Ease of interpretation
  - Bias-variance tradeoff
  - Analytical derivation of sampling distributions (next few weeks)
  - We can make the model more flexible, even in a linear framework (e.g. we can add polynomials, use log transformations, etc.)
- Perhaps the biggest reason is that it extends easily to the case where X is a vector of random variables.

Interpretation of the regression slope

Interpretation of the regression slope

- When we model the regression function as a line, we can interpret the parameters of the line in appealing ways:

Interpretation of the regression slope

- When we model the regression function as a line, we can interpret the parameters of the line in appealing ways:
  1. **Intercept**: the average outcome among units with $X = 0$ is β_0:

$$E[Y|X = 0] \approx$$

Interpretation of the regression slope

- When we model the regression function as a line, we can interpret the parameters of the line in appealing ways:
    1. **Intercept**: the average outcome among units with $X = 0$ is β_0:

$$E[Y|X = 0] \approx$$

Interpretation of the regression slope

- When we model the regression function as a line, we can interpret the parameters of the line in appealing ways:
  1. **Intercept**: the average outcome among units with $X = 0$ is β_0:

$$E[Y|X = 0] \approx \beta_0 + \beta_1 0 =$$

Interpretation of the regression slope

- When we model the regression function as a line, we can interpret the parameters of the line in appealing ways:
  1. **Intercept**: the average outcome among units with $X = 0$ is β_0:

$$E[Y|X = 0] \approx \beta_0 + \beta_1 0 = \beta_0$$

Interpretation of the regression slope

- When we model the regression function as a line, we can interpret the parameters of the line in appealing ways:
  1. **Intercept**: the average outcome among units with $X = 0$ is β_0:

$$E[Y|X = 0] \approx \beta_0 + \beta_1 0 = \beta_0$$

  2. **Slope**: a one-unit change in X changes our prediction of Y by β_1

$$E[Y|X = x + 1] - E[Y|X = x] \approx (\beta_0 + \beta_1(x + 1)) - (\beta_0 + \beta_1 x)$$

Interpretation of the regression slope

- When we model the regression function as a line, we can interpret the parameters of the line in appealing ways:
  1. **Intercept**: the average outcome among units with $X = 0$ is β_0:
  $$E[Y|X = 0] \approx \beta_0 + \beta_1 0 = \beta_0$$
  2. **Slope**: a one-unit change in X changes our prediction of Y by β_1
  $$E[Y|X = x + 1] - E[Y|X = x] \approx (\beta_0 + \beta_1(x + 1)) - (\beta_0 + \beta_1 x)$$

Interpretation of the regression slope

- When we model the regression function as a line, we can interpret the parameters of the line in appealing ways:
  1. **Intercept**: the average outcome among units with $X = 0$ is β_0:

$$E[Y|X=0] \approx \beta_0 + \beta_1 0 = \beta_0$$

  2. **Slope**: a one-unit change in X changes our prediction of Y by β_1

$$\begin{aligned} E[Y|X=x+1] - E[Y|X=x] &\approx (\beta_0 + \beta_1(x+1)) - (\beta_0 + \beta_1 x) \\ &= \beta_0 + \beta_1 x + \beta_1 - \beta_0 - \beta_1 x \end{aligned}$$

Interpretation of the regression slope

- When we model the regression function as a line, we can interpret the parameters of the line in appealing ways:
  1. **Intercept**: the average outcome among units with $X = 0$ is β_0:

     $$E[Y|X=0] \approx \beta_0 + \beta_1 0 = \beta_0$$

  2. **Slope**: a one-unit change in X changes our prediction of Y by β_1

     $$\begin{aligned} E[Y|X=x+1] - E[Y|X=x] &\approx (\beta_0 + \beta_1(x+1)) - (\beta_0 + \beta_1 x) \\ &= \beta_0 + \beta_1 x + \beta_1 - \beta_0 - \beta_1 x \\ &= \beta_1 \end{aligned}$$

Linear Regression as a Descriptive Model

- A lot of social science reports regression as saying that a one unit change in X is associated with a β unit change in Y.

Linear Regression as a Descriptive Model

- A lot of social science reports regression as saying that a one unit change in X is associated with a β unit change in Y.
- This leans suggestively towards a causal interpretation that if we intervened to change X then Y would change in some way. We will talk more about the assumptions we would need to draw this conclusion later in the semester.

Linear Regression as a Descriptive Model

- A lot of social science reports regression as saying that a one unit change in X is associated with a β unit change in Y.
- This leans suggestively towards a causal interpretation that if we intervened to change X then Y would change in some way. We will talk more about the assumptions we would need to draw this conclusion later in the semester.
- For now, I find it helpful to think of regression descriptively as talking about different groups of people. Our guess of the mean for units with $x = 2$ is β higher than our guess of the mean for the units with $x = 1$.

Linear Regression as a Descriptive Model

- A lot of social science reports regression as saying that a one unit change in X is associated with a β unit change in Y.
- This leans suggestively towards a causal interpretation that if we intervened to change X then Y would change in some way. We will talk more about the assumptions we would need to draw this conclusion later in the semester.
- For now, I find it helpful to think of regression descriptively as talking about different groups of people. Our guess of the mean for units with $x = 2$ is β higher than our guess of the mean for the units with $x = 1$.
- This helps clear that it is an approximation to the CEF and that the units being described are different.

Linear Regression as a Predictive Model

Linear Regression as a Predictive Model

- Linear regression can also be used to predict new observations

Linear Regression as a Predictive Model

- Linear regression can also be used to predict new observations
- Basic idea:

Linear Regression as a Predictive Model

- Linear regression can also be used to predict new observations
- Basic idea:
  - Find estimates $\hat{\beta}_0, \hat{\beta}_1$ of β_0, β_1 based on the in-sample data

Linear Regression as a Predictive Model

- Linear regression can also be used to predict new observations
- Basic idea:
  - Find estimates $\hat{\beta}_0, \hat{\beta}_1$ of β_0, β_1 based on the in-sample data
  - To find the expected value of Y for an out-of-sample data drawn from the same population we can use information about $X = x_{new}$ to calculate:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_{new}$$

Linear Regression as a Predictive Model

- Linear regression can also be used to predict new observations
- Basic idea:
  - Find estimates $\hat{\beta}_0, \hat{\beta}_1$ of β_0, β_1 based on the in-sample data
  - To find the expected value of Y for an out-of-sample data drawn from the same population we can use information about $X = x_{new}$ to calculate:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_{new}$$

- This prediction is the best in the space of lines in terms of the squared error.

Linear Regression as a Predictive Model

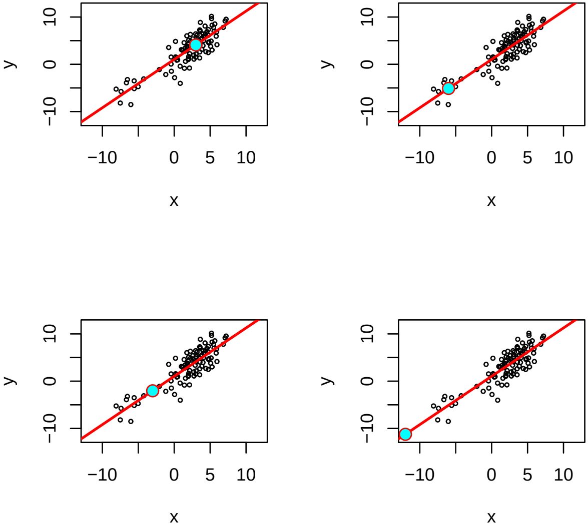
- Linear regression can also be used to predict new observations
- Basic idea:
  - Find estimates $\hat{\beta}_0, \hat{\beta}_1$ of β_0, β_1 based on the in-sample data
  - To find the expected value of Y for an out-of-sample data drawn from the same population we can use information about $X = x_{new}$ to calculate:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_{new}$$

- This prediction is the best in the space of lines in terms of the squared error.
- While the line is defined over all regions of the data we may be concerned about:
  - interpolation
  - extrapolation
  - predicting in ranges of X with sparse data

Which Predictions Do You Trust?

Which Predictions Do You Trust?



Example: Tatem, et al. Sprinters Data

Example: Tatem, et al. Sprinters Data

In a 2004 *Nature* article, Tatem et al. use linear regression to conclude that in the year 2156 the winner of the women's Olympic 100 meter sprint may likely have a faster time than the winner of the men's Olympic 100 meter sprint.

Example: Tatem, et al. Sprinters Data

In a 2004 *Nature* article, Tatem et al. use linear regression to conclude that in the year 2156 the winner of the women's Olympic 100 meter sprint may likely have a faster time than the winner of the men's Olympic 100 meter sprint.

How do the authors make this conclusion?

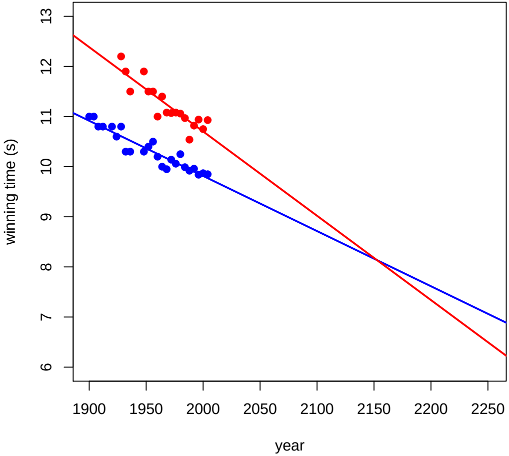
Example: Tatem, et al. Sprinters Data

In a 2004 *Nature* article, Tatem et al. use linear regression to conclude that in the year 2156 the winner of the women's Olympic 100 meter sprint may likely have a faster time than the winner of the men's Olympic 100 meter sprint.

How do the authors make this conclusion?

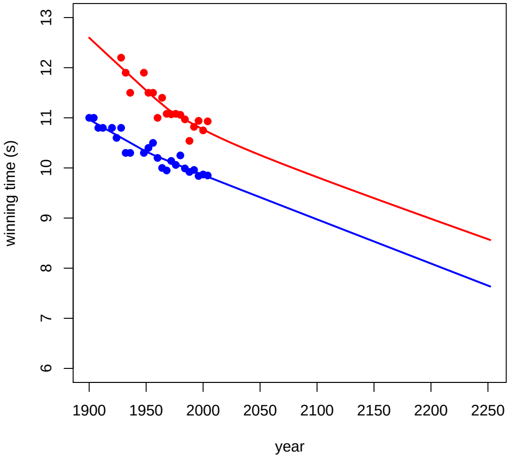
Using data from 1900 to 2004, they fit linear regression models of the winning 100 meter time on year for both men and women. They then use the estimates from these models to extrapolate 152 years into the future.

Tatem et al. Extrapolation

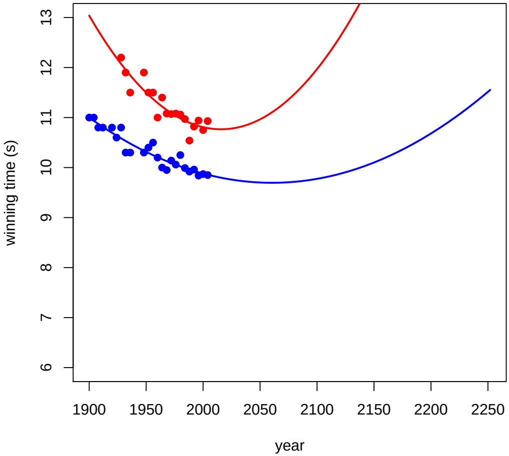


Tatem et al.'s predictions. Men's times are in blue, women's times are in red.

Alternate Models Fit Well, Yield Different Predictions



Alternate Models Fit Well, Yield Different Predictions



The Trouble with Extrapolation

The Trouble with Extrapolation

- The model only gives the best fitting line where we have data, it says little about the shape where there isn't any data.

The Trouble with Extrapolation

- The model only gives the best fitting line where we have data, it says little about the shape where there isn't any data.
- We can always ask illogical questions and the model gives answers.

The Trouble with Extrapolation

- The model only gives the best fitting line where we have data, it says little about the shape where there isn't any data.
- We can always ask illogical questions and the model gives answers.
  - For example, when will women finish the sprint in negative time?

The Trouble with Extrapolation

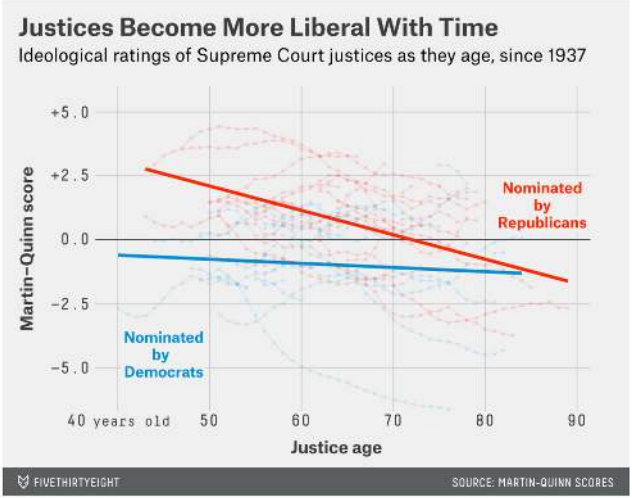
- The model only gives the best fitting line where we have data, it says little about the shape where there isn't any data.
- We can always ask illogical questions and the model gives answers.
  - For example, when will women finish the sprint in negative time?
- Fundamentally we are assuming that data outside the sample looks like data inside the sample, and the further away it is the less likely that is to hold.

The Trouble with Extrapolation

- The model only gives the best fitting line where we have data, it says little about the shape where there isn't any data.
- We can always ask illogical questions and the model gives answers.
  - For example, when will women finish the sprint in negative time?
- Fundamentally we are assuming that data outside the sample looks like data inside the sample, and the further away it is the less likely that is to hold.
- This problem gets much harder in high dimensions

A More Subtle Example

A More Subtle Example



A More Subtle Example

Nate Silver @NateSilver538 · Oct 5

So, basically, John Roberts is going to be Ruth Bader Ginsburg by 2036.

53eig.ht/1GsI2u6

Supreme Court Justices Get More Liberal As They Get Older

The Supreme Court justices are back from vacation. They've picked up their robes from the cleaners — Alito's had a pesky mustard stain — and are ...

Regression as a Causal Model (A Preview)

- Can regression be also used for causal inference?

Regression as a Causal Model (A Preview)

- Can regression be also used for causal inference?
- Answer: A very qualified yes

Regression as a Causal Model (A Preview)

- Can regression be also used for causal inference?
- Answer: A very qualified yes
- For example, can we say that sending that social pressure caused people to vote?

Regression as a Causal Model (A Preview)

- Can regression be also used for causal inference?
- Answer: A very qualified yes
- For example, can we say that sending that social pressure caused people to vote?
- To interpret β as a causal effect of X on Y, we need very specific and often unrealistic assumptions:

Regression as a Causal Model (A Preview)

- Can regression be also used for causal inference?
- Answer: A very qualified yes
- For example, can we say that sending that social pressure caused people to vote?
- To interpret β as a causal effect of X on Y, we need very specific and often unrealistic assumptions:
  (1) $E[Y|X]$ is correctly specified as a linear function (linearity)

Regression as a Causal Model (A Preview)

- Can regression be also used for causal inference?
- Answer: A very qualified yes
- For example, can we say that sending that social pressure caused people to vote?
- To interpret β as a causal effect of X on Y, we need very specific and often unrealistic assumptions:
  (1) $E[Y|X]$ is correctly specified as a linear function (linearity)
  (2) There are no other variables that affect both X and Y (exogeneity)

Regression as a Causal Model (A Preview)

- Can regression be also used for causal inference?
- Answer: A very qualified yes
- For example, can we say that sending that social pressure caused people to vote?
- To interpret β as a causal effect of X on Y, we need very specific and often unrealistic assumptions:
  (1) $E[Y|X]$ is correctly specified as a linear function (linearity)
  (2) There are no other variables that affect both X and Y (exogeneity)

  (1) can be relaxed by:

Regression as a Causal Model (A Preview)

- Can regression be also used for causal inference?
- Answer: A very qualified yes
- For example, can we say that sending that social pressure caused people to vote?
- To interpret β as a causal effect of X on Y, we need very specific and often unrealistic assumptions:
  (1) $E[Y|X]$ is correctly specified as a linear function (linearity)
  (2) There are no other variables that affect both X and Y (exogeneity)

  (1) can be relaxed by:
    - ⋆ Using a flexible nonlinear or nonparametric method

Regression as a Causal Model (A Preview)

- Can regression be also used for causal inference?
- Answer: A very qualified yes
- For example, can we say that sending that social pressure caused people to vote?
- To interpret β as a causal effect of X on Y, we need very specific and often unrealistic assumptions:
  (1) $E[Y|X]$ is correctly specified as a linear function (linearity)
  (2) There are no other variables that affect both X and Y (exogeneity)

  (1) can be relaxed by:
    - Using a flexible nonlinear or nonparametric method
    - "Preprocessing" data to make analysis robust to misspecification

Regression as a Causal Model (A Preview)

- Can regression be also used for causal inference?
- Answer: A very qualified yes
- For example, can we say that sending that social pressure caused people to vote?
- To interpret β as a causal effect of X on Y, we need very specific and often unrealistic assumptions:
  (1) $E[Y|X]$ is correctly specified as a linear function (linearity)
  (2) There are no other variables that affect both X and Y (exogeneity)

  (1) can be relaxed by:
    - Using a flexible nonlinear or nonparametric method
    - "Preprocessing" data to make analysis robust to misspecification

  (2) can be made plausible by:

Regression as a Causal Model (A Preview)

- Can regression be also used for causal inference?
- Answer: A very qualified yes
- For example, can we say that sending that social pressure caused people to vote?
- To interpret β as a causal effect of X on Y, we need very specific and often unrealistic assumptions:
  (1) $E[Y|X]$ is correctly specified as a linear function (linearity)
  (2) There are no other variables that affect both X and Y (exogeneity)

  (1) can be relaxed by:
    - Using a flexible nonlinear or nonparametric method
    - "Preprocessing" data to make analysis robust to misspecification

  (2) can be made plausible by:
    - Including carefully-selected control variables in the model

Regression as a Causal Model (A Preview)

- Can regression be also used for causal inference?
- Answer: A very qualified yes
- For example, can we say that sending that social pressure caused people to vote?
- To interpret β as a causal effect of X on Y, we need very specific and often unrealistic assumptions:
  (1) $E[Y|X]$ is correctly specified as a linear function (linearity)
  (2) There are no other variables that affect both X and Y (exogeneity)

  (1) can be relaxed by:
    - Using a flexible nonlinear or nonparametric method
    - "Preprocessing" data to make analysis robust to misspecification

  (2) can be made plausible by:
    - Including carefully-selected control variables in the model
    - Choosing a clever research design to rule out confounding

Regression as a Causal Model (A Preview)

- Can regression be also used for causal inference?
- Answer: A very qualified yes
- For example, can we say that sending that social pressure caused people to vote?
- To interpret β as a causal effect of X on Y, we need very specific and often unrealistic assumptions:
  (1) $E[Y|X]$ is correctly specified as a linear function (linearity)
  (2) There are no other variables that affect both X and Y (exogeneity)

  (1) can be relaxed by:
    - ⋆ Using a flexible nonlinear or nonparametric method
    - ⋆ "Preprocessing" data to make analysis robust to misspecification

  (2) can be made plausible by:
    - ⋆ Including carefully-selected control variables in the model
    - ⋆ Choosing a clever research design to rule out confounding

Regression as a Causal Model (A Preview)

- Can regression be also used for causal inference?
- Answer: A very qualified yes
- For example, can we say that sending that social pressure caused people to vote?
- To interpret β as a causal effect of X on Y, we need very specific and often unrealistic assumptions:
  (1) $E[Y|X]$ is correctly specified as a linear function (linearity)
  (2) There are no other variables that affect both X and Y (exogeneity)

  (1) can be relaxed by:
    - Using a flexible nonlinear or nonparametric method
    - "Preprocessing" data to make analysis robust to misspecification

  (2) can be made plausible by:
    - Including carefully-selected control variables in the model
    - Choosing a clever research design to rule out confounding
- We will return to this later in the course

Regression as a Causal Model (A Preview)

- Can regression be also used for causal inference?
- Answer: A very qualified yes
- For example, can we say that sending that social pressure caused people to vote?
- To interpret β as a causal effect of X on Y, we need very specific and often unrealistic assumptions:
  1. $E[Y|X]$ is correctly specified as a linear function (linearity)
  2. There are no other variables that affect both X and Y (exogeneity)

  (1) can be relaxed by:
    - ⋆ Using a flexible nonlinear or nonparametric method
    - ⋆ "Preprocessing" data to make analysis robust to misspecification

  (2) can be made plausible by:
    - ⋆ Including carefully-selected control variables in the model
    - ⋆ Choosing a clever research design to rule out confounding
- We will return to this later in the course
- For now, it is safest to treat β as a purely descriptive/predictive quantity

Fun with Linearity

"The Siren's Song of Linearity"

Fun with Linearity

Psychonomic Bulletin & Review
2007, 14 (2), 288-294

Iterated learning: Intergenerational knowledge transmission reveals inductive biases

Michael L. Kalish
University of Louisiana, Lafayette, Louisiana

Thomas L. Griffiths
University of California, Berkeley, California

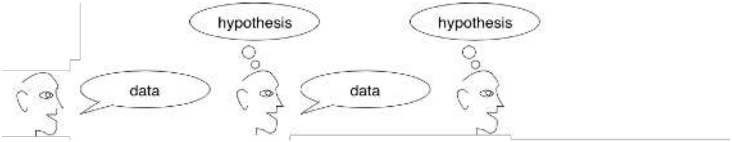
and

Stephan Lewandowsky
University of Western Australia, Perth, Australia

Cultural transmission of information plays a central role in shaping human knowledge. Some of the most complex knowledge that people acquire, such as languages or cultural norms, can only be learned from other people, who themselves learned from previous generations. The prevalence of this process of "iterated learning" as a mode of cultural transmission raises the question of how it affects the information being transmitted. Analyses of iterated learning utilizing the assumption that the learners are Bayesian agents predict that this process should converge to an equilibrium that reflects the inductive biases of the learners. An experiment in iterated function learning with human participants confirmed this prediction, providing insight into the consequences of intergenerational knowledge transmission and a method for discovering the inductive biases that guide human inferences.

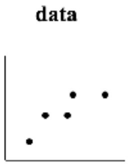
Images on following slides courtesy of Tom Griffiths

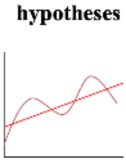
Fun with Linearity



The Design

The Design

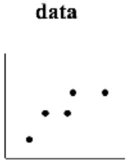


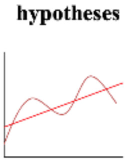


The Design

- Each learner sees a set of (x, y) pairs

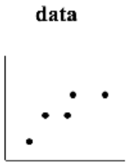
The Design

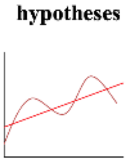




- Each learner sees a set of (x, y) pairs
- Makes predictions of y for new x values

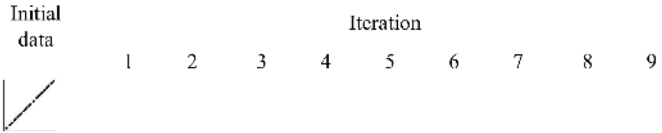
The Design



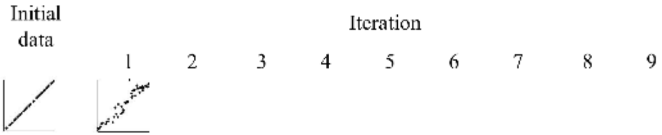


- Each learner sees a set of (x, y) pairs
- Makes predictions of y for new x values
- Predictions are data for the next learner

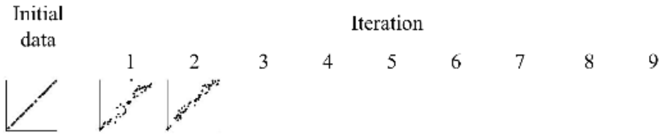
Results



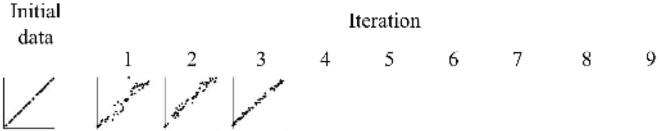
Results



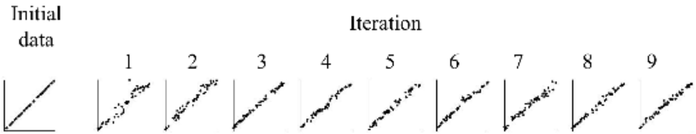
Results



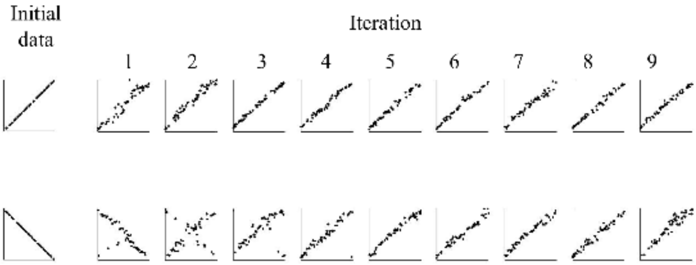
Results



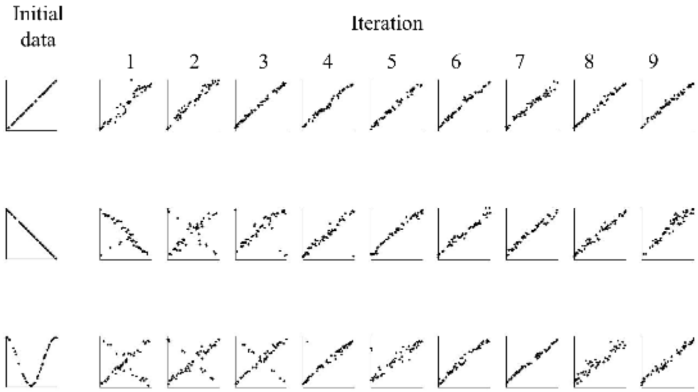
Results



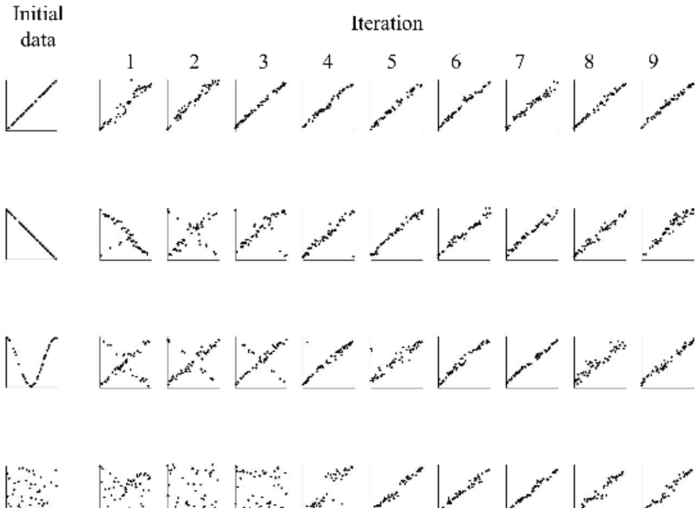
Results



Results



Results



This Week in Review

- Hypothesis Testing!
- P-Values!
- What is Regression?
- Interpreting Regression!

This Week in Review

- Hypothesis Testing!
- P-Values!
- What is Regression?
- Interpreting Regression!

Going Deeper:

Aronow and Miller (2019) *Foundations of Agnostic Statistics*. Cambridge University Press. Chapter 4.

This Week in Review

- Hypothesis Testing!
- P-Values!
- What is Regression?
- Interpreting Regression!

Going Deeper:

Aronow and Miller (2019) *Foundations of Agnostic Statistics*. Cambridge University Press. Chapter 4.

Next week: Properties of Linear Regression with One Explanatory Variable.