Week 6: Estimating Conditional Expectation Functions

Brandon Stewart[1]

Princeton

October 8–10, 2024

[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
  - causal inference
  - stratification

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

- Last Week
  - causal inference
  - stratification
- This Week

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

- Last Week
  - causal inference
  - stratification
- This Week
  - what is regression?

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

- Last Week
  - causal inference
  - stratification
- This Week
  - what is regression?
  - non-parametric and linear regression

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

- Last Week
  - causal inference
  - stratification
- This Week
  - what is regression?
  - non-parametric and linear regression
- Next Week
  - finite-sample and asymptotic theory for regression

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

- Last Week
  - causal inference
  - stratification
- This Week
  - what is regression?
  - non-parametric and linear regression
- Next Week
  - finite-sample and asymptotic theory for regression
- Long Run
  - probability $\rightarrow$ inference $\rightarrow$ regression $\rightarrow$ causal inference

1. What is Regression?
2. Best Linear Predictor
3. Fun With Linearity
4. Estimation of the BLP
5. Additional Techniques
   - R^2
   - LOESS
   - Log Transformations
   - Bootstrap

1. What is Regression?
2. Best Linear Predictor
3. Fun With Linearity
4. Estimation of the BLP
5. Additional Techniques
   - R^2
   - LOESS
   - Log Transformations
   - Bootstrap

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 more assumptions we can talk about how intervening to change the value of X will change Y.

The Conditional Expectation Function

Recall, a conditional expectation, for random variables X and Y:
Discrete with joint PMF f:

$$E[Y \mid X = x] = \sum_y y f_{Y|X}(y|x), \forall x \in \text{Supp}(X)$$

Continuous with joint PDF f:

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

The Conditional Expectation Function

Then define the conditional expectation function as:

The Conditional Expectation Function

Then define the conditional expectation function as:

For random variables X and Y with joint PMF/PDF f, the conditional expectation function of Y given $X = x$ is:

$$\mu(\mathbf{x}) = E[Y \mid X = x], \forall x \in \text{Supp}(X)$$

The Conditional Expectation Function

Then define the conditional expectation function as:

For random variables X and Y with joint PMF/PDF f, the conditional expectation function of Y given $X = x$ is:

$$\mu(\mathbf{x}) = E[Y \mid X = x], \forall x \in \text{Supp}(X)$$

Which is a function that that maps x to $E[Y \mid X = x]$ (not simply the value of $E[Y \mid X = x]$ at some specific x).

The Conditional Expectation Function

Then define the conditional expectation function as:

For random variables X and Y with joint PMF/PDF f, the conditional expectation function of Y given $X = x$ is:

$$\mu(\mathbf{x}) = E[Y \mid X = x], \forall x \in \text{Supp}(X)$$

Which is a function that that maps x to $E[Y \mid X = x]$ (not simply the value of $E[Y \mid X = x]$ at some specific x).

It is a univariate function of x, not a function of the random variable Y.

The Conditional Expectation Function

Then define the conditional expectation function as:

For random variables X and Y with joint PMF/PDF f, the conditional expectation function of Y given $X = x$ is:

$$\mu(\mathbf{x}) = E[Y \mid X = x], \forall x \in \text{Supp}(X)$$

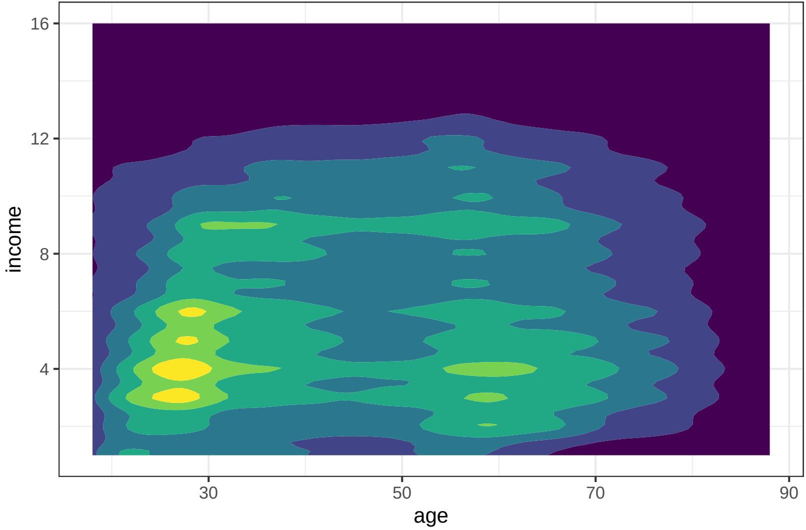
Which is a function that that maps x to $E[Y \mid X = x]$ (not simply the value of $E[Y \mid X = x]$ at some specific x).

It is a univariate function of x, not a function of the random variable Y.

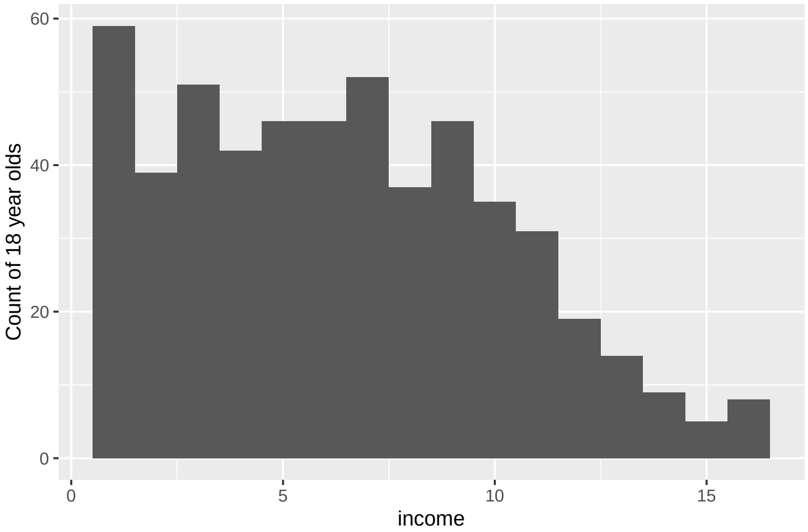
We will generally denote it $E[Y \mid X]$ or $\mu(\mathbf{x})$.

Review of CEF Concepts

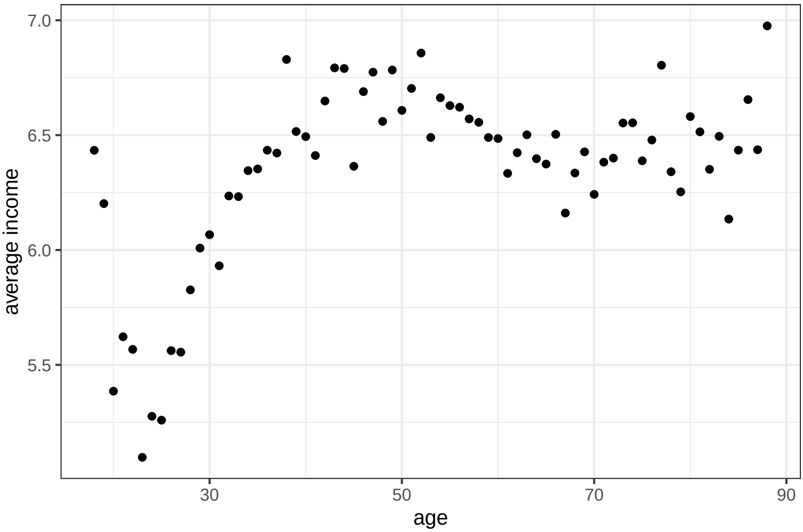
Review of CEF Concepts



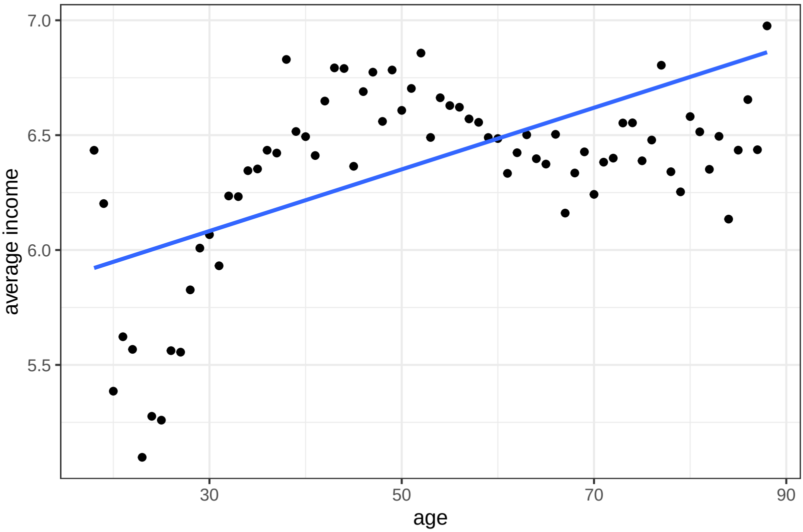
Review of CEF Concepts



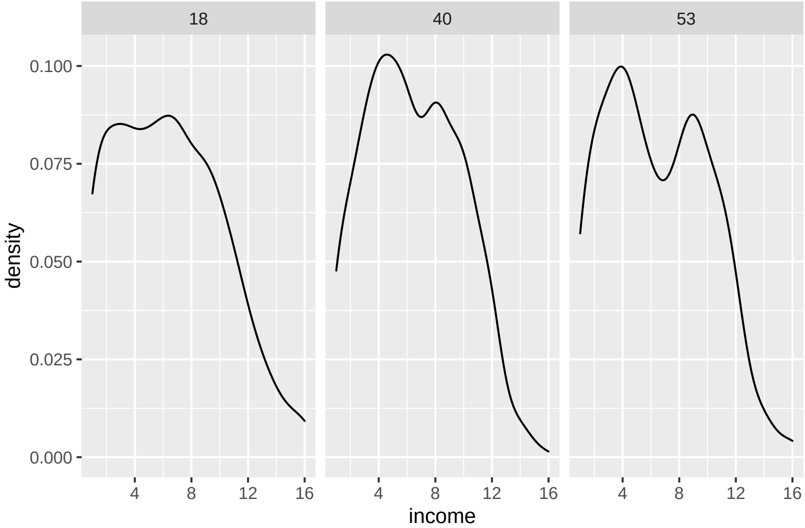
Review of CEF Concepts



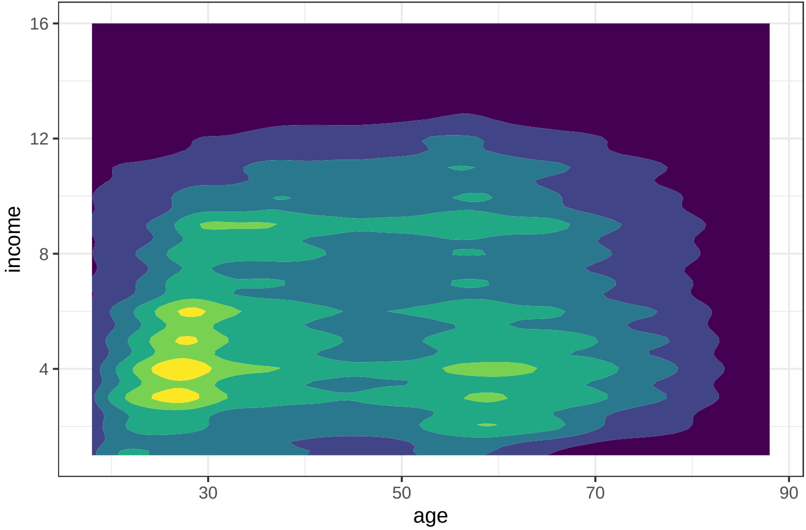
Review of CEF Concepts



Review of CEF Concepts



Review of CEF Concepts



Properties of the CEF

The CEF is very powerful. Let X and Y be random variables and let $\epsilon = Y - E[Y \mid X]$, then:

1. Average deviation at any value of X is zero: $E[\epsilon \mid X] = 0$
    - By linearity in expectations

Properties of the CEF

The CEF is very powerful. Let X and Y be random variables and let $\epsilon = Y - E[Y \mid X]$, then:

1. Average deviation at any value of X is zero: $E[\epsilon \mid X] = 0$
   - By linearity in expectations
2. Average deviation is zero: $E[\epsilon] = 0$
   - By the Law of Iterated Expectations

Properties of the CEF

The CEF is very powerful. Let X and Y be random variables and let $\epsilon = Y - E[Y \mid X]$, then:

1. Average deviation at any value of X is zero: $E[\epsilon \mid X] = 0$
   - By linearity in expectations
2. Average deviation is zero: $E[\epsilon] = 0$
   - By the Law of Iterated Expectations
3. Orthogonal to X (and functions thereof): If g is a function of X, then $\text{Cov}\big[g(X), \epsilon\big] = 0$
   - By def. of covariance and property (2)

Properties of the CEF

The CEF is very powerful. Let X and Y be random variables and let $\epsilon = Y - E[Y \mid X]$, then:

1. Average deviation at any value of X is zero: $E[\epsilon \mid X] = 0$
   - By linearity in expectations
2. Average deviation is zero: $E[\epsilon] = 0$
   - By the Law of Iterated Expectations
3. Orthogonal to X (and functions thereof): If g is a function of X, then $\text{Cov}\big[g(X), \epsilon\big] = 0$
   - By def. of covariance and property (2)
4. The conditional variance of deviations is the conditional variance of Y: $V[\epsilon \mid X] = V[Y \mid X]$
   - By def. of variance and property (1)

Properties of the CEF

The CEF is very powerful. Let X and Y be random variables and let $\epsilon = Y - E[Y \mid X]$, then:

1. Average deviation at any value of X is zero: $E[\epsilon \mid X] = 0$
   - By linearity in expectations
2. Average deviation is zero: $E[\epsilon] = 0$
   - By the Law of Iterated Expectations
3. Orthogonal to X (and functions thereof): If g is a function of X, then $\text{Cov}\big[g(X), \epsilon\big] = 0$
   - By def. of covariance and property (2)
4. The conditional variance of deviations is the conditional variance of Y: $V[\epsilon \mid X] = V[Y \mid X]$
   - By def. of variance and property (1)
5. $V[\epsilon] = E\big[V[Y \mid X]\big]$
   - By the Law of Total Variance

Properties of the CEF

Most importantly:

- The CEF is the best (lowest MSE) predictor of Y given X

Properties of the CEF

Most importantly:

- The CEF is the best (lowest MSE) predictor of Y given X
  - The CEF yields the best approximation of Y conditional on the observed value of X.

Properties of the CEF

Most importantly:

- The CEF is the best (lowest MSE) predictor of Y given X
  - The CEF yields the best approximation of Y conditional on the observed value of X.
  - There is no better way (in terms of MSE) to approximate $Y \mid X$ than with the CEF

Properties of the CEF

Most importantly:

- The CEF is the best (lowest MSE) predictor of Y given X
  - The CEF yields the best approximation of Y conditional on the observed value of X.
  - There is no better way (in terms of MSE) to approximate $Y \mid X$ than with the CEF
  - If we know the CEF, we know *a lot* about how X relates to Y

Properties of the CEF

Most importantly:

- The CEF is the best (lowest MSE) predictor of Y given X
  - The CEF yields the best approximation of Y conditional on the observed value of X.
  - There is no better way (in terms of MSE) to approximate $Y \mid X$ than with the CEF
  - If we know the CEF, we know *a lot* about how X relates to Y
  - But, just because it is the best predictor doesn't mean that it is simple
    - ⋆ It could take on *any* shape!

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:
  - $\hat{\mu}(1) = \hat{E}[Y|X = 1] = \frac{1}{n_{x=1}} \sum_{i:X_i=1} Y_i$
  - $\hat{\mu}(2) = \hat{E}[Y|X = 2] = \frac{1}{n_{x=2}} \sum_{i:X_i=2} Y_i$
  - $\ldots$
- 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:
  - $\hat{\mu}(1) = \hat{E}[Y|X = 1] = \frac{1}{n_{x=1}} \sum_{i:X_i=1} Y_i$
  - $\hat{\mu}(2) = \hat{E}[Y|X = 2] = \frac{1}{n_{x=2}} \sum_{i:X_i=2} Y_i$
  - ...
- 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

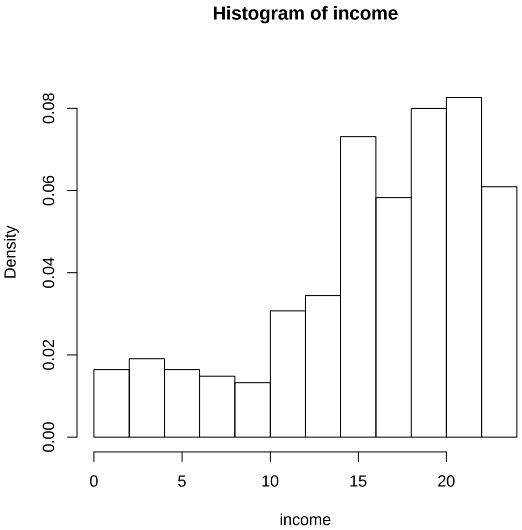
$\vdots$

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

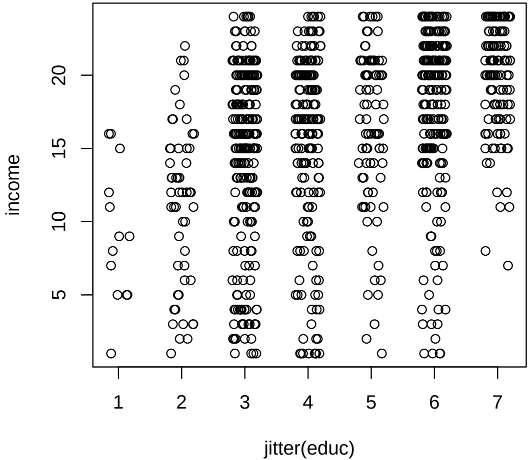
$\vdots$

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

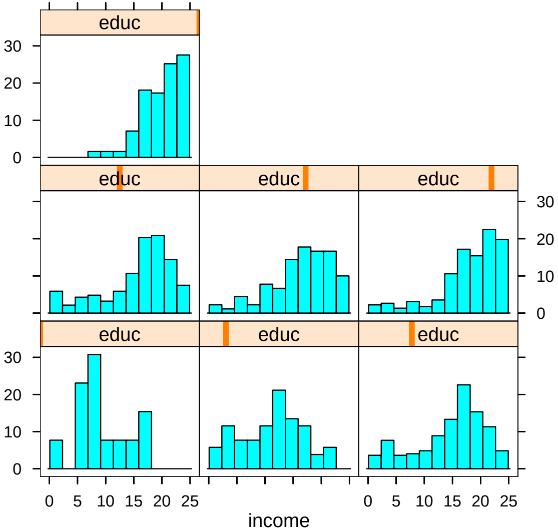
Marginal Distribution of Y (income)



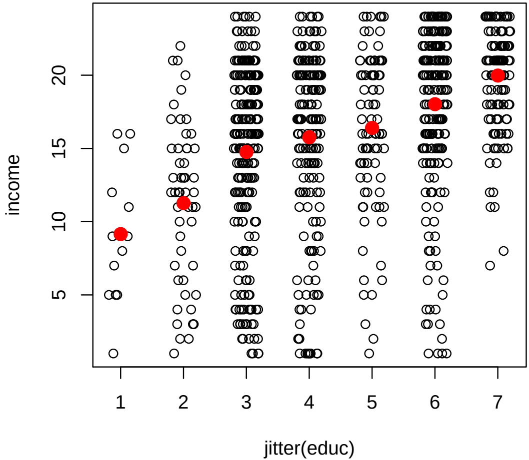
Joint Distribution of X and Y (Income and Education)



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

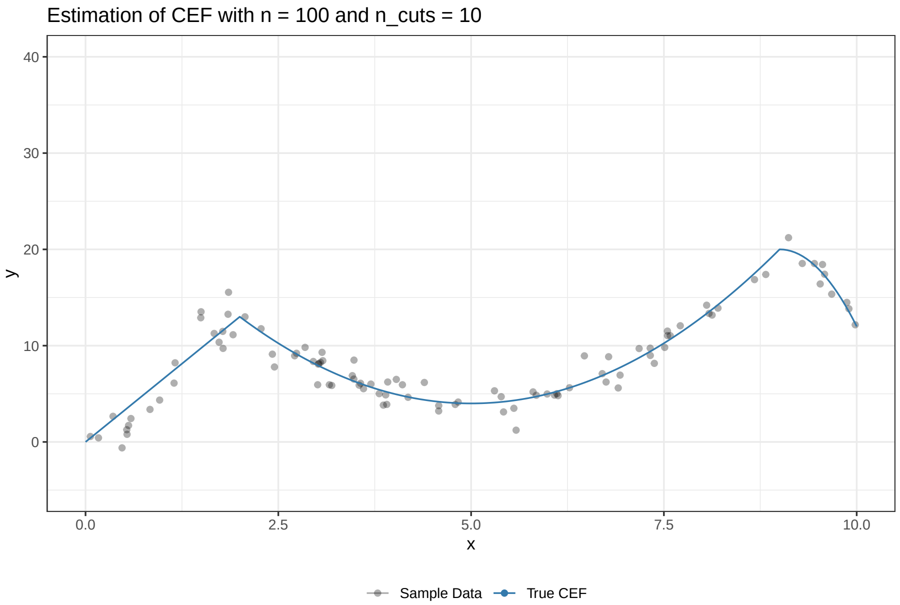
Nonparametric Regression

- 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?

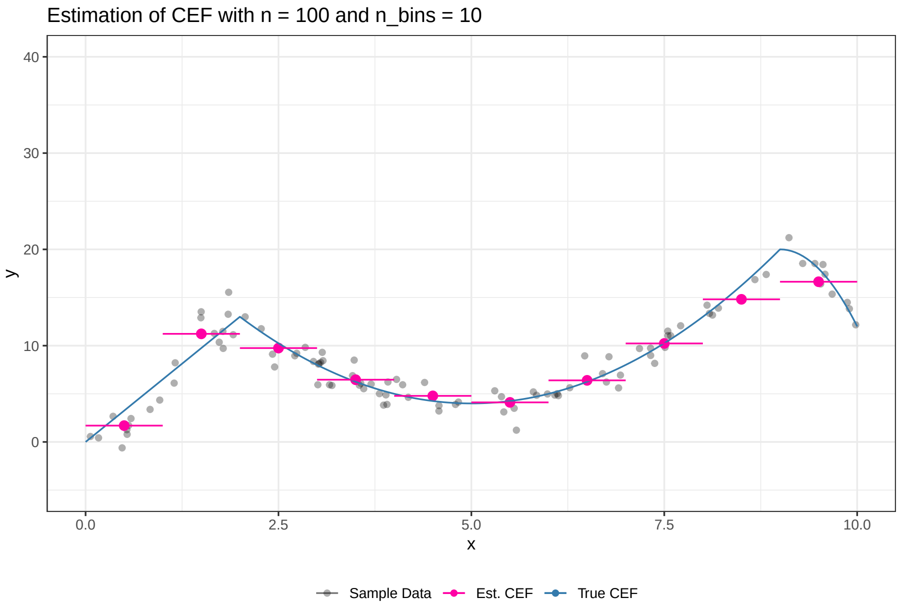
Nonparametric Regression

- 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.

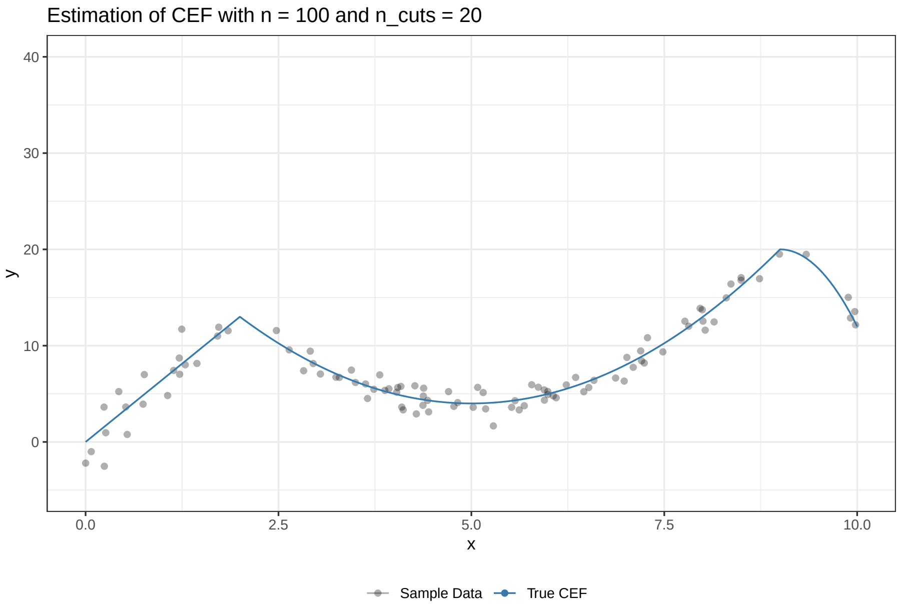
A Binning Approach to CEF Estimation



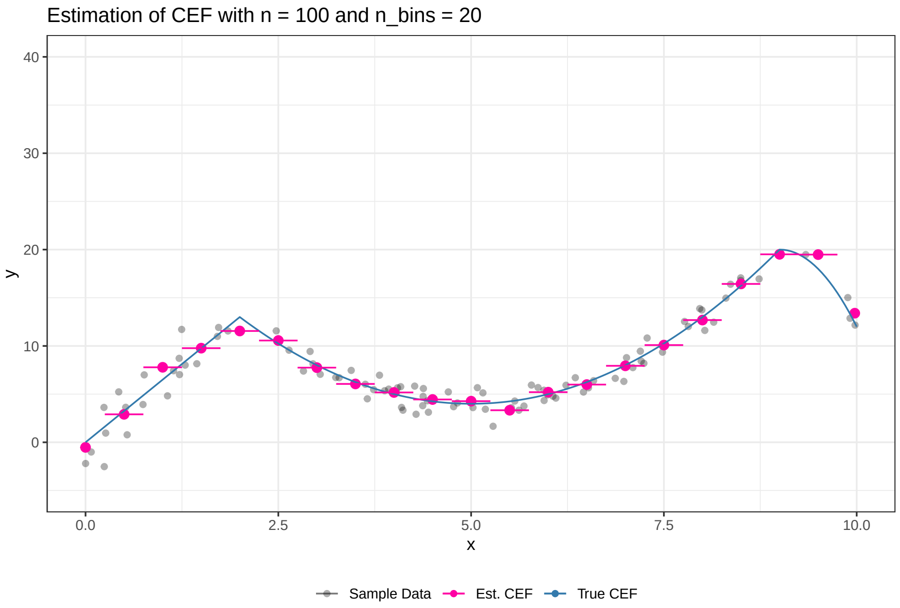
A Binning Approach to CEF Estimation



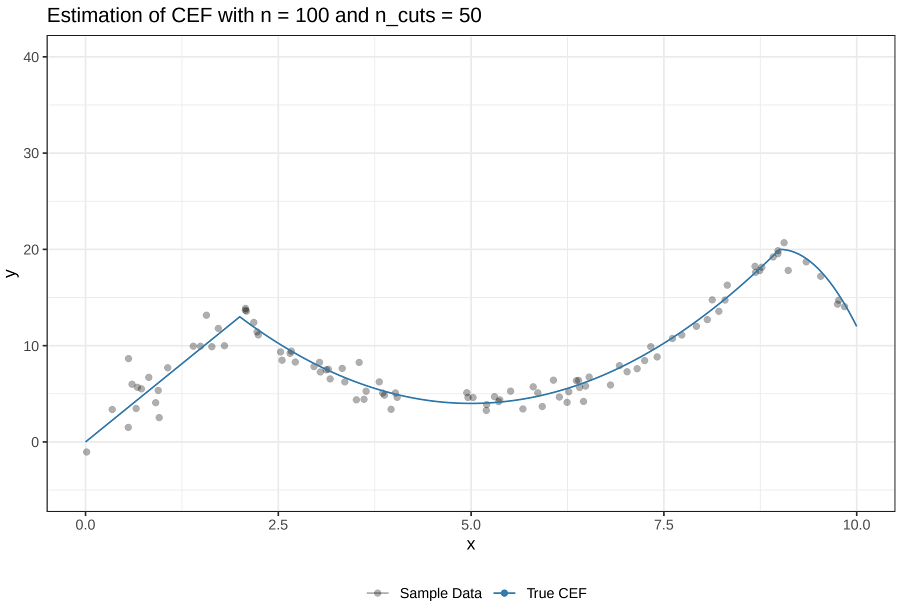
A Binning Approach to CEF Estimation



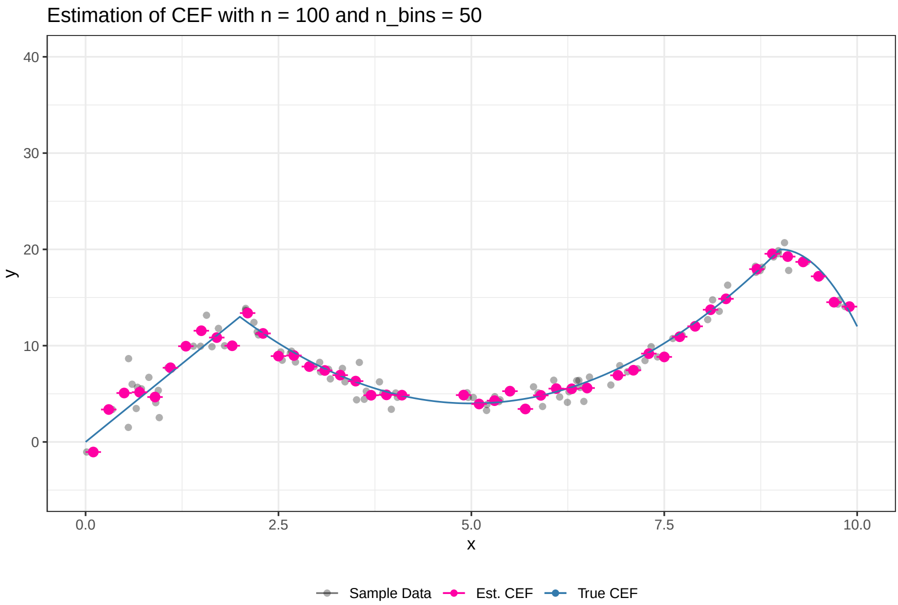
A Binning Approach to CEF Estimation



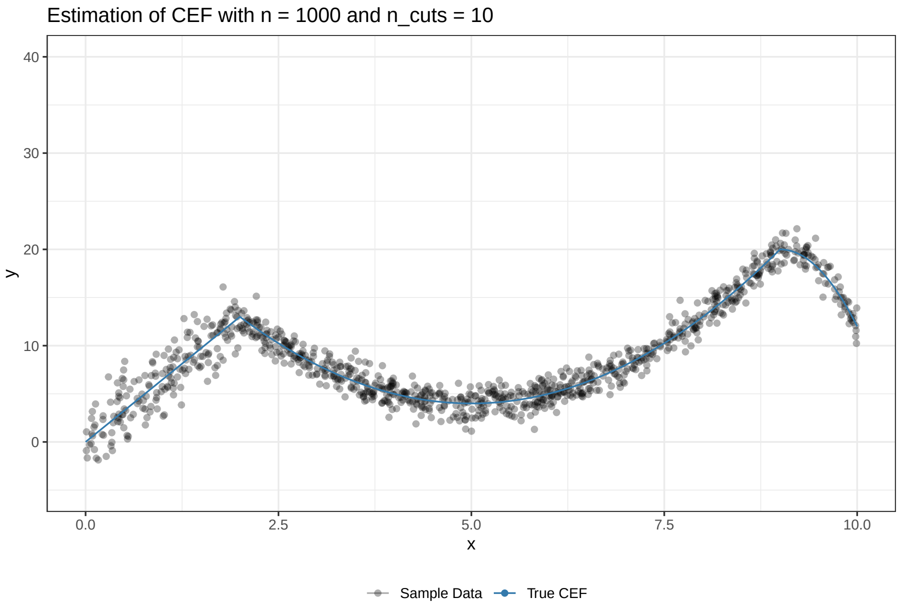
A Binning Approach to CEF Estimation



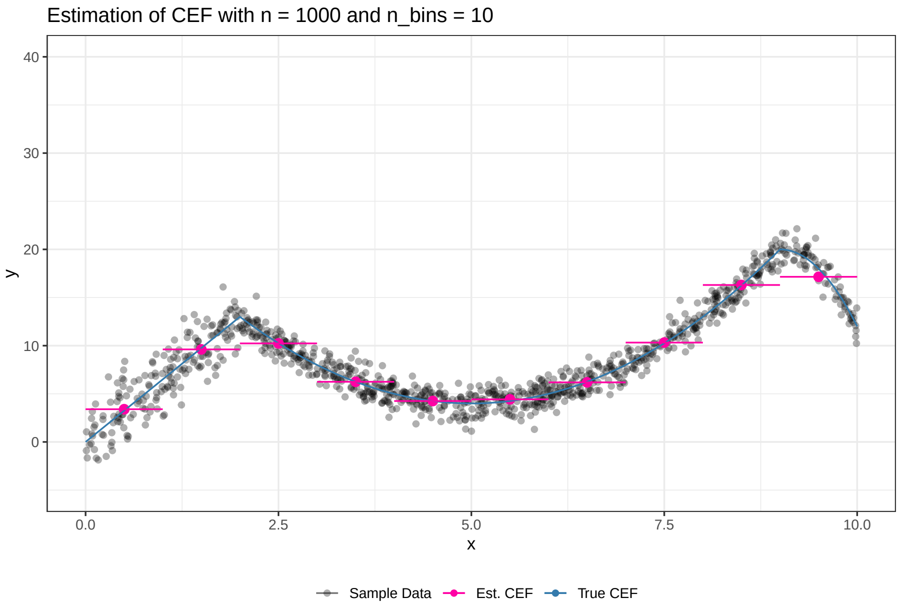
A Binning Approach to CEF Estimation



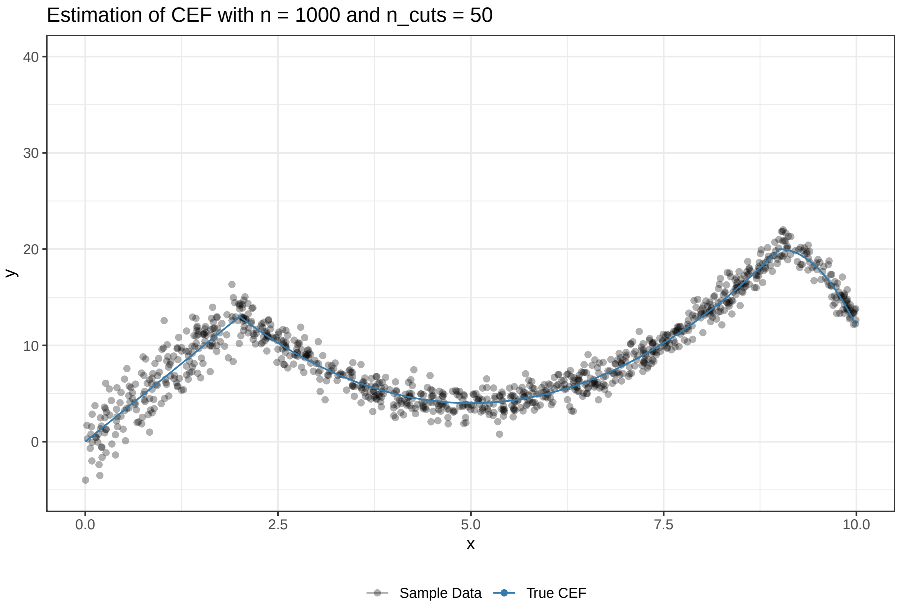
A Binning Approach to CEF Estimation



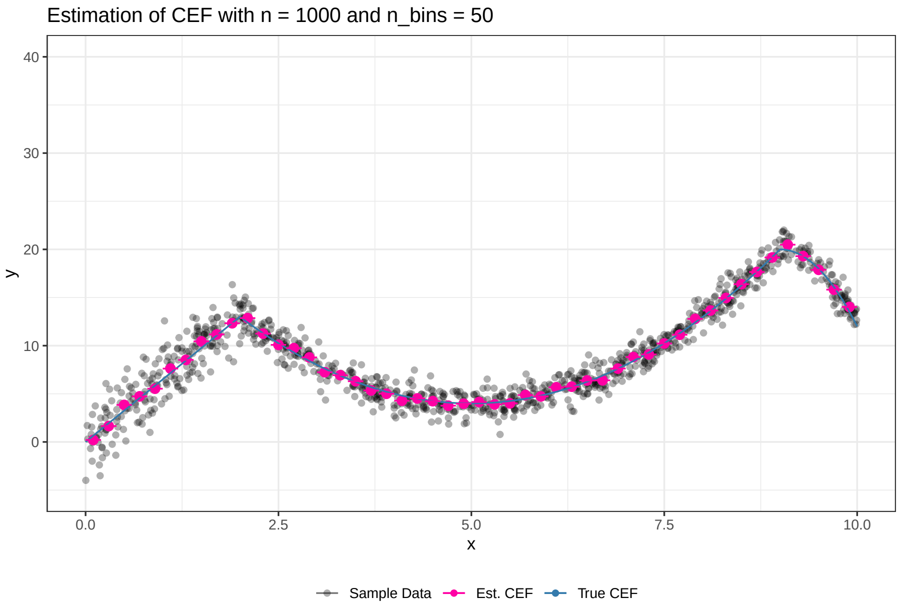
A Binning Approach to CEF Estimation



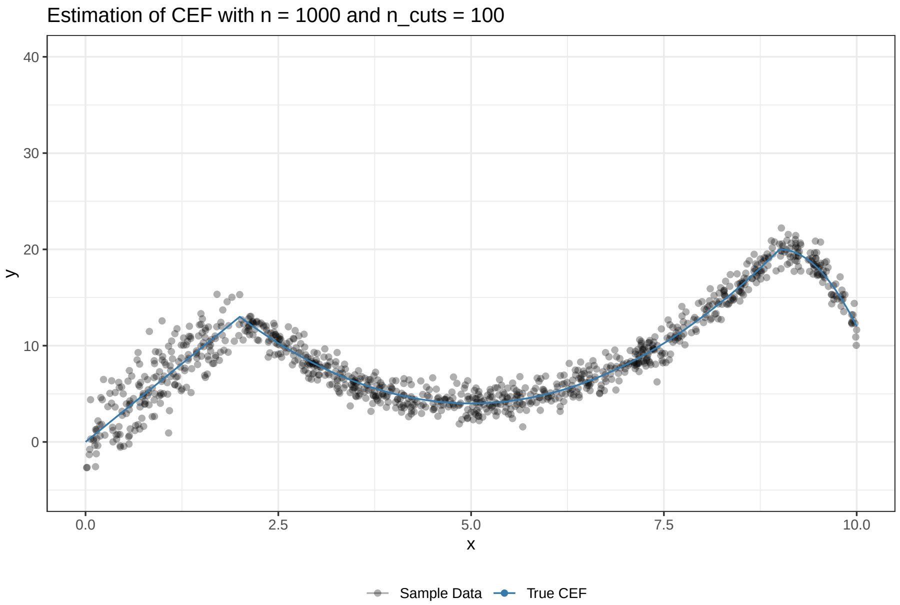
A Binning Approach to CEF Estimation



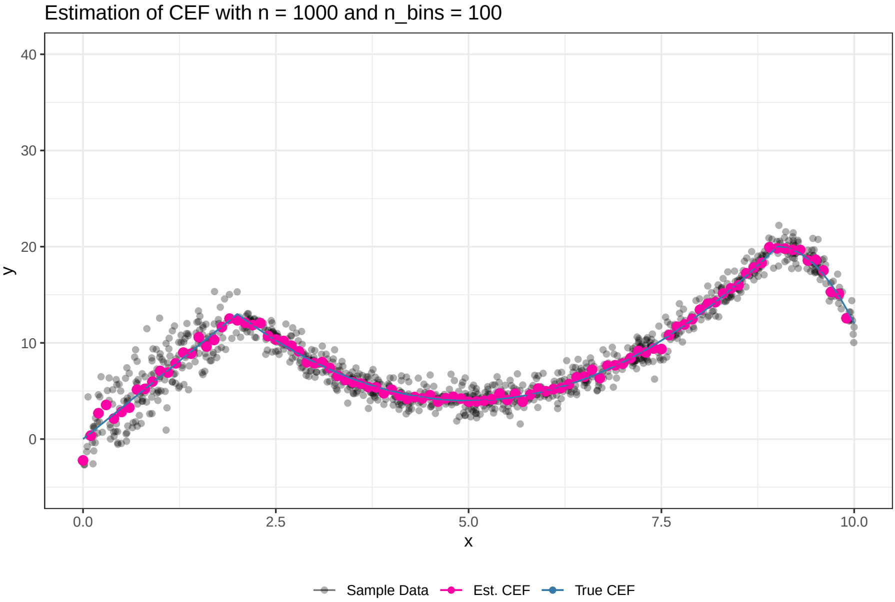
A Binning Approach to CEF Estimation



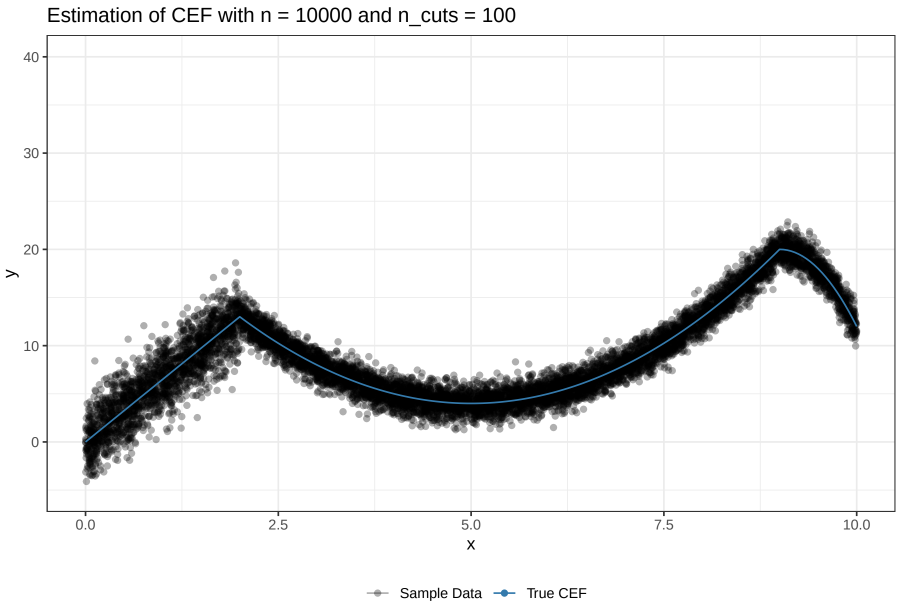
A Binning Approach to CEF Estimation



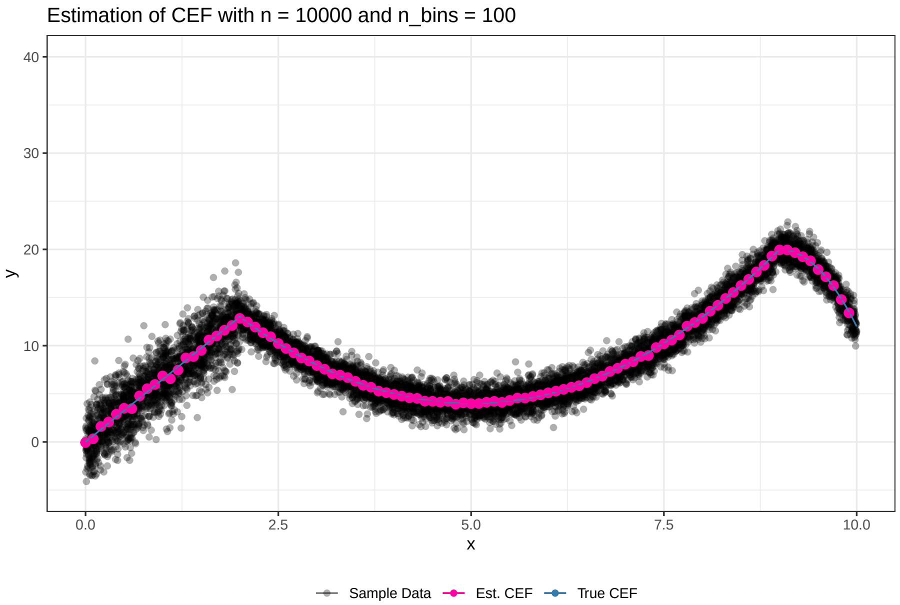
A Binning Approach to CEF Estimation



A Binning Approach to CEF Estimation



A Binning Approach to CEF Estimation



Uniform Kernel Regression: Simple Local Averages

Uniform Kernel Regression: Simple Local Averages

- Dividing into discrete bins can get pretty noisy.

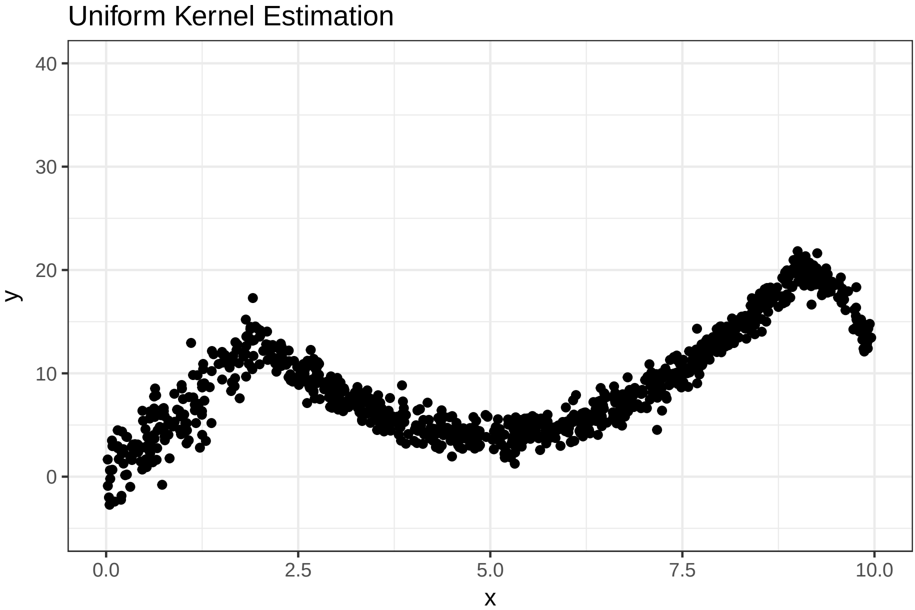
Uniform Kernel Regression: Simple Local Averages

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

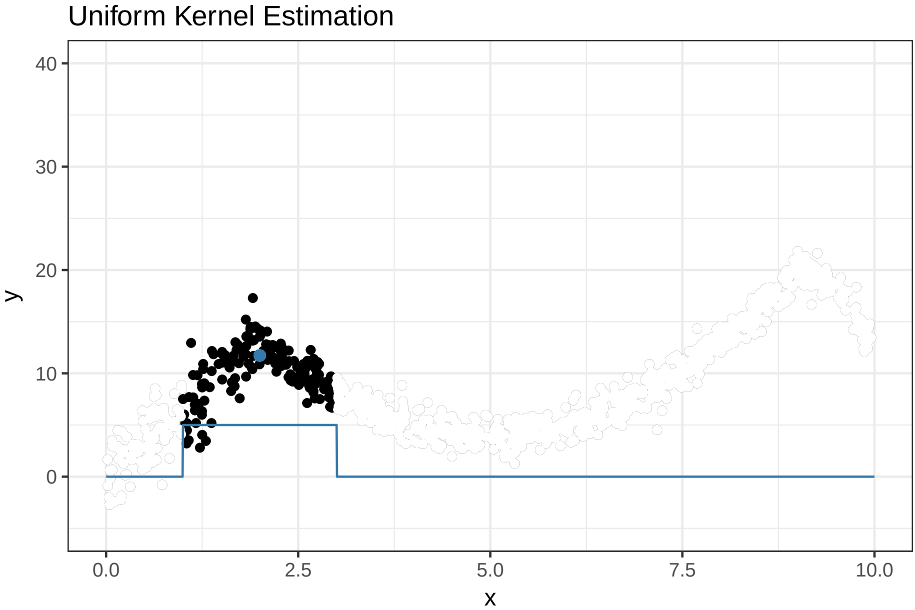
Uniform Kernel Regression: Simple Local Averages

- 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.

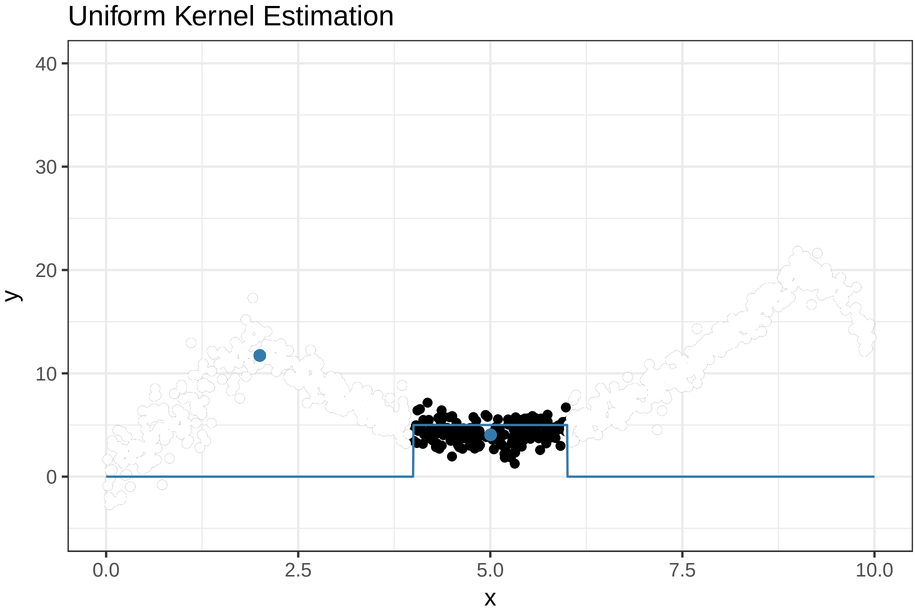
Example of Kernel CEF Estimation



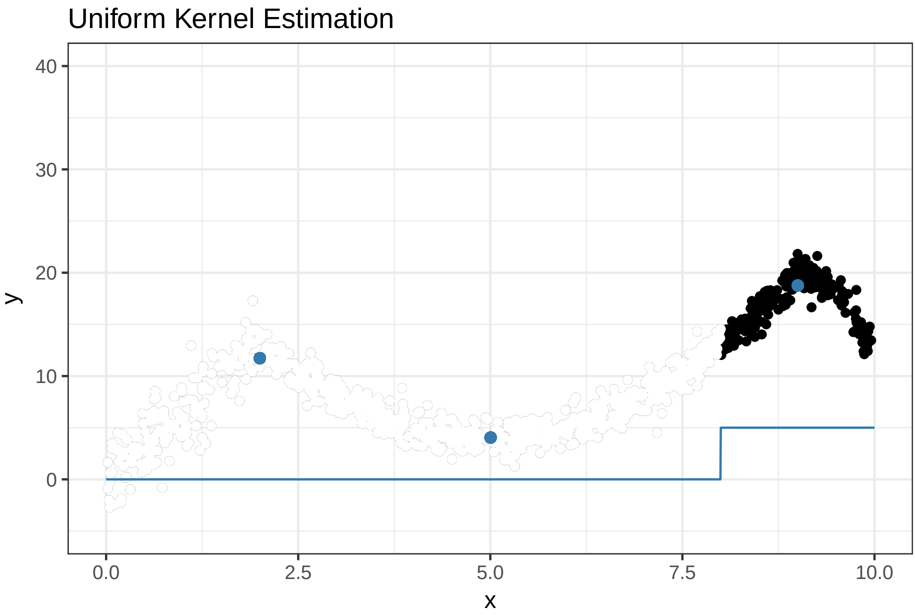
Example of Kernel CEF Estimation



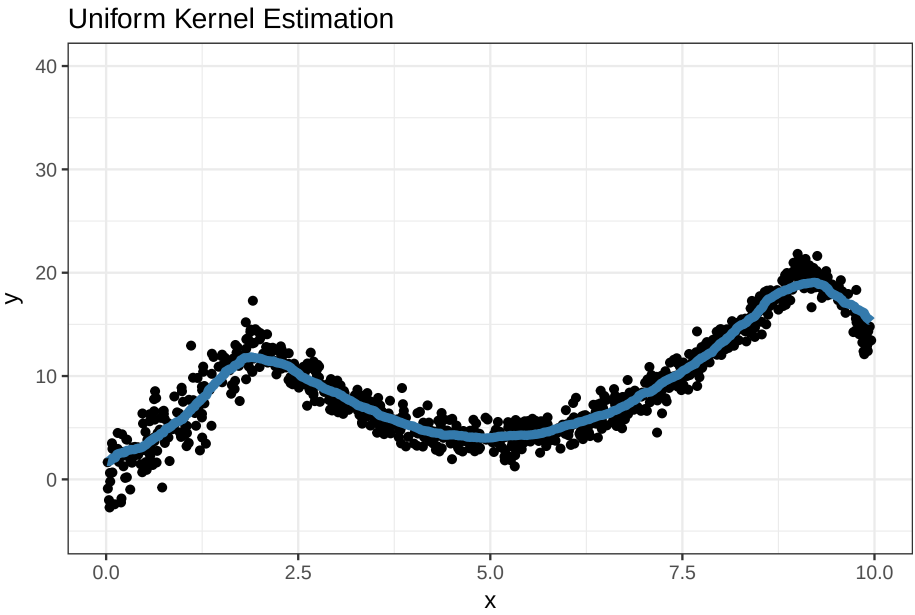
Example of Kernel CEF Estimation



Example of Kernel CEF Estimation



Example of Kernel CEF Estimation



Uniform Kernel Regression: Simple Local Averages

Uniform Kernel Regression: Simple Local Averages

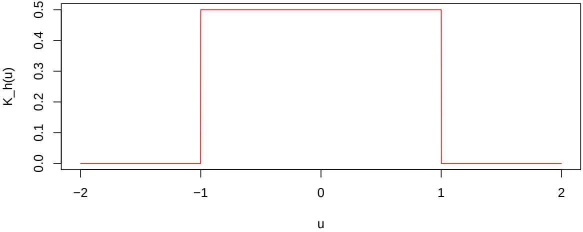
- Calculate the average of the observed y points that have x values in the interval $[x_0 - h\,,\, x_0 + h]$

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)

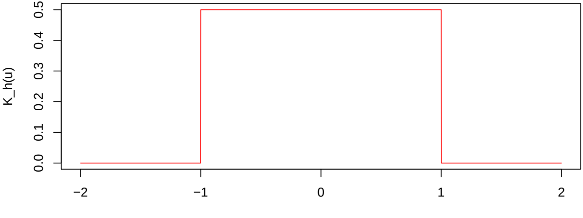
Uniform Kernel Regression: Simple Local Averages

- 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

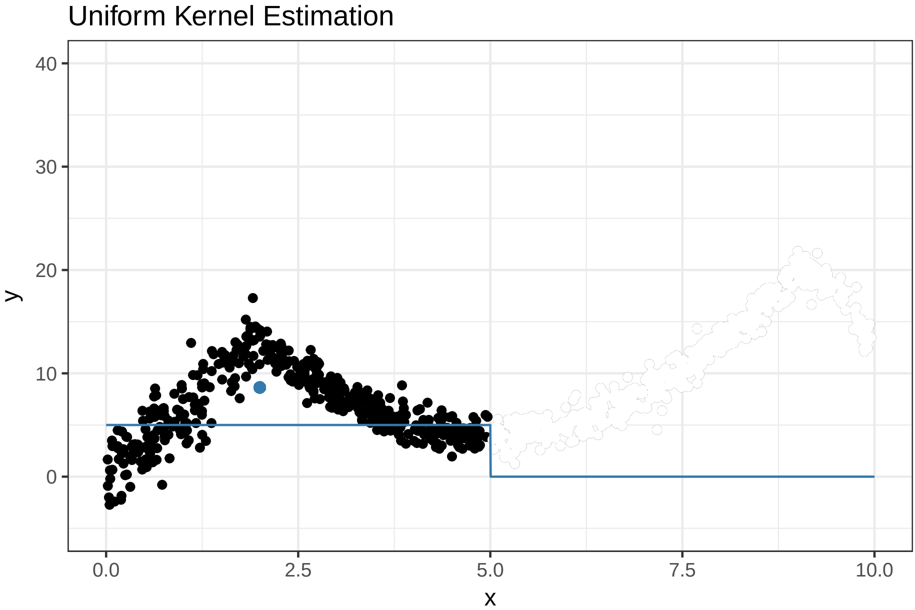
- Calculate the average of the observed y points that have x values in the interval $[x_0 - h\,,\, x_0 + h]$
- $h =$ some positive number (called the bandwidth)
- Uniform kernel: every observation in the interval is equally weighted



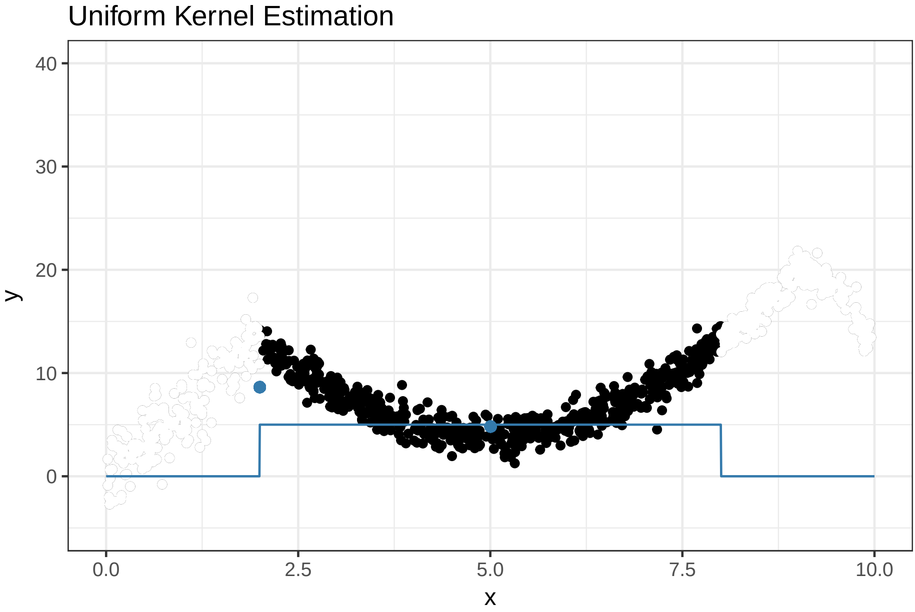
- This gives the uniform kernel regression:

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

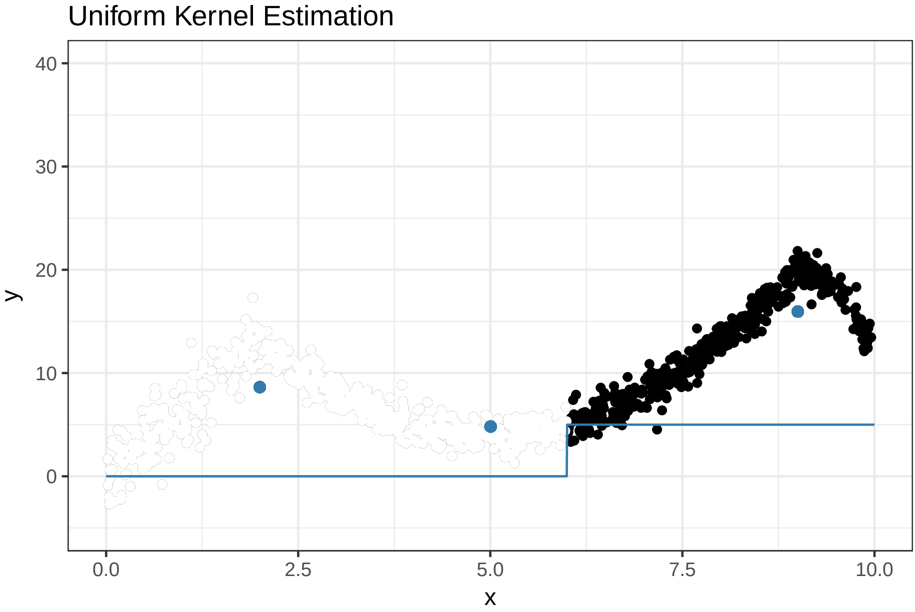
Changing the Bandwidth



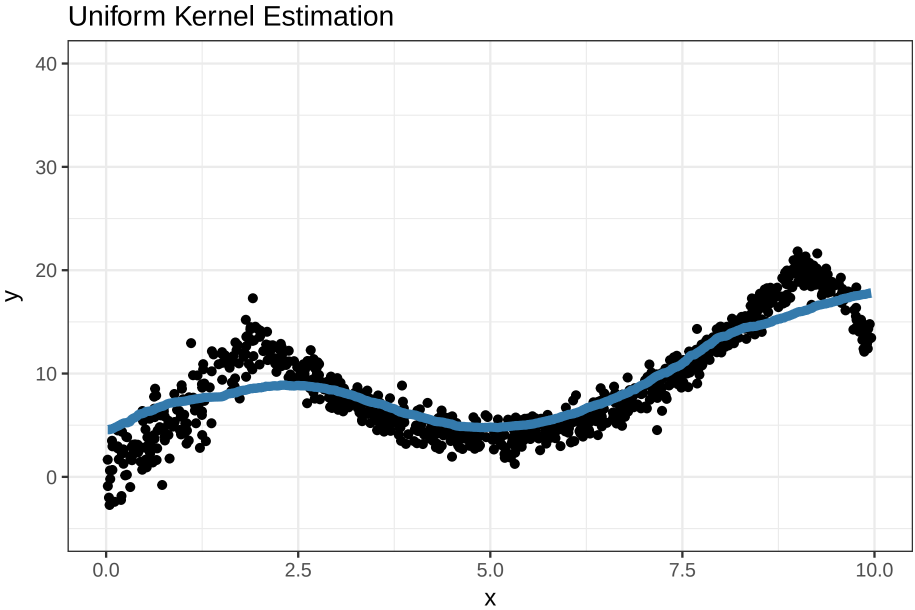
Changing the Bandwidth



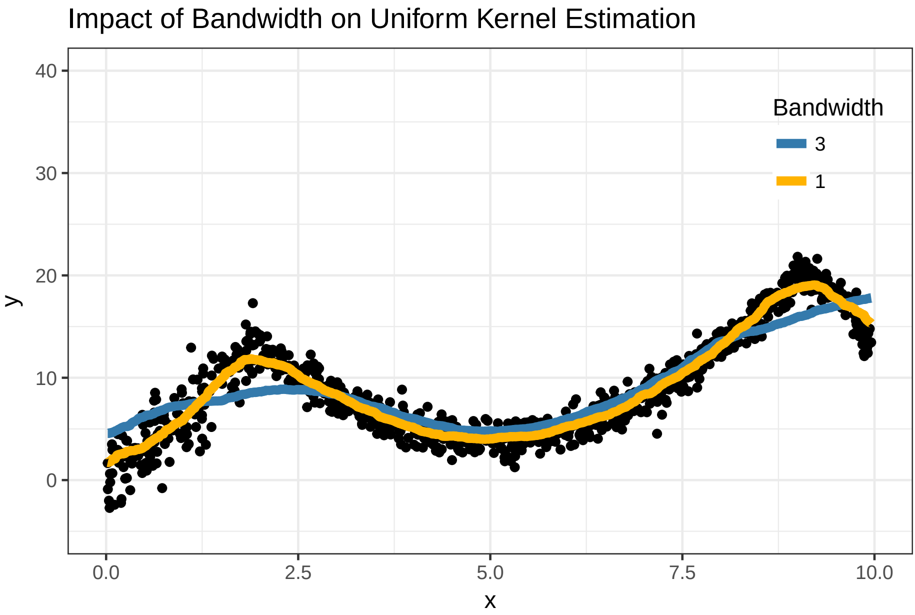
Changing the Bandwidth



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

Let $K : \mathbb{R} \rightarrow \mathbb{R}$ 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

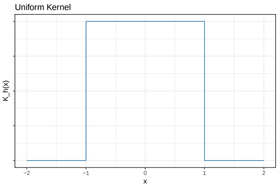
- We can create other estimators by filling in other kernels.

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

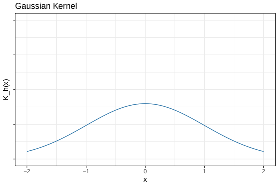
- We can create other estimators by filling in other kernels.
- Generally the idea is to weight things further from the focal point by less than those closer to the focal point.
- Uniform Kernel
- Calculate the average of the observed y points that have x values in the interval $[x_0 - h, x_0 + h]$
-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



Kernel Estimation

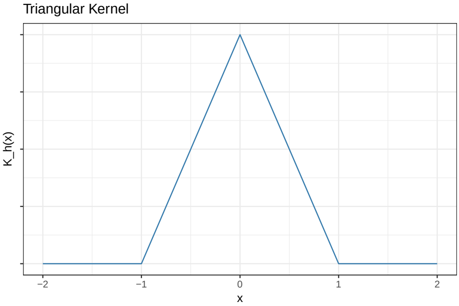
- We can create other estimators by filling in other kernels.
- Generally the idea is to weight things further from the focal point by less than those closer to the focal point.
- Gaussian Kernel
- Distance weighted by how far from x_0 following the normal density

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



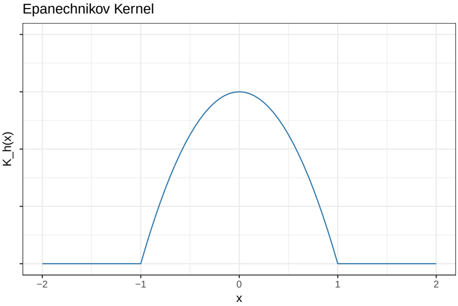
Kernel Estimation

- 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

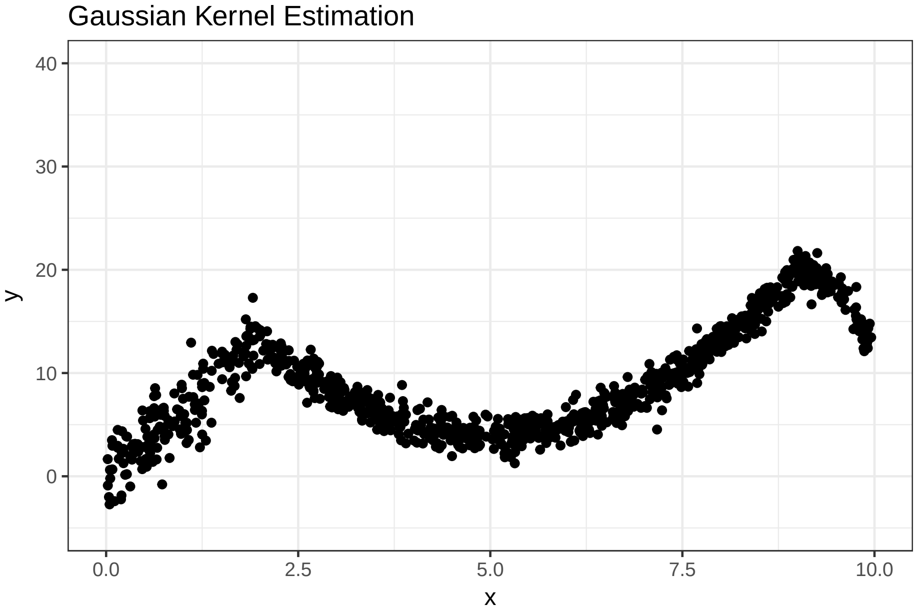


Kernel Estimation

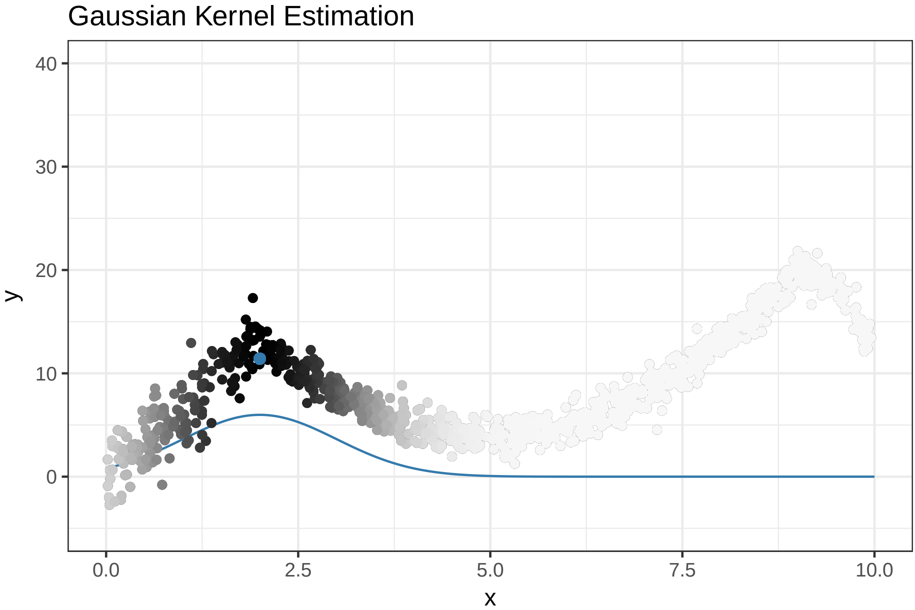
- We can create other estimators by filling in other kernels.
- Generally the idea is to weight things further from the focal point by less than those closer to the focal point.
- Epanechnikov
- Distance weighted by how far from x_0 using a parabolic function



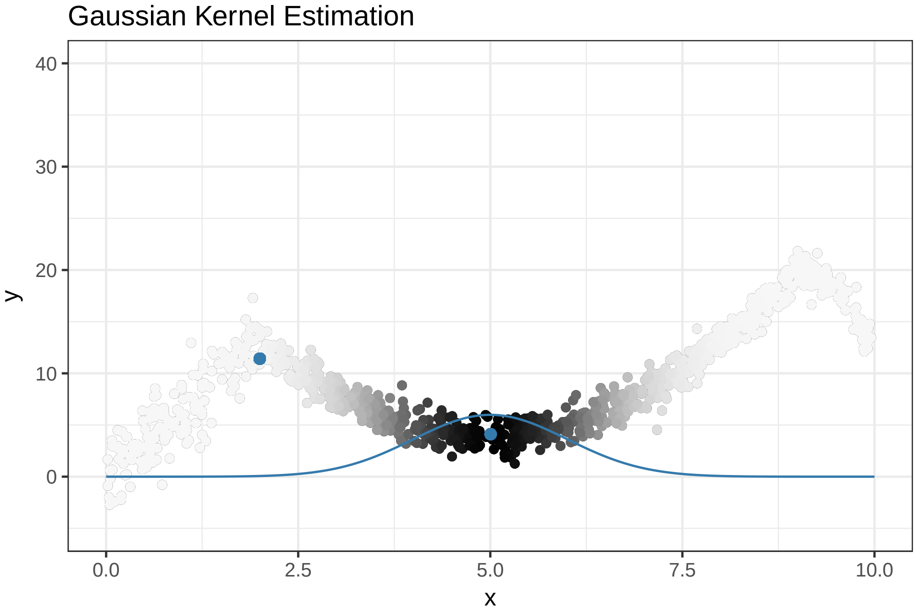
Example of Gaussian Kernel



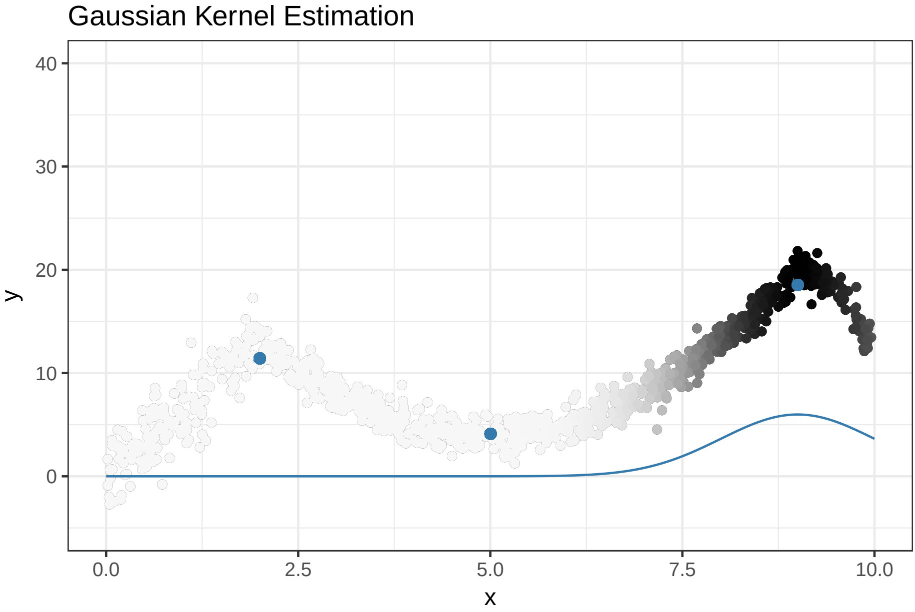
Example of Gaussian Kernel



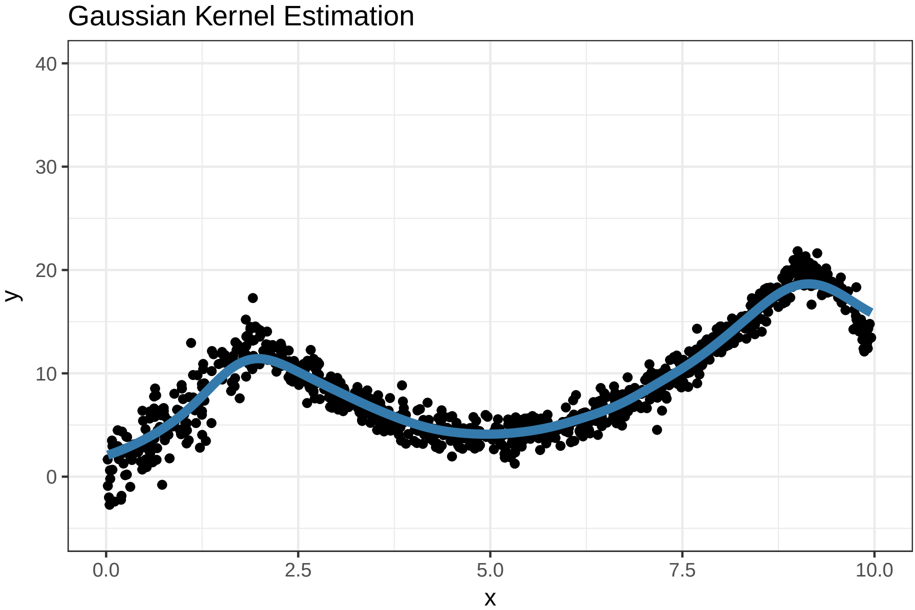
Example of Gaussian Kernel



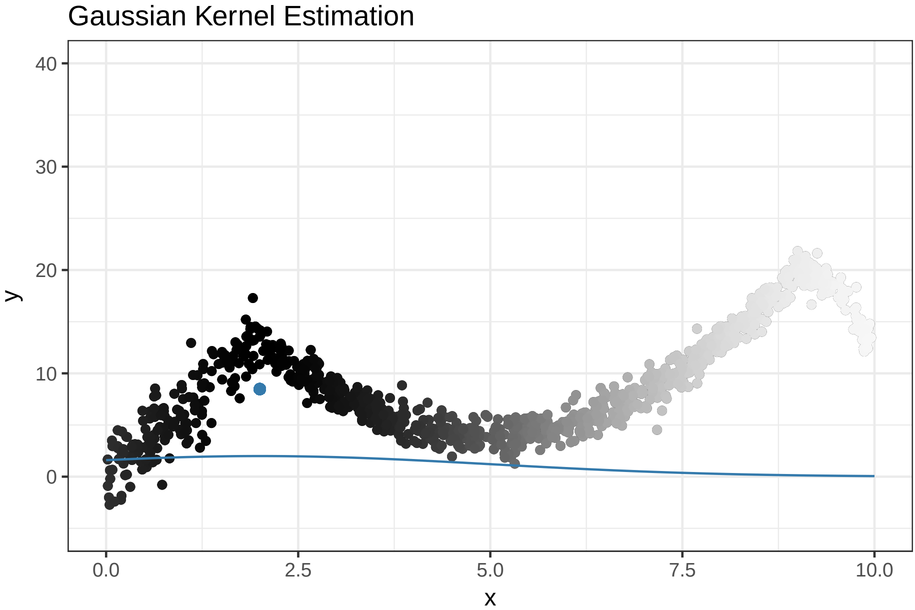
Example of Gaussian Kernel



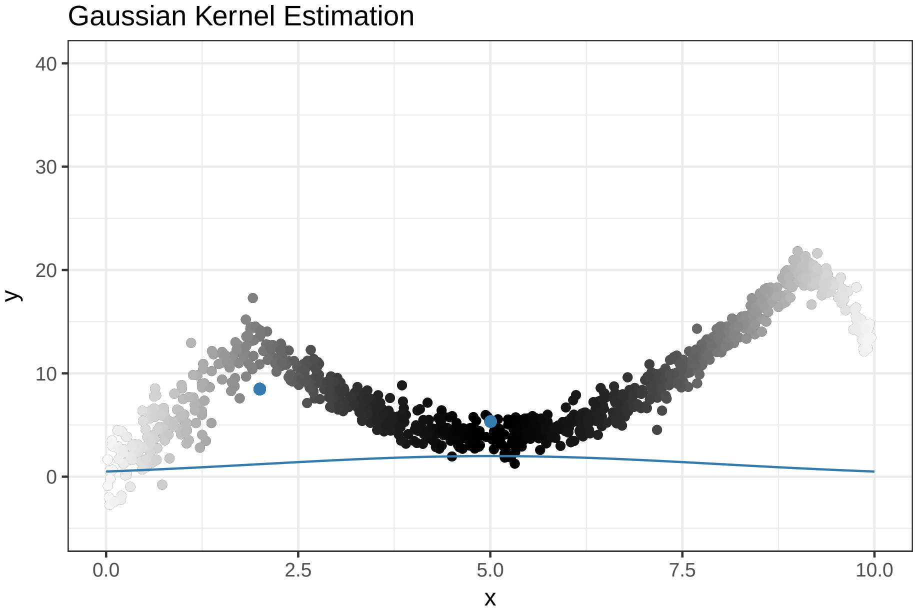
Example of Gaussian Kernel



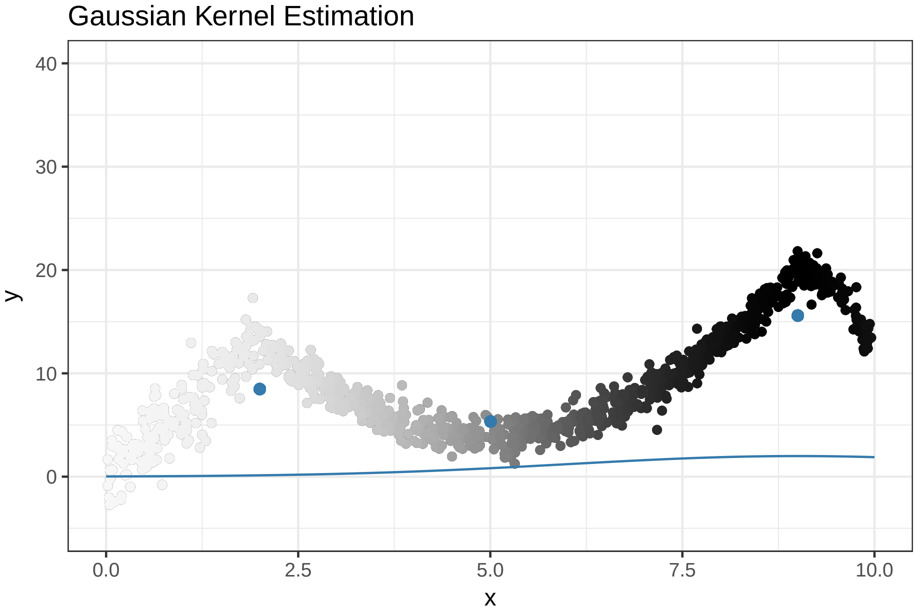
Example of Gaussian Kernel



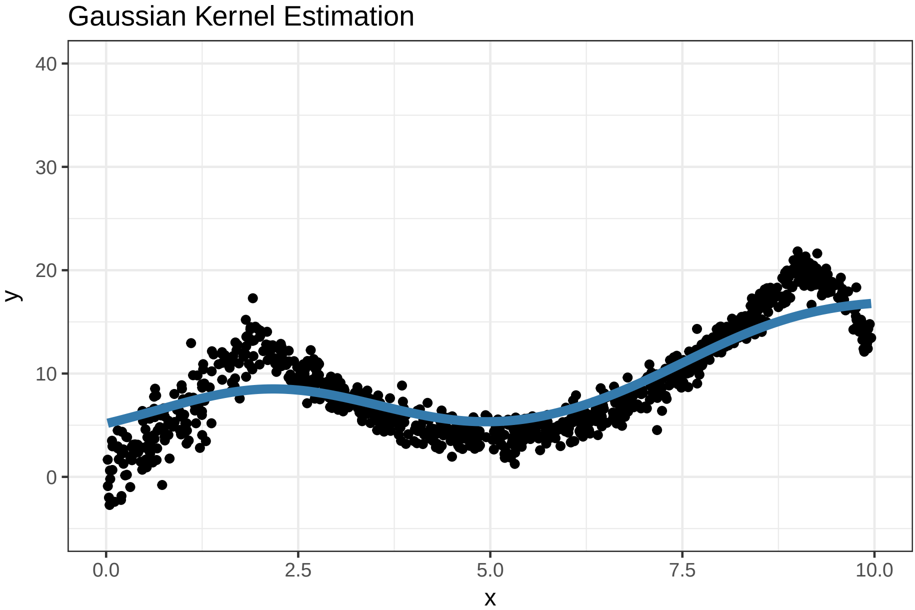
Example of Gaussian Kernel



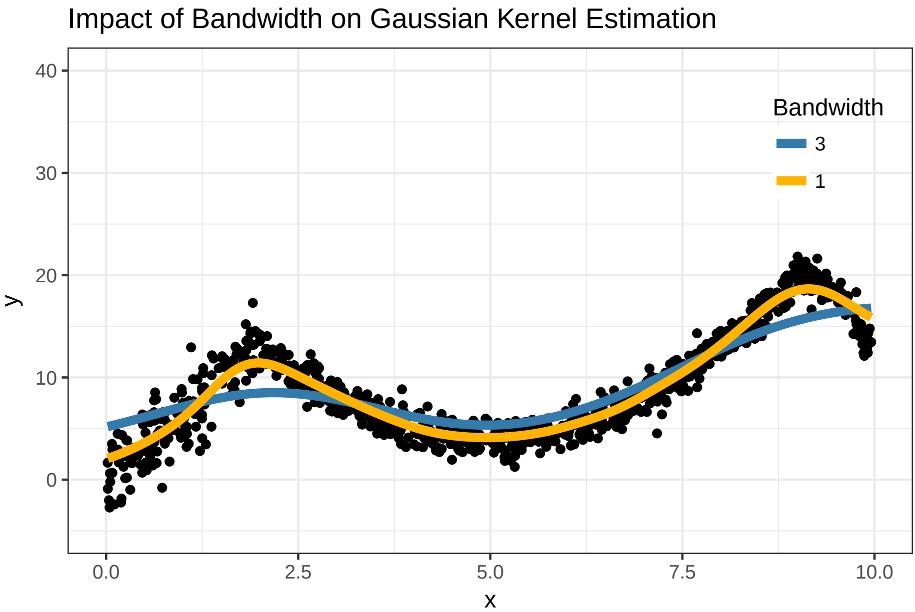
Example of Gaussian Kernel



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,

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

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

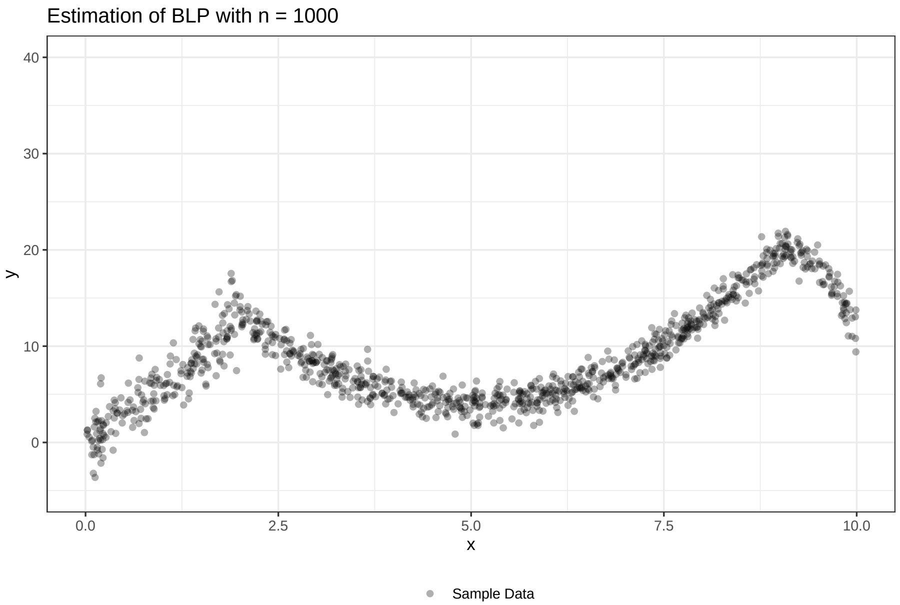
Corollary

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

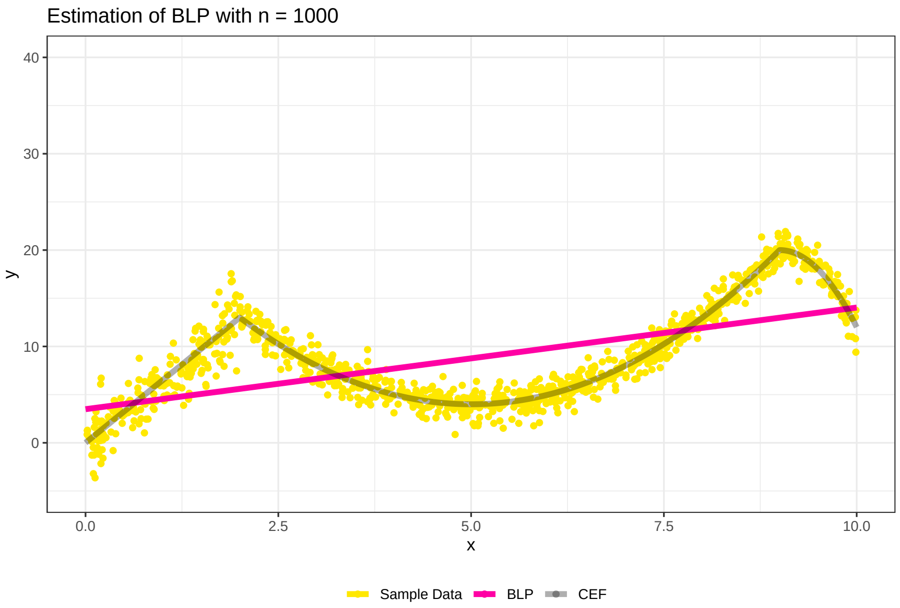
$$E\left[(E[Y \mid X] - (a + bX))^2\right]$$

- *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.

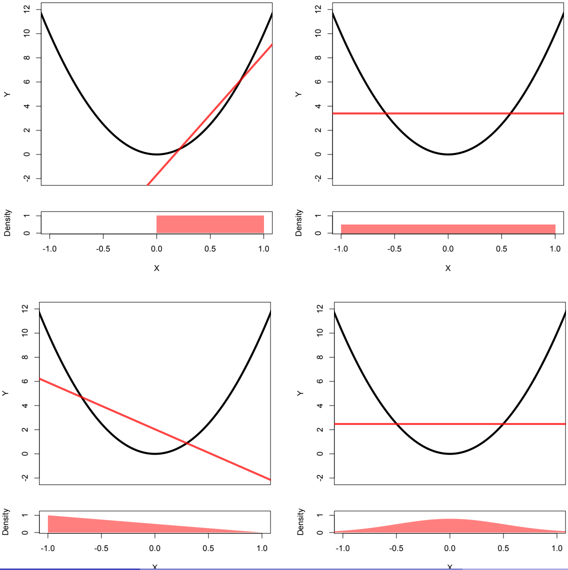
Linear Approximations

- 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:

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

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:

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

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

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

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

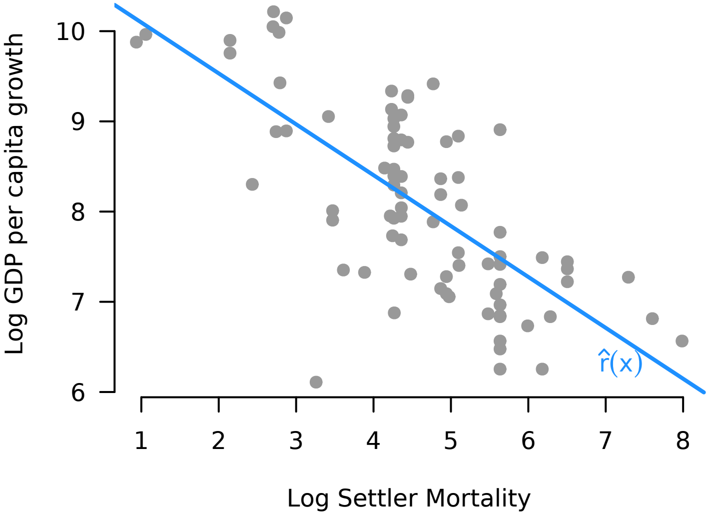
Noting we can rewrite $\frac{Cov[X,Y]}{V[X]} = \frac{E[XY]-E[X]E[Y]}{E[X^2]-E[X]^2}$, and using the plug-in principle, we can estimate the parameters of the BLP as:

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

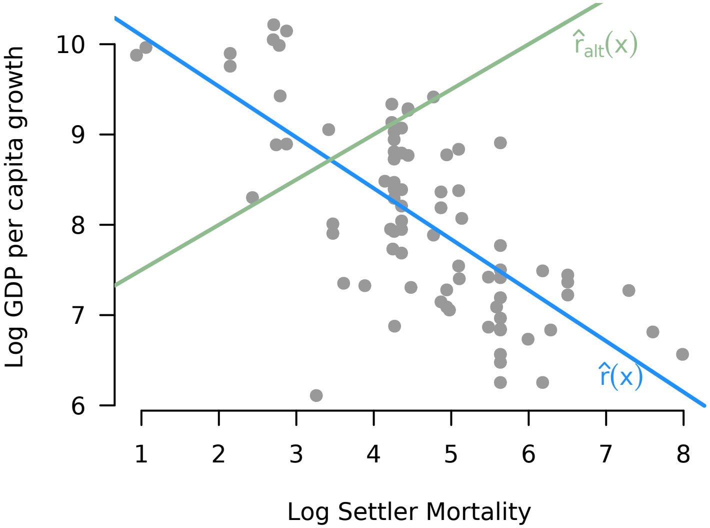
$$\hat{\beta} = \frac{\overline{XY} - \overline{X}\cdot\overline{Y}}{\overline{X^2} - \overline{X}^2} = \frac{\widehat{Cov}(X,Y)}{\widehat{V}(X)} = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

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:

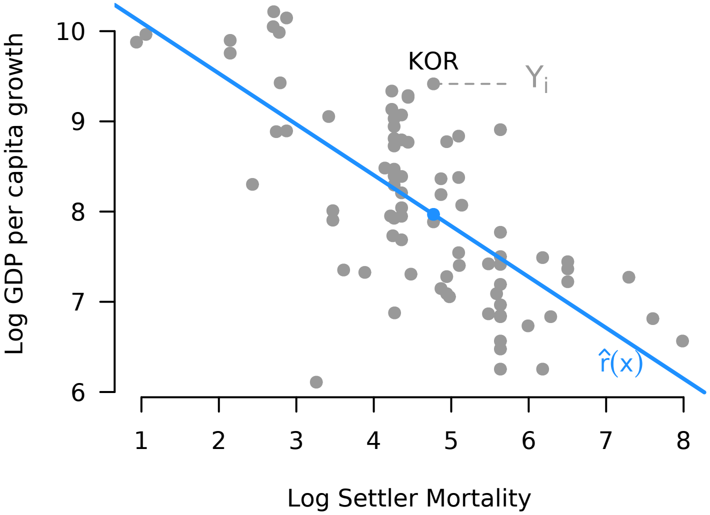
$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$$

Definition (Residual)

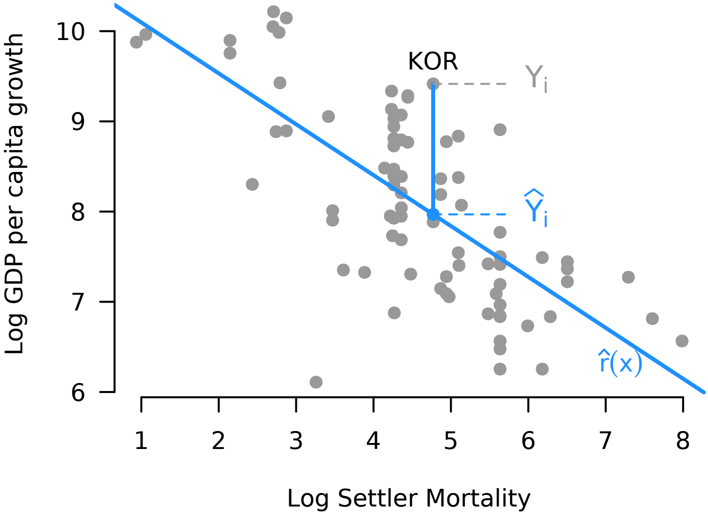
The **residual** is the difference between the actual value of Y_i and the predicted value, $\widehat{Y}_i$:

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

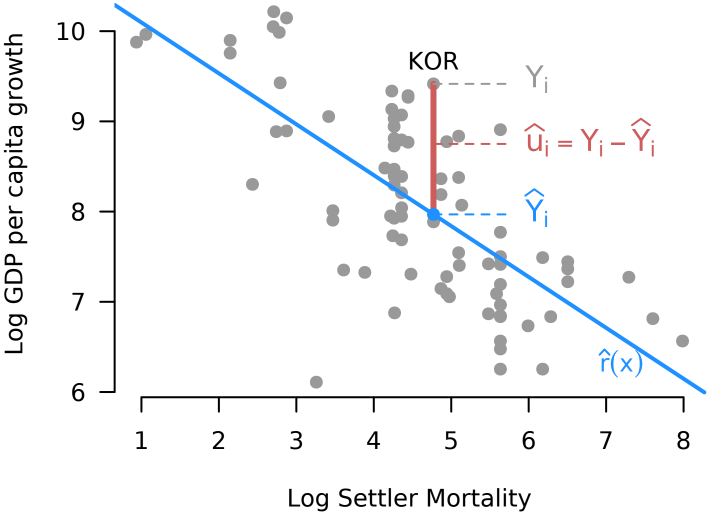
Fitted linear CEF/regression function



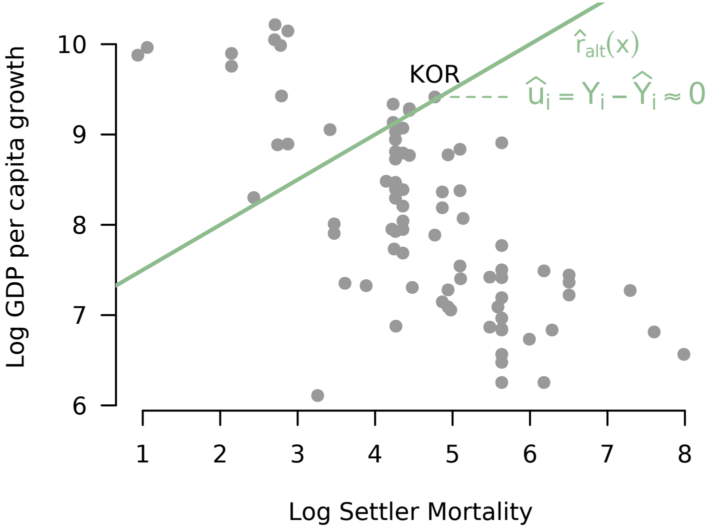
Fitted linear CEF/regression function



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

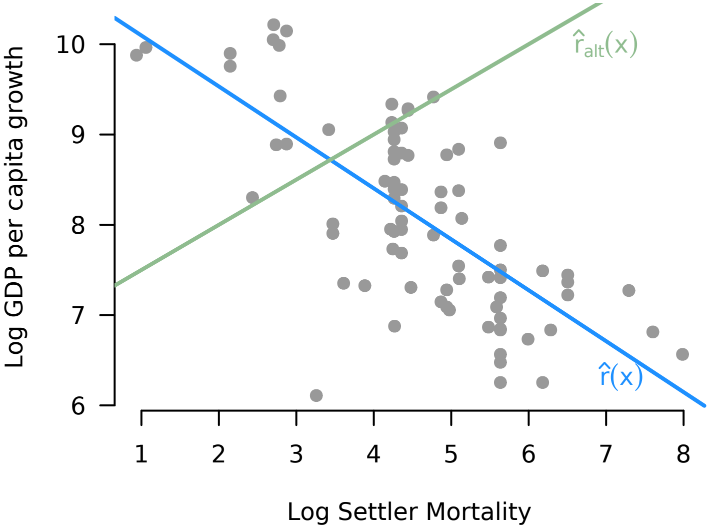
Minimize the residuals

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

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.

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.
- Let's start by thinking descriptively: different groups of units. 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.
- Let's start by thinking descriptively: different groups of units. 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 <u>in-sample data</u>

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

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?

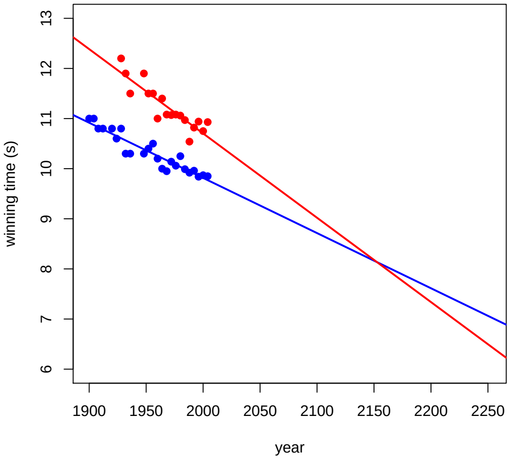
Example: Tatem, et al. Sprinters Data

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?

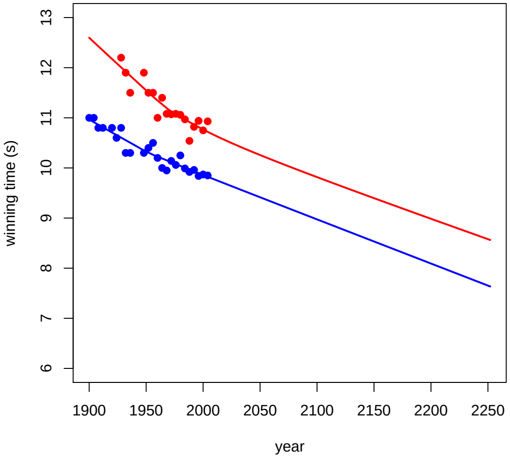
Using data from 1900 to 2004, they fit linear regression models of the winning 100 meter time on year for both men and women. They then use the estimates from these models to extrapolate 152 years into the future.

Tatem et al. Extrapolation

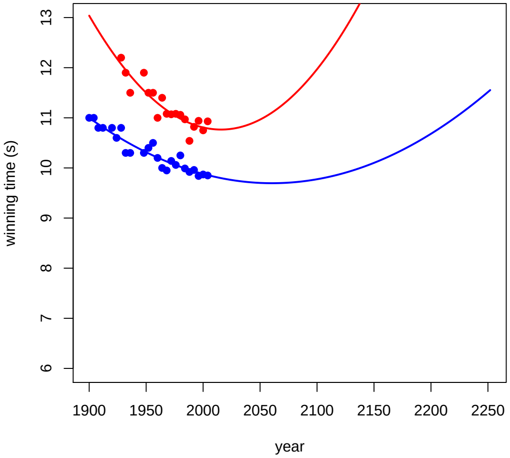


Tatem et al.'s predictions. Men's times are in blue, women's times are in red.

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

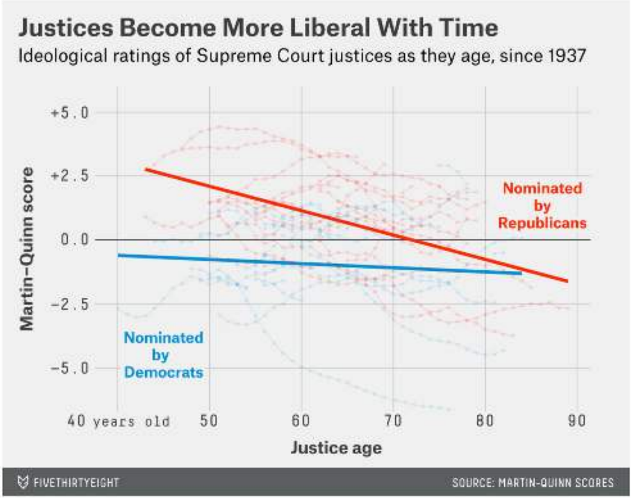
- The model only gives the best fitting line where we have data, it says little about the shape where there isn't any data.
- We can always ask illogical questions and the model gives answers.
  - For example, when will women finish the sprint in negative time?
- Fundamentally we are assuming that data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

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

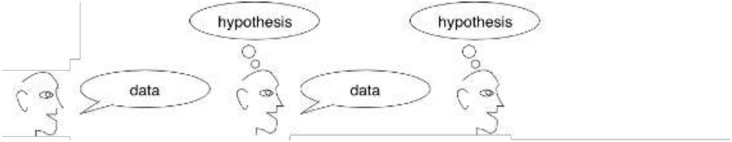
AND

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.

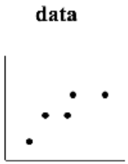
Images on following slides courtesy of Tom Griffiths

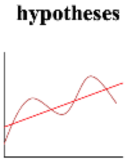
Fun with Linearity



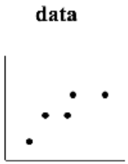
The Design

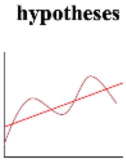
The Design





The Design





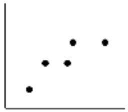
- Each learner sees a set of (x, y) pairs

The Design

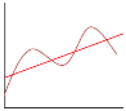
- Each learner sees a set of (x, y) pairs
- Makes predictions of y for new x values

The Design

data

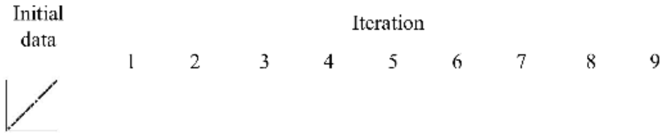


hypotheses

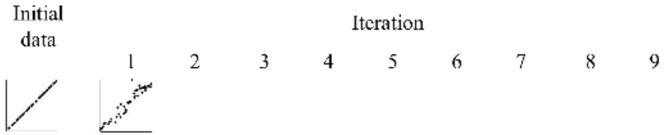


- Each learner sees a set of (x, y) pairs
- Makes predictions of y for new x values
- Predictions are data for the next learner

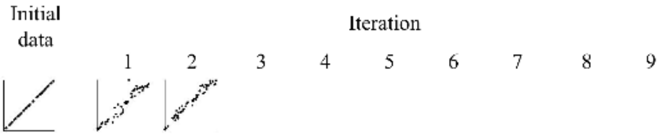
Results



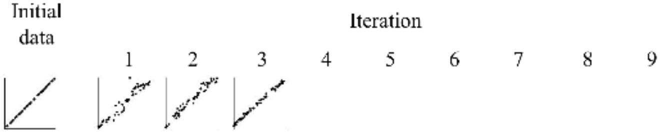
Results



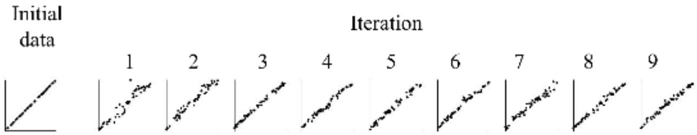
Results



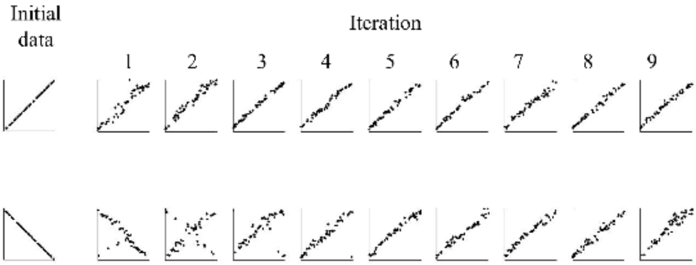
Results



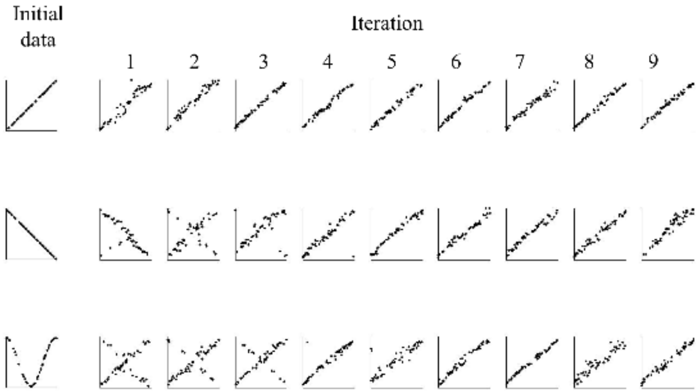
Results



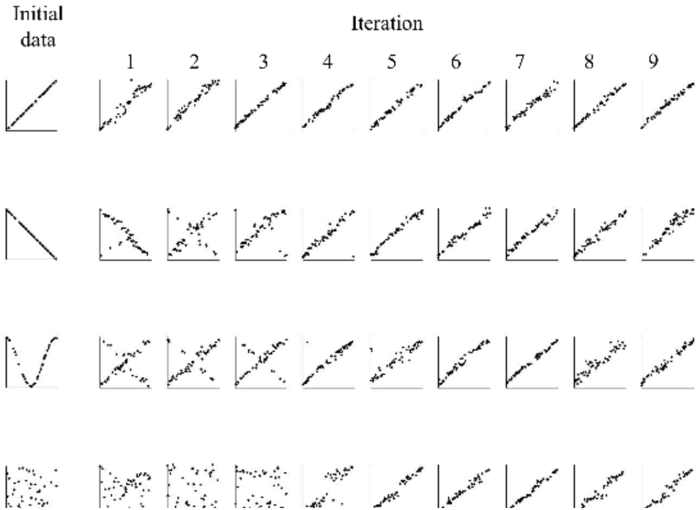
Results



Results



Results



Deriving the OLS estimator

Deriving the OLS estimator

- Let's think about n pairs of sample observations: $(Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n)$

Deriving the OLS estimator

- Let's think about n pairs of sample observations: $(Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n)$
- Let $\{b_0, b_1\}$ be possible values for $\{\beta_0, \beta_1\}$

Deriving the OLS estimator

- Let's think about n pairs of sample observations: $(Y_1, X_1), (Y_2, X_2), \dots, (Y_n, X_n)$
- Let $\{b_0, b_1\}$ be possible values for $\{\beta_0, \beta_1\}$
- Define the least squares objective function:

$$S(b_0, b_1) = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2.$$

Deriving the OLS estimator

- Let's think about n pairs of sample observations: $(Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n)$
- Let $\{b_0, b_1\}$ be possible values for $\{\beta_0, \beta_1\}$
- Define the least squares objective function:

$$S(b_0, b_1) = \sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)^2.$$

- How do we derive the LS estimators for β_0 and β_1? We want to minimize this function, which is actually a very well-defined calculus problem.

Deriving the OLS estimator

- Let's think about n pairs of sample observations: $(Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n)$
- Let $\{b_0, b_1\}$ be possible values for $\{\beta_0, \beta_1\}$
- Define the least squares objective function:

$$S(b_0, b_1) = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2.$$

- How do we derive the LS estimators for β_0 and β_1? We want to minimize this function, which is actually a very well-defined calculus problem.
    1. Take partial derivatives of S with respect to b_0 and b_1.

Deriving the OLS estimator

- Let's think about n pairs of sample observations: $(Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n)$
- Let $\{b_0, b_1\}$ be possible values for $\{\beta_0, \beta_1\}$
- Define the least squares objective function:

$$S(b_0, b_1) = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2.$$

- How do we derive the LS estimators for β_0 and β_1? We want to minimize this function, which is actually a very well-defined calculus problem.
    1. Take partial derivatives of S with respect to b_0 and b_1.
    2. Set each of the partial derivatives to 0

Deriving the OLS estimator

- Let's think about n pairs of sample observations: $(Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n)$
- Let $\{b_0, b_1\}$ be possible values for $\{\beta_0, \beta_1\}$
- Define the least squares objective function:

$$S(b_0, b_1) = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2.$$

- How do we derive the LS estimators for β_0 and β_1? We want to minimize this function, which is actually a very well-defined calculus problem.
  1. Take partial derivatives of S with respect to b_0 and b_1.
  2. Set each of the partial derivatives to 0
  3. Solve for $\{b_0, b_1\}$ and replace them with the solutions

Deriving the OLS estimator

- Let's think about n pairs of sample observations: $(Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n)$
- Let $\{b_0, b_1\}$ be possible values for $\{\beta_0, \beta_1\}$
- Define the least squares objective function:

$$S(b_0, b_1) = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2.$$

- How do we derive the LS estimators for β_0 and β_1? We want to minimize this function, which is actually a very well-defined calculus problem.
    1. Take partial derivatives of S with respect to b_0 and b_1.
    2. Set each of the partial derivatives to 0
    3. Solve for $\{b_0, b_1\}$ and replace them with the solutions
- Derivations are on the next few slides if you are interested.

Step 1: Take Partial Derivatives

$$S(b_0, b_1) = \sum_{i=1}^{n} (Y_i - b_0 - X_i b_1)^2$$

Step 1: Take Partial Derivatives

$$
\begin{aligned}
S(b_0, b_1) &= \sum_{i=1}^{n} (Y_i - b_0 - X_i b_1)^2 \\
&= \sum_{i=1}^{n} (Y_i^2 - 2Y_i b_0 - 2Y_i b_1 X_i + b_0^2 + 2b_0 b_1 X_i + b_1^2 X_i^2)
\end{aligned}
$$

Step 1: Take Partial Derivatives

$$
\begin{aligned}
S(b_0, b_1) &= \sum_{i=1}^{n} (Y_i - b_0 - X_i b_1)^2 \\
&= \sum_{i=1}^{n} (Y_i^2 - 2Y_i b_0 - 2Y_i b_1 X_i + b_0^2 + 2b_0 b_1 X_i + b_1^2 X_i^2) \\
\frac{\partial S(b_0, b_1)}{\partial b_0} &= \sum_{i=1}^{n} (-2Y_i + 2b_0 + 2b_1 X_i)
\end{aligned}
$$

Step 1: Take Partial Derivatives

$$
\begin{aligned}
S(b_0, b_1) &= \sum_{i=1}^{n} (Y_i - b_0 - X_i b_1)^2 \\
&= \sum_{i=1}^{n} (Y_i^2 - 2Y_i b_0 - 2Y_i b_1 X_i + b_0^2 + 2b_0 b_1 X_i + b_1^2 X_i^2) \\
\frac{\partial S(b_0, b_1)}{\partial b_0} &= \sum_{i=1}^{n} (-2Y_i + 2b_0 + 2b_1 X_i) \\
&= -2 \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)
\end{aligned}
$$

Step 1: Take Partial Derivatives

$$
\begin{aligned}
S(b_0, b_1) &= \sum_{i=1}^{n}(Y_i - b_0 - X_i b_1)^2 \\
&= \sum_{i=1}^{n}(Y_i^2 - 2Y_i b_0 - 2Y_i b_1 X_i + b_0^2 + 2b_0 b_1 X_i + b_1^2 X_i^2) \\
\frac{\partial S(b_0, b_1)}{\partial b_0} &= \sum_{i=1}^{n}(-2Y_i + 2b_0 + 2b_1 X_i) \\
&= -2\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i) \\
\frac{\partial S(b_0, b_1)}{\partial b_1} &= \sum_{i=1}^{n}(-2Y_i X_i + 2b_0 X_i + 2b_1 X_i^2)
\end{aligned}
$$

Step 1: Take Partial Derivatives

$$
\begin{aligned}
S(b_0, b_1) &= \sum_{i=1}^{n} (Y_i - b_0 - X_i b_1)^2 \\
&= \sum_{i=1}^{n} (Y_i^2 - 2Y_i b_0 - 2Y_i b_1 X_i + b_0^2 + 2b_0 b_1 X_i + b_1^2 X_i^2) \\
\frac{\partial S(b_0, b_1)}{\partial b_0} &= \sum_{i=1}^{n} (-2Y_i + 2b_0 + 2b_1 X_i) \\
&= -2 \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i) \\
\frac{\partial S(b_0, b_1)}{\partial b_1} &= \sum_{i=1}^{n} (-2Y_i X_i + 2b_0 X_i + 2b_1 X_i^2) \\
&= -2 \sum_{i=1}^{n} X_i (Y_i - b_0 - b_1 X_i)
\end{aligned}
$$

Solving for the Intercept

$$\frac{\partial S(b_0, b_1)}{\partial b_0} = -2\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)$$

Solving for the Intercept

$$\frac{\partial S(b_0, b_1)}{\partial b_0} = -2\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)$$

$$0 = -2\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)$$

Solving for the Intercept

$$\frac{\partial S(b_0, b_1)}{\partial b_0} = -2\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)$$

$$0 = -2\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)$$

$$0 = \sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)$$

Solving for the Intercept

$$\begin{aligned}
\frac{\partial S(b_0, b_1)}{\partial b_0} &= -2\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i) \\
0 &= -2\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i) \\
0 &= \sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i) \\
0 &= \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} b_0 - \sum_{i=1}^{n} b_1 X_i
\end{aligned}$$

Solving for the Intercept

$$
\begin{aligned}
\frac{\partial S(b_0, b_1)}{\partial b_0} &= -2\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i) \\
0 &= -2\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i) \\
0 &= \sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i) \\
0 &= \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} b_0 - \sum_{i=1}^{n} b_1 X_i \\
b_0 n &= \left(\sum_{i=1}^{n} Y_i\right) - b_1 \left(\sum_{i=1}^{n} X_i\right)
\end{aligned}
$$

Solving for the Intercept

$$
\begin{aligned}
\frac{\partial S(b_0, b_1)}{\partial b_0} &= -2 \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i) \\
0 &= -2 \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i) \\
0 &= \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i) \\
0 &= \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} b_0 - \sum_{i=1}^{n} b_1 X_i \\
b_0 n &= \left(\sum_{i=1}^{n} Y_i \right) - b_1 \left(\sum_{i=1}^{n} X_i \right) \\
b_0 &= \overline{Y} - b_1 \overline{X}
\end{aligned}
$$

A Helpful Lemma on Deviations from Means

Lemmas are like helper results that are often invoked repeatedly.

A Helpful Lemma on Deviations from Means

Lemmas are like helper results that are often invoked repeatedly.

Lemma (Deviations from the Mean Sum to 0)

$$\sum_{i=1}^{n}(X_i - \bar{X}) \quad =$$

A Helpful Lemma on Deviations from Means

Lemmas are like helper results that are often invoked repeatedly.

Lemma (Deviations from the Mean Sum to 0)

$$\sum_{i=1}^{n}(X_i - \bar{X}) \;\; = \;\; \left(\sum_{i=1}^{n} X_i\right) - n\bar{X}$$

A Helpful Lemma on Deviations from Means

Lemmas are like helper results that are often invoked repeatedly.

Lemma (Deviations from the Mean Sum to 0)

$$
\begin{aligned}
\sum_{i=1}^{n}(X_i - \bar{X}) &= \left(\sum_{i=1}^{n} X_i\right) - n\bar{X} \\
&= \left(\sum_{i=1}^{n} X_i\right) - n\sum_{i=1}^{n} X_i/n
\end{aligned}
$$

A Helpful Lemma on Deviations from Means

Lemmas are like helper results that are often invoked repeatedly.

Lemma (Deviations from the Mean Sum to 0)

$$
\begin{aligned}
\sum_{i=1}^{n}(X_i - \bar{X}) &= \left(\sum_{i=1}^{n} X_i\right) - n\bar{X} \\
&= \left(\sum_{i=1}^{n} X_i\right) - n\sum_{i=1}^{n} X_i/n \\
&= \left(\sum_{i=1}^{n} X_i\right) - \sum_{i=1}^{n} X_i
\end{aligned}
$$

A Helpful Lemma on Deviations from Means

Lemmas are like helper results that are often invoked repeatedly.

Lemma (Deviations from the Mean Sum to 0)

$$
\begin{aligned}
\sum_{i=1}^{n}(X_i - \bar{X}) &= \left(\sum_{i=1}^{n} X_i\right) - n\bar{X} \\
&= \left(\sum_{i=1}^{n} X_i\right) - n\sum_{i=1}^{n} X_i/n \\
&= \left(\sum_{i=1}^{n} X_i\right) - \sum_{i=1}^{n} X_i \\
&= 0
\end{aligned}
$$

Solving for the Slope

$$0 = -2\sum_{i=1}^{n} X_i(Y_i - b_0 - b_1 X_i)$$

Solving for the Slope

$$0 = -2\sum_{i=1}^{n} X_i(Y_i - b_0 - b_1 X_i)$$

$$0 = \sum_{i=1}^{n} X_i(Y_i - b_0 - b_1 X_i)$$

Solving for the Slope

$$0 = -2\sum_{i=1}^{n} X_i(Y_i - b_0 - b_1 X_i)$$

$$0 = \sum_{i=1}^{n} X_i(Y_i - b_0 - b_1 X_i)$$

$$0 = \sum_{i=1}^{n} X_i(Y_i - (\overline{Y} - b_1\overline{X}) - b_1 X_i) \qquad \text{(sub in } b_0\text{)}$$

Solving for the Slope

$$
\begin{aligned}
0 &= -2\sum_{i=1}^{n} X_i(Y_i - b_0 - b_1 X_i) \\
0 &= \sum_{i=1}^{n} X_i(Y_i - b_0 - b_1 X_i) \\
0 &= \sum_{i=1}^{n} X_i(Y_i - (\overline{Y} - b_1\overline{X}) - b_1 X_i) && \text{(sub in } b_0\text{)} \\
0 &= \sum_{i=1}^{n} X_i(Y_i - \overline{Y} - b_1(X_i - \overline{X}))
\end{aligned}
$$

Solving for the Slope

$$0 = -2\sum_{i=1}^{n} X_i(Y_i - b_0 - b_1 X_i)$$

$$0 = \sum_{i=1}^{n} X_i(Y_i - b_0 - b_1 X_i)$$

$$0 = \sum_{i=1}^{n} X_i(Y_i - (\overline{Y} - b_1\overline{X}) - b_1 X_i) \qquad \text{(sub in } b_0\text{)}$$

$$0 = \sum_{i=1}^{n} X_i(Y_i - \overline{Y} - b_1(X_i - \overline{X}))$$

$$0 = \sum_{i=1}^{n} X_i(Y_i - \overline{Y}) - b_1\sum_{i=1}^{n} X_i(X_i - \overline{X})$$

Solving for the Slope

$$0 = -2\sum_{i=1}^{n} X_i(Y_i - b_0 - b_1 X_i)$$

$$0 = \sum_{i=1}^{n} X_i(Y_i - b_0 - b_1 X_i)$$

$$0 = \sum_{i=1}^{n} X_i(Y_i - (\overline{Y} - b_1\overline{X}) - b_1 X_i) \qquad \text{(sub in } b_0\text{)}$$

$$0 = \sum_{i=1}^{n} X_i(Y_i - \overline{Y} - b_1(X_i - \overline{X}))$$

$$0 = \sum_{i=1}^{n} X_i(Y_i - \overline{Y}) - b_1 \sum_{i=1}^{n} X_i(X_i - \overline{X})$$

$$b_1 \sum_{i=1}^{n} X_i(X_i - \overline{X}) = \sum_{i=1}^{n} X_i(Y_i - \overline{Y}) - \overline{X}\sum_{i=1}(Y_i - \overline{Y}) \qquad \text{(add 0)}$$

Solving for the Slope

$$b_1 \sum_{i=1}^{n} X_i(X_i - \overline{X}) = \sum_{i=1}^{n} X_i(Y_i - \overline{Y}) - \overline{X} \sum_{i=1}(Y_i - \overline{Y})$$

Solving for the Slope

$$b_1 \sum_{i=1}^{n} X_i(X_i - \overline{X}) = \sum_{i=1}^{n} X_i(Y_i - \overline{Y}) - \overline{X} \sum_{i=1}^{n} (Y_i - \overline{Y})$$

$$b_1 \sum_{i=1}^{n} X_i(X_i - \overline{X}) = \sum_{i=1}^{n} X_i(Y_i - \overline{Y}) - \sum_{i=1}^{n} \overline{X}(Y_i - \overline{Y})$$

Solving for the Slope

$$b_1 \sum_{i=1}^{n} X_i(X_i - \overline{X}) = \sum_{i=1}^{n} X_i(Y_i - \overline{Y}) - \overline{X} \sum_{i=1}^{n}(Y_i - \overline{Y})$$

$$b_1 \sum_{i=1}^{n} X_i(X_i - \overline{X}) = \sum_{i=1}^{n} X_i(Y_i - \overline{Y}) - \sum_{i=1}^{n} \overline{X}(Y_i - \overline{Y})$$

$$b_1 \left(\sum_{i=1}^{n} X_i(X_i - \overline{X}) - \sum_{i=1} \overline{X}(X_i - \overline{X}) \right) = \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y}) \qquad \text{add 0}$$

Solving for the Slope

$$b_1 \sum_{i=1}^{n} X_i(X_i - \overline{X}) = \sum_{i=1}^{n} X_i(Y_i - \overline{Y}) - \overline{X}\sum_{i=1}(Y_i - \overline{Y})$$

$$b_1 \sum_{i=1}^{n} X_i(X_i - \overline{X}) = \sum_{i=1}^{n} X_i(Y_i - \overline{Y}) - \sum_{i=1} \overline{X}(Y_i - \overline{Y})$$

$$b_1 \left(\sum_{i=1}^{n} X_i(X_i - \overline{X}) - \sum_{i=1} \overline{X}(X_i - \overline{X})\right) = \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y}) \qquad \text{add 0}$$

$$b_1 \sum_{i=1}^{n} (X_i - \overline{X})(X_i - \overline{X}) = \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})$$

Solving for the Slope

$$b_1 \sum_{i=1}^{n} X_i(X_i - \overline{X}) = \sum_{i=1}^{n} X_i(Y_i - \overline{Y}) - \overline{X} \sum_{i=1} (Y_i - \overline{Y})$$

$$b_1 \sum_{i=1}^{n} X_i(X_i - \overline{X}) = \sum_{i=1}^{n} X_i(Y_i - \overline{Y}) - \sum_{i=1} \overline{X}(Y_i - \overline{Y})$$

$$b_1 \left(\sum_{i=1}^{n} X_i(X_i - \overline{X}) - \sum_{i=1} \overline{X}(X_i - \overline{X}) \right) = \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y}) \qquad \text{add } 0$$

$$b_1 \sum_{i=1}^{n} (X_i - \overline{X})(X_i - \overline{X}) = \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})$$

$$b_1 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$$

The OLS estimator

- Now we're done! Here are the **OLS estimators**:

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}$$

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

Intuition of the OLS estimator

Intuition of the OLS estimator

- The slope for the regression line can be written as the following:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{\text{Sample Covariance between } X \text{ and } Y}{\text{Sample Variance of } X}$$

Intuition of the OLS estimator

- The slope for the regression line can be written as the following:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{\text{Sample Covariance between } X \text{ and } Y}{\text{Sample Variance of } X}$$

- The higher the covariance between X and Y, the higher the slope will be.

Intuition of the OLS estimator

- The slope for the regression line can be written as the following:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{\text{Sample Covariance between } X \text{ and } Y}{\text{Sample Variance of } X}$$

- The higher the covariance between X and Y, the higher the slope will be.
- Negative covariances $\rightarrow$ negative slopes;
  positive covariances $\rightarrow$ positive slopes

Intuition of the OLS estimator

- The slope for the regression line can be written as the following:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{\text{Sample Covariance between } X \text{ and } Y}{\text{Sample Variance of } X}$$

- The higher the covariance between X and Y, the higher the slope will be.
- Negative covariances $\rightarrow$ negative slopes;
  positive covariances $\rightarrow$ positive slopes
- If X_i doesn't vary, the denominator is undefined.
- If Y_i doesn't vary, you get a flat line.

Mechanical properties of OLS

Mechanical properties of OLS

- Later we'll see that under certain assumptions, OLS will have nice statistical properties.

Mechanical properties of OLS

- Later we'll see that under certain assumptions, OLS will have nice statistical properties.
- But some properties are mechanical since they can be derived from the first order conditions of OLS.

Mechanical properties of OLS

- Later we'll see that under certain assumptions, OLS will have nice statistical properties.
- But some properties are mechanical since they can be derived from the first order conditions of OLS.

1. The sample mean of the residuals will be zero:

$$\frac{1}{n}\sum_{i=1}^{n}\widehat{u}_i = 0$$

Mechanical properties of OLS

- Later we'll see that under certain assumptions, OLS will have nice statistical properties.
- But some properties are mechanical since they can be derived from the first order conditions of OLS.

1. The sample mean of the residuals will be zero:

$$\frac{1}{n}\sum_{i=1}^{n}\widehat{u}_i = 0$$

2. The residuals will be uncorrelated with the predictor ($\widehat{\text{Cov}}$ is the sample covariance):

$$\sum_{i=1}^{n} X_i\widehat{u}_i = 0 \implies \widehat{\text{Cov}}(X_i, \widehat{u}_i) = 0$$

Mechanical properties of OLS

- Later we'll see that under certain assumptions, OLS will have nice statistical properties.
- But some properties are mechanical since they can be derived from the first order conditions of OLS.

1. The sample mean of the residuals will be zero:

$$\frac{1}{n}\sum_{i=1}^{n}\widehat{u}_i = 0$$

2. The residuals will be uncorrelated with the predictor ($\widehat{\text{Cov}}$ is the sample covariance):

$$\sum_{i=1}^{n} X_i\widehat{u}_i = 0 \implies \widehat{\text{Cov}}(X_i, \widehat{u}_i) = 0$$

3. The residuals will be uncorrelated with the fitted values:

$$\sum_{i=1}^{n} \widehat{Y}_i\widehat{u}_i = 0 \implies \widehat{\text{Cov}}(\widehat{Y}_i, \widehat{u}_i) = 0$$

OLS slope as a weighted sum of the outcomes

OLS slope as a weighted sum of the outcomes

- One useful derivation is to write the OLS estimator for the slope as a weighted sum of the outcomes.

$$\widehat{\beta}_1 = \sum_{i=1}^{n} W_i Y_i$$

OLS slope as a weighted sum of the outcomes

- One useful derivation is to write the OLS estimator for the slope as a weighted sum of the outcomes.

$$\widehat{\beta}_1 = \sum_{i=1}^{n} W_i Y_i$$

- Where here we have the weights, W_i as:

$$W_i = \frac{(X_i - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

OLS slope as a weighted sum of the outcomes

- One useful derivation is to write the OLS estimator for the slope as a weighted sum of the outcomes.

$$\widehat{\beta}_1 = \sum_{i=1}^{n} W_i Y_i$$

- Where here we have the weights, W_i as:

$$W_i = \frac{(X_i - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

- This is important for two reasons. First, it'll make derivations later much easier. And second, it shows that is just the sum of a random variable. Therefore it is also a random variable.

Lemma 2: OLS as a Weighted Sum of Outcomes

Lemma (OLS as Weighted Sum of Outcomes)

Lemma 2: OLS as a Weighted Sum of Outcomes

Lemma (OLS as Weighted Sum of Outcomes)

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

Lemma 2: OLS as a Weighted Sum of Outcomes

Lemma (OLS as Weighted Sum of Outcomes)

$$\begin{aligned}
\widehat{\beta}_1 &= \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2} \\
&= \frac{\sum_{i=1}^n (X_i - \overline{X})Y_i}{\sum_{i=1}^n (X_i - \overline{X})^2} - \frac{\sum_{i=1}^n (X_i - \overline{X})\overline{Y}}{\sum_{i=1}^n (X_i - \overline{X})^2}
\end{aligned}$$

Lemma 2: OLS as a Weighted Sum of Outcomes

Lemma (OLS as Weighted Sum of Outcomes)

$$\begin{aligned}
\widehat{\beta}_1 &= \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2} \\
&= \frac{\sum_{i=1}^n (X_i - \overline{X})Y_i}{\sum_{i=1}^n (X_i - \overline{X})^2} - \frac{\sum_{i=1}^n (X_i - \overline{X})\overline{Y}}{\sum_{i=1}^n (X_i - \overline{X})^2} \\
&= \frac{\sum_{i=1}^n (X_i - \overline{X})Y_i}{\sum_{i=1}^n (X_i - \overline{X})^2}
\end{aligned}$$

Lemma 2: OLS as a Weighted Sum of Outcomes

Lemma (OLS as Weighted Sum of Outcomes)

$$
\begin{aligned}
\widehat{\beta}_1 &= \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2} \\
&= \frac{\sum_{i=1}^n (X_i - \overline{X})Y_i}{\sum_{i=1}^n (X_i - \overline{X})^2} - \frac{\sum_{i=1}^n (X_i - \overline{X})\overline{Y}}{\sum_{i=1}^n (X_i - \overline{X})^2} \\
&= \frac{\sum_{i=1}^n (X_i - \overline{X})Y_i}{\sum_{i=1}^n (X_i - \overline{X})^2} \\
&= \sum_{i=1}^n W_i Y_i
\end{aligned}
$$

Lemma 2: OLS as a Weighted Sum of Outcomes

Lemma (OLS as Weighted Sum of Outcomes)

$$
\begin{aligned}
\widehat{\beta}_1 &= \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2} \\
&= \frac{\sum_{i=1}^n (X_i - \overline{X})Y_i}{\sum_{i=1}^n (X_i - \overline{X})^2} - \frac{\sum_{i=1}^n (X_i - \overline{X})\overline{Y}}{\sum_{i=1}^n (X_i - \overline{X})^2} \\
&= \frac{\sum_{i=1}^n (X_i - \overline{X})Y_i}{\sum_{i=1}^n (X_i - \overline{X})^2} \\
&= \sum_{i=1}^n W_i Y_i
\end{aligned}
$$

Where the weights, W_i are:

$$
W_i = \frac{(X_i - \overline{X})}{\sum_{i=1}^n (X_i - \overline{X})^2}
$$

Sampling distribution of the OLS estimator

Sampling distribution of the OLS estimator

- Remember: OLS is an estimator—it's a machine that we plug samples into and we get out estimates.

Sampling distribution of the OLS estimator

- Remember: OLS is an estimator—it's a machine that we plug samples into and we get out estimates.

Sampling distribution of the OLS estimator

- Remember: OLS is an estimator—it's a machine that we plug samples into and we get out estimates.

OLS

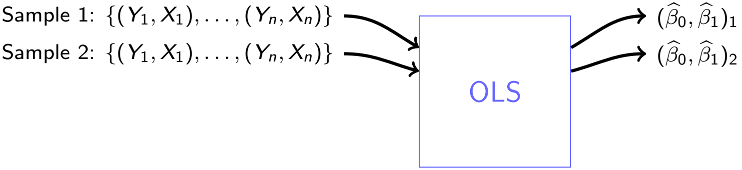
Sampling distribution of the OLS estimator

- Remember: OLS is an estimator—it's a machine that we plug samples into and we get out estimates.

Sample 1: $\{(Y_1, X_1), \dots, (Y_n, X_n)\} \longrightarrow \boxed{\text{OLS}} \longrightarrow (\widehat{\beta}_0, \widehat{\beta}_1)_1$

Sampling distribution of the OLS estimator

- Remember: OLS is an estimator—it's a machine that we plug samples into and we get out estimates.



Sampling distribution of the OLS estimator

- Remember: OLS is an estimator—it's a machine that we plug samples into and we get out estimates.

Sample 1: $\{(Y_1, X_1), \ldots, (Y_n, X_n)\}$ $\rightarrow$ OLS $\rightarrow$ $(\widehat{\beta}_0, \widehat{\beta}_1)_1$

Sample 2: $\{(Y_1, X_1), \ldots, (Y_n, X_n)\}$ $\rightarrow$ OLS $\rightarrow$ $(\widehat{\beta}_0, \widehat{\beta}_1)_2$

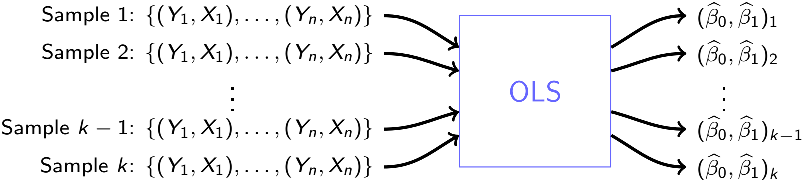
$\vdots$

Sample $k-1$: $\{(Y_1, X_1), \ldots, (Y_n, X_n)\}$ $\rightarrow$ OLS $\rightarrow$ $(\widehat{\beta}_0, \widehat{\beta}_1)_{k-1}$

Sample k: $\{(Y_1, X_1), \ldots, (Y_n, X_n)\}$ $\rightarrow$ OLS $\rightarrow$ $(\widehat{\beta}_0, \widehat{\beta}_1)_k$

Sampling distribution of the OLS estimator

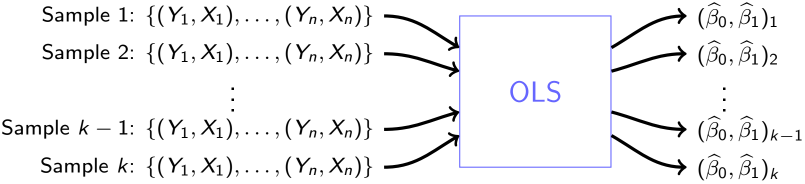
- Remember: OLS is an estimator—it's a machine that we plug samples into and we get out estimates.



- Just like the sample mean, sample difference in means, or the sample variance

Sampling distribution of the OLS estimator

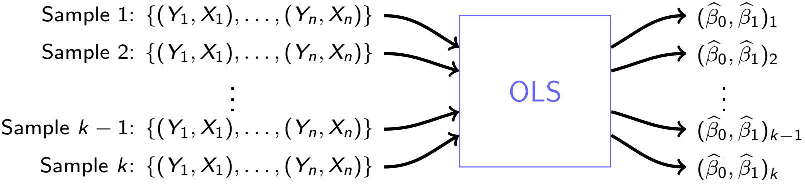
- Remember: OLS is an estimator—it's a machine that we plug samples into and we get out estimates.



- Just like the sample mean, sample difference in means, or the sample variance
- It has a sampling distribution, with a sampling variance/standard error, etc.

Sampling distribution of the OLS estimator

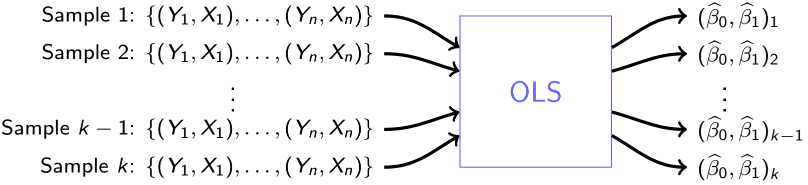
- Remember: OLS is an estimator—it's a machine that we plug samples into and we get out estimates.



- Just like the sample mean, sample difference in means, or the sample variance
- It has a sampling distribution, with a sampling variance/standard error, etc.
- Let's take a simulation approach to demonstrate:

Sampling distribution of the OLS estimator

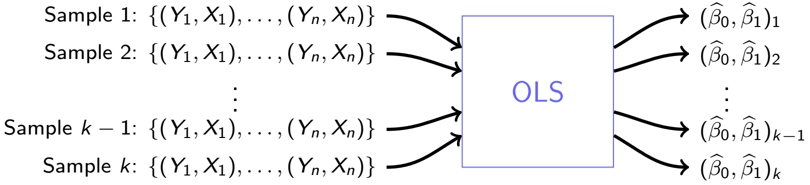
- Remember: OLS is an estimator—it's a machine that we plug samples into and we get out estimates.



- Just like the sample mean, sample difference in means, or the sample variance
- It has a sampling distribution, with a sampling variance/standard error, etc.
- Let's take a simulation approach to demonstrate:
  - Let's use some data from Acemoglu, Daron, Simon Johnson, and James A. Robinson. "The colonial origins of comparative development: An empirical investigation." 2000

Sampling distribution of the OLS estimator

- Remember: OLS is an estimator—it's a machine that we plug samples into and we get out estimates.



- Just like the sample mean, sample difference in means, or the sample variance
- It has a sampling distribution, with a sampling variance/standard error, etc.
- Let's take a simulation approach to demonstrate:
    - Let's use some data from Acemoglu, Daron, Simon Johnson, and James A. Robinson. "The colonial origins of comparative development: An empirical investigation." 2000
    - See how the line varies from sample to sample

Simulation procedure

Simulation procedure

1. Draw a random sample of size $n = 30$ with replacement using `sample()`

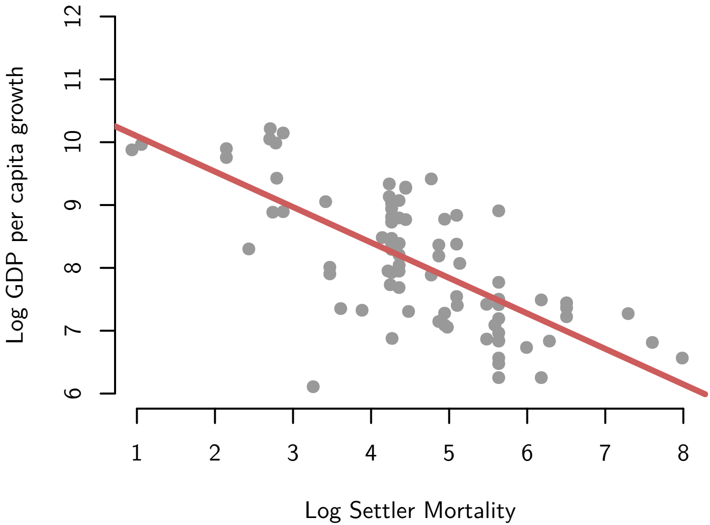
Simulation procedure

1. Draw a random sample of size $n = 30$ with replacement using `sample()`
2. Use `lm()` to calculate the OLS estimates of the slope and intercept

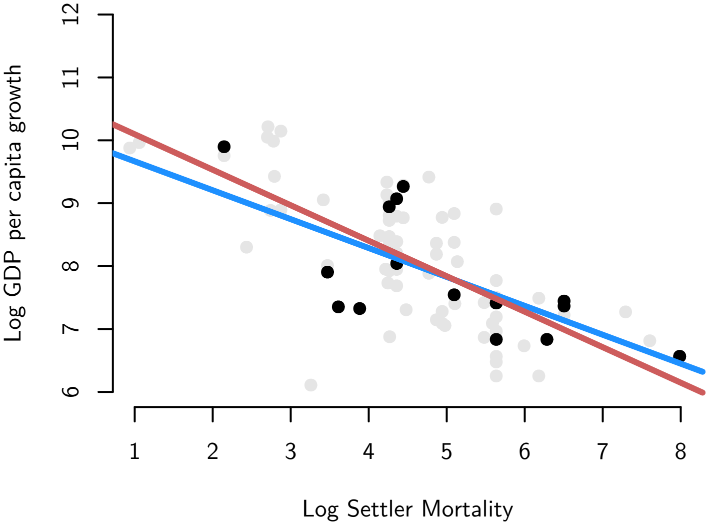
Simulation procedure

1. Draw a random sample of size $n = 30$ with replacement using `sample()`
2. Use `lm()` to calculate the OLS estimates of the slope and intercept
3. Plot the estimated regression line

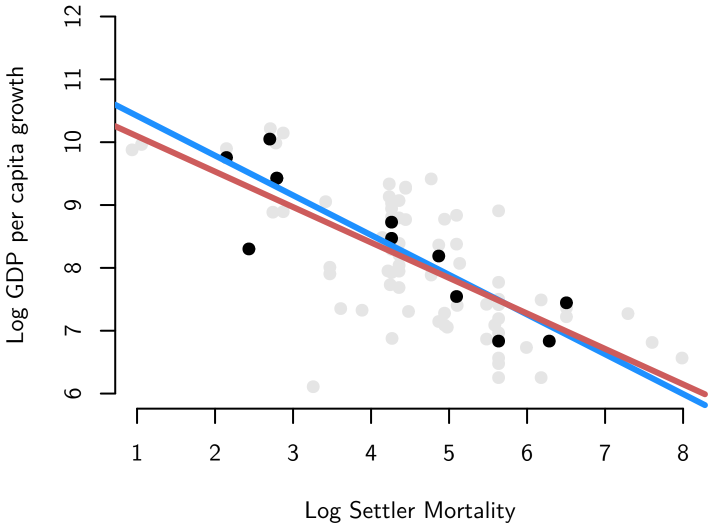
Population Regression



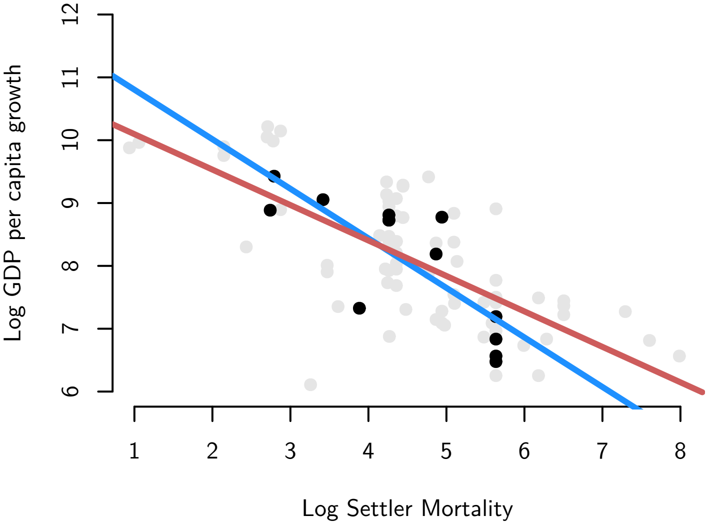
Randomly sample from AJR



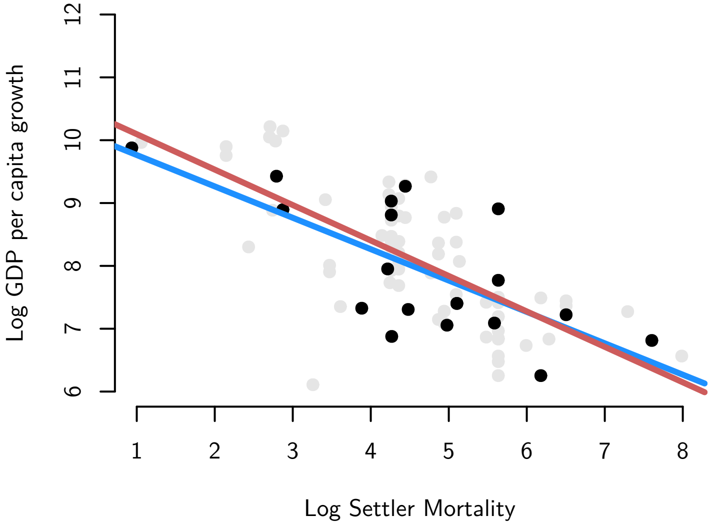
Randomly sample from AJR



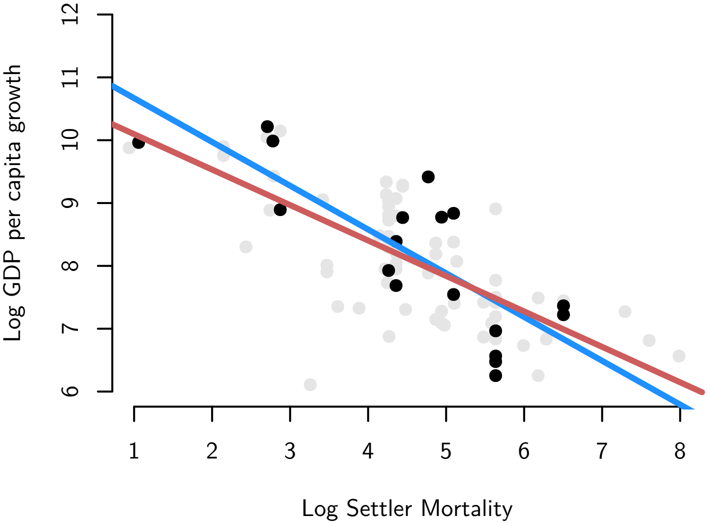
Randomly sample from AJR



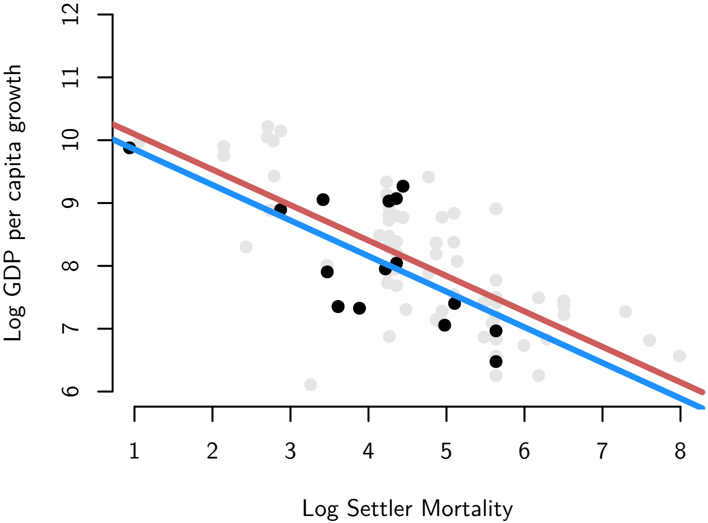
Randomly sample from AJR



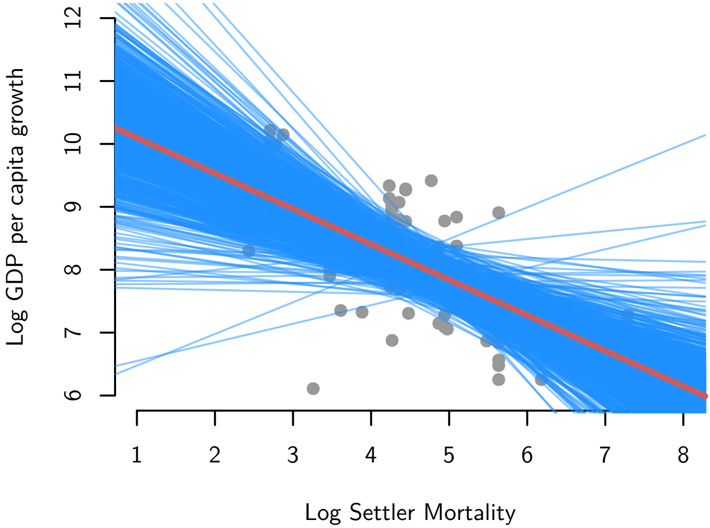
Randomly sample from AJR



Randomly sample from AJR



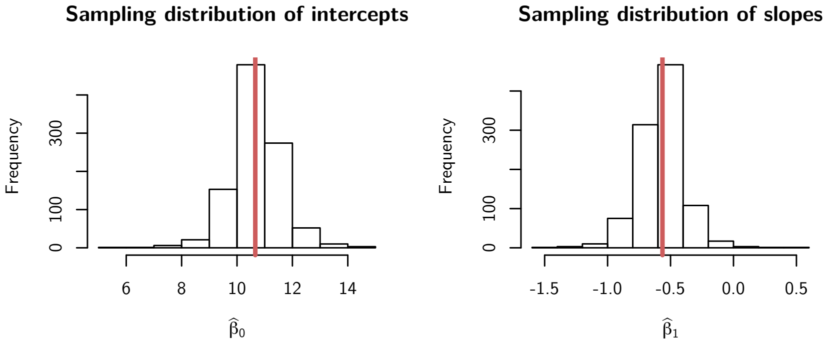
Randomly sample from AJR



Sampling distribution of OLS

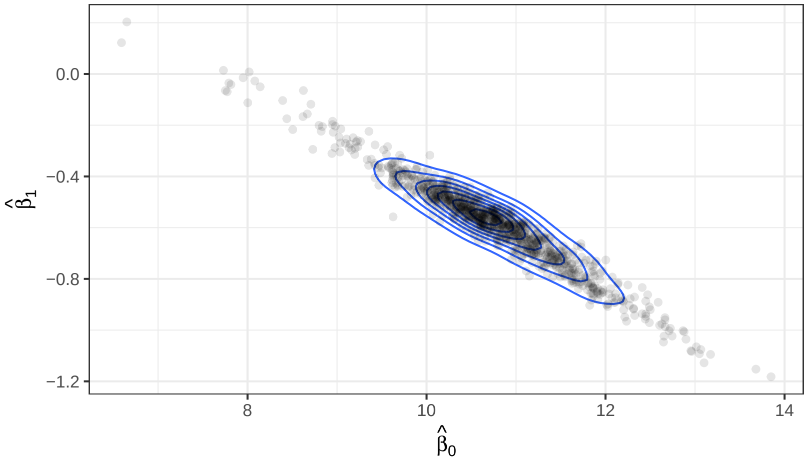
Sampling distribution of OLS

- You can see that the estimated slopes and intercepts vary from sample to sample, but that the "average" of the lines looks about right.



The Sampling Distribution is a Joint Distribution!

While both the intercept and the slope vary, they vary together.



1. What is Regression?
2. Best Linear Predictor
3. Fun With Linearity
4. Estimation of the BLP
5. Additional Techniques
   - R^2
   - LOESS
   - Log Transformations
   - Bootstrap

Goodness of Fit

- How do we judge how well a line fits the data?

Goodness of Fit

- How do we judge how well a line fits the data?
- One way is to find out how much better we do at "predicting" Y once we include X into the regression model.

Goodness of Fit

- How do we judge how well a line fits the data?
- One way is to find out how much better we do at "predicting" Y once we include X into the regression model.
- Prediction errors without X: best prediction is the mean, so our squared errors, or the total sum of squares (SS_{tot}) would be:

$$SS_{tot} = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

Goodness of Fit

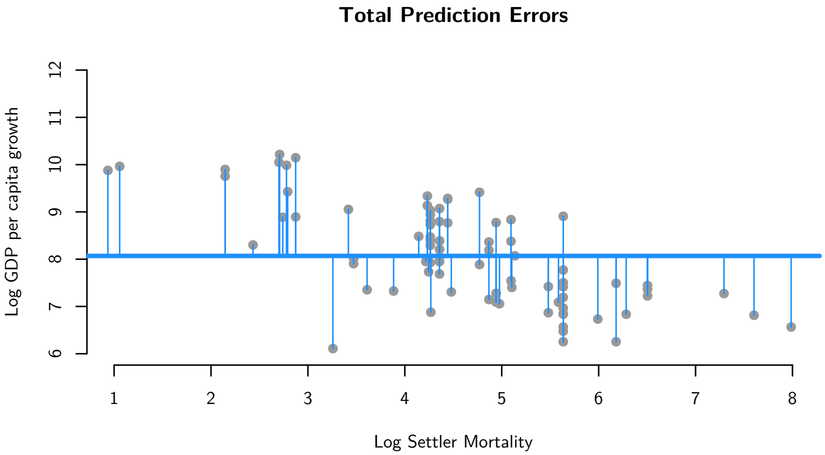
- How do we judge how well a line fits the data?
- One way is to find out how much better we do at "predicting" Y once we include X into the regression model.
- Prediction errors without X: best prediction is the mean, so our squared errors, or the total sum of squares (SS_{tot}) would be:

$$SS_{tot} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

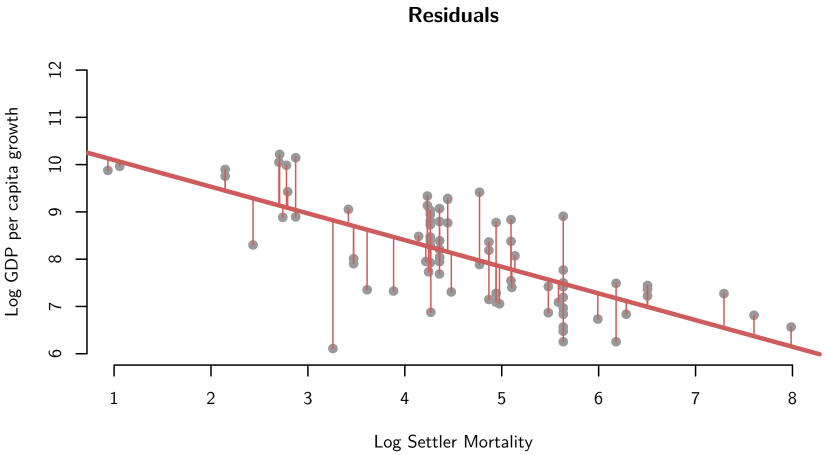
- Once we have estimated our model, we have new prediction errors, which are just the sum of the squared residuals or SS_{res}:

$$SS_{res} = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$$

Sum of Squares



Sum of Squares



R-square

- By definition, the residuals have to be smaller than the deviations from the mean, so we might ask the following: how much lower is the SS_{res} compared to the SS_{tot}?

R-square

- By definition, the residuals have to be smaller than the deviations from the mean, so we might ask the following: how much lower is the SS_{res} compared to the SS_{tot}?
- We quantify this question with the coefficient of determination or R^2. This is the following:

$$R^2 = \frac{SS_{tot} - SS_{res}}{SS_{tot}} = 1 - \frac{SS_{res}}{SS_{tot}}$$

R-square

- By definition, the residuals have to be smaller than the deviations from the mean, so we might ask the following: how much lower is the SS_{res} compared to the SS_{tot}?
- We quantify this question with the coefficient of determination or R^2. This is the following:

$$R^2 = \frac{SS_{tot} - SS_{res}}{SS_{tot}} = 1 - \frac{SS_{res}}{SS_{tot}}$$

- This is the fraction of the total prediction error eliminated by providing information on X.

R-square

- By definition, the residuals have to be smaller than the deviations from the mean, so we might ask the following: how much lower is the SS_{res} compared to the SS_{tot}?
- We quantify this question with the coefficient of determination or R^2. This is the following:

$$R^2 = \frac{SS_{tot} - SS_{res}}{SS_{tot}} = 1 - \frac{SS_{res}}{SS_{tot}}$$

- This is the fraction of the total prediction error eliminated by providing information on X.
- Alternatively, this is the fraction of the variation in Y is "explained by" X.

R-square

- By definition, the residuals have to be smaller than the deviations from the mean, so we might ask the following: how much lower is the SS_{res} compared to the SS_{tot}?
- We quantify this question with the coefficient of determination or R^2. This is the following:

$$R^2 = \frac{SS_{tot} - SS_{res}}{SS_{tot}} = 1 - \frac{SS_{res}}{SS_{tot}}$$

- This is the fraction of the total prediction error eliminated by providing information on X.
- Alternatively, this is the fraction of the variation in Y is "explained by" X.
- $R^2 = 0$ means no relationship

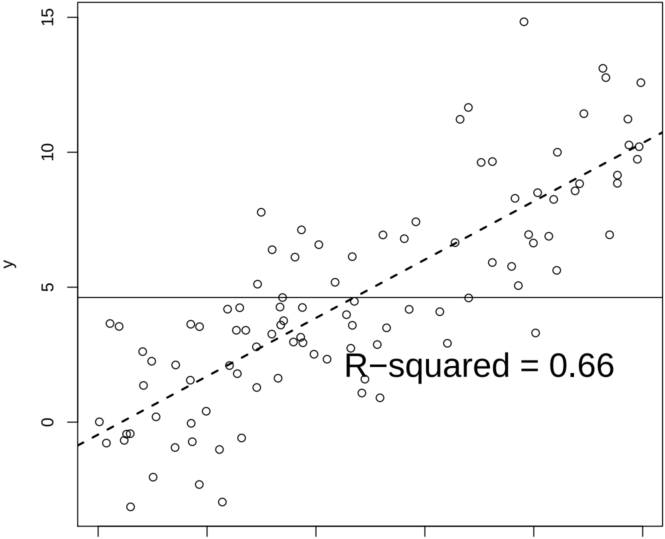
R-square

- By definition, the residuals have to be smaller than the deviations from the mean, so we might ask the following: how much lower is the SS_{res} compared to the SS_{tot}?
- We quantify this question with the coefficient of determination or R^2. This is the following:

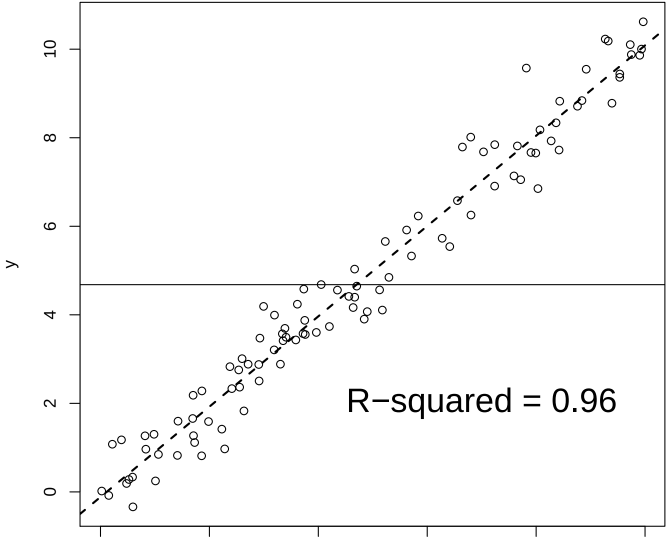
$$R^2 = \frac{SS_{tot} - SS_{res}}{SS_{tot}} = 1 - \frac{SS_{res}}{SS_{tot}}$$

- This is the fraction of the total prediction error eliminated by providing information on X.
- Alternatively, this is the fraction of the variation in Y is "explained by" X.
- $R^2 = 0$ means no relationship
- $R^2 = 1$ implies perfect linear fit

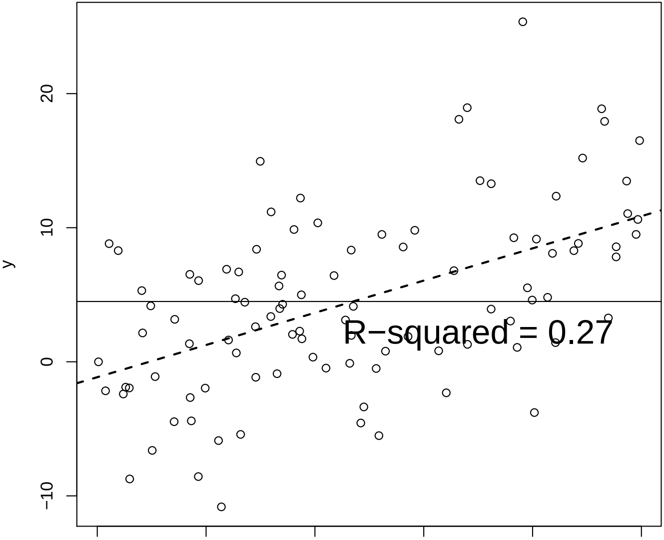
Is R-squared useful?



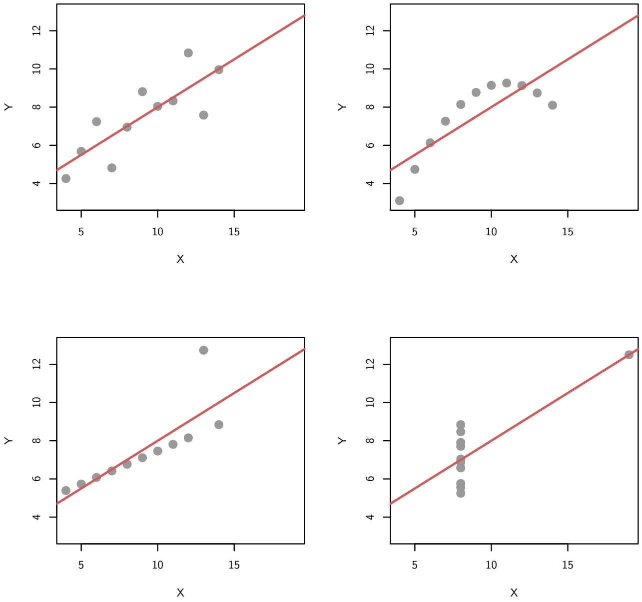
Is R-squared useful?



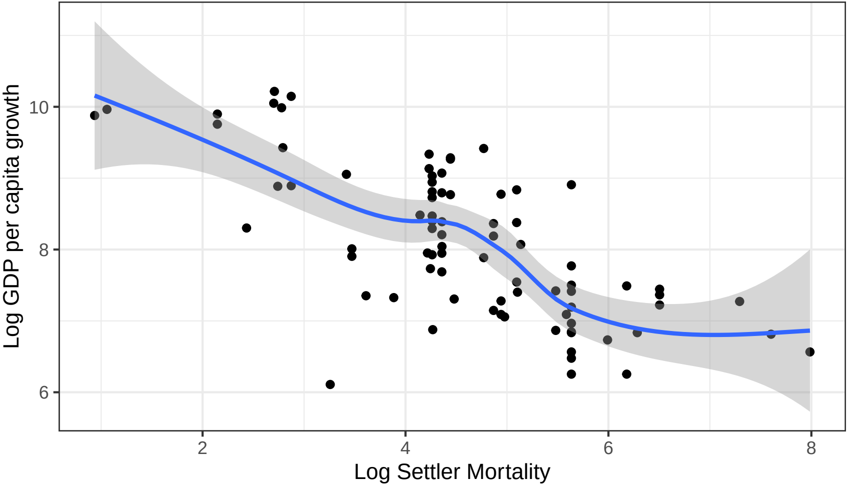
Is R-squared useful?



Is R-squared useful?



So what is ggplot2 doing?



LOESS

LOESS

- We can combine the nonparametric kernel method idea of using only local data with a parametric model

LOESS

- We can combine the nonparametric kernel method idea of using only local data with a parametric model
- Idea: fit a linear regression within each band

LOESS

- We can combine the nonparametric kernel method idea of using only local data with a parametric model
- Idea: fit a linear regression within each band
- Locally weighted scatterplot smoothing (LOWESS or LOESS):

LOESS

- We can combine the nonparametric kernel method idea of using only local data with a parametric model
- Idea: fit a linear regression within each band
- Locally weighted scatterplot smoothing (LOWESS or LOESS):
    1. Pick a subset of the data that falls in the interval $[x - h, x + h]$

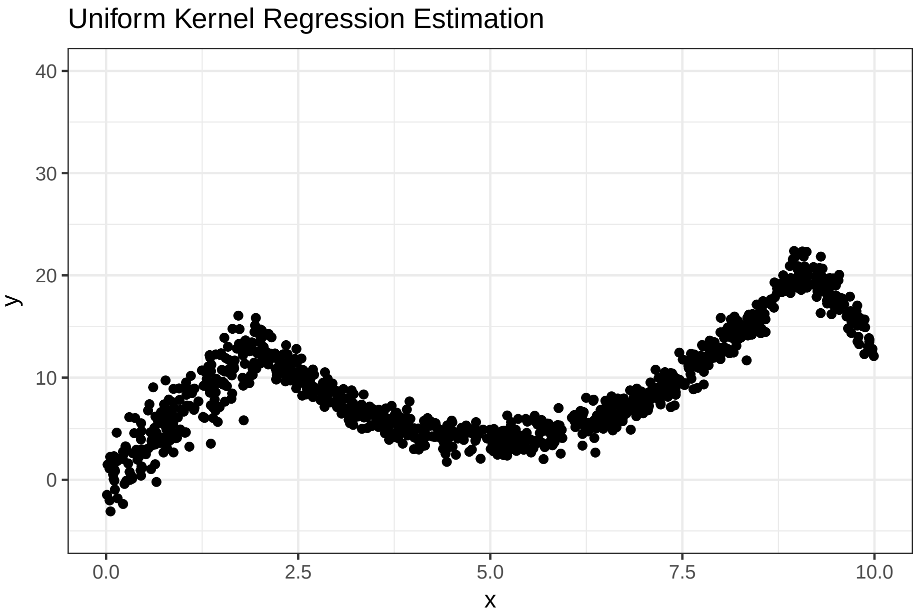
LOESS

- We can combine the nonparametric kernel method idea of using only local data with a parametric model
- Idea: fit a linear regression within each band
- Locally weighted scatterplot smoothing (LOWESS or LOESS):
  1. Pick a subset of the data that falls in the interval $[x - h, x + h]$
  2. Fit a line to this subset of the data (= local linear regression), weighting the points by their distance to x using a kernel function

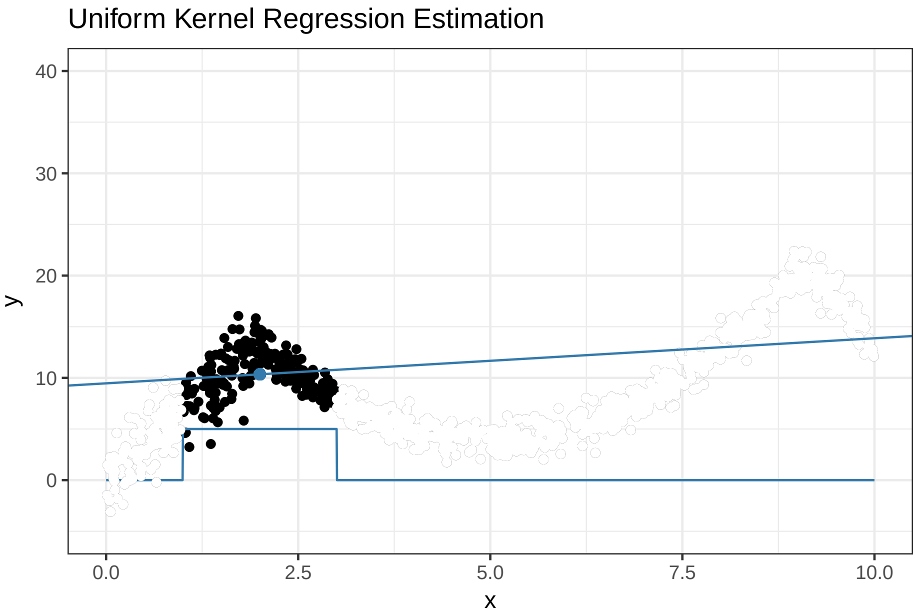
LOESS

- We can combine the nonparametric kernel method idea of using only local data with a parametric model
- Idea: fit a linear regression within each band
- Locally weighted scatterplot smoothing (LOWESS or LOESS):
  1. Pick a subset of the data that falls in the interval $[x - h, x + h]$
  2. Fit a line to this subset of the data (= local linear regression), weighting the points by their distance to x using a kernel function
  3. Use the fitted regression line to predict the expected value of $E[Y|X = x_0]$

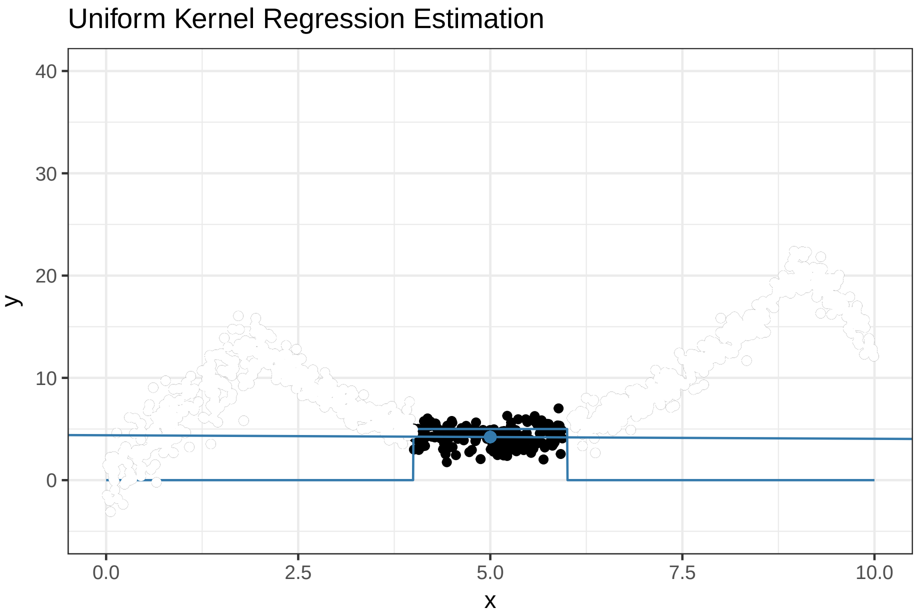
LOESS Example



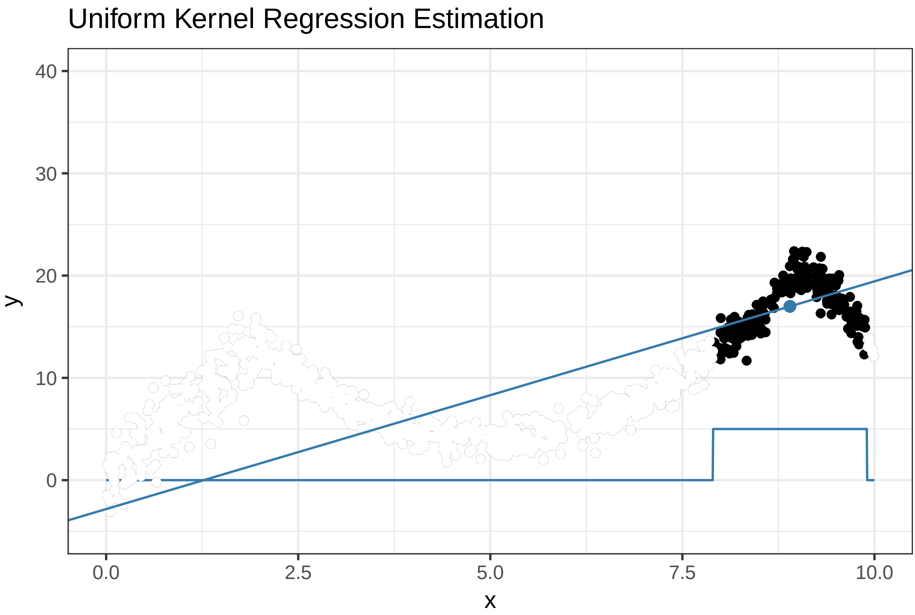
LOESS Example



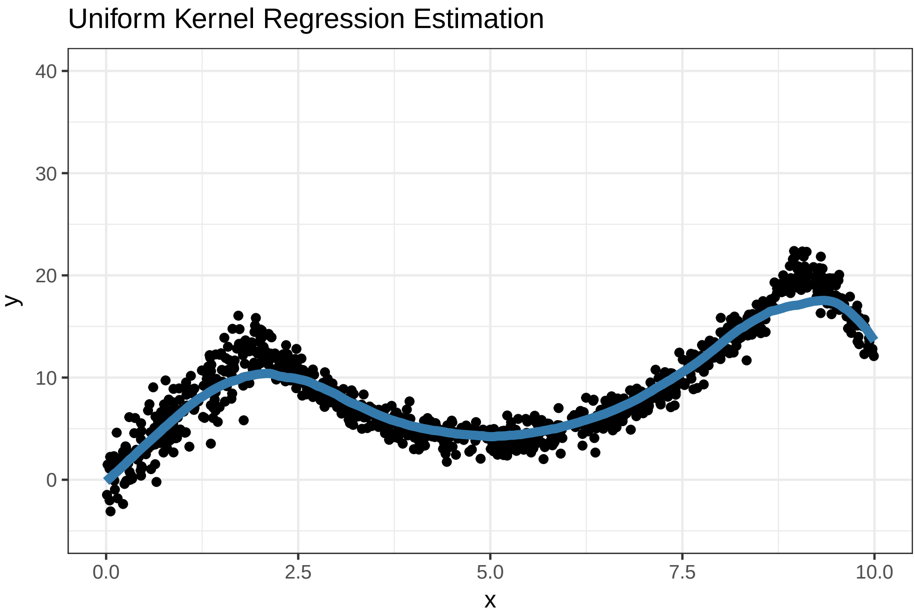
LOESS Example



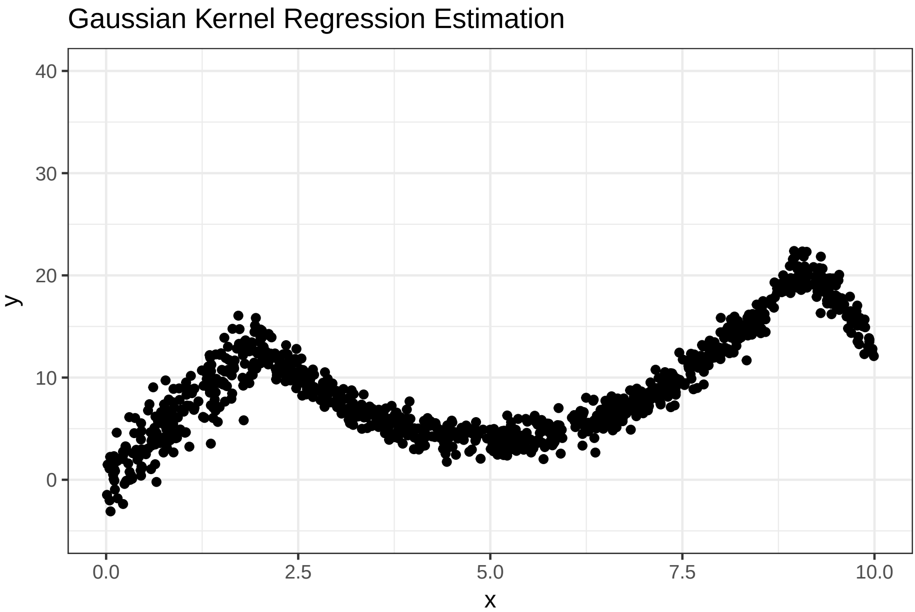
LOESS Example



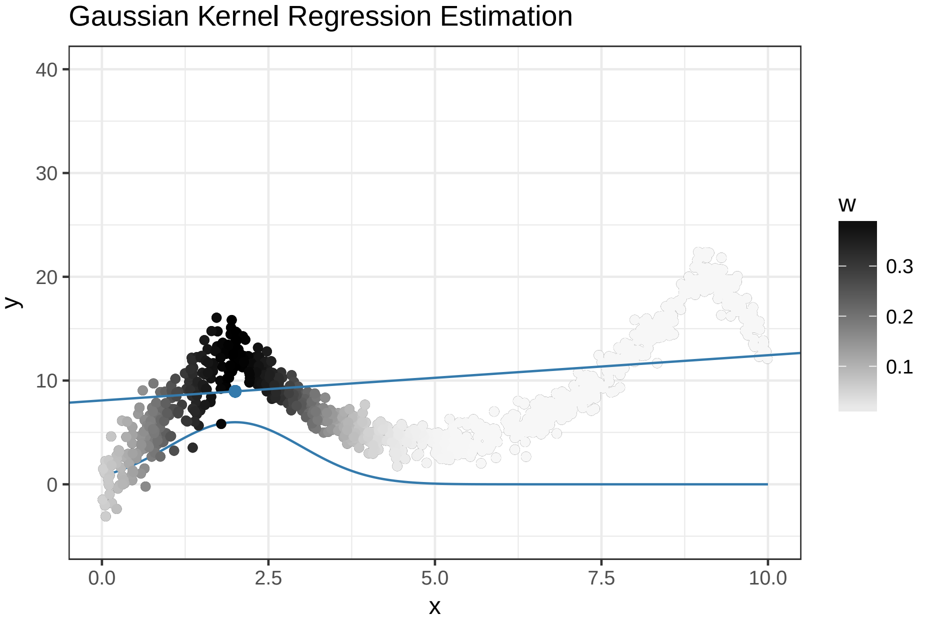
LOESS Example



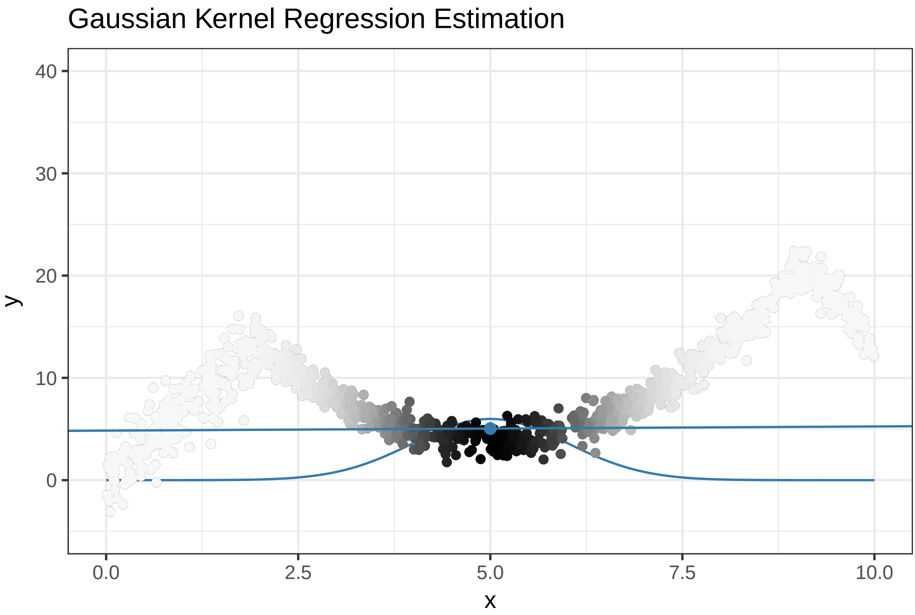
LOESS Example



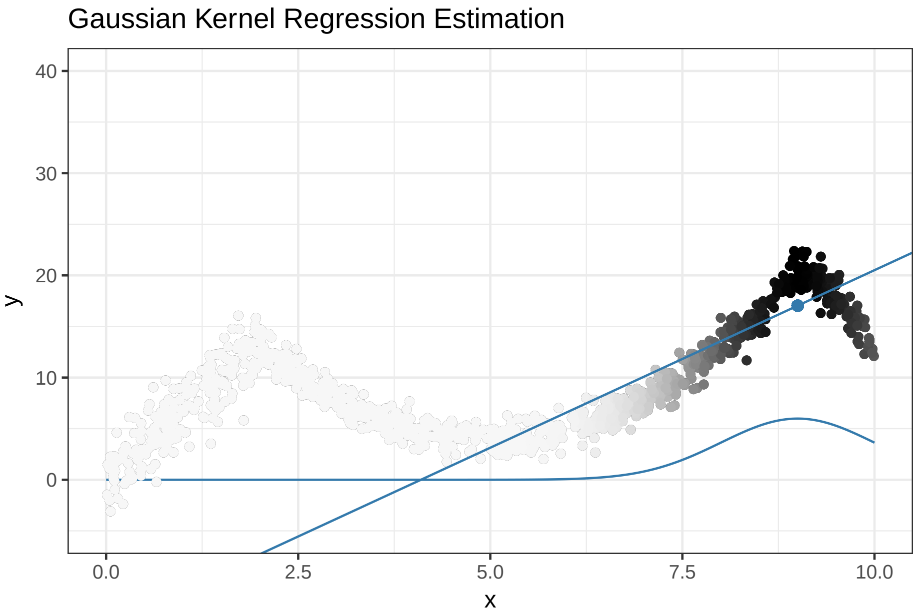
LOESS Example



LOESS Example



LOESS Example



Non-linear CEFs

Non-linear CEFs

- When we say that CEFs are linear with regression, we mean linear in parameters but by including transformations of our variables we can make non-linear shapes of pre-specified functional forms.

Non-linear CEFs

- When we say that CEFs are linear with regression, we mean linear in parameters but by including transformations of our variables we can make non-linear shapes of pre-specified functional forms.
- Many of these non-linear transformations are made by creating multiple variables out of a single X and so will have to wait for future weeks.

Non-linear CEFs

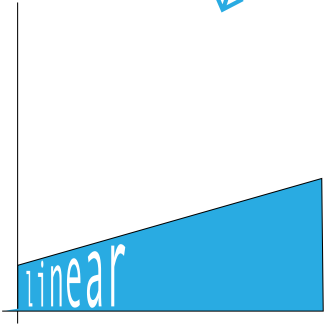
- When we say that CEFs are linear with regression, we mean linear in parameters but by including transformations of our variables we can make non-linear shapes of pre-specified functional forms.
- Many of these non-linear transformations are made by creating multiple variables out of a single X and so will have to wait for future weeks.
- The function $\log(\cdot)$ is one common transformation that has only one parameter.

Non-linear CEFs

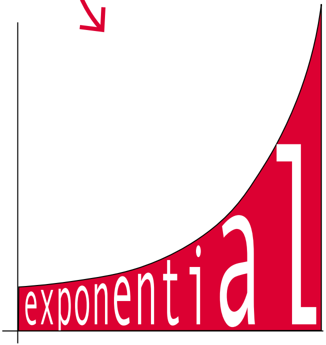
- When we say that CEFs are linear with regression, we mean linear in parameters but by including transformations of our variables we can make non-linear shapes of pre-specified functional forms.
- Many of these non-linear transformations are made by creating multiple variables out of a single X and so will have to wait for future weeks.
- The function $\log(\cdot)$ is one common transformation that has only one parameter.
- This is particularly useful for positive and right-skewed variables.

Why does everyone keep logging stuff??

Logs linearize exponential growth.



grows by a fixed amount.



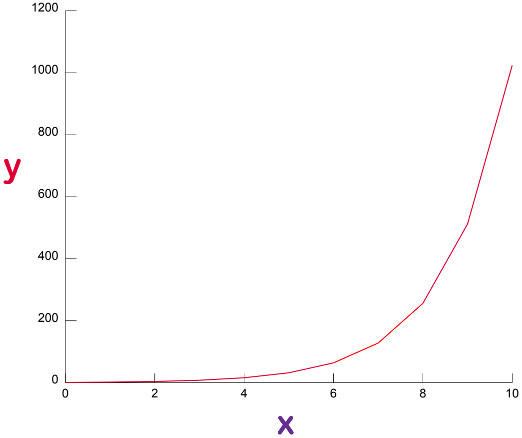
grows by a fixed percent.

How? Let's look.

First, here's a graph showing exponential growth.

We're going to use $y = 2^x$, but any other exponent will work

x	$y=(2^x)$
0	1
1	2
2	4
3	8
4	16
5	32
6	64
7	128
8	256
9	512
10	1024

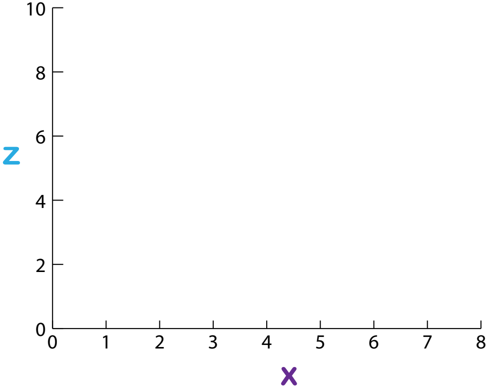


What happens when we take the log of y?

$\log y = z \qquad e^{z} = y$

We're going to use $y = 2^x$, but any other exponent will work

x	y	z
0	1	
1	2	
2	4	
3	8	
4	16	
5	32	
6	64	
7	128	
8	256	
9	512	
10	1024	

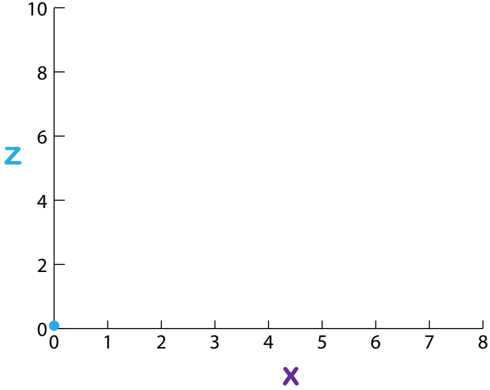


What happens when we take the log of y?

$$\log 1 = 0 \qquad e^0 = 1$$

We're going to use $y = 2^x$, but any other exponent will work

x	y	z
0	1	0
1	2	
2	4	
3	8	
4	16	
5	32	
6	64	
7	128	
8	256	
9	512	
10	1024	

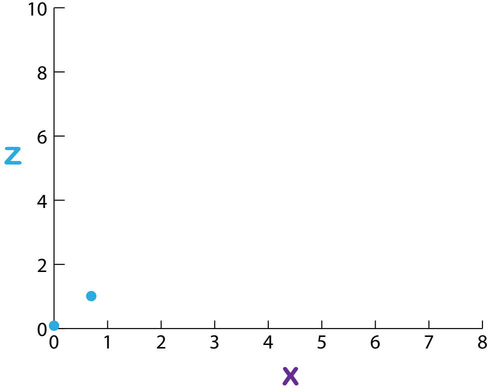


What happens when we take the log of y?

$$\log 2 = .69 \qquad e^{.69} = 2$$

We're going to use $y = 2^x$, but any other exponent will work

x	y	z
0	1	0
1	2	.69
2	4	
3	8	
4	16	
5	32	
6	64	
7	128	
8	256	
9	512	
10	1024	

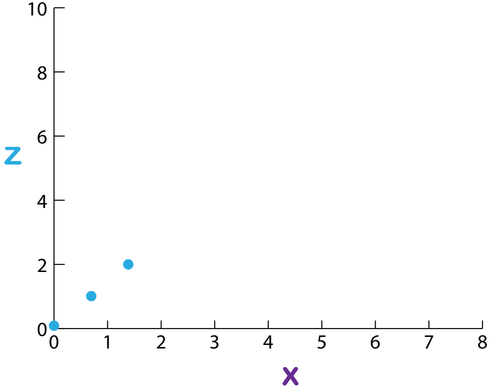


What happens when we take the log of y?

$$\log 4 = 1.39 \qquad e^{1.39} = 4$$

We're going to use $y = 2^x$, but any other exponent will work

x	y	z
0	1	0
1	2	.69
2	4	1.39
3	8	
4	16	
5	32	
6	64	
7	128	
8	256	
9	512	
10	1024	

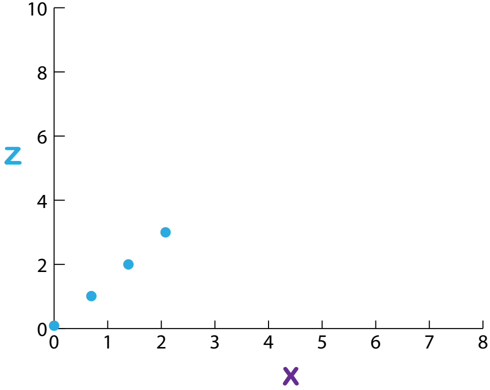


What happens when we take the log of y?

$\log 8 = 2.08 \qquad e^{2.08} = 8$

We're going to use $y = 2^x$, but any other exponent will work

x	y	z
0	1	0
1	2	.69
2	4	1.39
3	8	2.08
4	16	
5	32	
6	64	
7	128	
8	256	
9	512	
10	1024	

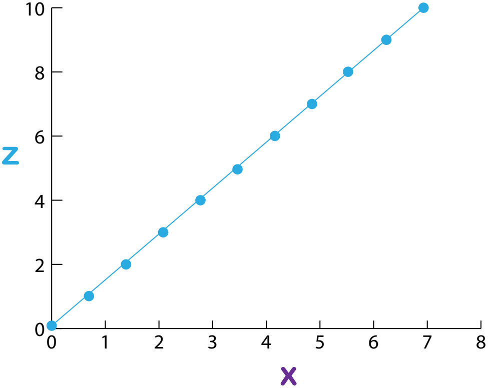


What happens when we take the log of y?

$\log y = z \qquad e^{z} = y$

We're going to use $y = 2^x$, but any other exponent will work

x	y	z
0	1	0
1	2	.69
2	4	1.39
3	8	2.08
4	16	2.77
5	32	3.47
6	64	4.16
7	128	4.85
8	256	5.55
9	512	6.24
10	1024	6.93



Interpretation

Interpretation

The log transformation changes the interpretation of β_1:

Interpretation

The log transformation changes the interpretation of β_1:

- Regress $\log(Y)$ on $X \longrightarrow \beta_1$ approximates percent increase in our prediction of Y associated with one unit increase in X.

Interpretation

The log transformation changes the interpretation of β_1:

- Regress $\log(Y)$ on $X \longrightarrow \beta_1$ approximates percent increase in our prediction of Y associated with one unit increase in X.
- Regress Y on $\log(X) \longrightarrow \beta_1$ approximates increase in Y associated with a percent increase in X.

Interpretation

The log transformation changes the interpretation of β_1:

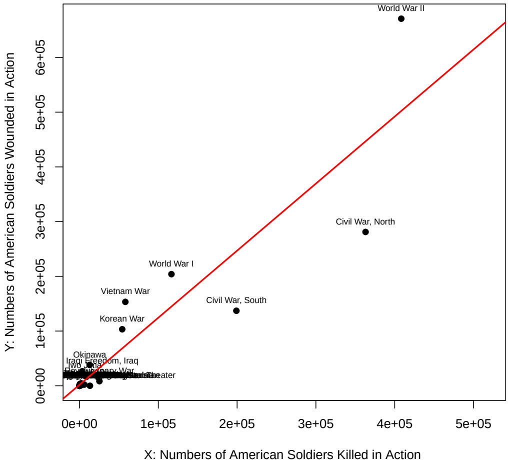
- Regress $\log(Y)$ on $X \longrightarrow \beta_1$ approximates percent increase in our prediction of Y associated with one unit increase in X.
- Regress Y on $\log(X) \longrightarrow \beta_1$ approximates increase in Y associated with a percent increase in X.
- Note that these approximations work only for small increments.

Interpretation

The log transformation changes the interpretation of β_1:

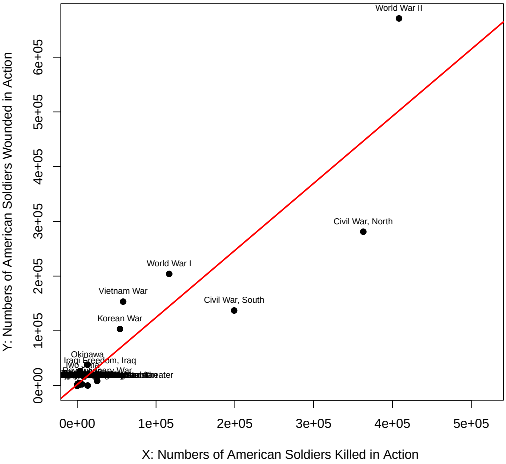
- Regress $\log(Y)$ on $X \longrightarrow \beta_1$ approximates percent increase in our prediction of Y associated with one unit increase in X.
- Regress Y on $\log(X) \longrightarrow \beta_1$ approximates increase in Y associated with a percent increase in X.
- Note that these approximations work only for small increments.
- In particular, they do not work when X is a discrete random variable.

Example from the American War Library



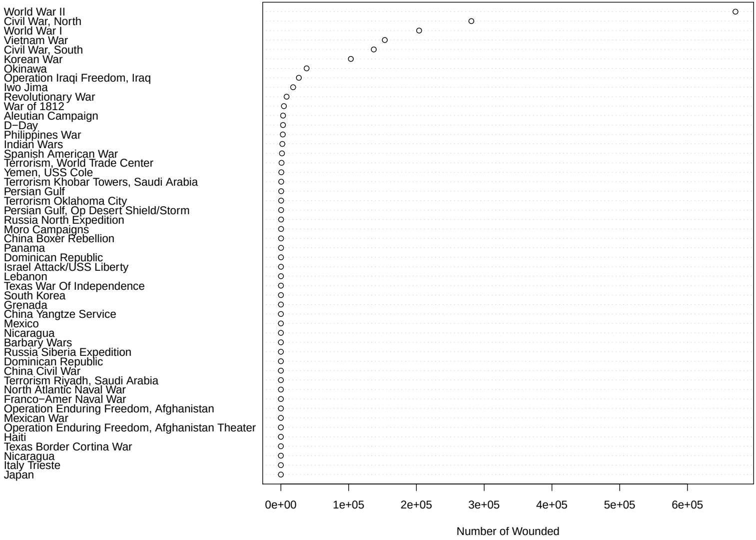
$\hat{\beta}_1 = 1.23 \longrightarrow$

Example from the American War Library

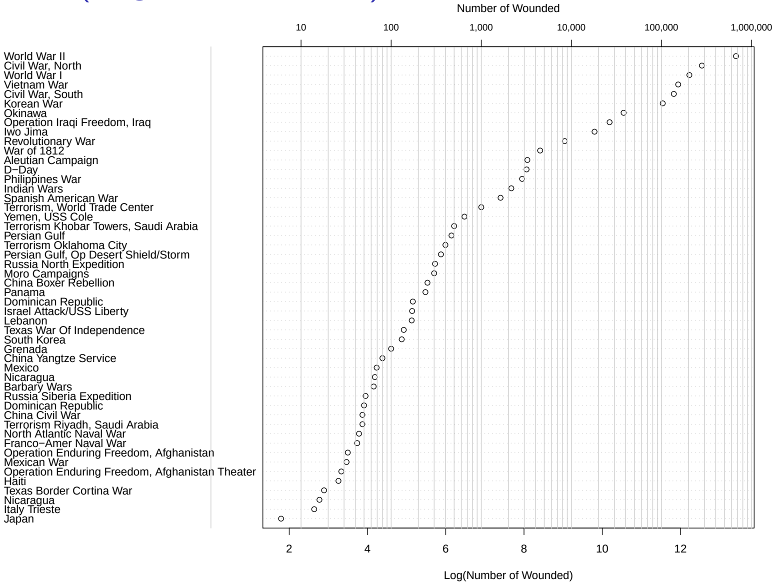


$\hat{\beta}_1 = 1.23 \longrightarrow$ One additional soldier killed predicts 1.23 additional soldiers wounded

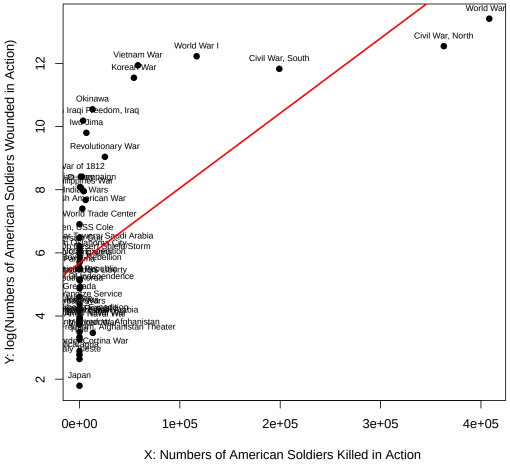
Wounded (Scale in Levels)



Wounded (Logarithmic Scale)

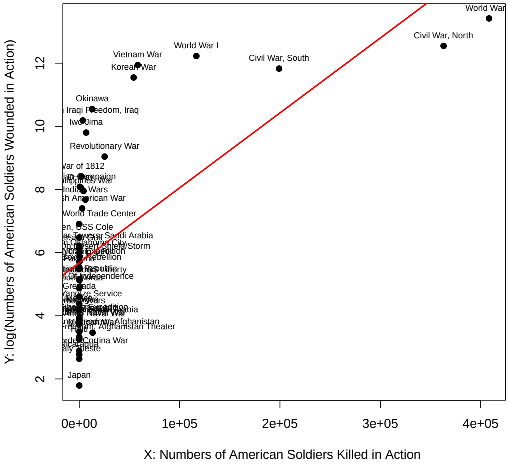


Regression: Log-Level



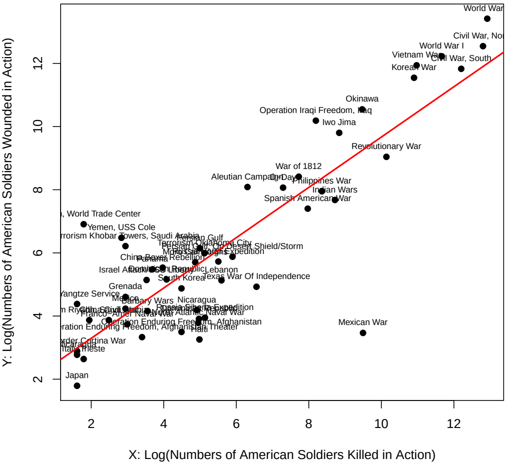
$\hat{\beta}_1 = 0.0000237 \longrightarrow$

Regression: Log-Level



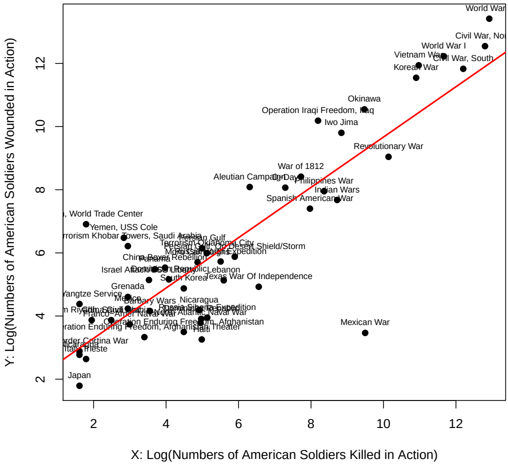
$\hat{\beta}_1 = 0.0000237 \longrightarrow$ One additional soldier killed predicts 0.0023 percent increase in the number of soldiers wounded

Regression: Log-Log



$\hat{\beta}_1 = 0.797 \longrightarrow$

Regression: Log-Log



$\hat{\beta}_1 = 0.797 \longrightarrow$ A percent increase in deaths predicts 0.797 percent increase in the wounded

Four Most Commonly Used Models

Model	Equation	β_1 Interpretation
Level-Level	$Y = \beta_0 + \beta_1 X$	$\Delta Y = \beta_1 \Delta X$
Log-Level	$log(Y) = \beta_0 + \beta_1 X$	$\% \Delta Y = 100 \beta_1 \Delta X$
Level-Log	$Y = \beta_0 + \beta_1 log(X)$	$\Delta Y = (\beta_1 / 100) \% \Delta X$
Log-Log	$log(Y) = \beta_0 + \beta_1 log(X)$	$\% \Delta Y = \beta_1 \% \Delta X$

Why Does This Approximation Work?

Why Does This Approximation Work?

A useful thing to know is that for small x,

$$\begin{aligned} \log(1+x) &\approx x \\ \exp(x) &\approx 1+x \end{aligned}$$

Why Does This Approximation Work?

A useful thing to know is that for small x,

$$\begin{aligned} \log(1+x) &\approx x \\ \exp(x) &\approx 1+x \end{aligned}$$

This can be derived from a series expansion of the log function. Numerically, when $|x| \leq .1$, the approximation is within 0.001.

Why Does This Approximation Work?

Take two numbers $a > b > 0$. The percentage difference between a and b is

$$p = 100\left(\frac{a-b}{b}\right)$$

Why Does This Approximation Work?

Take two numbers $a > b > 0$. The percentage difference between a and b is

$$p = 100\left(\frac{a-b}{b}\right)$$

We can rewrite this as

$$\frac{a}{b} = 1 + \frac{p}{100}$$

Why Does This Approximation Work?

Take two numbers $a > b > 0$. The percentage difference between a and b is

$$p = 100\left(\frac{a-b}{b}\right)$$

We can rewrite this as

$$\frac{a}{b} = 1 + \frac{p}{100}$$

Taking natural logs

$$\log(a) - \log(b) = \log\left(1 + \frac{p}{100}\right)$$

Why Does This Approximation Work?

Take two numbers $a > b > 0$. The percentage difference between a and b is

$$p = 100\left(\frac{a-b}{b}\right)$$

We can rewrite this as

$$\frac{a}{b} = 1 + \frac{p}{100}$$

Taking natural logs

$$\log(a) - \log(b) = \log\left(1 + \frac{p}{100}\right)$$

Applying our approximation and multiplying by 100 we find,

$$p \approx 100\left(\log(a) - \log(b)\right)$$

Be Careful: Log-Level with binary X

Assume we have: $log(Y) = \beta_0 + \beta_1 X$ where X is binary with values 1 or 0. Assume $\beta_1 > .2$. What is the problem with saying that a one unit increase in X is associated with a $\beta_1 \cdot 100$ percent change in Y?

Be Careful: Log-Level with binary X

Assume we have: $log(Y) = \beta_0 + \beta_1 X$ where X is binary with values 1 or 0. Assume $\beta_1 > .2$. What is the problem with saying that a one unit increase in X is associated with a $\beta_1 \cdot 100$ percent change in Y?

Log approximation is inaccurate for large changes like going from $X = 0$ to $X = 1$.

Be Careful: Log-Level with binary X

Assume we have: $log(Y) = \beta_0 + \beta_1 X$ where X is binary with values 1 or 0. Assume $\beta_1 > .2$. What is the problem with saying that a one unit increase in X is associated with a $\beta_1 \cdot 100$ percent change in Y?

Log approximation is inaccurate for large changes like going from $X = 0$ to $X = 1$. Instead the percent change in Y when X goes from 0 to 1 needs to be computed using:

$$100(\hat{Y}_{X=1} - \hat{Y}_{X=0})/\hat{Y}_{X=0} \quad = \quad 100((\hat{Y}_{X=1}/\hat{Y}_{X=0}) - 1)$$

Be Careful: Log-Level with binary X

Assume we have: $log(Y) = \beta_0 + \beta_1 X$ where X is binary with values 1 or 0. Assume $\beta_1 > .2$. What is the problem with saying that a one unit increase in X is associated with a $\beta_1 \cdot 100$ percent change in Y?

Log approximation is inaccurate for large changes like going from $X = 0$ to $X = 1$. Instead the percent change in Y when X goes from 0 to 1 needs to be computed using:

$$
\begin{aligned}
100(\hat{Y}_{X=1} - \hat{Y}_{X=0})/\hat{Y}_{X=0} &= 100((\hat{Y}_{X=1}/\hat{Y}_{X=0}) - 1) \\
&= 100((\hat{Y}_{X=1}/\hat{Y}_{X=0}) - 1)
\end{aligned}
$$

Be Careful: Log-Level with binary X

Assume we have: $log(Y) = \beta_0 + \beta_1 X$ where X is binary with values 1 or 0. Assume $\beta_1 > .2$. What is the problem with saying that a one unit increase in X is associated with a $\beta_1 \cdot 100$ percent change in Y?

Log approximation is inaccurate for large changes like going from $X = 0$ to $X = 1$. Instead the percent change in Y when X goes from 0 to 1 needs to be computed using:

$$
\begin{aligned}
100(\hat{Y}_{X=1} - \hat{Y}_{X=0})/\hat{Y}_{X=0} &= 100((\hat{Y}_{X=1}/\hat{Y}_{X=0}) - 1) \\
&= 100((\hat{Y}_{X=1}/\hat{Y}_{X=0}) - 1) \\
&= 100(exp(\hat{\beta}_1) - 1)
\end{aligned}
$$

Be Careful: Log-Level with binary X

Assume we have: $log(Y) = \beta_0 + \beta_1 X$ where X is binary with values 1 or 0. Assume $\beta_1 > .2$. What is the problem with saying that a one unit increase in X is associated with a $\beta_1 \cdot 100$ percent change in Y?

Log approximation is inaccurate for large changes like going from $X = 0$ to $X = 1$. Instead the percent change in Y when X goes from 0 to 1 needs to be computed using:

$$\begin{aligned} 100(\hat{Y}_{X=1} - \hat{Y}_{X=0})/\hat{Y}_{X=0} &= 100((\hat{Y}_{X=1}/\hat{Y}_{X=0}) - 1) \\ &= 100((\hat{Y}_{X=1}/\hat{Y}_{X=0}) - 1) \\ &= 100(exp(\hat{\beta}_1) - 1) \end{aligned}$$

Recall: $log(\hat{Y}_{X=1}) - log(\hat{Y}_{X=0}) = log(\hat{Y}_{X=1}/\hat{Y}_{X=0}) = \hat{\beta}_1$.

Be Careful: Log-Level with binary X

Assume we have: $log(Y) = \beta_0 + \beta_1 X$ where X is binary with values 1 or 0. Assume $\beta_1 > .2$. What is the problem with saying that a one unit increase in X is associated with a $\beta_1 \cdot 100$ percent change in Y?

Log approximation is inaccurate for large changes like going from $X = 0$ to $X = 1$. Instead the percent change in Y when X goes from 0 to 1 needs to be computed using:

$$
\begin{aligned}
100(\hat{Y}_{X=1} - \hat{Y}_{X=0})/\hat{Y}_{X=0} &= 100((\hat{Y}_{X=1}/\hat{Y}_{X=0}) - 1) \\
&= 100((\hat{Y}_{X=1}/\hat{Y}_{X=0}) - 1) \\
&= 100(exp(\hat{\beta}_1) - 1)
\end{aligned}
$$

Recall: $log(\hat{Y}_{X=1}) - log(\hat{Y}_{X=0}) = log(\hat{Y}_{X=1}/\hat{Y}_{X=0}) = \hat{\beta}_1$.

Be Careful: Log-Level with binary X

Assume we have: $log(Y) = \beta_0 + \beta_1 X$ where X is binary with values 1 or 0. Assume $\beta_1 > .2$. What is the problem with saying that a one unit increase in X is associated with a $\beta_1 \cdot 100$ percent change in Y?

Log approximation is inaccurate for large changes like going from $X = 0$ to $X = 1$. Instead the percent change in Y when X goes from 0 to 1 needs to be computed using:

$$\begin{aligned} 100(\hat{Y}_{X=1} - \hat{Y}_{X=0})/\hat{Y}_{X=0} &= 100((\hat{Y}_{X=1}/\hat{Y}_{X=0}) - 1) \\ &= 100((\hat{Y}_{X=1}/\hat{Y}_{X=0}) - 1) \\ &= 100(exp(\hat{\beta}_1) - 1) \end{aligned}$$

Recall: $log(\hat{Y}_{X=1}) - log(\hat{Y}_{X=0}) = log(\hat{Y}_{X=1}/\hat{Y}_{X=0}) = \hat{\beta}_1$.

A one unit change in X (ie. going from 0 to 1) creates $100(exp(\hat{\beta}_1) - 1)$ percent increase in our prediction $\hat{Y}$.

Interpreting a Logged Outcome

Interpreting a Logged Outcome

- On the last few slides, there was a bit that was a little dodgy.

Interpreting a Logged Outcome

- On the last few slides, there was a bit that was a little dodgy.
- When we log the outcome, we are no longer approximating $E[Y|X]$ we are approximating $E[\log(Y)|X]$.

Interpreting a Logged Outcome

- On the last few slides, there was a bit that was a little dodgy.
- When we log the outcome, we are no longer approximating $E[Y|X]$ we are approximating $E[\log(Y)|X]$.
- Jensen's inequality gives us information on this relation: $f(E[X]) \leq E[f(X)]$ for any convex function $f()$.

Interpreting a Logged Outcome

- On the last few slides, there was a bit that was a little dodgy.
- When we log the outcome, we are no longer approximating $E[Y|X]$ we are approximating $E[\log(Y)|X]$.
- Jensen's inequality gives us information on this relation: $f(E[X]) \leq E[f(X)]$ for any convex function $f()$.
- In practice, this means we are no longer characterizing the expectation of Y and it is technically inaccurate to talk about Y 'on average' changing in a certain way.

Interpreting a Logged Outcome

- On the last few slides, there was a bit that was a little dodgy.
- When we log the outcome, we are no longer approximating $E[Y|X]$ we are approximating $E[\log(Y)|X]$.
- Jensen's inequality gives us information on this relation: $f(E[X]) \leq E[f(X)]$ for any convex function $f()$.
- In practice, this means we are no longer characterizing the expectation of Y and it is technically inaccurate to talk about Y 'on average' changing in a certain way.
- What are we characterizing? The geometric mean.

Geometric Mean

$$\exp(E(\log(Y))) = \exp\left(\frac{1}{N}\sum_{i=1}^{N}\log(Y_i)\right)$$

Geometric Mean

$$\exp(E(\log(Y))) = \exp\left(\frac{1}{N}\sum_{i=1}^{N}\log(Y_i)\right)$$
$$= \exp\left(\frac{1}{N}\log(\prod_{i=1}^{N} Y_i)\right)$$

Geometric Mean

$$
\begin{aligned}
\exp(E(\log(Y))) &= \exp\left(\frac{1}{N}\sum_{i=1}^{N}\log(Y_i)\right) \\
&= \exp\left(\frac{1}{N}\log(\prod_{i=1}^{N} Y_i)\right) \\
&= \exp\left(\log\left(\left(\prod_{i=1}^{N} Y_i\right)^{\frac{1}{N}}\right)\right)
\end{aligned}
$$

Geometric Mean

$$\begin{aligned}
\exp(E(\log(Y))) &= \exp\left(\frac{1}{N}\sum_{i=1}^{N}\log(Y_i)\right) \\
&= \exp\left(\frac{1}{N}\log(\prod_{i=1}^{N}Y_i)\right) \\
&= \exp\left(\log\left(\left(\prod_{i=1}^{N}Y_i\right)^{\frac{1}{N}}\right)\right) \\
&= \left(\prod_{i=1}^{N}Y_i\right)^{\frac{1}{N}}
\end{aligned}$$

Geometric Mean

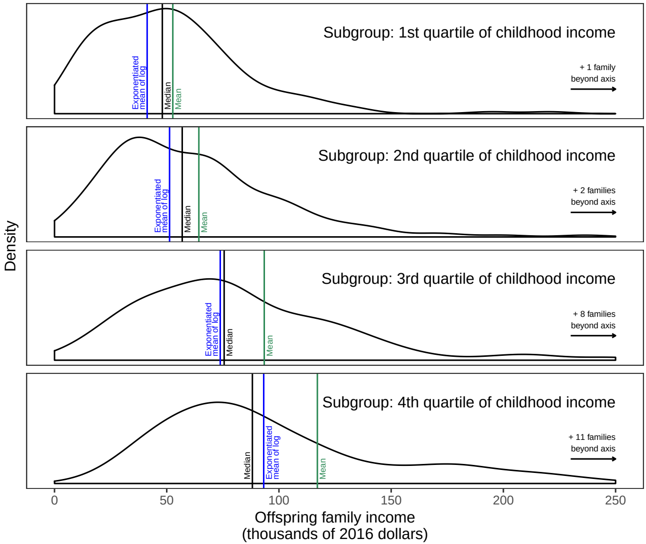
$$
\begin{aligned}
\exp(E(\log(Y))) &= \exp\left(\frac{1}{N}\sum_{i=1}^{N}\log(Y_i)\right) \\
&= \exp\left(\frac{1}{N}\log(\prod_{i=1}^{N}Y_i)\right) \\
&= \exp\left(\log\left(\left(\prod_{i=1}^{N}Y_i\right)^{\frac{1}{N}}\right)\right) \\
&= \left(\prod_{i=1}^{N}Y_i\right)^{\frac{1}{N}} \\
&= \text{Geometric Mean}(Y)
\end{aligned}
$$

Geometric Mean

$$
\begin{aligned}
\exp(E(\log(Y))) &= \exp\left(\frac{1}{N}\sum_{i=1}^{N}\log(Y_i)\right) \\
&= \exp\left(\frac{1}{N}\log(\prod_{i=1}^{N}Y_i)\right) \\
&= \exp\left(\log\left(\left(\prod_{i=1}^{N}Y_i\right)^{\frac{1}{N}}\right)\right) \\
&= \left(\prod_{i=1}^{N}Y_i\right)^{\frac{1}{N}} \\
&= \text{Geometric Mean}(Y)
\end{aligned}
$$

The geometric mean is a robust measure of central tendency.

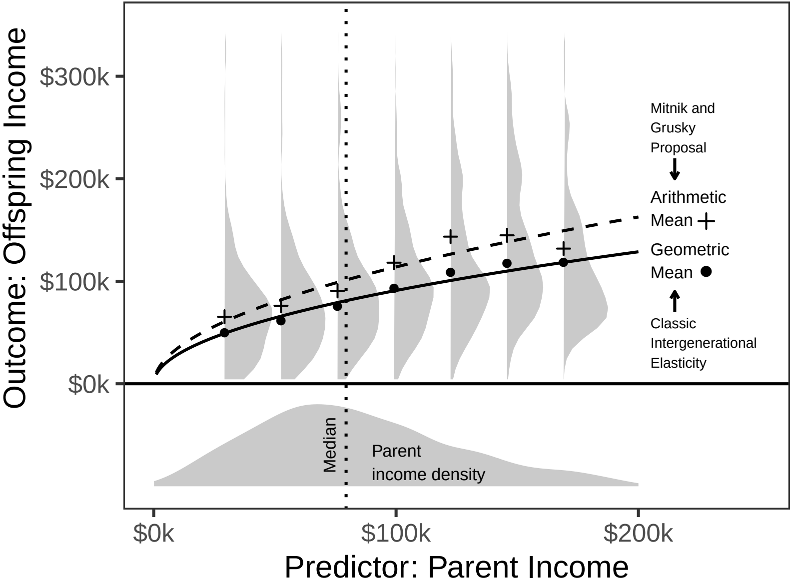
Geometric Mean is Closer to the Median Than the Mean



(Lundberg and Stewart, 2020)

Lundberg and Stewart, 2020

Lundberg and Stewart, 2020



Bootstrapped Sampling Distributions

What if there was a way to replace thinking with computers?

Bootstrapped Sampling Distributions

What if there was a way to replace thinking with computers?

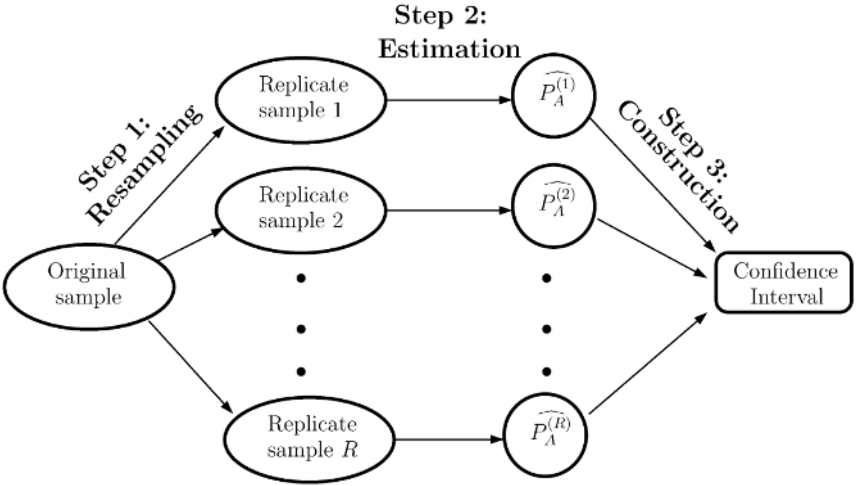
What if there was a way to replacing analytical derivations, which can be hard, with computer simulations which are easy?

Bootstrapped Sampling Distributions

What if there was a way to replace thinking with computers?

What if there was a way to replacing analytical derivations, which can be hard, with computer simulations which are easy?

The plug-in principle gives us a way forward.



Source: Salganik (2006)

This works for almost* any estimator

*basically it works when plug-in estimation works

Statistical Science
1986, Vol. 1, No. 1, 54–77

Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy

B. Efron and R. Tibshirani

Efron and Tibshirani (1986), `http://www.jstor.org/stable/2245500`

The Bootstrap

The Bootstrap

Bootstrap: Use the eCDF as a plug-in for the CDF, and resample from that. I.e. we are pretending our sample eCDF looks sufficiently close to our true CDF, and so we're sampling from the eCDF as an approximation to repeated sampling from the true CDF. This is called a resampling method.

The Bootstrap

Bootstrap: Use the eCDF as a plug-in for the CDF, and resample from that. I.e. we are pretending our sample eCDF looks sufficiently close to our true CDF, and so we're sampling from the eCDF as an approximation to repeated sampling from the true CDF. This is called a resampling method.

1. Take a with replacement sample of size n from our sample.

The Bootstrap

Bootstrap: Use the eCDF as a plug-in for the CDF, and resample from that. I.e. we are pretending our sample eCDF looks sufficiently close to our true CDF, and so we're sampling from the eCDF as an approximation to repeated sampling from the true CDF. This is called a resampling method.

1. Take a with replacement sample of size n from our sample.
2. Calculate our would-be estimate using this bootstrap sample.

The Bootstrap

Bootstrap: Use the eCDF as a plug-in for the CDF, and resample from that. I.e. we are pretending our sample eCDF looks sufficiently close to our true CDF, and so we're sampling from the eCDF as an approximation to repeated sampling from the true CDF. This is called a resampling method.

1. Take a with replacement sample of size n <u>from our sample</u>.
2. Calculate our would-be estimate using this bootstrap sample.
3. Repeat steps 1 and 2 many (B) times.

The Bootstrap

Bootstrap: Use the eCDF as a plug-in for the CDF, and resample from that. I.e. we are pretending our sample eCDF looks sufficiently close to our true CDF, and so we're sampling from the eCDF as an approximation to repeated sampling from the true CDF. This is called a resampling method.

1. Take a with replacement sample of size n from our sample.
2. Calculate our would-be estimate using this bootstrap sample.
3. Repeat steps 1 and 2 many (B) times.
4. Using the resulting collection of bootstrap estimates, calculate the standard deviation of the bootstrap distribution of our estimator. This serves our estimate of the standard deviation of the sampling distribution

Example of a Bootstrap

```r
samp <- c(9.7, 4.99, 5.9, 3.58, 8.15, 5.54, 4.77, 5.01, 4.89,
          3.42, 8.63, 7.17, 8.93, 7.5, 4.93, 8.6, 6.26, 7.31,
          8.96, 3.95)

obs_mean = mean(samp)

obs_mean
```

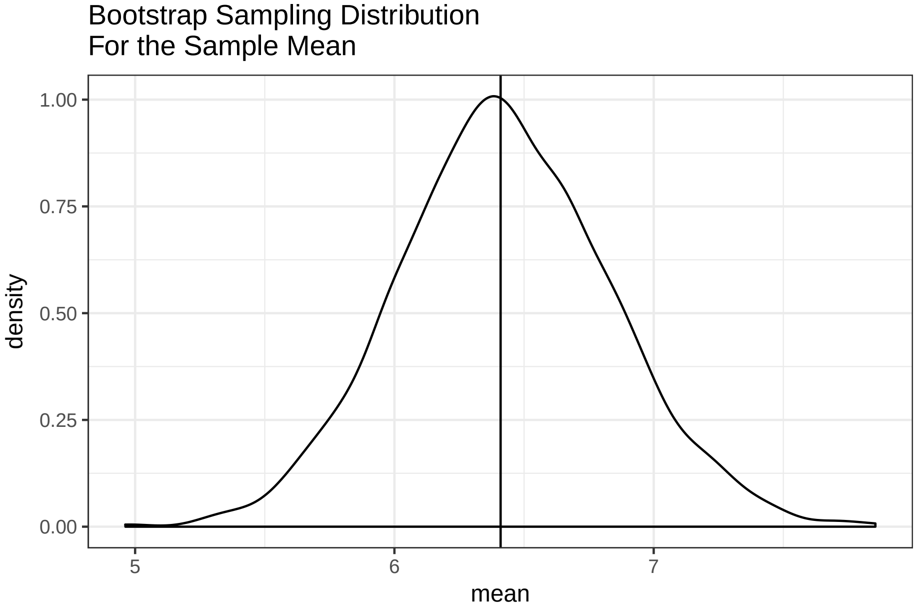
```
## [1] 6.4095
```

Example of a Bootstrap

```r
# resample WITH REPLACEMENT reps times
# recalculate the mean within each bootstrap replicate
boot_samp_dist <- replicate(2000, {
      mean(samp[sample.int(length(samp), replace = TRUE)])
    })
```

```r
ggplot(tibble(boot_samp_dist = boot_samp_dist),
       aes(x = boot_samp_dist)) +
  geom_density() +
  geom_vline(xintercept = obs_mean) +
  theme_bw() + ggtitle("Bootstrap Sampling Distribution
                        For the Sample Mean") +
  xlab("mean")
```

Example of a Bootstrap



Two ways to calculate bootstrap intervals

Two ways to calculate bootstrap intervals

1) Using normal approximation intervals, use the estimates from step 4.

$$\left[\overline{X} - \Phi^{-1}(1-\alpha/2) * \hat{\sigma}_{\text{boot}}, \overline{X} + \Phi^{-1}(1-\alpha/2) * \hat{\sigma}_{\text{boot}}\right]$$

Two ways to calculate bootstrap intervals

1) Using normal approximation intervals, use the estimates from step 4.

$$\left[\overline{X} - \Phi^{-1}(1-\alpha/2) * \hat{\sigma}_{\text{boot}}, \overline{X} + \Phi^{-1}(1-\alpha/2) * \hat{\sigma}_{\text{boot}}\right]$$

- Note here that the standard error is just the standard deviation of the boostrap replicates. There is no square root of n. Why?

Two ways to calculate bootstrap intervals

1) Using normal approximation intervals, use the estimates from step 4.

$$\left[\overline{X} - \Phi^{-1}(1-\alpha/2) * \hat{\sigma}_{\text{boot}}, \overline{X} + \Phi^{-1}(1-\alpha/2) * \hat{\sigma}_{\text{boot}}\right]$$

- Note here that the standard error is just the standard deviation of the boostrap replicates. There is no square root of n. Why?

2) Percentile method for the CI: Sort B bootstrap estimates from smallest to largest. α interval is constructed as

$$CI_{1-\alpha} = \left[\alpha/2 * B \text{ sample}, (1-\alpha/2) * B \text{ sample}\right]$$

- Percentile method does not rely on normal approximation, and behaves better with small n.

This Week in Review

- CEFs!
- OLS!
- Additional Techniques You Will See!

This Week in Review

- CEFs!
- OLS!
- Additional Techniques You Will See!

Going Deeper:

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

This Week in Review

- CEFs!
- OLS!
- Additional Techniques You Will See!

Going Deeper:

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

Next week: Statistical Theory for Linear Regression