Week 5: Simple Linear Regression

Brandon Stewart[1]

Princeton

October 3–5, 2022

[1] Many of these slides are adapted from Matt Blackwell, Adam Glynn, Erin Hartman and Jens Hainmueller. Illustrations by Shay O'Brien.

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

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

- Last Week
  - hypothesis testing

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

- Last Week
  - hypothesis testing
- This Week
  - conditional expectation functions and estimation

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

- Last Week
  - hypothesis testing
- This Week
  - conditional expectation functions and estimation
  - mechanics and properties of simple linear regression

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

- Last Week
  - hypothesis testing
- This Week
  - conditional expectation functions and estimation
  - mechanics and properties of simple linear regression
  - inference and measures of model fit

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

- Last Week
  - hypothesis testing
- This Week
  - conditional expectation functions and estimation
  - mechanics and properties of simple linear regression
  - inference and measures of model fit
- Next Week
  - mechanics with two regressors
  - omitted variables, multicollinearity

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

- Last Week
  - hypothesis testing
- This Week
  - conditional expectation functions and estimation
  - mechanics and properties of simple linear regression
  - inference and measures of model fit
- Next Week
  - mechanics with two regressors
  - omitted variables, multicollinearity
- Long Run
  - probability $\rightarrow$ inference $\rightarrow$ regression $\rightarrow$ causal inference

Narrow Goal for the Week: Understand `lm()` Output

```
Call:
lm(formula = sr ~ pop15, data = LifeCycleSavings)

Residuals:
   Min     1Q Median     3Q    Max
-8.637 -2.374  0.349  2.022 11.155

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.49660    2.27972   7.675 6.85e-10 ***
pop15       -0.22302    0.06291  -3.545 0.000887 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.03 on 48 degrees of freedom
Multiple R-squared:  0.2075,Adjusted R-squared:  0.191
F-statistic: 12.57 on 1 and 48 DF,  p-value: 0.0008866
```

Macrostructure—This Semester

Macrostructure—This Semester

The next few weeks,

- Linear Regression with Two Regressors
- Break Week
- Multiple Linear Regression
- What Can Go Wrong and How to Fix It
- Frameworks for Causal Inference
- Causality with Measured Confounding
- Sensitivity Analysis / Thanksgiving
- Unmeasured Confounding and Instrumental Variables
- Repeated Observations, Panel Data and Wrap-up

Reminder of Definitions

Definition (Conditional Expectation (Discrete))

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

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

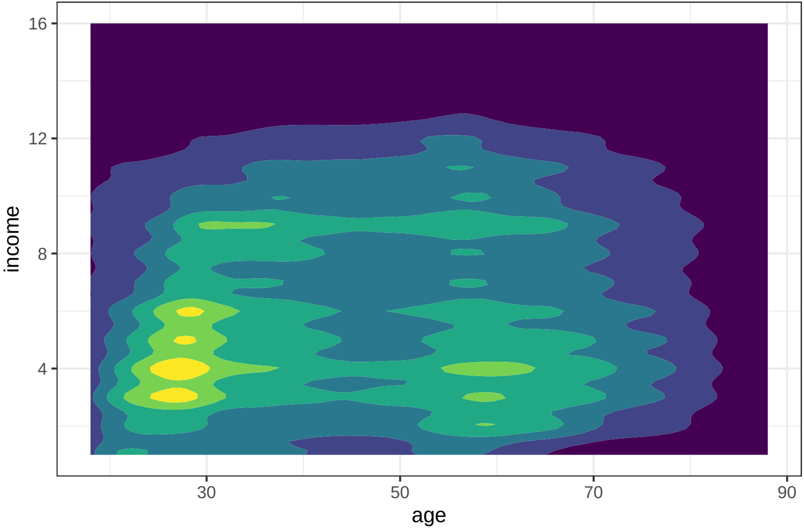
Definition (Conditional Expectation (Continuous))

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

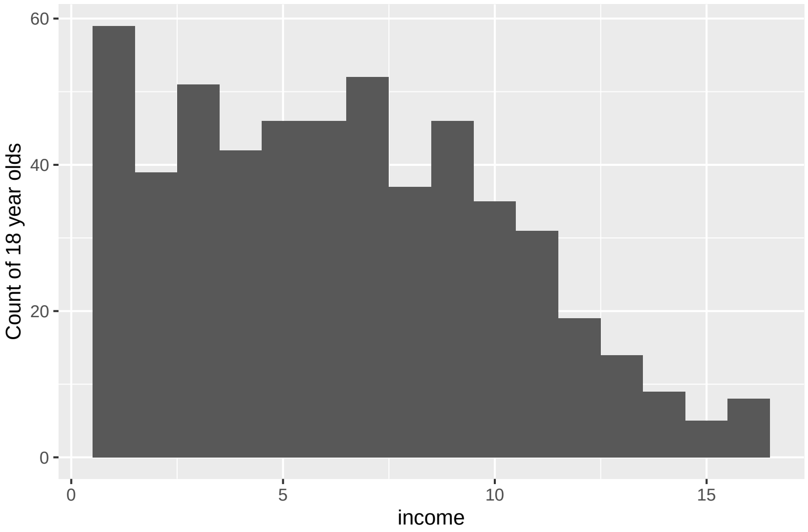
$$E[Y|X = x] = \int_{-\infty}^{\infty} y\, f_{Y|X}(y|x) dy$$

Review of CEF Concepts

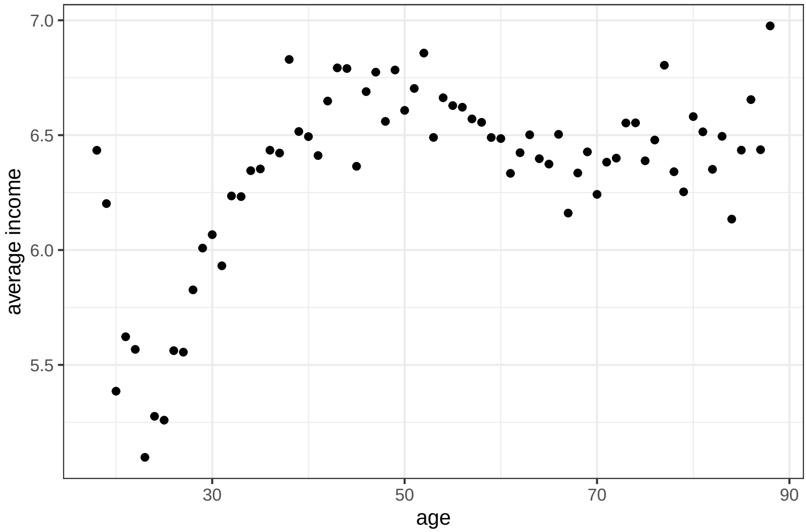
Review of CEF Concepts



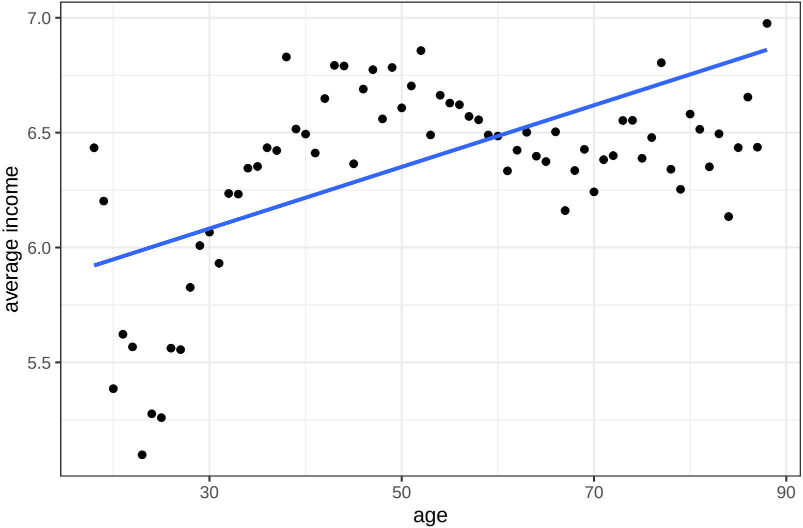
Review of CEF Concepts



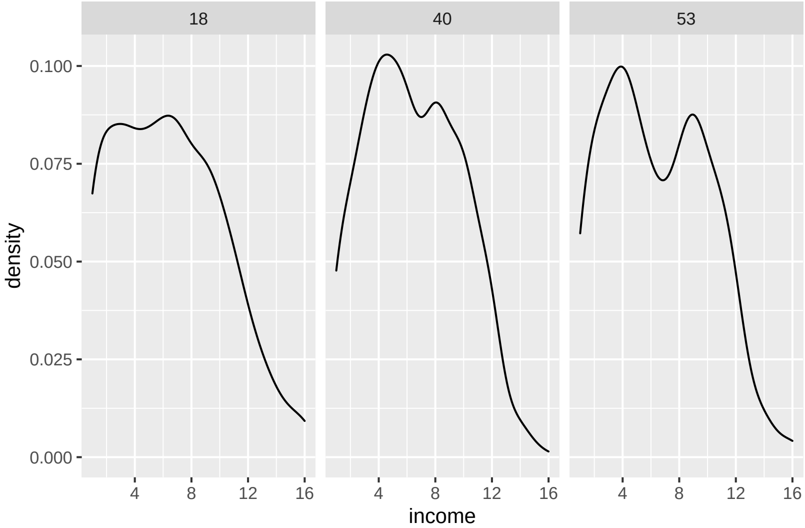
Review of CEF Concepts



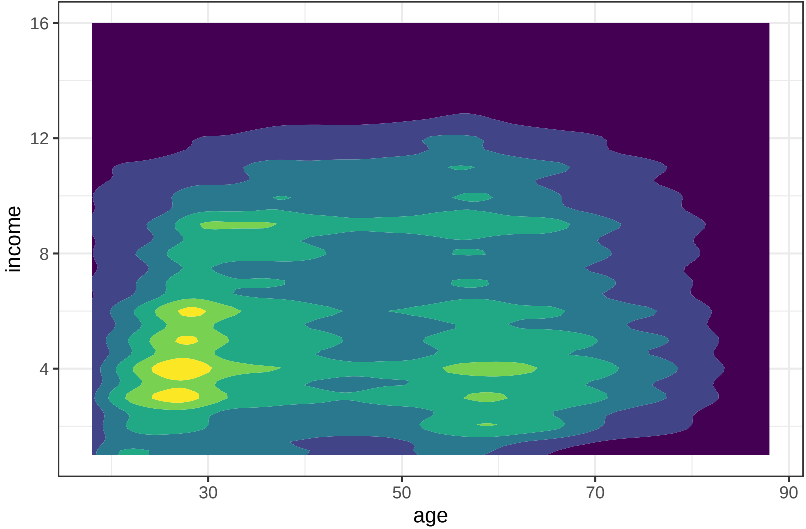
Review of CEF Concepts



Review of CEF Concepts



Review of CEF Concepts



Implications of the CEF Definition

Let X and Y be random variables and $\epsilon = Y - E[Y|X]$, then

Implications of the CEF Definition

Let X and Y be random variables and $\epsilon = Y - E[Y|X]$, then

- the expectation of the error doesn't depend on X

$$E[\epsilon|X] = E[\epsilon] = 0$$

Implications of the CEF Definition

Let X and Y be random variables and $\epsilon = Y - E[Y|X]$, then

- the expectation of the error doesn't depend on X

$$E[\epsilon|X] = E[\epsilon] = 0$$

- the error is uncorrelated with any function g of X

$$Cov[g(X), \epsilon] = 0$$

Implications of the CEF Definition

Let X and Y be random variables and $\epsilon = Y - E[Y|X]$, then

- the expectation of the error doesn't depend on X

$$E[\epsilon|X] = E[\epsilon] = 0$$

- the error is uncorrelated with any function g of X

$$Cov[g(X), \epsilon] = 0$$

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

$$V[\epsilon|X] = V[Y|X]$$

Implications of the CEF Definition

Let X and Y be random variables and $\epsilon = Y - E[Y|X]$, then

- the expectation of the error doesn't depend on X

$$E[\epsilon|X] = E[\epsilon] = 0$$

- the error is uncorrelated with any function g of X

$$Cov[g(X), \epsilon] = 0$$

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

$$V[\epsilon|X] = V[Y|X]$$

- and the CEF is the lowest mean squared error predictor of Y given X.

For proofs see Aronow and Miller Thm 2.2.19, 2.2.20

CEF for binary covariates

CEF for binary covariates

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

CEF for binary covariates

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

CEF for binary covariates

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

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

CEF for binary covariates

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

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

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

CEF for binary covariates

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

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

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

Non-binary CEFs

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

Non-binary CEFs

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

Non-binary CEFs

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

Non-binary CEFs

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

Nonparametric Regression with Discrete X

Nonparametric Regression with Discrete X

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

Nonparametric Regression with Discrete X

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

Nonparametric Regression with Discrete X

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

Nonparametric Regression with Discrete X

Nonparametric Regression with Discrete X

`educ`: Respondent's education:

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

Nonparametric Regression with Discrete X

`income`: Respondent's family income:

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

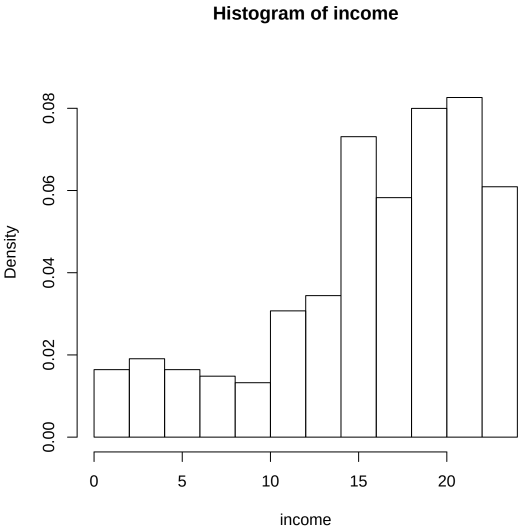
  ⋮

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

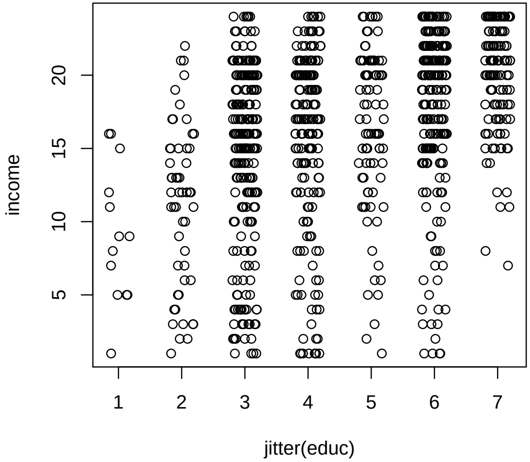
  ⋮

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

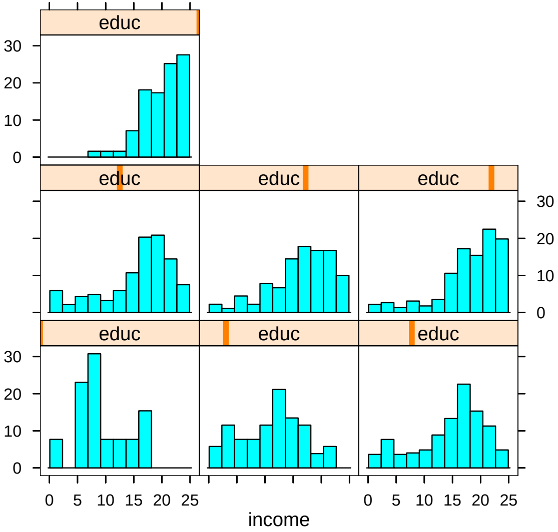
Marginal Distribution of Y (income)



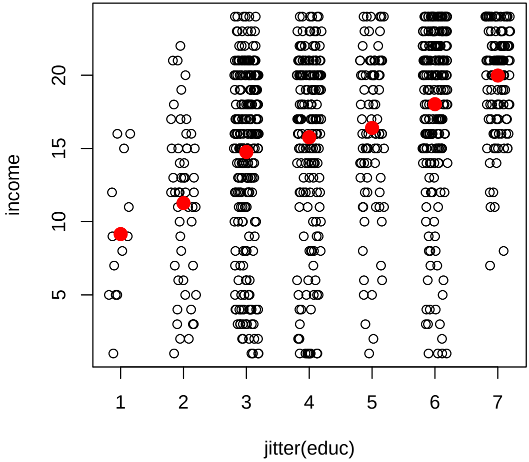
Joint Distribution of X and Y (Income and Education)



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.

Coming up with Estimators

Coming up with Estimators

We now know how to study some properties of estimators, but how do we come up with candidate estimators?

Coming up with Estimators

We now know how to study some properties of estimators, but how do we come up with candidate estimators?

- The simplest way is to use the sample analog.

Coming up with Estimators

We now know how to study some properties of estimators, but how do we come up with candidate estimators?

- The simplest way is to use the sample analog.
- Ex: If we're interested in the population mean, we use the sample mean

Coming up with Estimators

We now know how to study some properties of estimators, but how do we come up with candidate estimators?

- The simplest way is to use the sample analog.
- Ex: If we're interested in the population mean, we use the sample mean
- This is justified because of the plug-in principle.

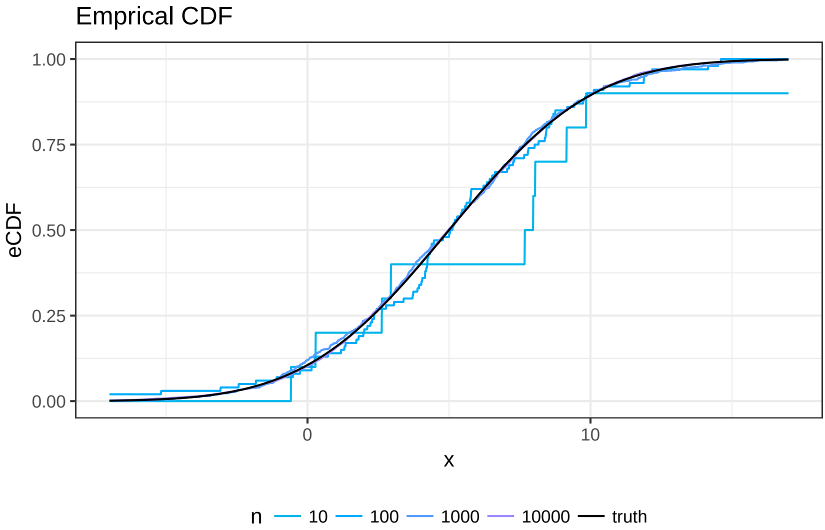
Coming up with Estimators

We now know how to study some properties of estimators, but how do we come up with candidate estimators?

- The simplest way is to use the sample analog.
- Ex: If we're interested in the population mean, we use the sample mean
- This is justified because of the plug-in principle.
- The Weak Law of Large Numbers tells us that the empirical CDF is a good sample analog of the true CDF (which fully describes a distribution).

The Plug-in Principle in Action

Say we have a $\mathcal{N}(5, 4)$ distribution



The Plug-in Principle

The Plug-in Principle

Note that the CDF is:

$$F(x) = P(X \leq x) = E[\mathbb{I}(X \leq x)]$$

The Plug-in Principle

Note that the CDF is:

$$F(x) = P(X \leq x) = E[\mathbb{I}(X \leq x)]$$

we define the empirical CDF (eCDF) as:

$$\hat{F}(x) = \overline{\mathbb{I}(X \leq x)}, \forall x \in \mathbb{R}$$

The Plug-in Principle

Note that the CDF is:

$$F(x) = P(X \leq x) = E[\mathbb{I}(X \leq x)]$$

we define the empirical CDF (eCDF) as:

$$\hat{F}(x) = \overline{\mathbb{I}(X \leq x)}, \forall x \in \mathbb{R}$$

- WLLN tells us that the eCDF will be unbiased and consistently estimated. Any given sample will, on average, look representative of the true distribution.

The Plug-in Principle

Note that the CDF is:

$$F(x) = P(X \leq x) = E[\mathbb{I}(X \leq x)]$$

we define the empirical CDF (eCDF) as:

$$\hat{F}(x) = \overline{\mathbb{I}(X \leq x)}, \forall x \in \mathbb{R}$$

- WLLN tells us that the eCDF will be unbiased and consistently estimated. Any given sample will, on average, look representative of the true distribution.

For iid random variables $X_1, X_2, \ldots, X_n$ with common CDF F, the plug-in estimator of $\theta = T(F)$ is:

$$\hat{\theta} = T(\hat{F})$$

The Plug-in Principle

Note that the CDF is:

$$F(x) = P(X \leq x) = E[\mathbb{I}(X \leq x)]$$

we define the empirical CDF (eCDF) as:

$$\hat{F}(x) = \overline{\mathbb{I}(X \leq x)}, \forall x \in \mathbb{R}$$

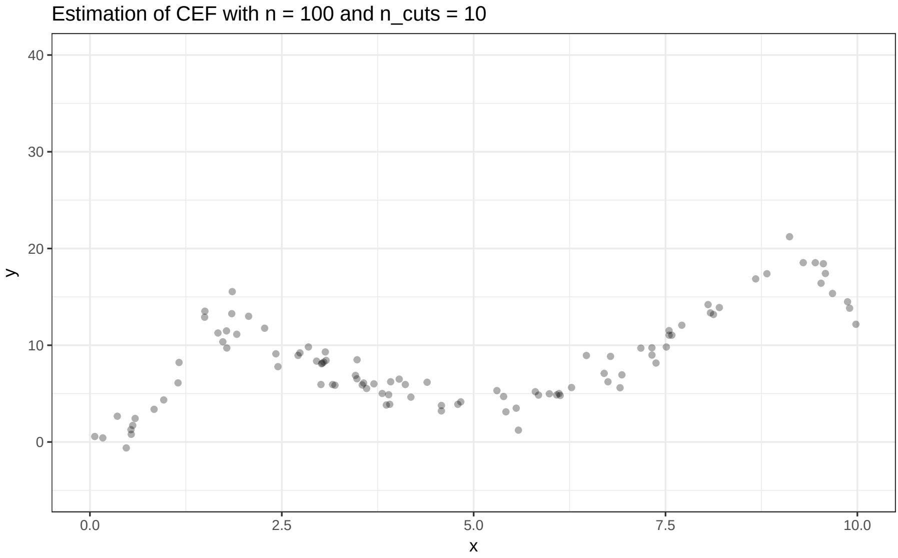
- WLLN tells us that the eCDF will be unbiased and consistently estimated. Any given sample will, on average, look representative of the true distribution.

For iid random variables $X_1, X_2, \ldots, X_n$ with common CDF F, the plug-in estimator of $\theta = T(F)$ is:

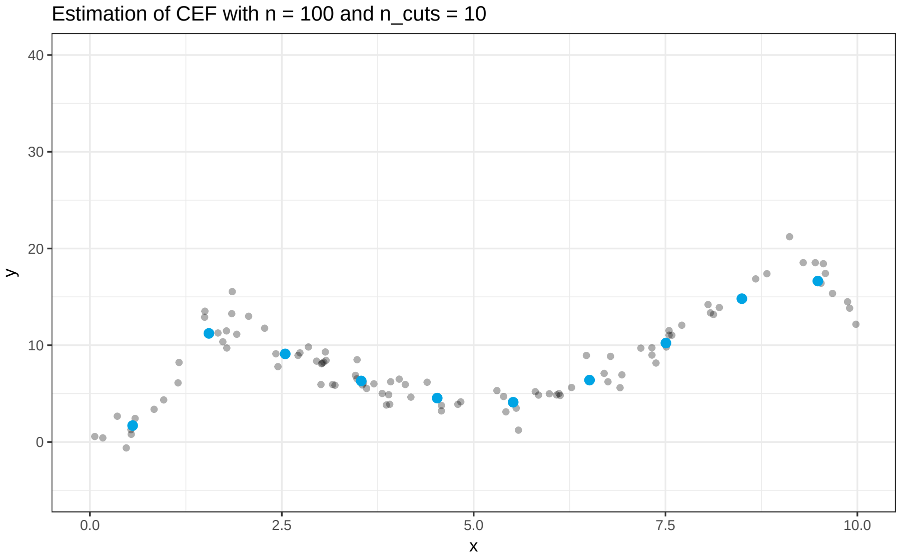
$$\hat{\theta} = T(\hat{F})$$

if T is well-behaved, then $\hat{\theta}$ is also asymptotically normal.

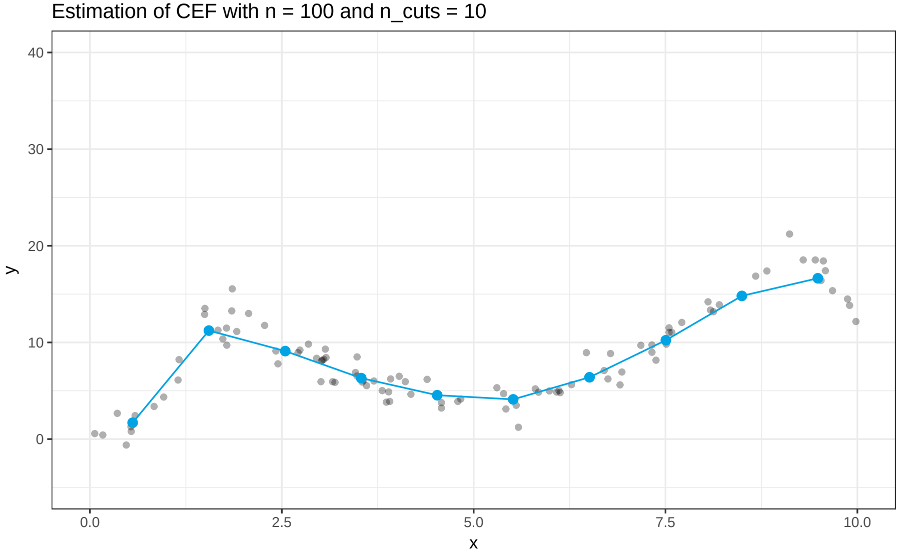
A Binning Approach to CEF for continuous random variables



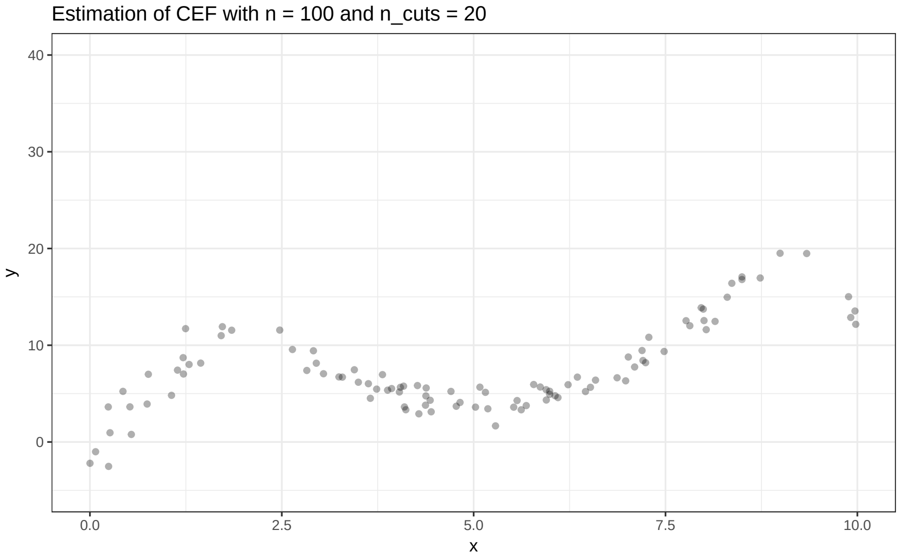
A Binning Approach to CEF for continuous random variables



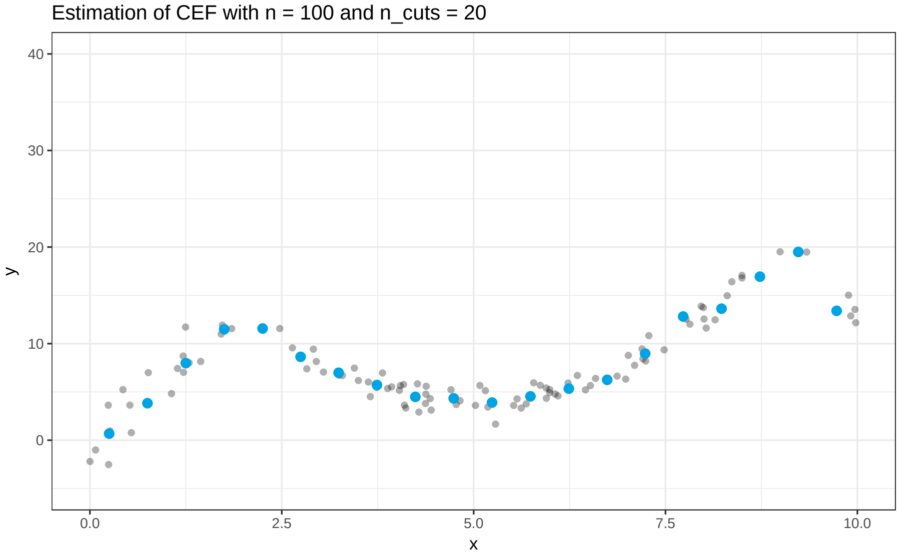
A Binning Approach to CEF for continuous random variables



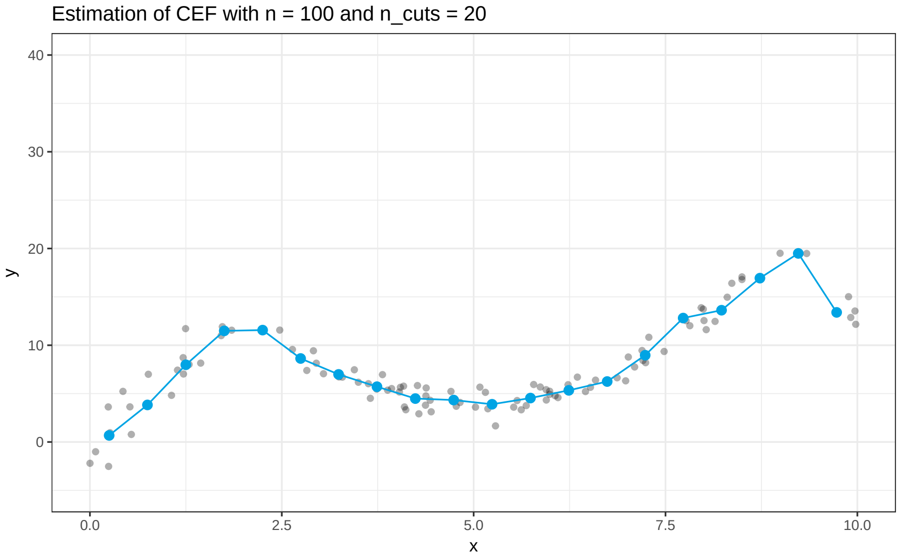
A Binning Approach to CEF for continuous random variables



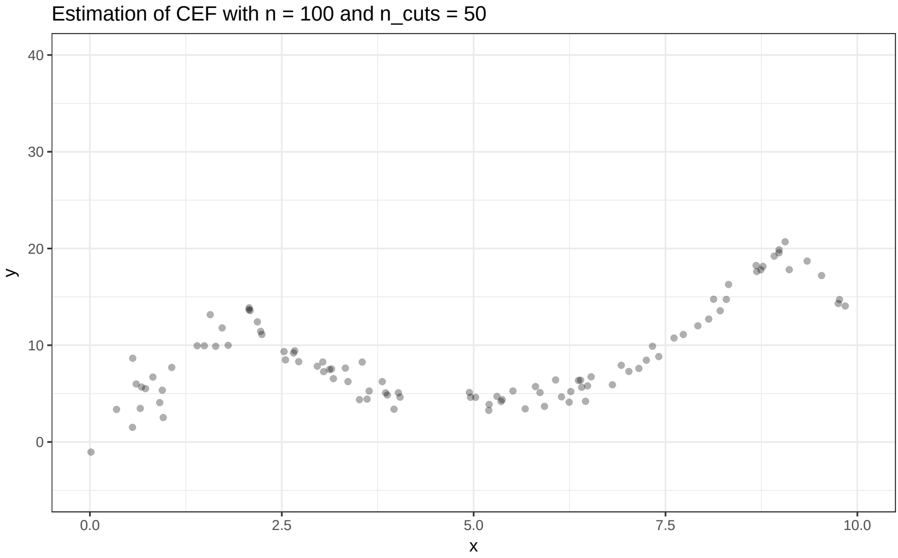
A Binning Approach to CEF for continuous random variables



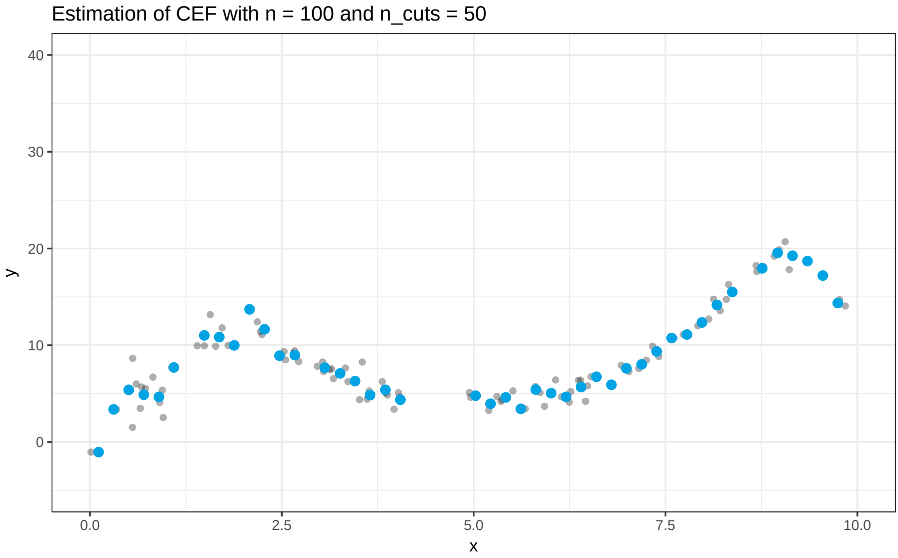
A Binning Approach to CEF for continuous random variables



A Binning Approach to CEF for continuous random variables

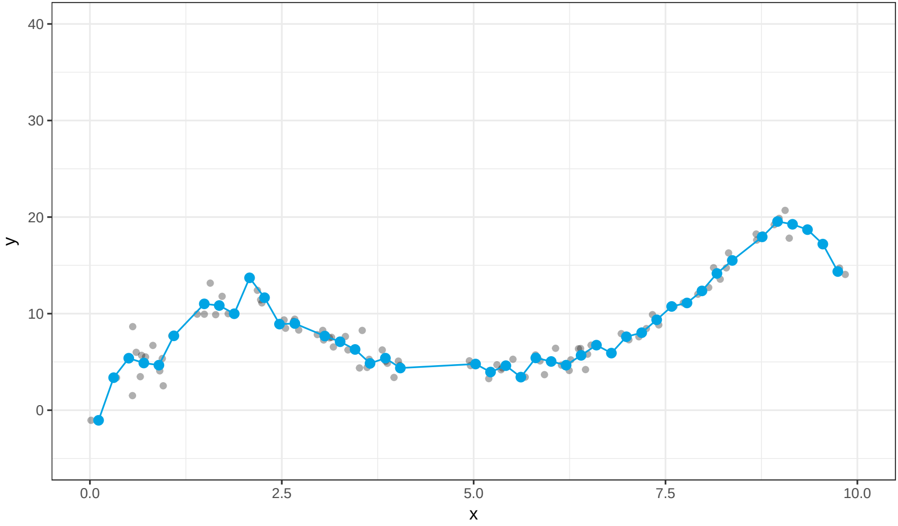


A Binning Approach to CEF for continuous random variables

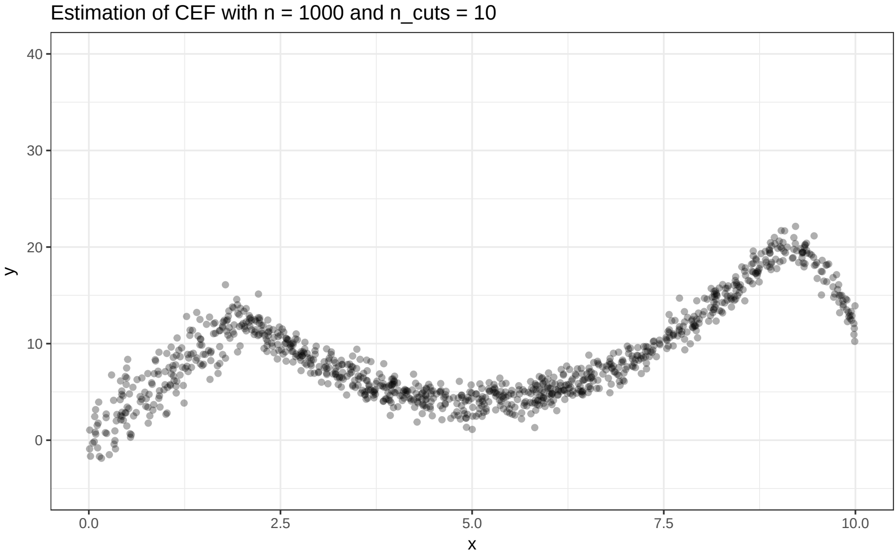


A Binning Approach to CEF for continuous random variables

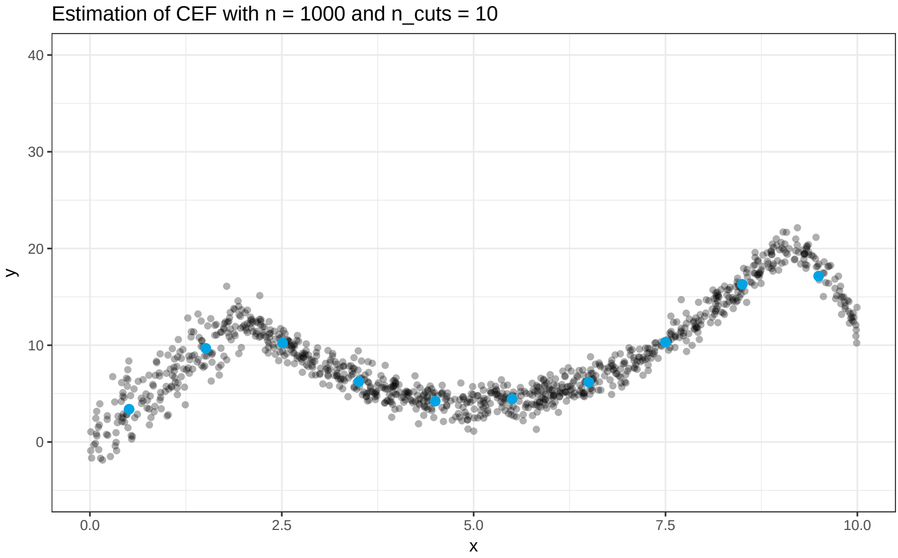
Estimation of CEF with n = 100 and n_cuts = 50



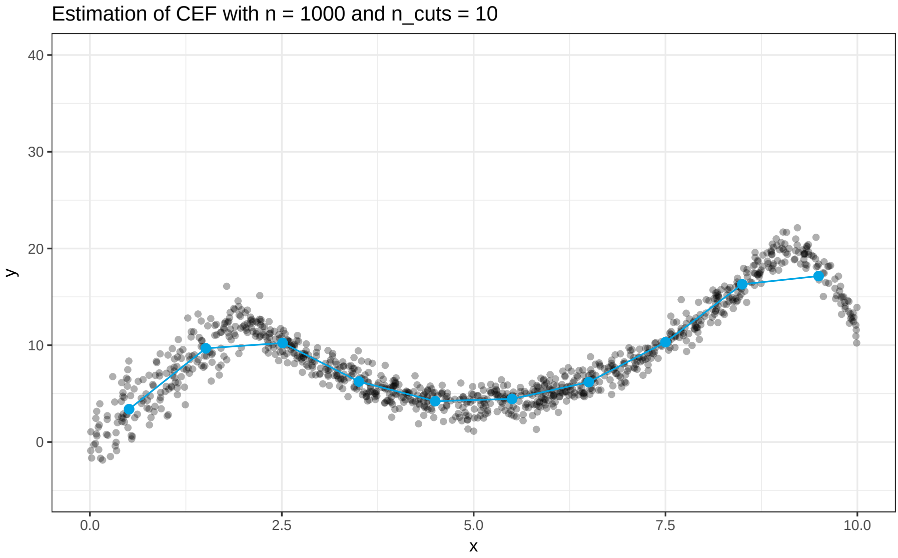
A Binning Approach to CEF for continuous random variables



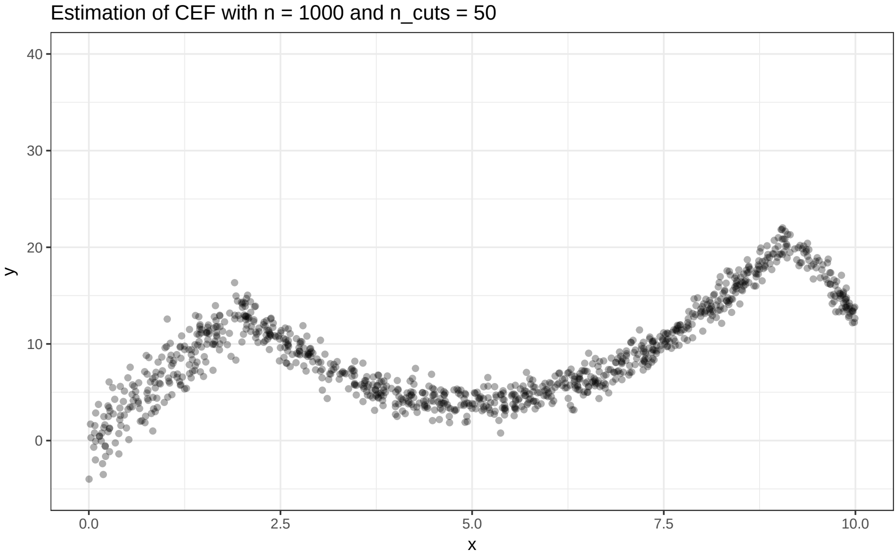
A Binning Approach to CEF for continuous random variables



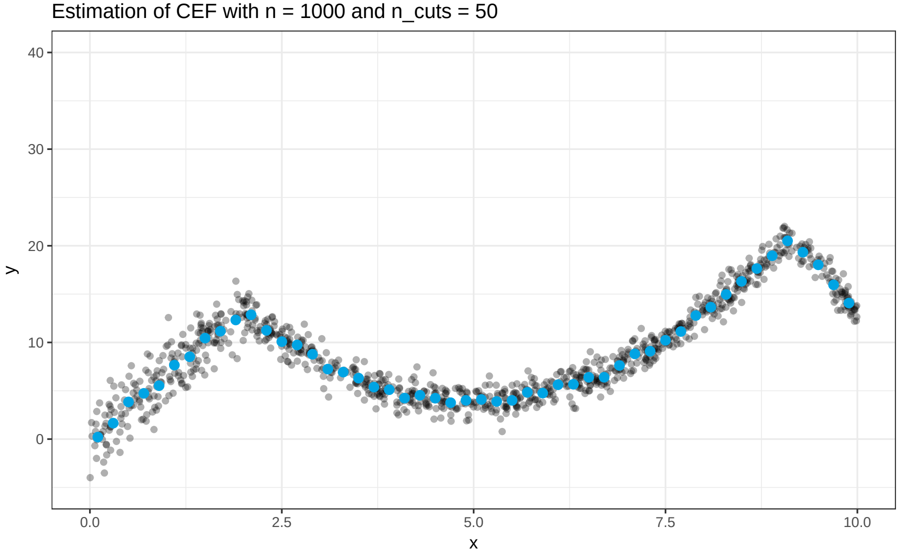
A Binning Approach to CEF for continuous random variables



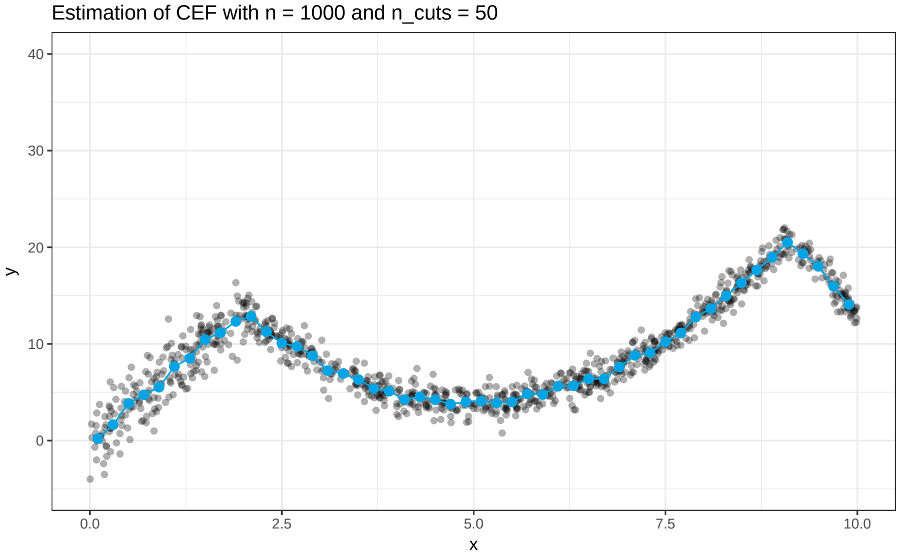
A Binning Approach to CEF for continuous random variables



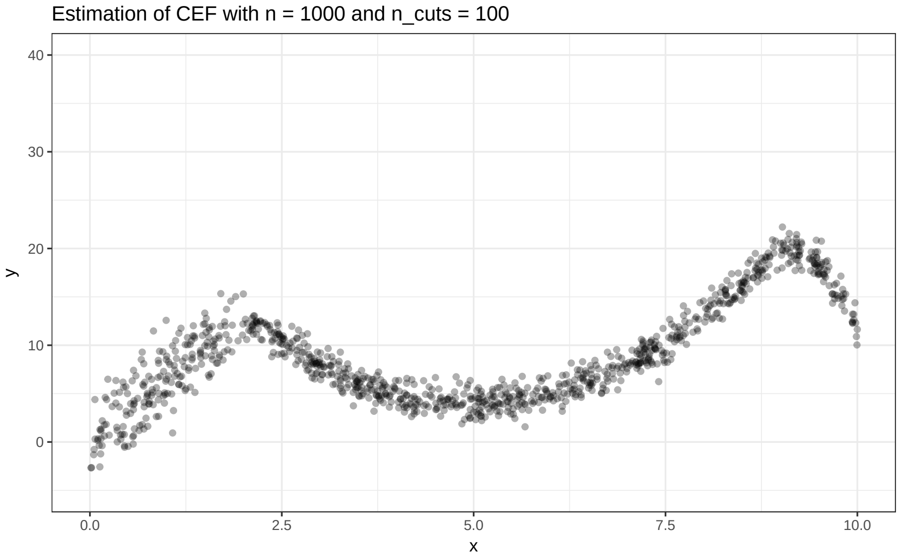
A Binning Approach to CEF for continuous random variables



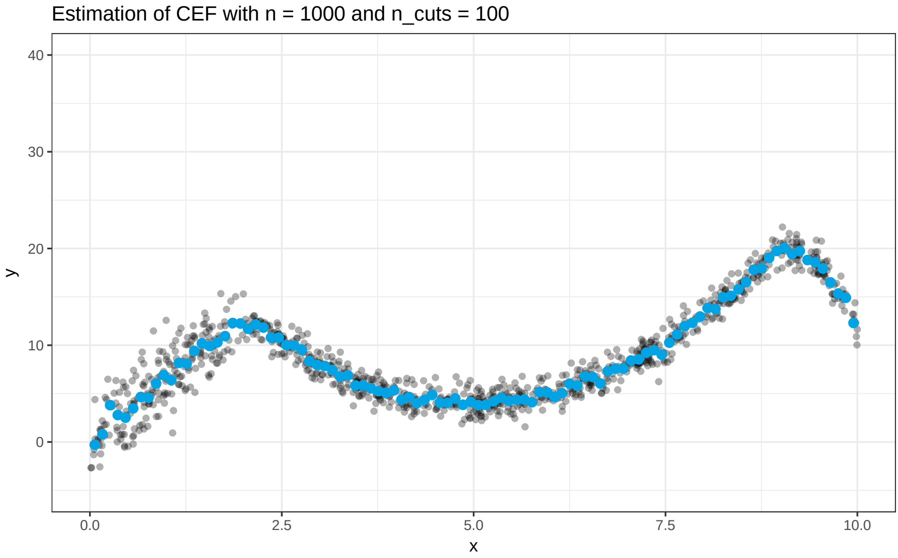
A Binning Approach to CEF for continuous random variables



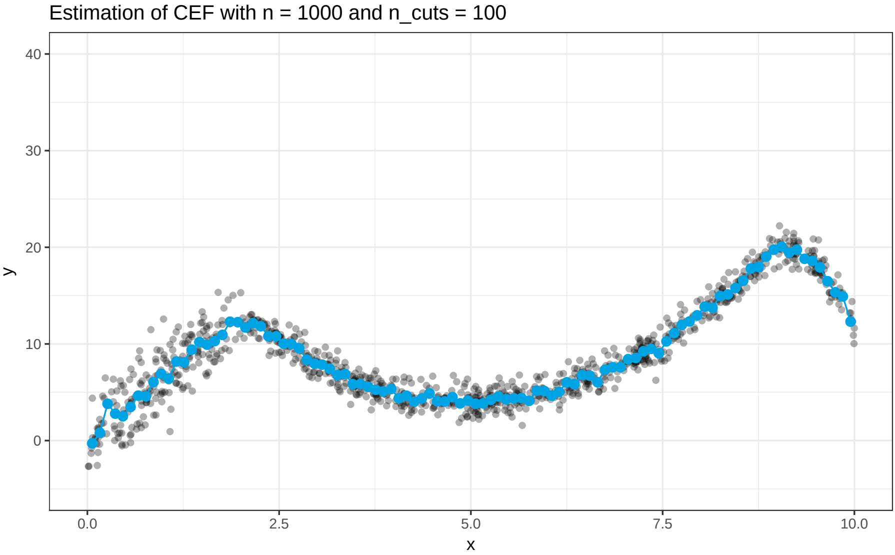
A Binning Approach to CEF for continuous random variables



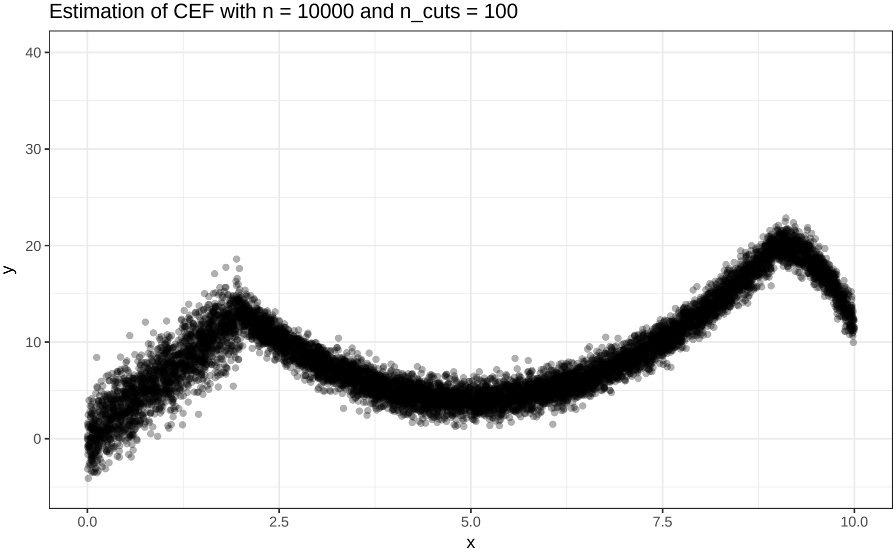
A Binning Approach to CEF for continuous random variables



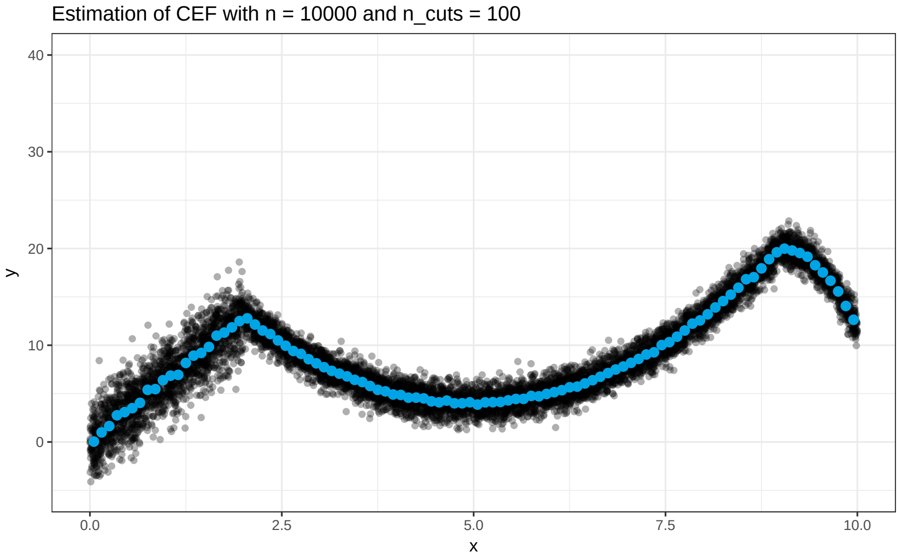
A Binning Approach to CEF for continuous random variables



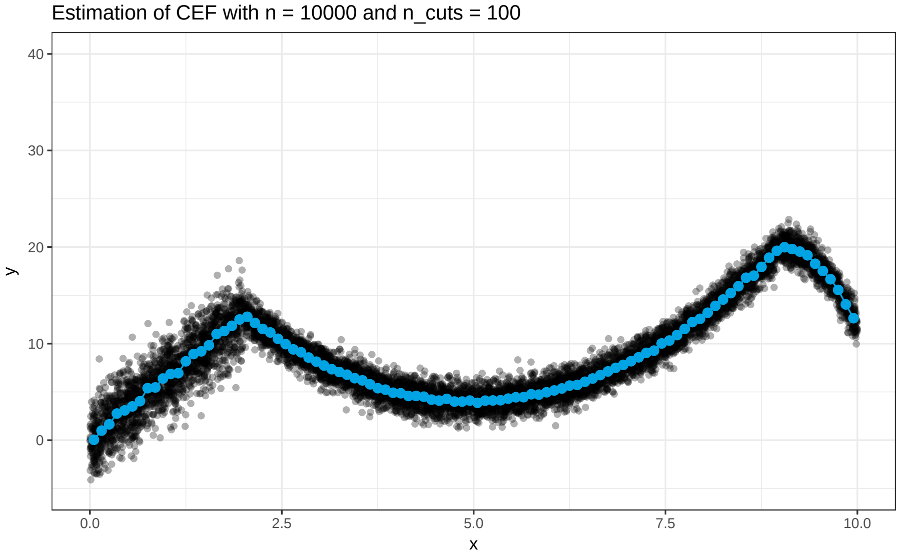
A Binning Approach to CEF for continuous random variables



A Binning Approach to CEF for continuous random variables



A Binning Approach to CEF for continuous random variables



Uniform Kernel Regression: Simple Local Averages

Uniform Kernel Regression: Simple Local Averages

- Dividing into discrete bins can get pretty noisy.

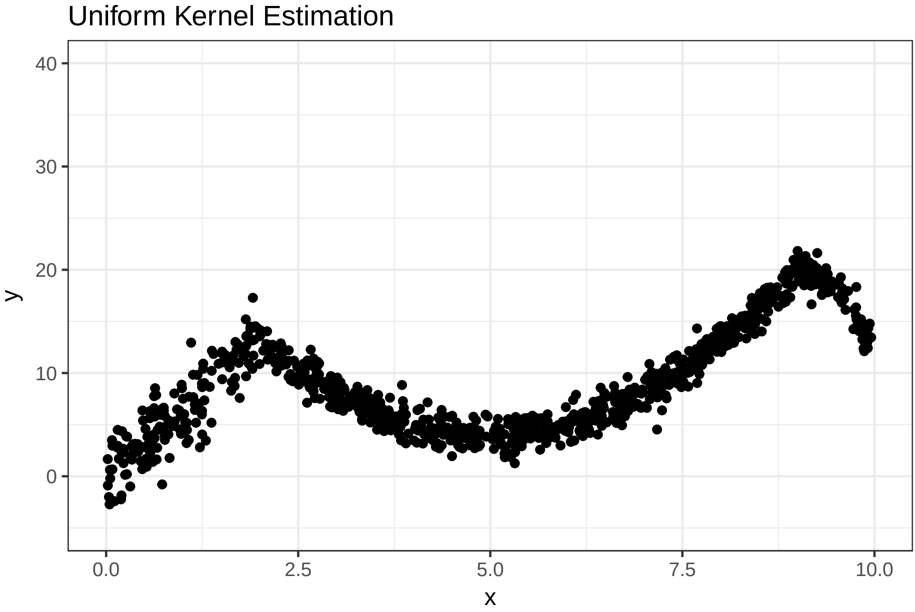
Uniform Kernel Regression: Simple Local Averages

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

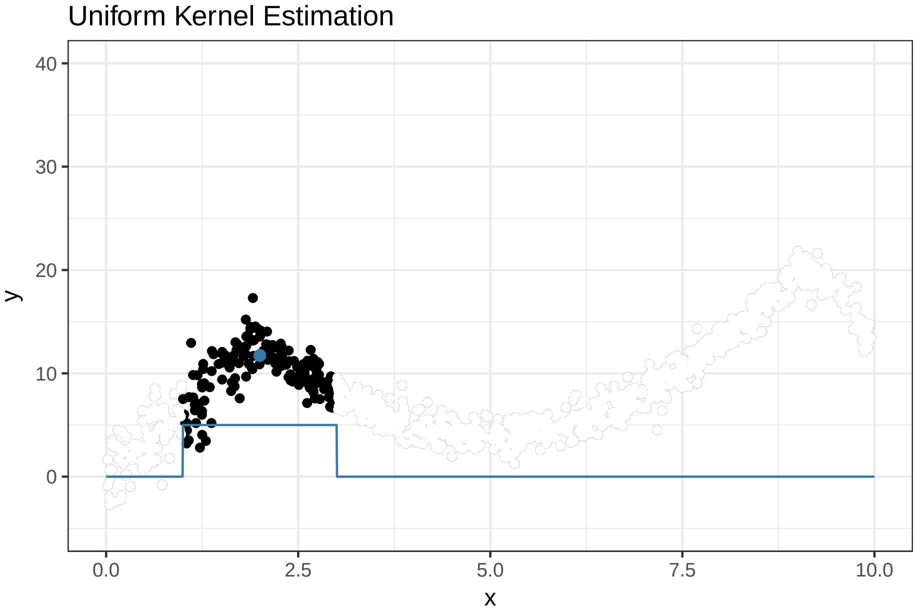
Uniform Kernel Regression: Simple Local Averages

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

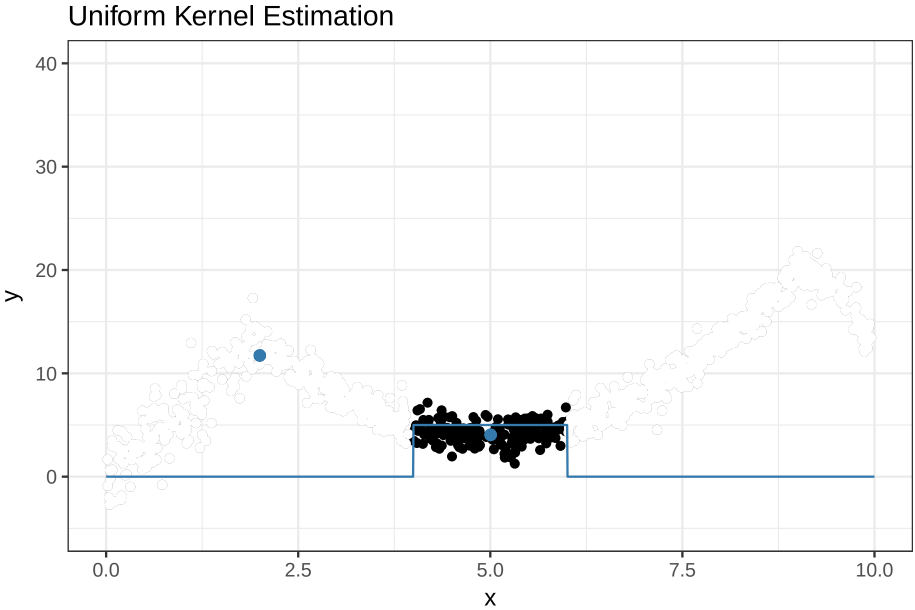
Example of Kernel CEF Estimation



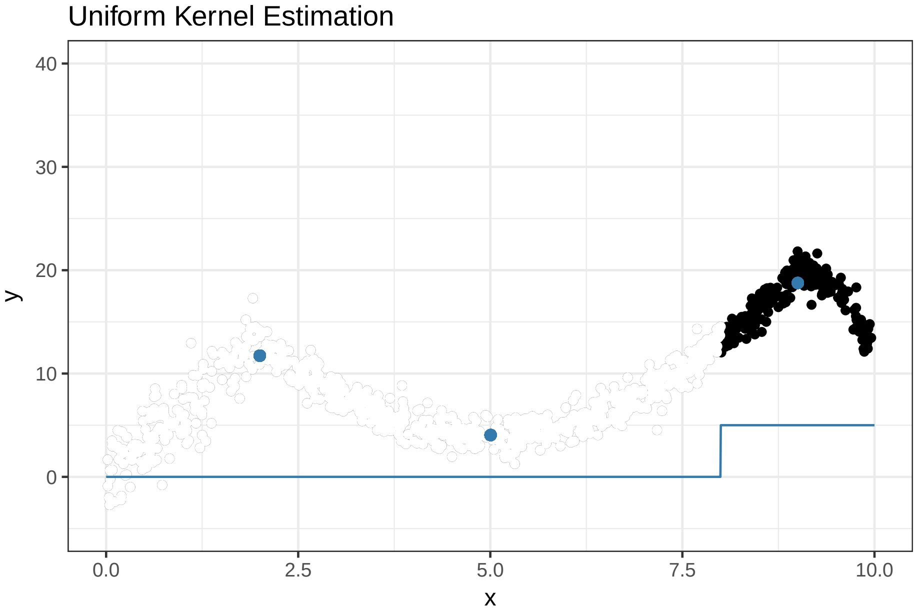
Example of Kernel CEF Estimation



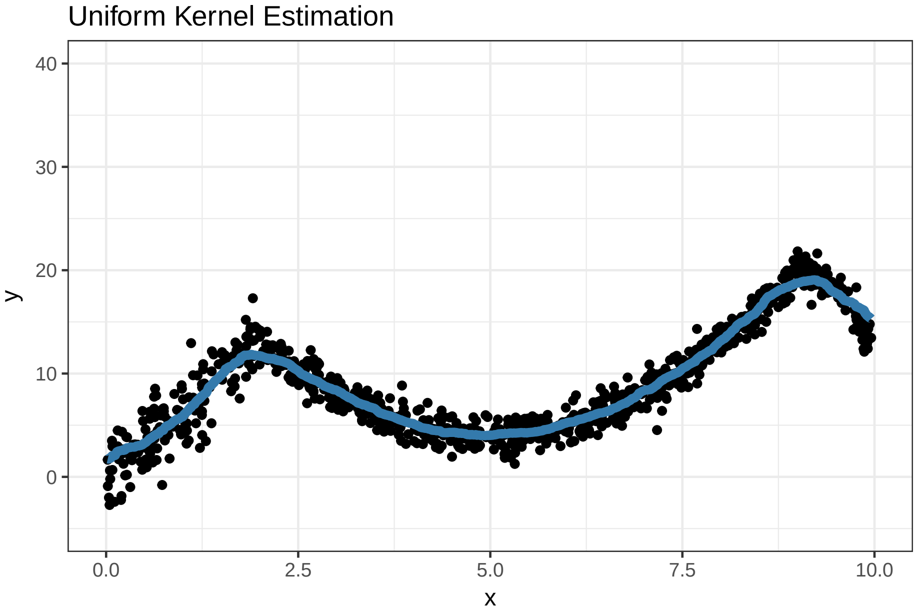
Example of Kernel CEF Estimation



Example of Kernel CEF Estimation



Example of Kernel CEF Estimation



Uniform Kernel Regression: Simple Local Averages

Uniform Kernel Regression: Simple Local Averages

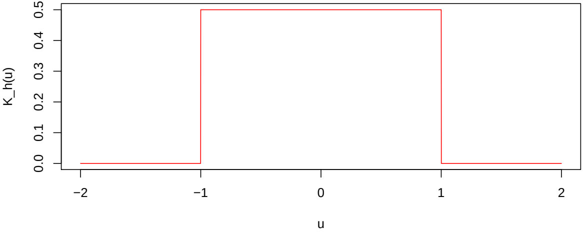
- Calculate the average of the observed y points that have x values in the interval $[x_0 - h\,,\, x_0 + h]$

Uniform Kernel Regression: Simple Local Averages

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

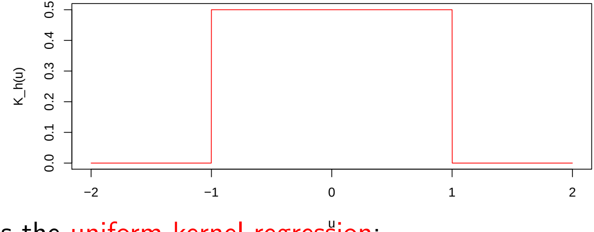
Uniform Kernel Regression: Simple Local Averages

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



Uniform Kernel Regression: Simple Local Averages

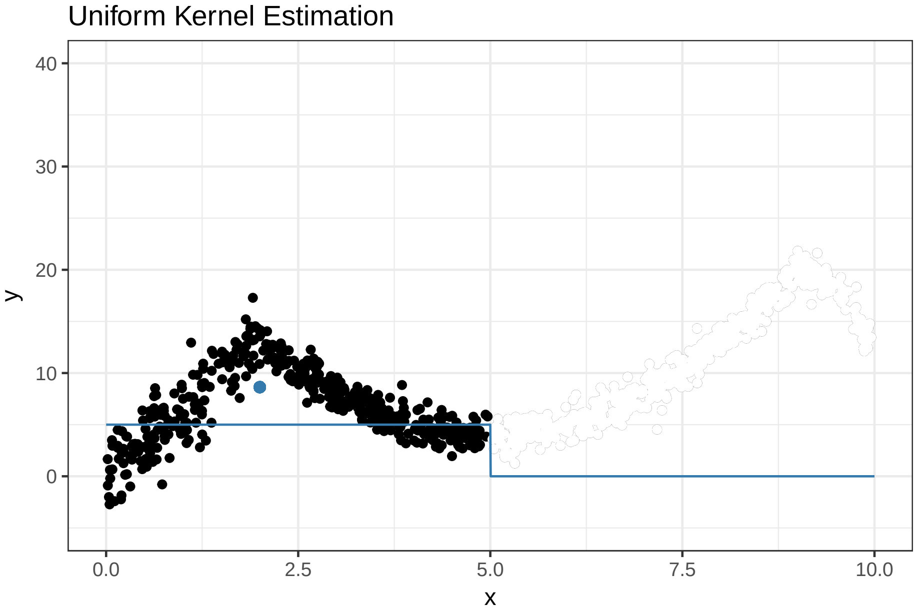
- Calculate the average of the observed y points that have x values in the interval $[x_0 - h\,,\, x_0 + h]$
- $h =$ some positive number (called the bandwidth)
- Uniform kernel: every observation in the interval is equally weighted



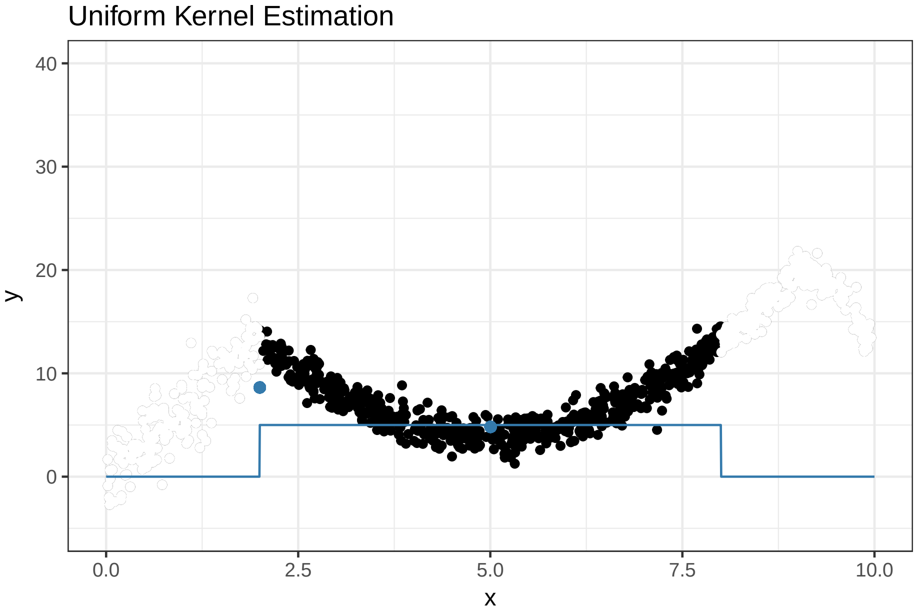
- This gives the uniform kernel regression:

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

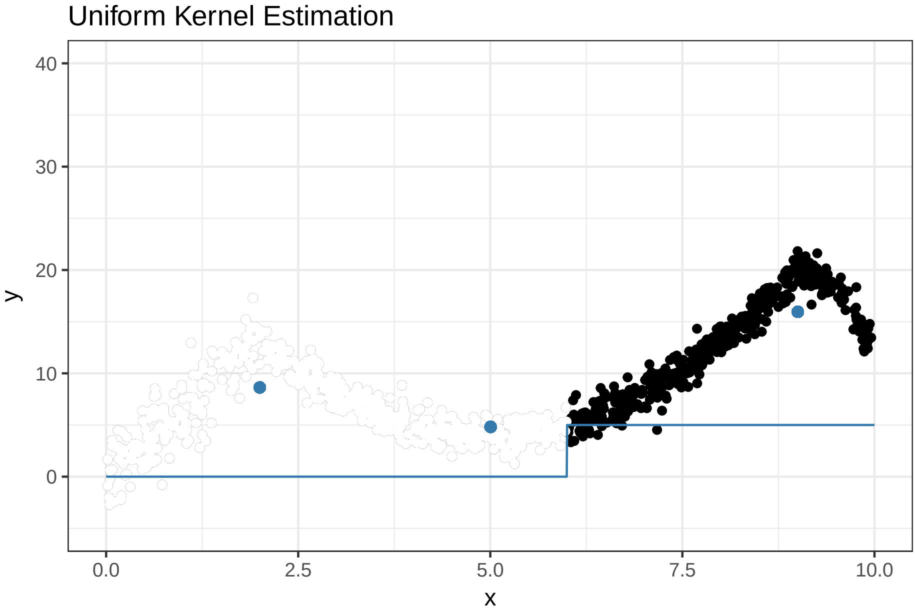
Changing the Bandwidth



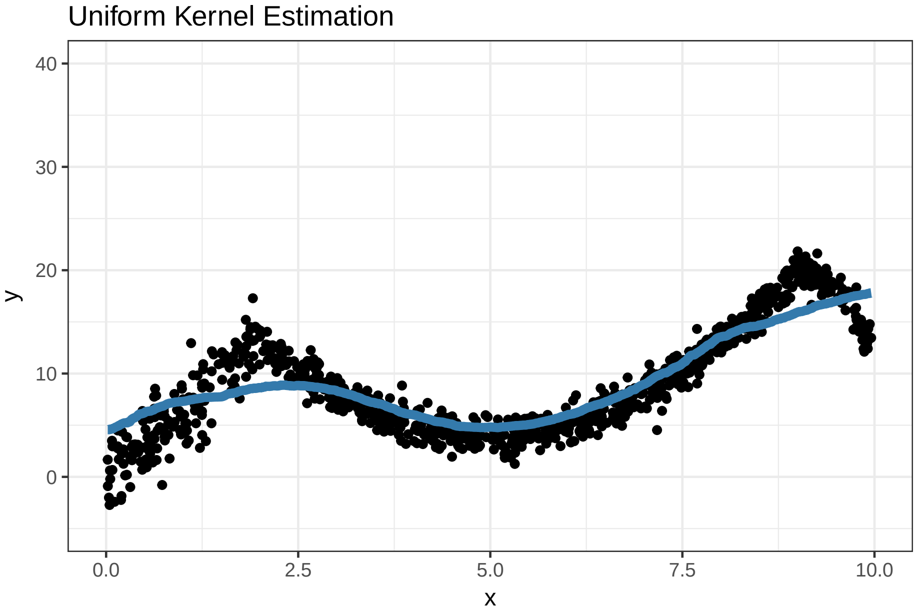
Changing the Bandwidth



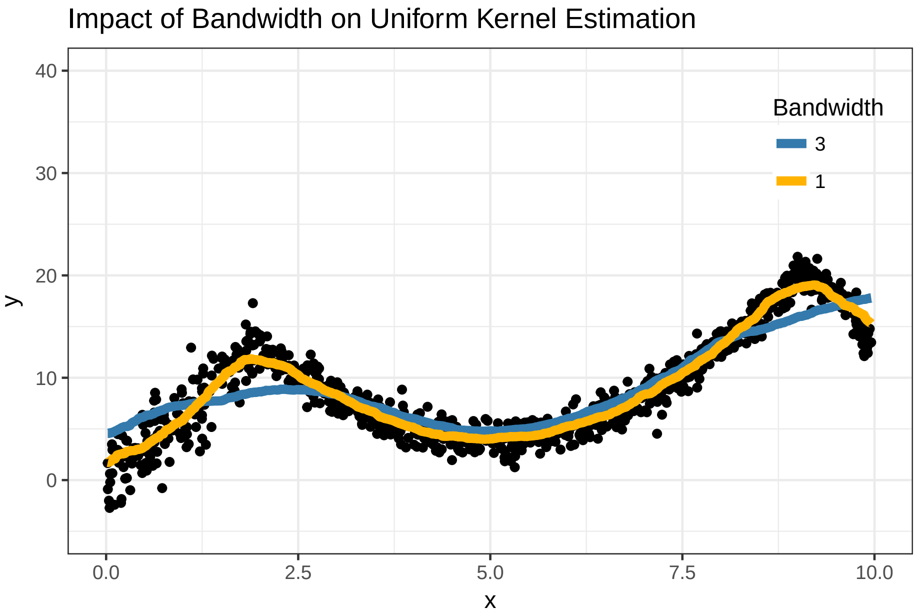
Changing the Bandwidth



Changing the Bandwidth



Changing the Bandwidth



Uniform Kernel Regression: Properties

Theorem (Consistency of the Uniform Kernel Density Estimator)

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

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

$$\hat{f}_K(x) \operatorname{plim} f(x).$$

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

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

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

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

Kernel Estimation

Kernel Estimation

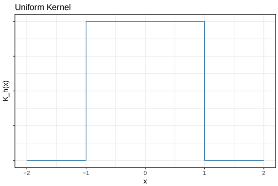
- We can create other estimators by filling in other kernels.

Kernel Estimation

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

Kernel Estimation

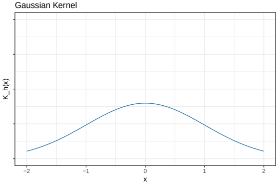
- We can create other estimators by filling in other kernels.
- Generally the idea is to weight things further from the focal point by less than those closer to the focal point.
- Uniform Kernel
- Calculate the average of the observed y points that have x values in the interval $[x_0 - h, x_0 + h]$
-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



Kernel Estimation

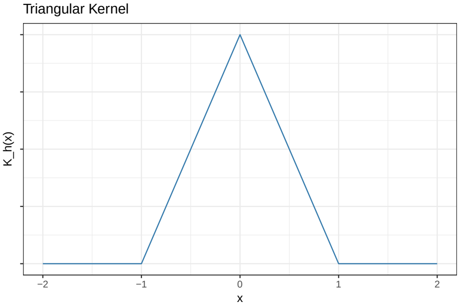
- We can create other estimators by filling in other kernels.
- Generally the idea is to weight things further from the focal point by less than those closer to the focal point.
- Gaussian Kernel
- Distance weighted by how far from x_0 following the normal density

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



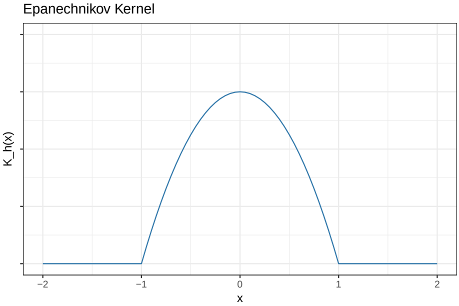
Kernel Estimation

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

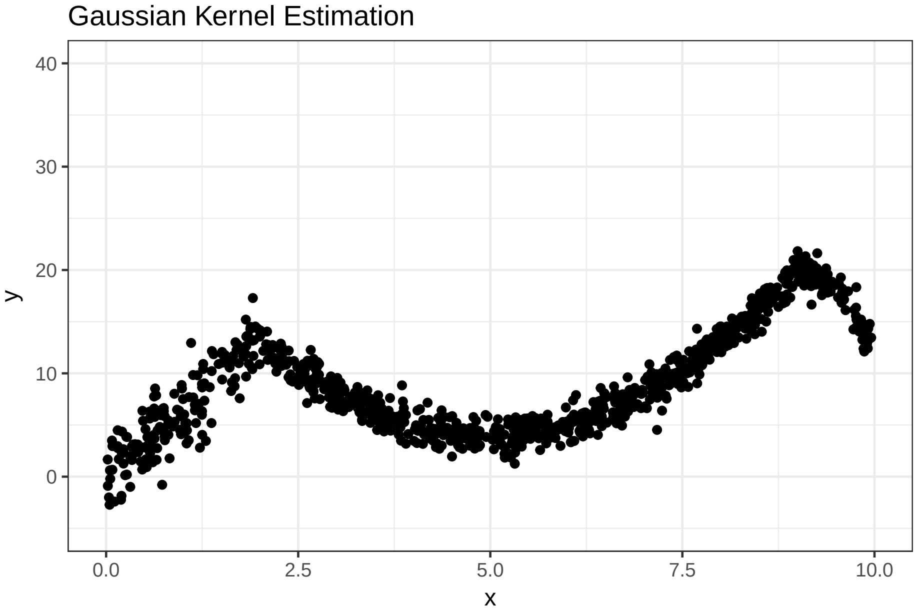


Kernel Estimation

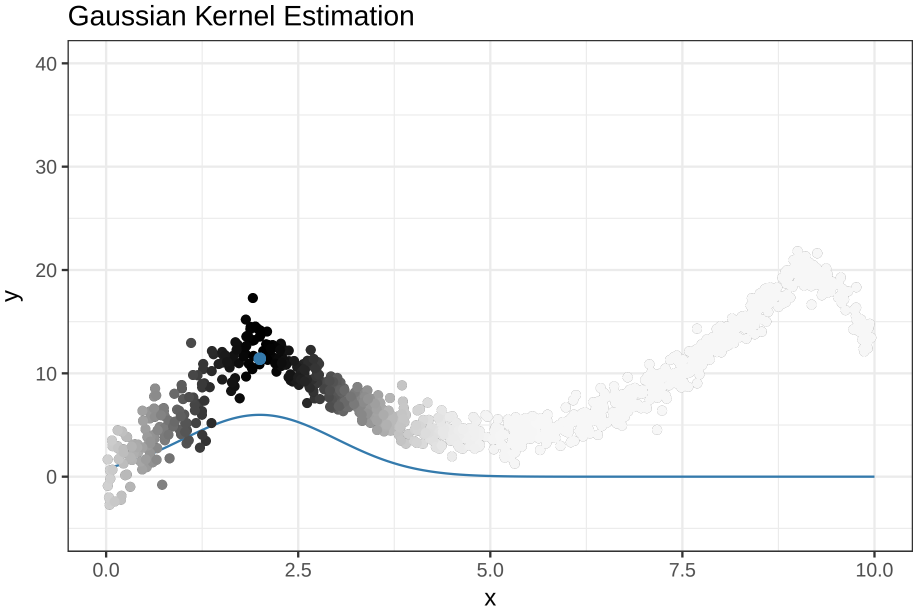
- We can create other estimators by filling in other kernels.
- Generally the idea is to weight things further from the focal point by less than those closer to the focal point.
- Epanechnikov
- Distance weighted by how far from x_0 using a parabolic function



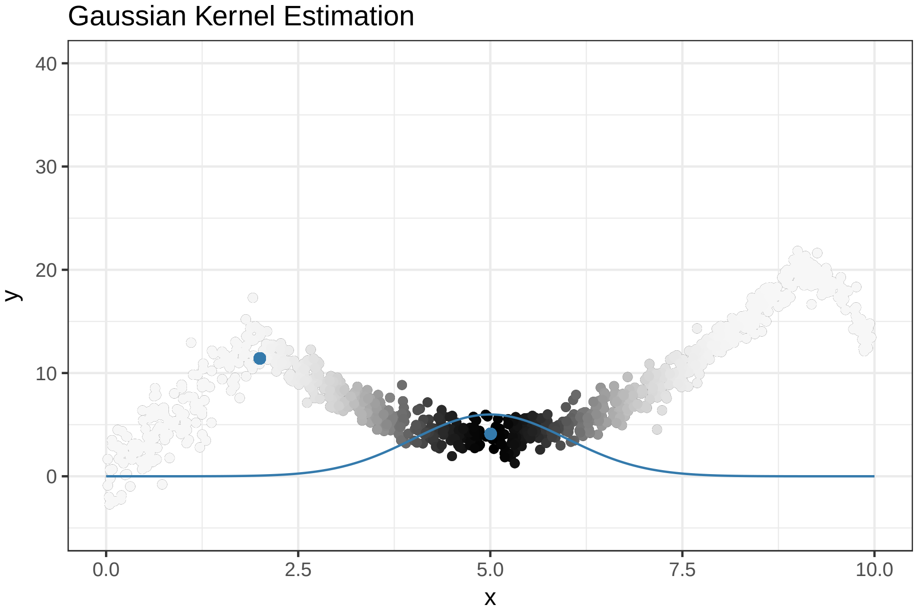
Example of Gaussian Kernel



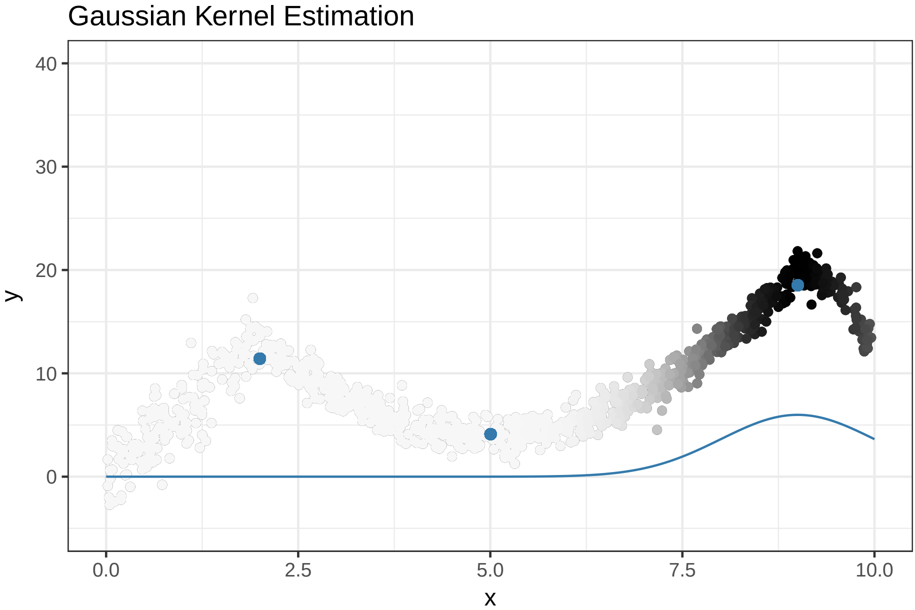
Example of Gaussian Kernel



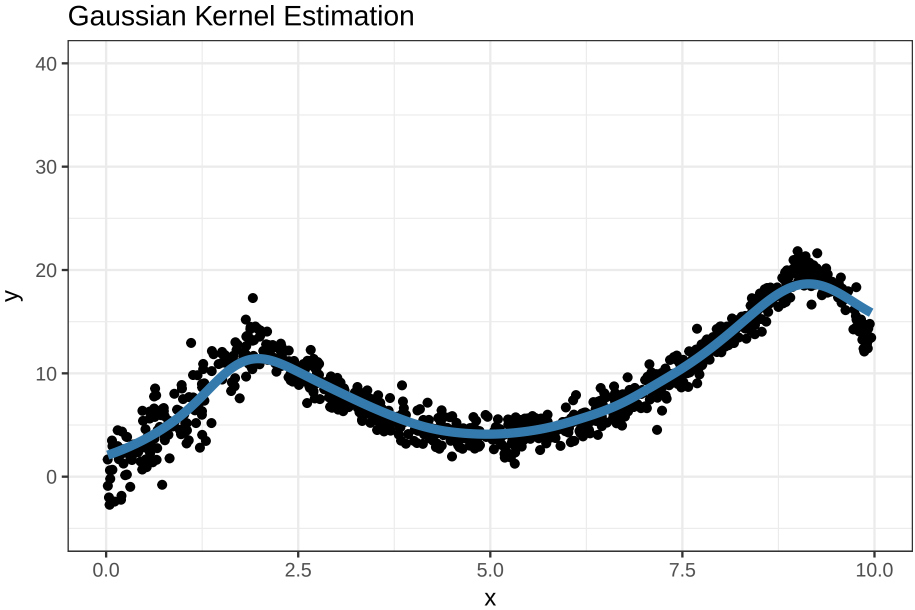
Example of Gaussian Kernel



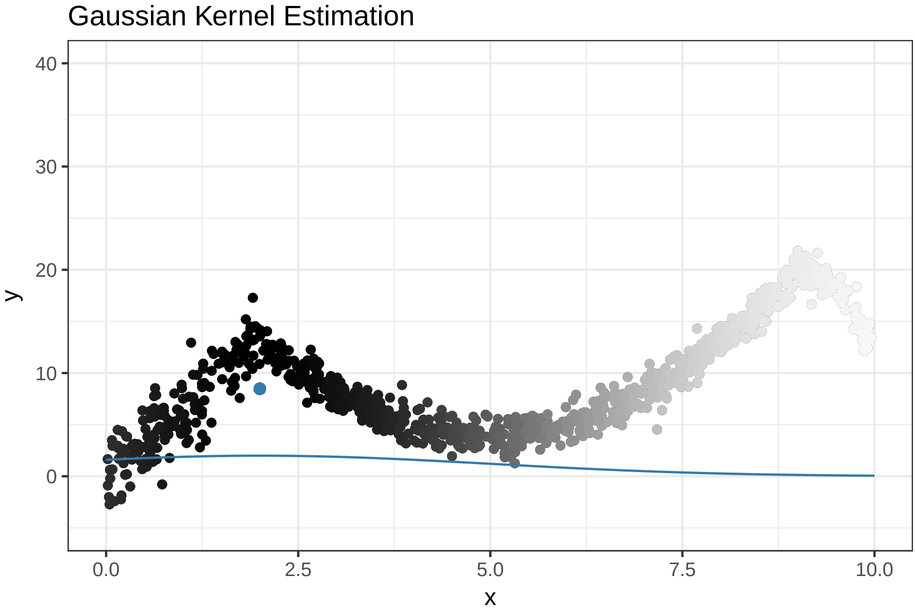
Example of Gaussian Kernel



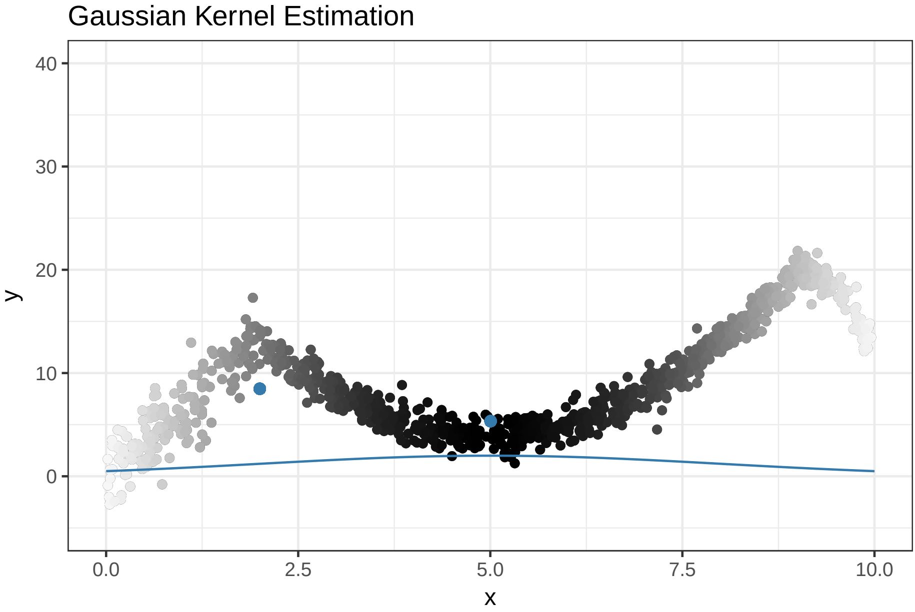
Example of Gaussian Kernel



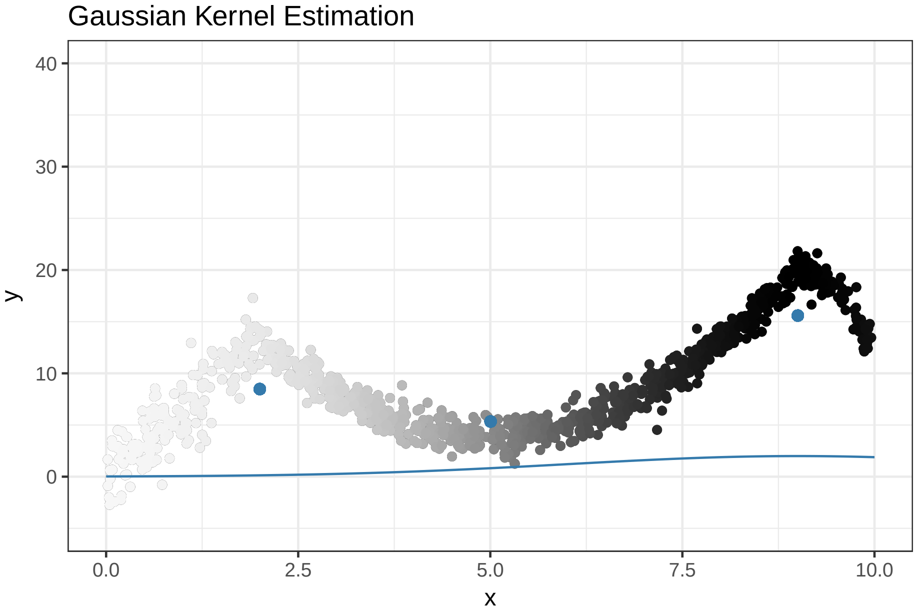
Example of Gaussian Kernel



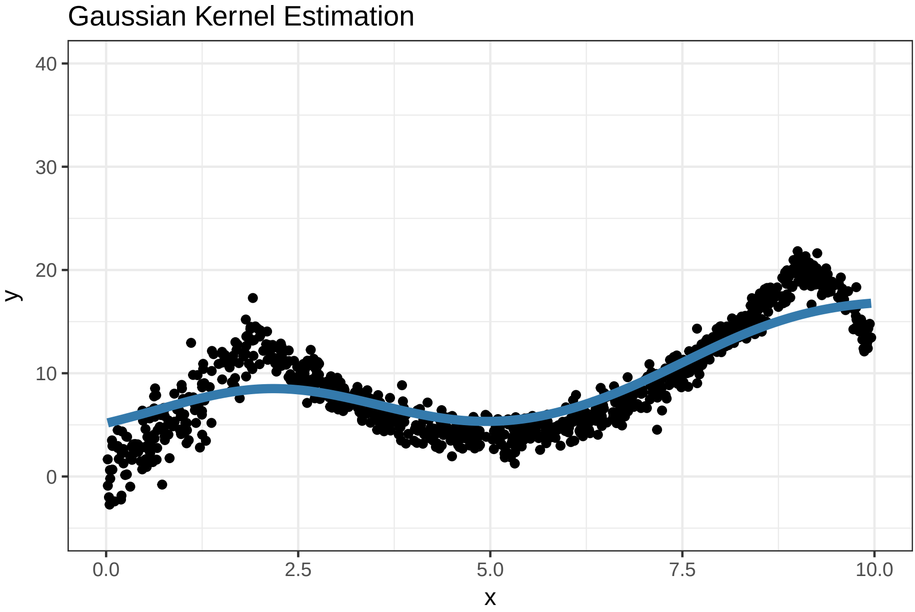
Example of Gaussian Kernel



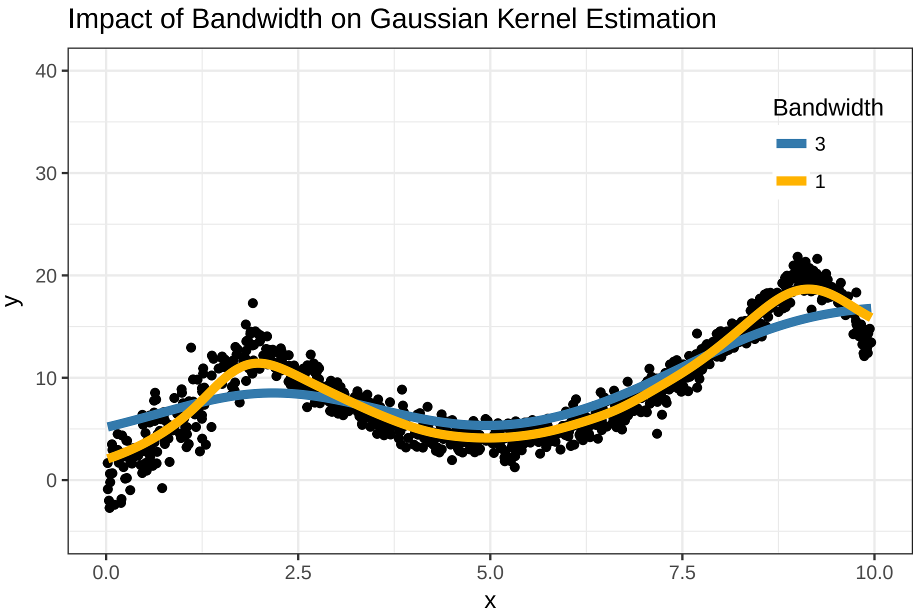
Example of Gaussian Kernel



Example of Gaussian Kernel



Example of Gaussian Kernel



Bias-Variance Tradeoff

Bias-Variance Tradeoff

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

Bias-Variance Tradeoff

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

Bias-Variance Tradeoff

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

Bias-Variance Tradeoff

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

Best Linear Predictor

Best Linear Predictor

- The CEF can be infinitely complex.

Best Linear Predictor

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

Best Linear Predictor

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

Best Linear Predictor

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

Best Linear Predictor

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

$$g(X) = b_0 + b_1 X$$

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:

$$g(X) = b_0 + b_1 X$$

- We call this the linear regression line (bonus: it extends well to a vector of X),

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:

$$g(X) = b_0 + b_1 X$$

- We call this the linear regression line (bonus: it extends well to a vector of X),
- What function minimizes MSE (in that sense is 'best'), among this class of functional forms?

Best Linear Predictor

Best Linear Predictor

Theorem (Aronow and Miller Thm 2.2.21)

For a random variable X and Y, if $V[X] > 0$, then the best (min MSE) linear predictor (BLP) of Y given X is $g(X) = \beta_0 + \beta_1 X$ where,

$$\beta_0 = E[Y] - \frac{Cov[X, Y]}{V[X]} E[X] \qquad\qquad \beta_1 = \frac{Cov(X, Y)}{V(X)}$$

Best Linear Predictor

Theorem (Aronow and Miller Thm 2.2.21)

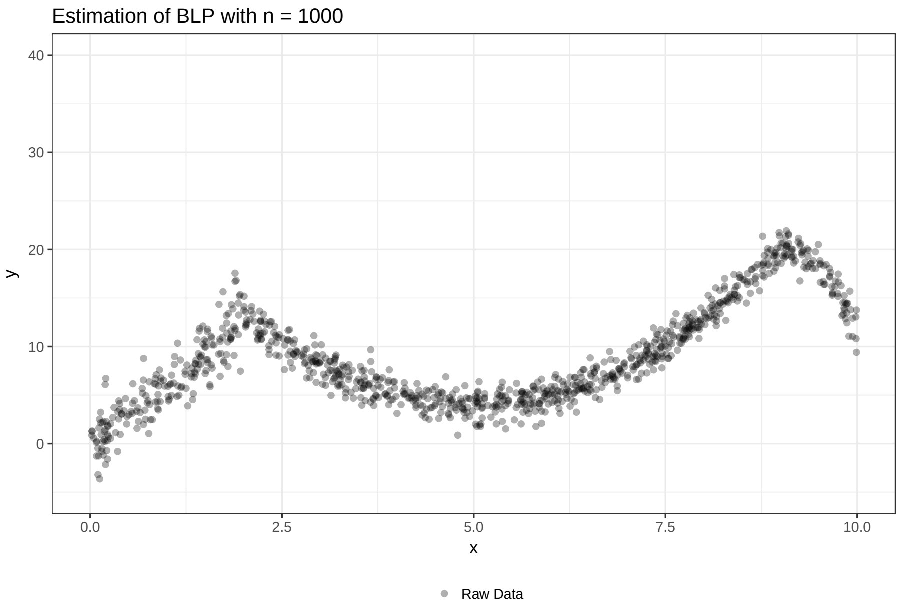
For a random variable X and Y, if $V[X] > 0$, then the best (min MSE) linear predictor (BLP) of Y given X is $g(X) = \beta_0 + \beta_1 X$ where,

$$\beta_0 = E[Y] - \frac{Cov[X, Y]}{V[X]} E[X] \qquad \beta_1 = \frac{Cov(X, Y)}{V(X)}$$

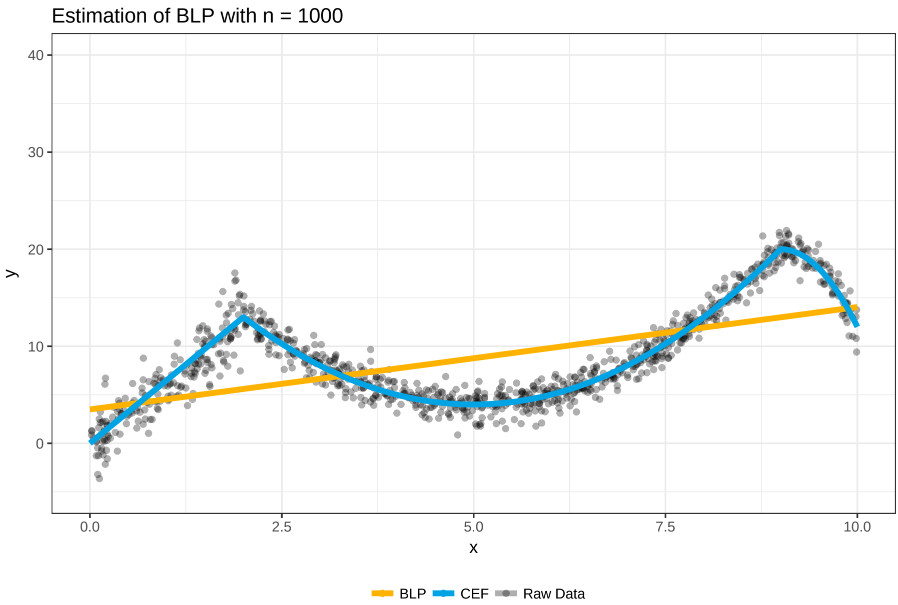
Corollary (Aronow and Miller Thm 2.2.22)

- *The BLP is the best linear predictor of the CEF. I.e. setting $b_0 = \beta_0$ and $b_1 = \beta_1$ minimizes $E\left[(E[Y \mid X] - (b_0 + b_1 X))^2\right]$*
- *If the CEF is linear, the CEF is the BLP*
- *Defining $\epsilon = Y - (\beta_0 + \beta_1 X)$,*
  - $E[\epsilon] = 0$
  - $E[X\epsilon] = 0$
  - $Cov[X, \epsilon] = 0$

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 onto X.

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 onto X.
- 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 onto X.
- In general, this is distinct from the CEF:
  - CEF is the best predictor of Y 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 onto X.
- In general, this is distinct from the CEF:
  - CEF is the best predictor of Y 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 onto X.
- In general, this is distinct from the CEF:
  - CEF is the best predictor of Y 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 $p(X)$

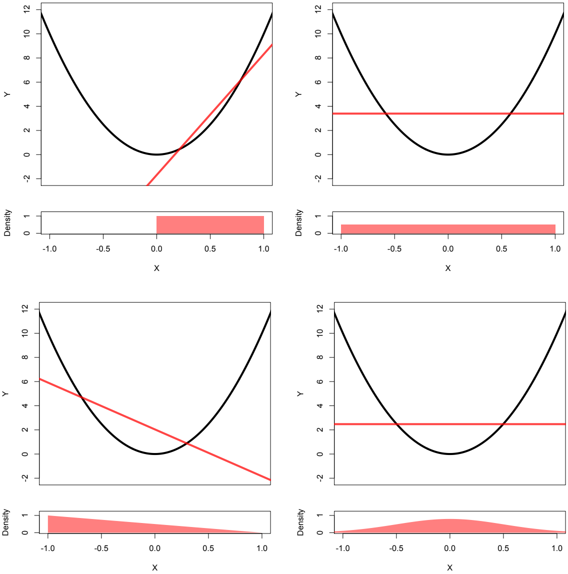


Figure 2.2.2: Plotting the CEF and the BLP over Different Distributions of X

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)

'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)

'If I fits, I sits'

The BLP is always a line regardless of the data.

BLP: Justification

BLP: Justification

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

BLP: Justification

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

BLP: 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

BLP: 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

BLP: 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

BLP: 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)

BLP: 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.)

BLP: 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 BLP slope

Interpretation of the BLP slope

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

Interpretation of the BLP slope

- When we model the CEF 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 BLP slope

- When we model the CEF 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 BLP slope

- When we model the CEF 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 BLP slope

- When we model the CEF 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**: for units one higher in X we predict Y to be higher 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 BLP slope

- When we model the CEF 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**: for units one higher in X we predict Y to be higher by β_1

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

Interpretation of the BLP slope

- When we model the CEF 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**: for units one higher in X we predict Y to be higher 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_1 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_1 unit change in Y.
- This leans suggestively towards a causal interpretation that if we intervened to change X then Y would change in some way. We will talk more about the assumptions we would need to draw this conclusion later in the semester.

Linear Regression as a Descriptive Model

- A lot of social science reports regression as saying that a one unit change in X is associated with a β_1 unit change in Y.
- This leans suggestively towards a causal interpretation that if we intervened to change X then Y would change in some way. We will talk more about the assumptions we would need to draw this conclusion later in the semester.
- For now, I find it helpful to think of regression descriptively as talking about different groups of units. Our guess of the mean for units with $x = 2$ is β_1 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_1 unit change in Y.
- This leans suggestively towards a causal interpretation that if we intervened to change X then Y would change in some way. We will talk more about the assumptions we would need to draw this conclusion later in the semester.
- For now, I find it helpful to think of regression descriptively as talking about different groups of units. Our guess of the mean for units with $x = 2$ is β_1 higher than our guess of the mean for the units with $x = 1$.
- This helps clarify that it is an approximation to the CEF and that the units being described are different.

Linear Regression as a Predictive Model

Linear Regression as a Predictive Model

- Linear regression can also be used to predict new observations

Linear Regression as a Predictive Model

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

Linear Regression as a Predictive Model

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

Linear Regression as a Predictive Model

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

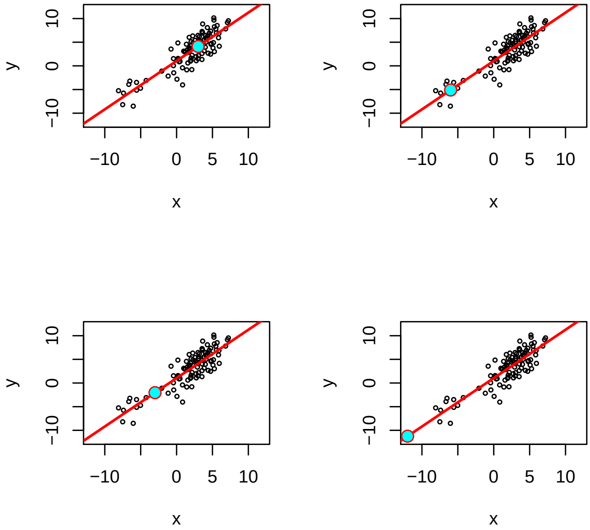
$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_{new}$$

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 a possibly new population we can use information about $X = x_{new}$ to calculate:
$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_{new}$$
- 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
  - changes in the distribution of X

Which Predictions Do You Trust?

Which Predictions Do You Trust?



Example: Tatem, et al. Sprinters Data

Example: Tatem, et al. Sprinters Data

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

Example: Tatem, et al. Sprinters Data

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

How do the authors make this conclusion?

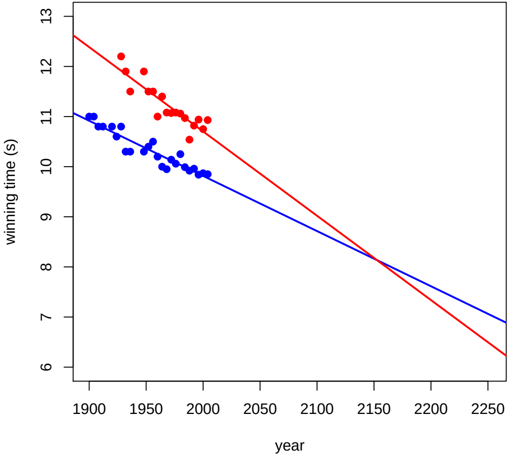
Example: Tatem, et al. Sprinters Data

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

How do the authors make this conclusion?

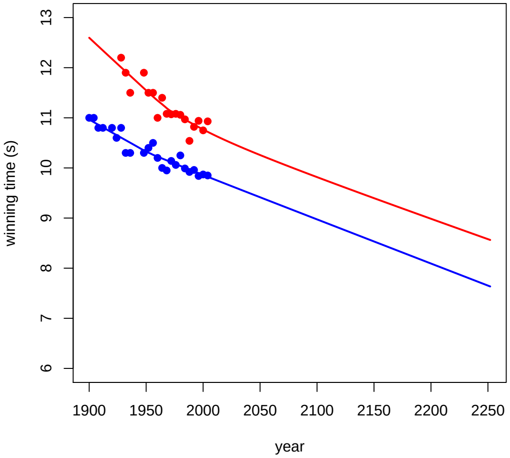
Using data from 1900 to 2004, they fit linear regression models of the winning 100 meter time on year for both men and women. They then use the estimates from these models to extrapolate 152 years into the future.

Tatem et al. Extrapolation

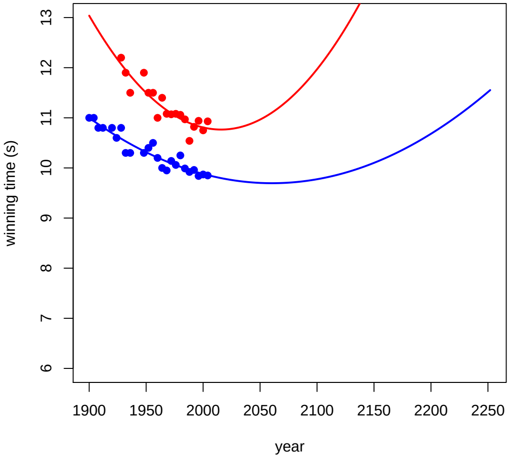


Tatem et al.'s predictions. Men's times are in blue, women's times are in red.

Alternate Models Fit Well, Yield Different Predictions



Alternate Models Fit Well, Yield Different Predictions



The Trouble with Extrapolation

The Trouble with Extrapolation

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

The Trouble with Extrapolation

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

The Trouble with Extrapolation

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

The Trouble with Extrapolation

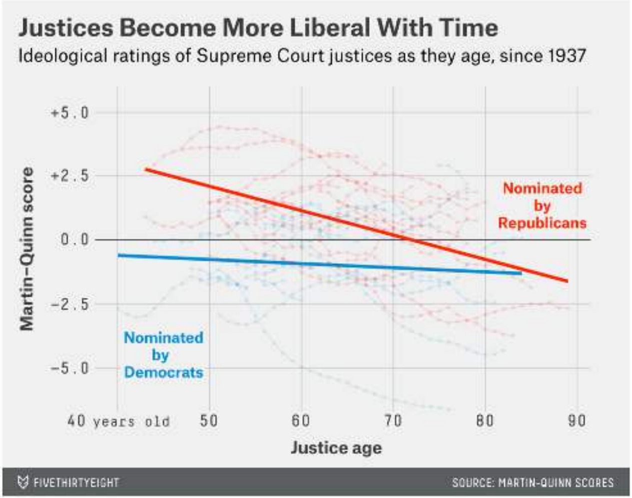
- The model only gives the best fitting line where we have data, it says little about the shape where there isn't any data.
- We can always ask illogical questions and the model gives answers.
  - For example, when will women finish the sprint in negative time?
- Fundamentally we are assuming that data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

The Trouble with Extrapolation

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

A More Subtle Example

A More Subtle Example



A More Subtle Example

Regression as a Causal Model (A Preview)

- Can regression be also used for causal inference?

Regression as a Causal Model (A Preview)

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

Regression as a Causal Model (A Preview)

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

Regression as a Causal Model (A Preview)

- Can regression be also used for causal inference?
- Answer: A very qualified yes
- For example, can we say that sending that social pressure caused people to vote?
- To interpret β_1 as a causal effect of X on Y, we need very specific and often unrealistic assumptions.

Regression as a Causal Model (A Preview)

- Can regression be also used for causal inference?
- Answer: A very qualified yes
- For example, can we say that sending that social pressure caused people to vote?
- To interpret β_1 as a causal effect of X on Y, we need very specific and often unrealistic assumptions.
- We will return to this later in the course

Regression as a Causal Model (A Preview)

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

- For example, can we say that sending that social pressure caused people to vote?

- To interpret β_1 as a causal effect of X on Y, we need very specific and often unrealistic assumptions.
- We will return to this later in the course
- For now, it is safest to treat β_1 as a purely descriptive/predictive quantity

Fun with Linearity

“The Siren’s Song of Linearity”

Fun with Linearity

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

Iterated learning: Intergenerational knowledge transmission reveals inductive biases

MICHAEL L. KALISH
University of Louisiana, Lafayette, Louisiana

THOMAS L. GRIFFITHS
University of California, Berkeley, California

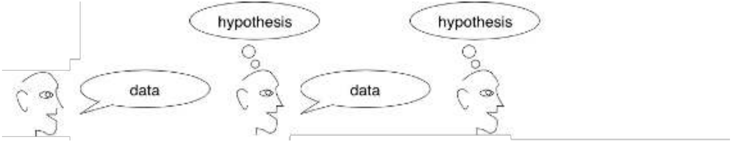
AND

STEPHAN LEWANDOWSKY
University of Western Australia, Perth, Australia

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

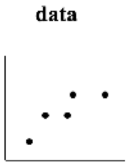
Images on following slides courtesy of Tom Griffiths

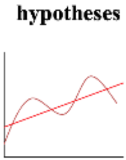
Fun with Linearity



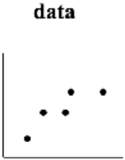
The Design

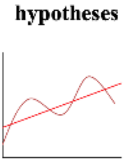
The Design





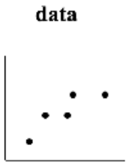
The Design

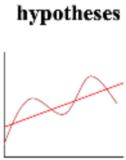




- Each learner sees a set of (x, y) pairs

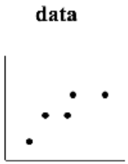
The Design

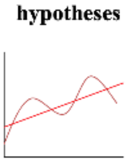




- Each learner sees a set of (x, y) pairs
- Makes predictions of y for new x values

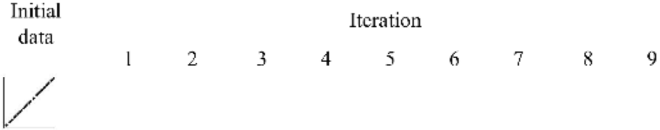
The Design



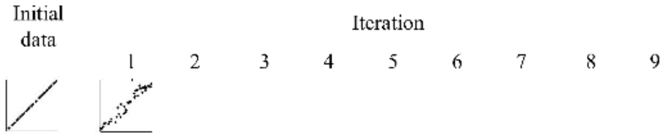


- Each learner sees a set of (x, y) pairs
- Makes predictions of y for new x values
- Predictions are data for the next learner

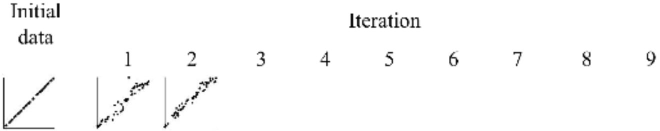
Results



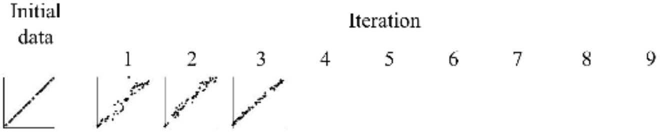
Results



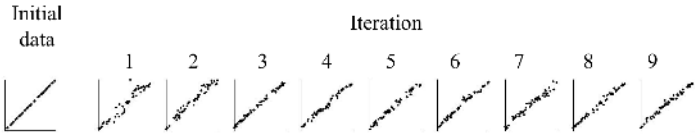
Results



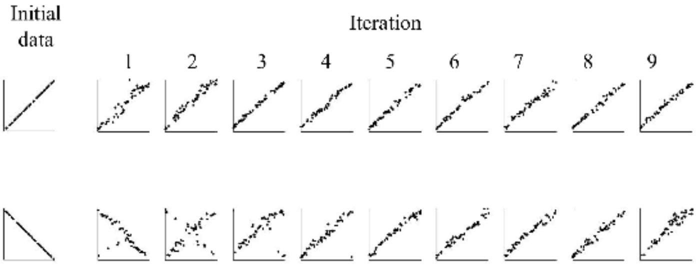
Results



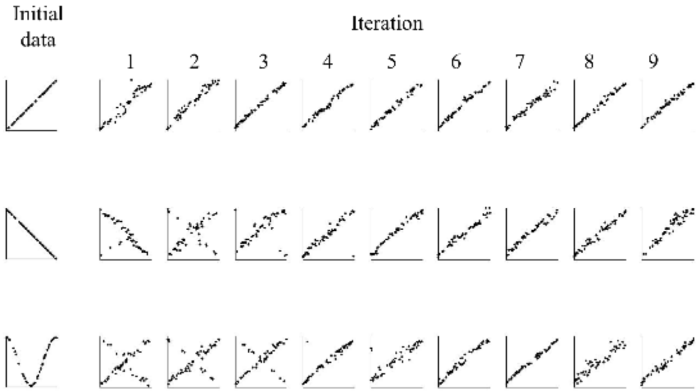
Results



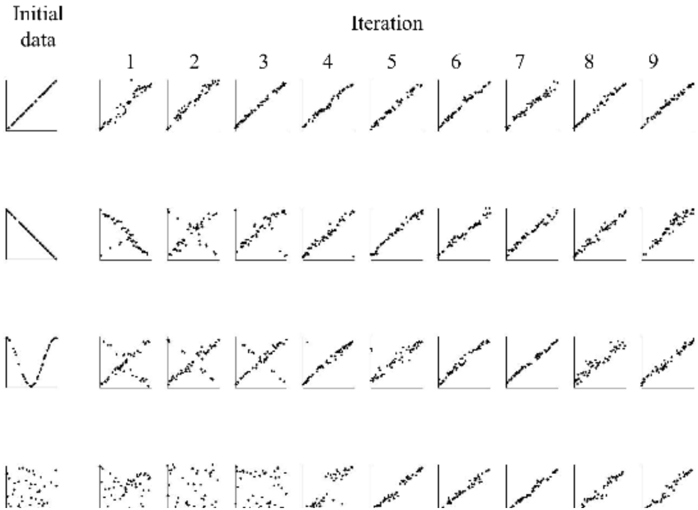
Results



Results



Results



Ordinary Least Squares

Ordinary Least Squares

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

Ordinary Least Squares

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

Ordinary Least Squares

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

Ordinary Least Squares

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

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

Ordinary Least Squares

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

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

- 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:

$$\hat{\beta}_0 = \overline{Y} - \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\overline{X^2} - \overline{X}^2}\overline{X} = \overline{Y} - \hat{\beta}_1 \overline{X}$$

$$\hat{\beta}_1 = \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\overline{X^2} - \overline{X}^2} = \frac{\widehat{Cov}(X, Y)}{\widehat{V}(X)} = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

Plug-in Estimation of the BLP

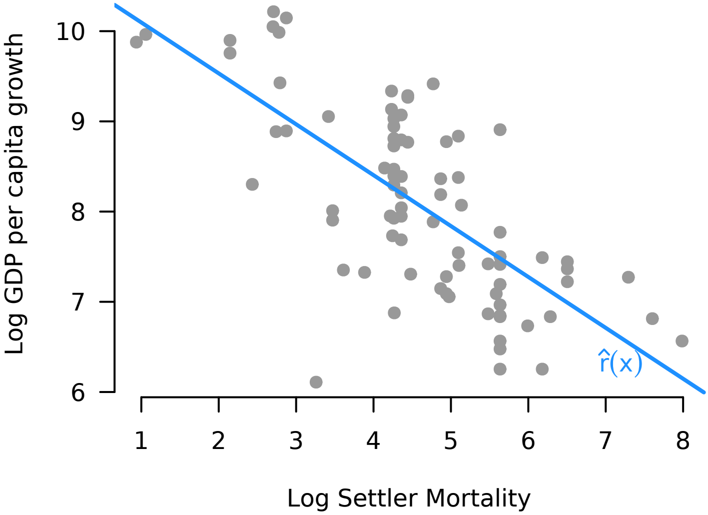
Noting we can rewrite $\frac{Cov[X,Y]}{V[X]} = \frac{E[XY]-E[X]E[Y]}{E[X^2]-E[X]^2}$, and using the plug-in principle, we can estimate the parameters of the BLP as:

$$\hat{\beta}_0 = \overline{Y} - \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\overline{X^2} - \overline{X}^2}\overline{X} = \overline{Y} - \hat{\beta}_1\overline{X}$$

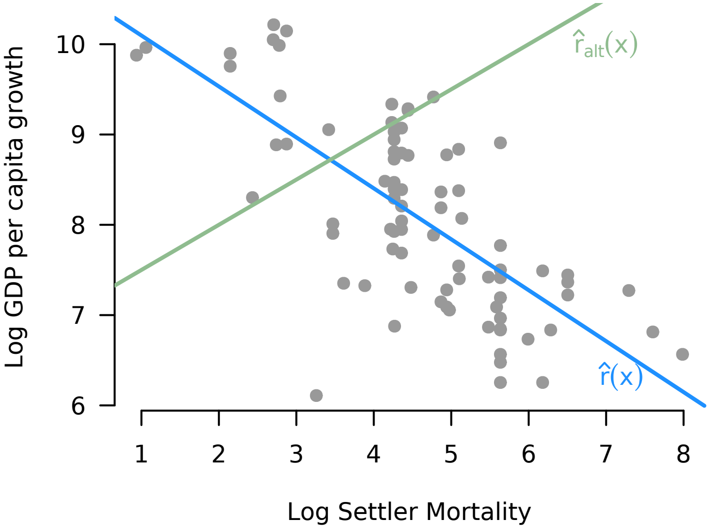
$$\hat{\beta}_1 = \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\overline{X^2} - \overline{X}^2} = \frac{\widehat{Cov}(X, Y)}{\widehat{V}(X)} = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

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:

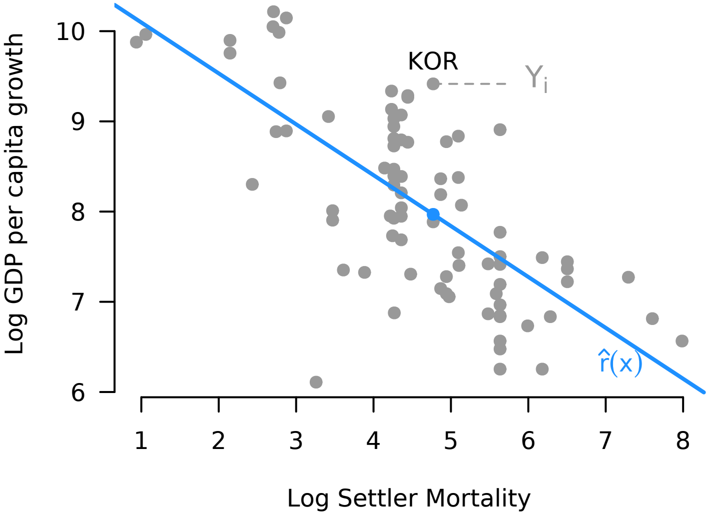
$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$$

Definition (Residual)

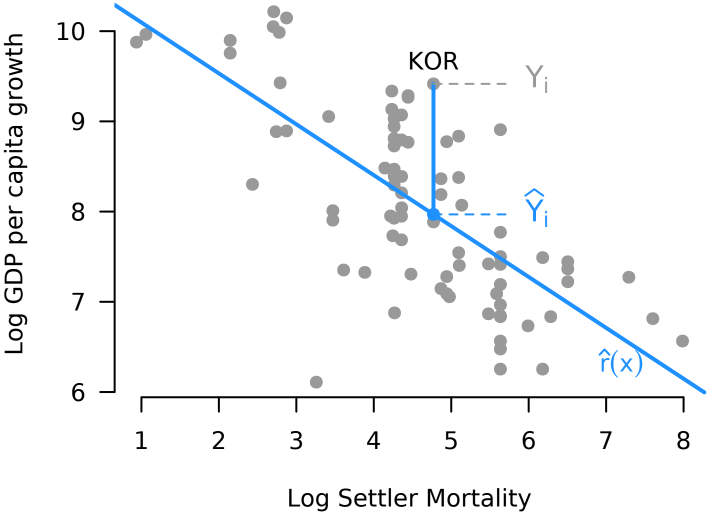
The **residual** is the difference between the actual value of Y_i and the predicted value, $\widehat{Y}_i$:

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

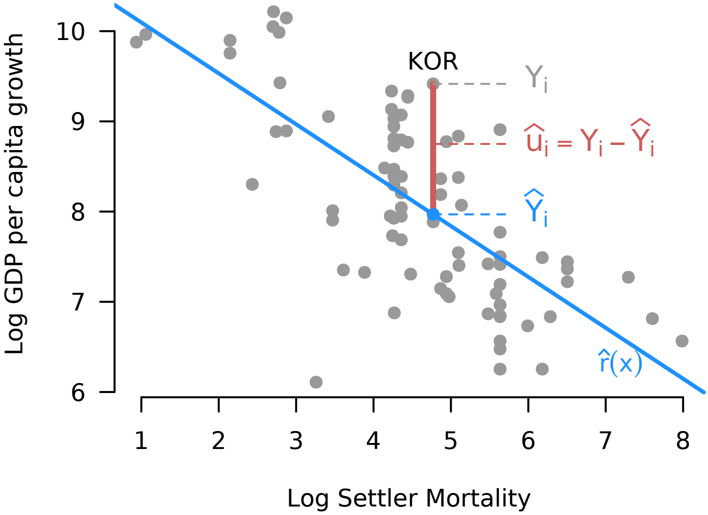
Fitted linear CEF/regression function



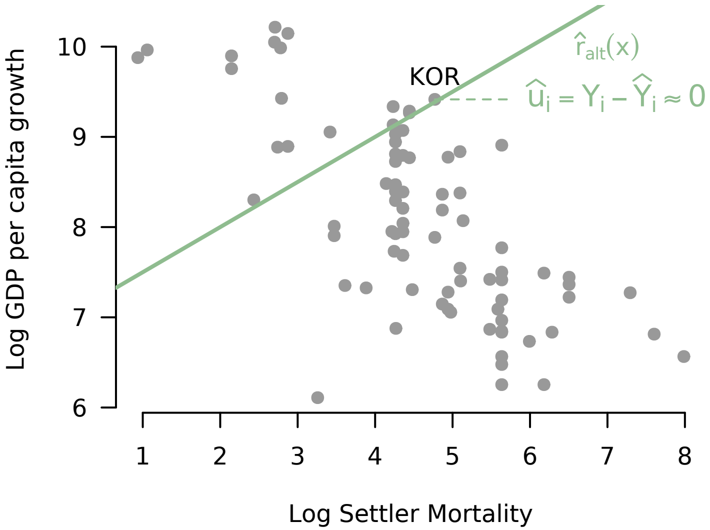
Fitted linear CEF/regression function



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

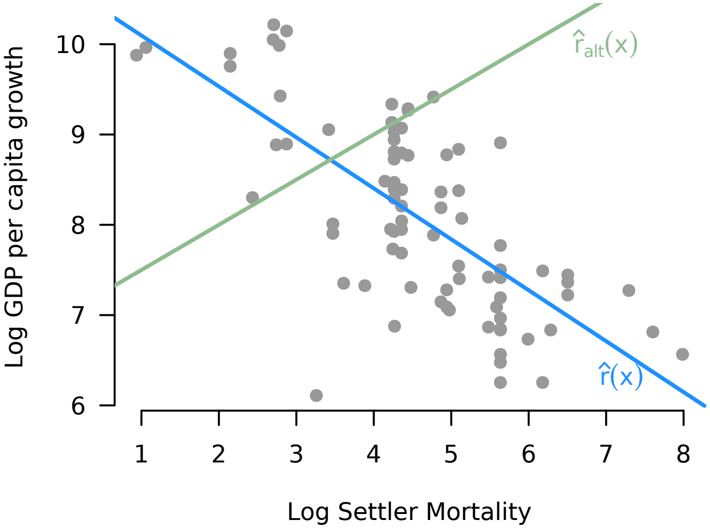
Minimize the residuals

- 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?



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: $(Y_1, X_1), (Y_2, X_2), \dots, (Y_n, X_n)$
- 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: $(Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n)$
- Let $\{b_0, b_1\}$ be possible values for $\{\beta_0, \beta_1\}$
- Define the least squares objective function:

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

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

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

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

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

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

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

$$b_0 = \overline{Y} - b_1 \overline{X}$$

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

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

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

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

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

Solving for the Slope

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

Solving for the Slope

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

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

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

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

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

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

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

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

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

Solving for the Slope

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

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

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

Solving for the Slope

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

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

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

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

Solving for the Slope

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

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

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

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

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

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 intercept equation tells us that the regression line goes through the point $(\overline{Y}, \overline{X})$:
$$\overline{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 \overline{X}$$

Intuition of the OLS estimator

- The intercept equation tells us that the regression line goes through the point $(\overline{Y}, \overline{X})$:

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

- 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 intercept equation tells us that the regression line goes through the point $(\overline{Y}, \overline{X})$:
$$\overline{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 \overline{X}$$
- 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 intercept equation tells us that the regression line goes through the point $(\overline{Y}, \overline{X})$:
$$\overline{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 \overline{X}$$
- 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 intercept equation tells us that the regression line goes through the point $(\overline{Y}, \overline{X})$:

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

- 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

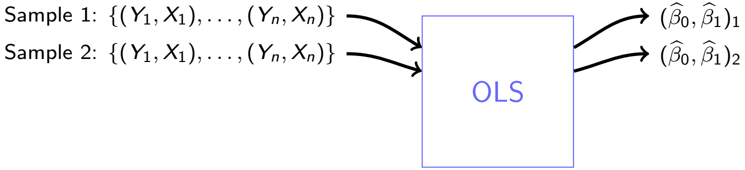
Sampling distribution of the OLS estimator

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

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

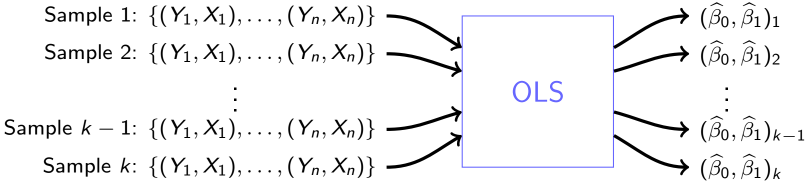
Sampling distribution of the OLS estimator

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



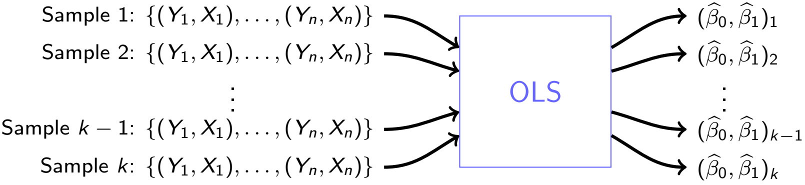
Sampling distribution of the OLS estimator

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



Sampling distribution of the OLS estimator

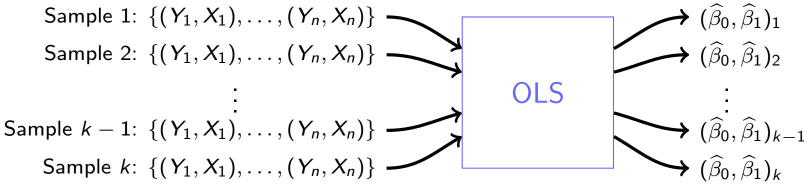
- Remember: OLS is an estimator—it's a machine that we plug samples into and we get out estimates.



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

Sampling distribution of the OLS estimator

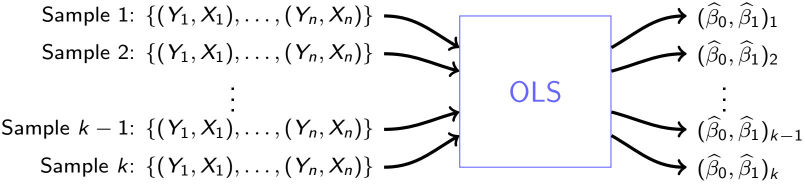
- Remember: OLS is an estimator—it's a machine that we plug samples into and we get out estimates.



- 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

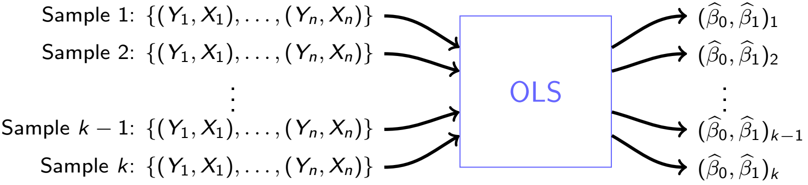
- Remember: OLS is an estimator—it's a machine that we plug samples into and we get out estimates.



- 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.

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

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

$\vdots$

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

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

- 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()`

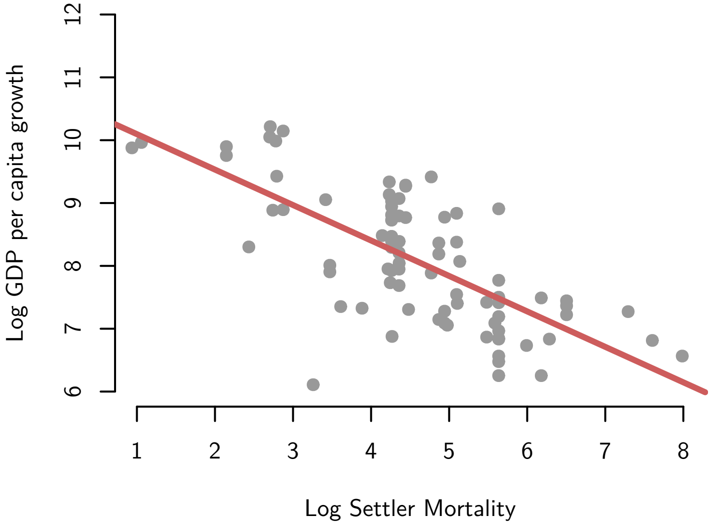
Simulation procedure

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

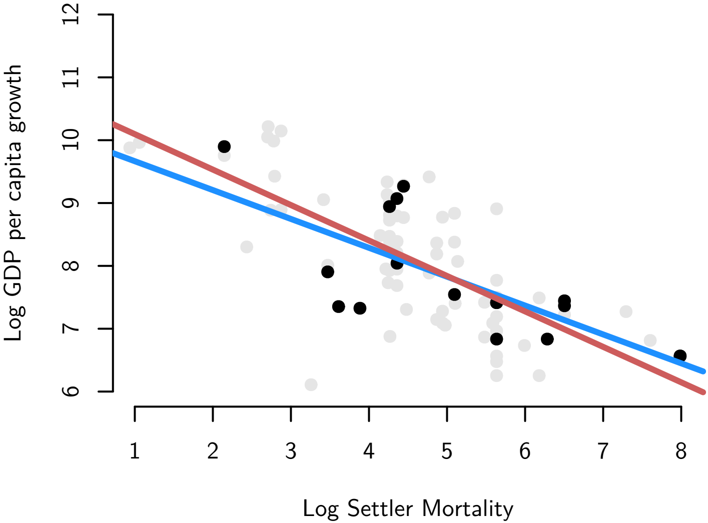
Simulation procedure

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

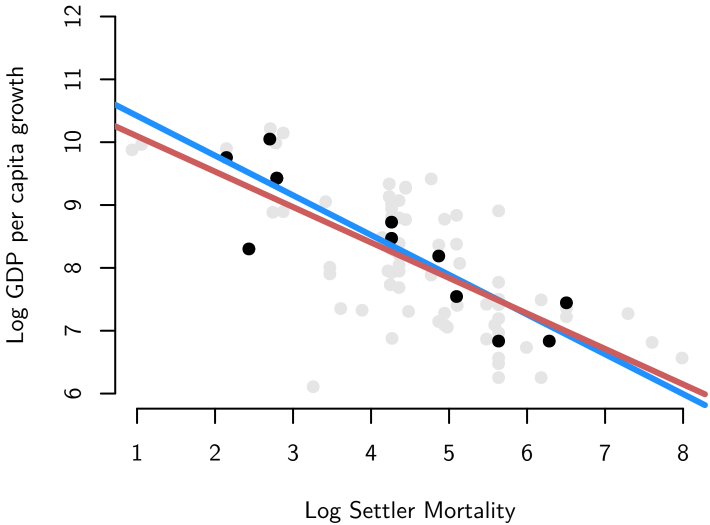
Population Regression



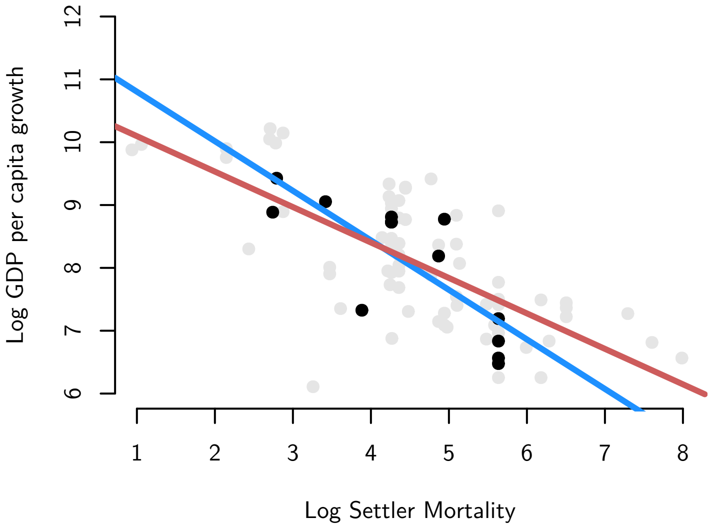
Randomly sample from AJR



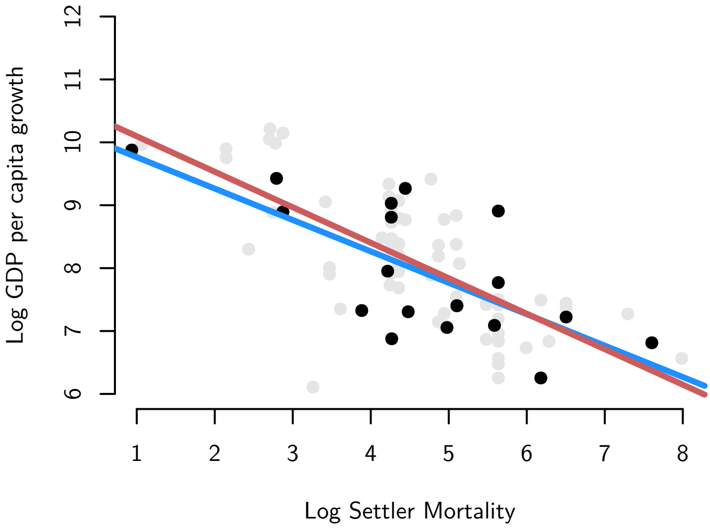
Randomly sample from AJR



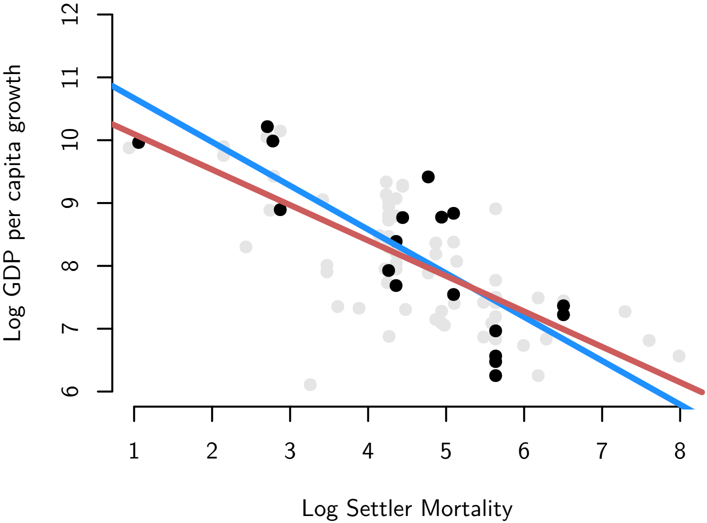
Randomly sample from AJR



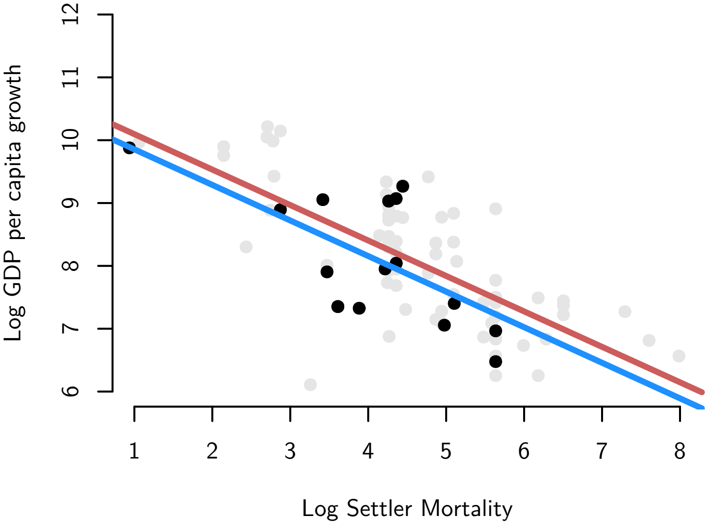
Randomly sample from AJR



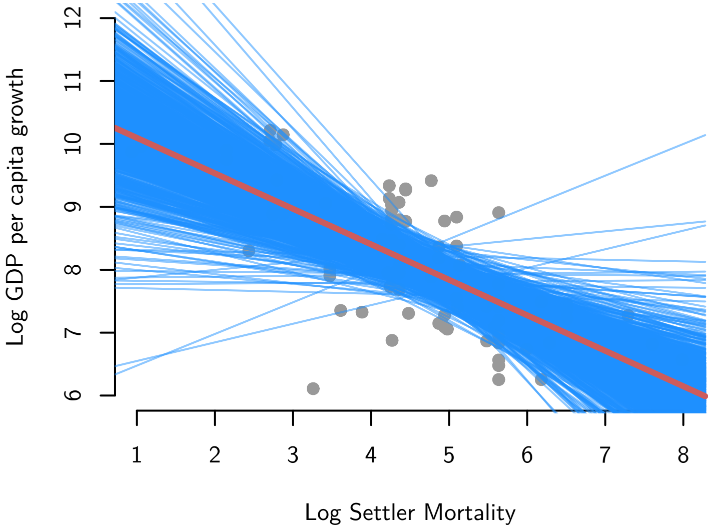
Randomly sample from AJR



Randomly sample from AJR



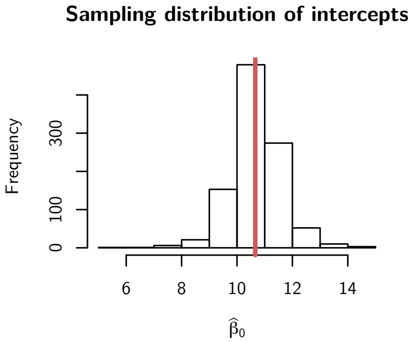
Randomly sample from AJR

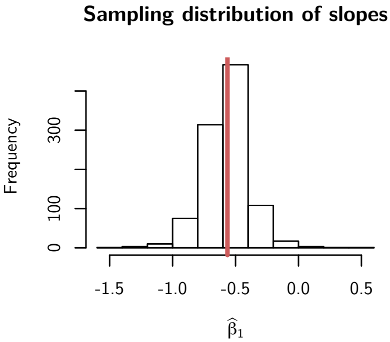


Sampling distribution of OLS

Sampling distribution of OLS

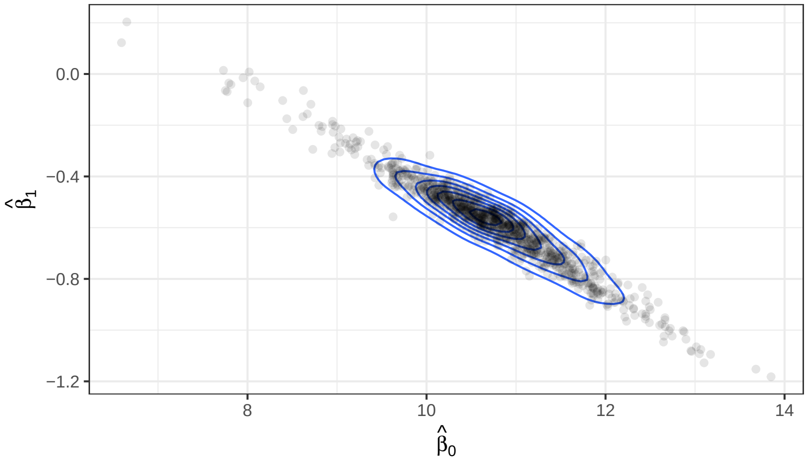
- You can see that the estimated slopes and intercepts vary from sample to sample, but that the "average" of the lines looks about right.





The Sampling Distribution is a Joint Distribution!

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



Sample Mean Properties Review

Sample Mean Properties Review

- In the last few weeks we derived the properties of the sampling distribution for the sample mean, $\bar{X}_n$.

Sample Mean Properties Review

- In the last few weeks we derived the properties of the sampling distribution for the sample mean, $\bar{X}_n$.
- Under essentially only the iid assumption (plus finite mean and variance) we derived the large sample distribution as

$$\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Sample Mean Properties Review

- In the last few weeks we derived the properties of the sampling distribution for the sample mean, $\bar{X}_n$.
- Under essentially only the iid assumption (plus finite mean and variance) we derived the large sample distribution as

$$\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

  - This means the estimator is unbiased for the population mean: $E[\bar{X}_n] = \mu$.

Sample Mean Properties Review

- In the last few weeks we derived the properties of the sampling distribution for the sample mean, $\bar{X}_n$.
- Under essentially only the iid assumption (plus finite mean and variance) we derived the large sample distribution as

$$\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

  - This means the estimator is unbiased for the population mean: $E[\bar{X}_n] = \mu$.
  - has sampling variance: σ^2/n

Sample Mean Properties Review

- In the last few weeks we derived the properties of the sampling distribution for the sample mean, $\bar{X}_n$.
- Under essentially only the iid assumption (plus finite mean and variance) we derived the large sample distribution as

$$\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

  - This means the estimator is unbiased for the population mean: $E[\bar{X}_n] = \mu$.
  - has sampling variance: σ^2/n
  - and standard error: $\sigma/\sqrt{n}$

Sample Mean Properties Review

- In the last few weeks we derived the properties of the sampling distribution for the sample mean, $\bar{X}_n$.
- Under essentially only the iid assumption (plus finite mean and variance) we derived the large sample distribution as

$$\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

  - This means the estimator is unbiased for the population mean: $E[\bar{X}_n] = \mu$.
  - has sampling variance: σ^2/n
  - and standard error: $\sigma/\sqrt{n}$
- This in turn gave us confidence intervals and hypothesis tests.

Sample Mean Properties Review

- In the last few weeks we derived the properties of the sampling distribution for the sample mean, $\bar{X}_n$.
- Under essentially only the iid assumption (plus finite mean and variance) we derived the large sample distribution as

$$\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

  - This means the estimator is unbiased for the population mean: $E[\bar{X}_n] = \mu$.
  - has sampling variance: σ^2/n
  - and standard error: $\sigma/\sqrt{n}$
- This in turn gave us confidence intervals and hypothesis tests.
- We will use the same strategy here!

Our goal

Our goal

- What is the sampling distribution of the OLS slope?

$$\widehat{\beta}_1 \sim ?(?, ?)$$

Our goal

- What is the sampling distribution of the OLS slope?

$$\widehat{\beta}_1 \sim ?(?, ?)$$

- We need fill in those ?s.

Our goal

- What is the sampling distribution of the OLS slope?

$$\widehat{\beta}_1 \sim ?(?, ?)$$

- We need fill in those ?s.
- It will turn out that if we maintain the BLP that OLS will have nice asymptotic properties: it will be consistent with an asymptotically normal sampling distribution such that,

$$\sqrt{n}\left(\hat{\beta} - \beta\right) \xrightarrow{d} \mathcal{N}(0, \mathbf{V}_{\hat{\beta}})$$

for a covariance matrix $\mathbf{V}_{\hat{\beta}}$ that we will define later.

Our goal

- What is the sampling distribution of the OLS slope?

$$\widehat{\beta}_1 \sim ?(?, ?)$$

- We need fill in those ?s.
- It will turn out that if we maintain the BLP that OLS will have nice asymptotic properties: it will be consistent with an asymptotically normal sampling distribution such that,

$$\sqrt{n}\left(\hat{\beta} - \beta\right) \xrightarrow{d} \mathcal{N}(0, \mathbf{V}_{\hat{\beta}})$$

  for a covariance matrix $\mathbf{V}_{\hat{\beta}}$ that we will define later.
- We will start though by deriving some properties under a very strong classical linear CEF model.

Our goal

- What is the sampling distribution of the OLS slope?

$$\widehat{\beta}_1 \sim ?(?, ?)$$

- We need fill in those ?s.
- It will turn out that if we maintain the BLP that OLS will have nice asymptotic properties: it will be consistent with an asymptotically normal sampling distribution such that,

$$\sqrt{n}\left(\hat{\beta} - \beta\right) \xrightarrow{d} \mathcal{N}(0, \mathbf{V}_{\hat{\beta}})$$

  for a covariance matrix $\mathbf{V}_{\hat{\beta}}$ that we will define later.
- We will start though by deriving some properties under a very strong classical linear CEF model.
- We will return to the agnostic BLP view in future weeks.

Classical Model: OLS Assumptions Preview

Classical Model: OLS Assumptions Preview

1. Linearity in Parameters: The population model is linear in its parameters and correctly specified

Classical Model: OLS Assumptions Preview

1. Linearity in Parameters: The population model is linear in its parameters and correctly specified
2. Random Sampling: The observed data represent a random sample from the population described by the model.

Classical Model: OLS Assumptions Preview

1. Linearity in Parameters: The population model is linear in its parameters and correctly specified
2. Random Sampling: The observed data represent a random sample from the population described by the model.
3. Variation in X: There is variation in the explanatory variable.

Classical Model: OLS Assumptions Preview

1. Linearity in Parameters: The population model is linear in its parameters and correctly specified
2. Random Sampling: The observed data represent a random sample from the population described by the model.
3. Variation in X: There is variation in the explanatory variable.
4. Zero conditional mean: Expected value of the error term is zero conditional on all values of the explanatory variable

Classical Model: OLS Assumptions Preview

1. Linearity in Parameters: The population model is linear in its parameters and correctly specified
2. Random Sampling: The observed data represent a random sample from the population described by the model.
3. Variation in X: There is variation in the explanatory variable.
4. Zero conditional mean: Expected value of the error term is zero conditional on all values of the explanatory variable
5. Homoskedasticity: The error term has the same variance conditional on all values of the explanatory variable.

Classical Model: OLS Assumptions Preview

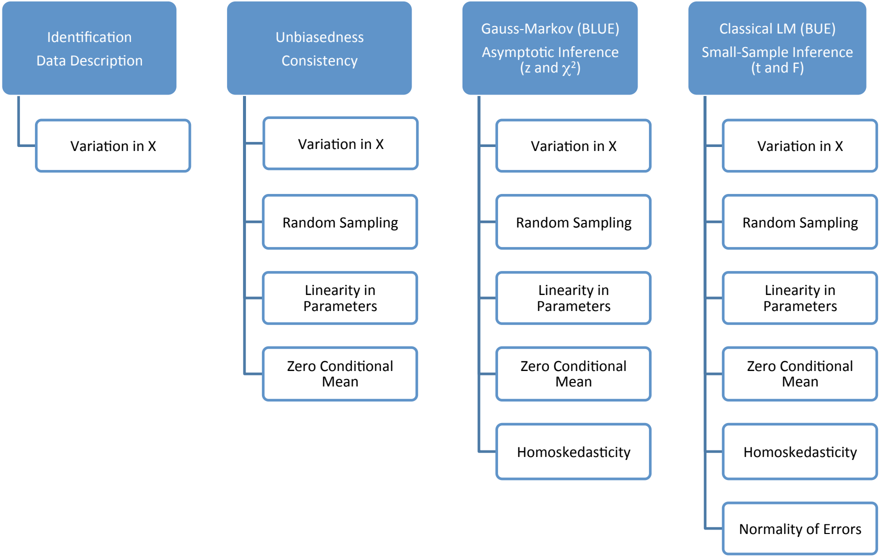
1. Linearity in Parameters: The population model is linear in its parameters and correctly specified
2. Random Sampling: The observed data represent a random sample from the population described by the model.
3. Variation in X: There is variation in the explanatory variable.
4. Zero conditional mean: Expected value of the error term is zero conditional on all values of the explanatory variable
5. Homoskedasticity: The error term has the same variance conditional on all values of the explanatory variable.
6. Normality: The error term is independent of the explanatory variables and normally distributed.

Classical Model: OLS Assumptions Preview

1. Linearity in Parameters: The population model is linear in its parameters and correctly specified
2. Random Sampling: The observed data represent a random sample from the population described by the model.
3. Variation in X: There is variation in the explanatory variable.
4. Zero conditional mean: Expected value of the error term is zero conditional on all values of the explanatory variable
5. Homoskedasticity: The error term has the same variance conditional on all values of the explanatory variable.
6. Normality: The error term is independent of the explanatory variables and normally distributed.

Hierarchy of OLS Assumptions

Hierarchy of OLS Assumptions



OLS Assumption I

OLS Assumption I

Assumption (I. Linearity in Parameters)

The population regression model is linear in its parameters and correctly specified as:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

OLS Assumption I

Assumption (I. Linearity in Parameters)

The population regression model is linear in its parameters and correctly specified as:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- Note that it can be nonlinear *in variables*

OLS Assumption I

Assumption (I. Linearity in Parameters)

The population regression model is linear in its parameters and correctly specified as:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- Note that it can be nonlinear *in variables*
  - OK: $Y_i = \beta_0 + \beta_1 X_i + u_i$ or
    $Y_i = \beta_0 + \beta_1 X_i^2 + u_i$ or
    $Y_i = \beta_0 + \beta_1 log(X_i) + u$

OLS Assumption I

Assumption (I. Linearity in Parameters)

The population regression model is linear in its parameters and correctly specified as:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- Note that it can be nonlinear *in variables*
  - OK: $Y_i = \beta_0 + \beta_1 X_i + u_i$ or
    $Y_i = \beta_0 + \beta_1 X_i^2 + u_i$ or
    $Y_i = \beta_0 + \beta_1 log(X_i) + u$
  - Not OK: $Y_i = \beta_0 + \beta_1^2 X_i + u_i$ or
    $Y_i = \beta_0 + exp(\beta_1) X_i + u_i$

OLS Assumption I

Assumption (I. Linearity in Parameters)

The population regression model is linear in its parameters and correctly specified as:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- Note that it can be nonlinear *in variables*
  - OK: $Y_i = \beta_0 + \beta_1 X_i + u_i$ or
    $Y_i = \beta_0 + \beta_1 X_i^2 + u_i$ or
    $Y_i = \beta_0 + \beta_1 log(X_i) + u$
  - Not OK: $Y_i = \beta_0 + \beta_1^2 X_i + u_i$ or
    $Y_i = \beta_0 + exp(\beta_1) X_i + u_i$
- β_0, β_1: Population parameters — fixed and unknown

OLS Assumption I

Assumption (I. Linearity in Parameters)

The population regression model is linear in its parameters and correctly specified as:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- Note that it can be nonlinear *in variables*
  - OK: $Y_i = \beta_0 + \beta_1 X_i + u_i$ or
    $Y_i = \beta_0 + \beta_1 X_i^2 + u_i$ or
    $Y_i = \beta_0 + \beta_1 log(X_i) + u$
  - Not OK: $Y_i = \beta_0 + \beta_1^2 X_i + u_i$ or
    $Y_i = \beta_0 + exp(\beta_1) X_i + u_i$

- β_0, β_1: Population parameters — fixed and unknown
- u_i: Unobserved random variable with $E[u_i] = 0$ — captures all other factors influencing Y_i other than X_i

OLS Assumption I

Assumption (I. Linearity in Parameters)

The population regression model is linear in its parameters and correctly specified as:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- Note that it can be nonlinear *in variables*
  - OK: $Y_i = \beta_0 + \beta_1 X_i + u_i$ or
    $Y_i = \beta_0 + \beta_1 X_i^2 + u_i$ or
    $Y_i = \beta_0 + \beta_1 log(X_i) + u$
  - Not OK: $Y_i = \beta_0 + \beta_1^2 X_i + u_i$ or
    $Y_i = \beta_0 + exp(\beta_1) X_i + u_i$
- β_0, β_1: Population parameters — fixed and unknown
- u_i: Unobserved random variable with $E[u_i] = 0$ — captures all other factors influencing Y_i other than X_i
- We assume this to be the structural model, i.e., the model describing the true process generating Y_i

OLS Assumption II

OLS Assumption II

Assumption (II. Random Sampling)

The observed data:

$$(y_i,\ x_i) \text{ for } i = 1, ..., n$$

represent an i.i.d. random sample of size n following the population model.

OLS Assumption II

Assumption (II. Random Sampling)

The observed data:

$$(y_i,\ x_i) \text{ for } i = 1, \dots, n$$

represent an i.i.d. random sample of size n following the population model.

Data examples consistent with this assumption:

OLS Assumption II

Assumption (II. Random Sampling)

The observed data:

$$(y_i,\ x_i) \text{ for } i = 1, ..., n$$

represent an i.i.d. random sample of size n following the population model.

Data examples consistent with this assumption:

- A cross-sectional survey where the units are sampled randomly

OLS Assumption II

Assumption (II. Random Sampling)

The observed data:

$$(y_i,\ x_i) \text{ for } i = 1, ..., n$$

represent an i.i.d. random sample of size n following the population model.

Data examples consistent with this assumption:

- A cross-sectional survey where the units are sampled randomly

Potential Violations:

OLS Assumption II

Assumption (II. Random Sampling)

The observed data:

$$(y_i,\ x_i) \text{ for } i = 1, ..., n$$

represent an i.i.d. random sample of size n following the population model.

Data examples consistent with this assumption:

- A cross-sectional survey where the units are sampled randomly

Potential Violations:

- Time series data (regressor values may exhibit persistence)

OLS Assumption II

Assumption (II. Random Sampling)

The observed data:

$$(y_i,\ x_i) \text{ for } i = 1, ..., n$$

represent an i.i.d. random sample of size n following the population model.

Data examples consistent with this assumption:

- A cross-sectional survey where the units are sampled randomly

Potential Violations:

- Time series data (regressor values may exhibit persistence)
- Sample selection problems (sample not representative of the population)

OLS Assumption III

OLS Assumption III

Assumption (III. Variation in X; a.k.a. No Perfect Collinearity)

The observed data:

$$x_i \text{ for } i = 1, ..., n$$

are not all the same value.

OLS Assumption III

Assumption (III. Variation in X; a.k.a. No Perfect Collinearity)

The observed data:

$$x_i \text{ for } i = 1, ..., n$$

are not all the same value.

Satisfied as long as there is some variation in the regressor X in the sample.

OLS Assumption III

Assumption (III. Variation in X; a.k.a. No Perfect Collinearity)

The observed data:

$$x_i \text{ for } i = 1, ..., n$$

are not all the same value.

Satisfied as long as there is some variation in the regressor X in the sample.

Why do we need this?

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

OLS Assumption III

Assumption (III. Variation in X; a.k.a. No Perfect Collinearity)

The observed data:

$$x_i \textit{ for } i = 1, ..., n$$

are not all the same value.

Satisfied as long as there is some variation in the regressor X in the sample.

Why do we need this?

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

This assumption is needed just to calculate $\hat{\beta}$.

OLS Assumption III

Assumption (III. Variation in X; a.k.a. No Perfect Collinearity)

The observed data:

$$x_i \text{ for } i = 1, ..., n$$

are not all the same value.

Satisfied as long as there is some variation in the regressor X in the sample.

Why do we need this?

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

This assumption is needed just to calculate $\hat{\beta}$.

Only assumption needed for using OLS as a pure data summary.

Stuck in a moment

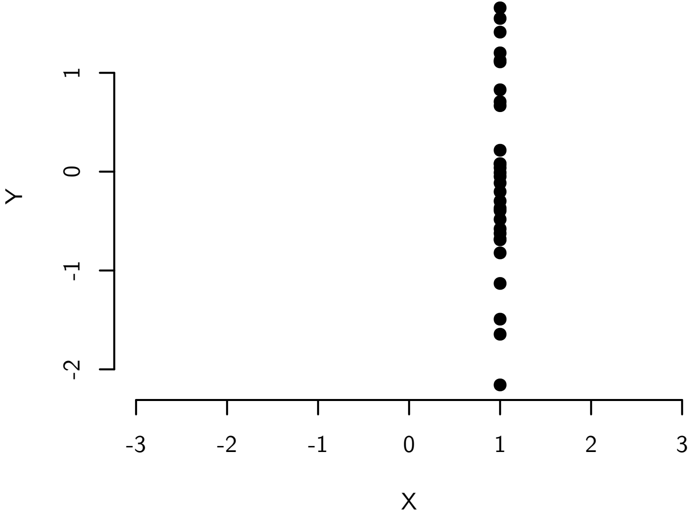
- Why does this matter?

Stuck in a moment

- Why does this matter?

Stuck in a moment

- Why does this matter? How would you draw the line of best fit through this scatterplot, which is a violation of this assumption?



OLS Assumption IV

OLS Assumption IV

Assumption (IV. Zero Conditional Mean)

The expected value of the error term is zero conditional on any value of the explanatory variable:

$$E[u_i|X_i = x] \ = \ 0$$

OLS Assumption IV

Assumption (IV. Zero Conditional Mean)

The expected value of the error term is zero conditional on any value of the explanatory variable:

$$E[u_i|X_i = x] \; = \; 0$$

- $E[u_i|X] = 0$ implies a slightly weaker condition $\text{Cov}(X, u) = 0$

OLS Assumption IV

Assumption (IV. Zero Conditional Mean)

The expected value of the error term is zero conditional on any value of the explanatory variable:

$$E[u_i|X_i = x] \ = \ 0$$

- $E[u_i|X] = 0$ implies a slightly weaker condition $\text{Cov}(X, u) = 0$
- Given random sampling, $E[u|X] = 0$ also implies $E[u_i|x_i] = 0$ for all i

OLS Assumption IV

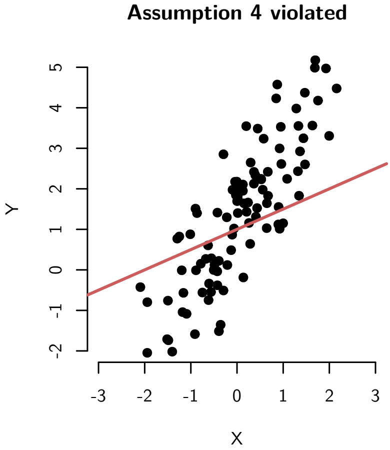
Assumption (IV. Zero Conditional Mean)

The expected value of the error term is zero conditional on any value of the explanatory variable:

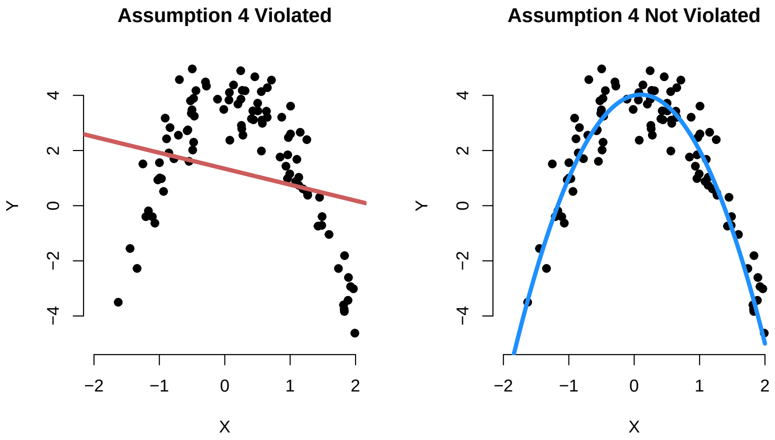
$$E[u_i|X_i = x] \ = \ 0$$

- $E[u_i|X] = 0$ implies a slightly weaker condition $\text{Cov}(X, u) = 0$
- Given random sampling, $E[u|X] = 0$ also implies $E[u_i|x_i] = 0$ for all i

Violating the zero conditional mean assumption



Violating the zero conditional mean assumption



Unbiasedness

With Assumptions 1-4, we can show that the OLS estimator for the slope is unbiased, that is $E[\widehat{\beta}_1] = \beta_1$.

Let's prove it!

Lemma 3: Weighted Combinations of X_i

Lemma ($\sum_i W_i X_i = 1$)

Lemma 3: Weighted Combinations of X_i

Lemma ($\sum_i W_i X_i = 1$)

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

Lemma 3: Weighted Combinations of X_i

Lemma ($\sum_i W_i X_i = 1$)

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

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

Lemma 3: Weighted Combinations of X_i

Lemma ($\sum_i W_i X_i = 1$)

$$\begin{aligned}
\sum_{i=1}^{n} W_i X_i &= \sum_{i=1}^{n} \frac{X_i(X_i - \overline{X})}{\sum_{j=1}^{n}(X_j - \overline{X})^2} \\
&= \frac{1}{\sum_{j=1}^{n}(X_j - \overline{X})^2} \sum_{i=1}^{n} X_i(X_i - \overline{X}) \\
&= \frac{1}{\sum_{j=1}^{n}(X_j - \overline{X})^2} \left[\sum_{i=1}^{n} X_i(X_i - \overline{X}) - \sum_{i=1}^{n} \overline{X}(X_i - \overline{X})\right]
\end{aligned}$$

Lemma 3: Weighted Combinations of X_i

Lemma ($\sum_i W_i X_i = 1$)

$$\begin{aligned}
\sum_{i=1}^{n} W_i X_i &= \sum_{i=1}^{n} \frac{X_i(X_i - \overline{X})}{\sum_{j=1}^{n}(X_j - \overline{X})^2} \\
&= \frac{1}{\sum_{j=1}^{n}(X_j - \overline{X})^2} \sum_{i=1}^{n} X_i(X_i - \overline{X}) \\
&= \frac{1}{\sum_{j=1}^{n}(X_j - \overline{X})^2} \left[\sum_{i=1}^{n} X_i(X_i - \overline{X}) - \sum_{i=1}^{n} \overline{X}(X_i - \overline{X}) \right] \\
&= \frac{1}{\sum_{j=1}^{n}(X_j - \overline{X})^2} \sum_{i=1}^{n} (X_i - \overline{X})(X_i - \overline{X})
\end{aligned}$$

Lemma 3: Weighted Combinations of X_i

Lemma ($\sum_i W_i X_i = 1$)

$$\begin{aligned}
\sum_{i=1}^{n} W_i X_i &= \sum_{i=1}^{n} \frac{X_i(X_i - \overline{X})}{\sum_{j=1}^{n}(X_j - \overline{X})^2} \\
&= \frac{1}{\sum_{j=1}^{n}(X_j - \overline{X})^2} \sum_{i=1}^{n} X_i(X_i - \overline{X}) \\
&= \frac{1}{\sum_{j=1}^{n}(X_j - \overline{X})^2} \left[\sum_{i=1}^{n} X_i(X_i - \overline{X}) - \sum_{i=1}^{n} \overline{X}(X_i - \overline{X}) \right] \\
&= \frac{1}{\sum_{j=1}^{n}(X_j - \overline{X})^2} \sum_{i=1}^{n} (X_i - \overline{X})(X_i - \overline{X}) \\
&= 1
\end{aligned}$$

Lemma 4: Estimation Error

Lemma

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

Lemma 4: Estimation Error

Lemma

$$
\begin{aligned}
\widehat{\beta}_1 &= \sum_{i=1}^{n} W_i Y_i \\
&= \sum_{i=1}^{n} W_i(\beta_0 + \beta_1 X_i + u_i)
\end{aligned}
$$

Lemma 4: Estimation Error

Lemma

$$
\begin{aligned}
\widehat{\beta}_1 &= \sum_{i=1}^{n} W_i Y_i \\
&= \sum_{i=1}^{n} W_i(\beta_0 + \beta_1 X_i + u_i) \\
&= \beta_0 \left(\sum_{i=1}^{n} W_i \right) + \beta_1 \left(\sum_{i=1}^{n} W_i X_i \right) + \sum_{i=1}^{n} W_i u_i
\end{aligned}
$$

Lemma 4: Estimation Error

Lemma

$$
\begin{aligned}
\widehat{\beta}_1 &= \sum_{i=1}^{n} W_i Y_i \\
&= \sum_{i=1}^{n} W_i(\beta_0 + \beta_1 X_i + u_i) \\
&= \beta_0 \left(\sum_{i=1}^{n} W_i \right) + \beta_1 \left(\sum_{i=1}^{n} W_i X_i \right) + \sum_{i=1}^{n} W_i u_i \\
&= \beta_1 + \sum_{i=1}^{n} W_i u_i
\end{aligned}
$$

Lemma 4: Estimation Error

Lemma

$$
\begin{aligned}
\widehat{\beta}_1 &= \sum_{i=1}^{n} W_i Y_i \\
&= \sum_{i=1}^{n} W_i(\beta_0 + \beta_1 X_i + u_i) \\
&= \beta_0 \left(\sum_{i=1}^{n} W_i \right) + \beta_1 \left(\sum_{i=1}^{n} W_i X_i \right) + \sum_{i=1}^{n} W_i u_i \\
&= \beta_1 + \sum_{i=1}^{n} W_i u_i \\
\hat{\beta}_1 - \beta_1 &= \sum_{i=1}^{n} W_i u_i
\end{aligned}
$$

Unbiasedness Proof

$$E[\hat{\beta}_1 - \beta_1|X] = E\left[\sum_{i=1}^{n} W_i u_i|X\right]$$

Unbiasedness Proof

$$\begin{aligned} E[\hat{\beta}_1 - \beta_1 | X] &= E\left[\sum_{i=1}^{n} W_i u_i | X\right] \\ &= \sum_{i=1}^{n} E[W_i u_i | X] \end{aligned}$$

Unbiasedness Proof

$$
\begin{aligned}
E[\hat{\beta}_1 - \beta_1 | X] &= E\left[\sum_{i=1}^{n} W_i u_i | X\right] \\
&= \sum_{i=1}^{n} E[W_i u_i | X] \\
&= \sum_{i=1} W_i E[u_i | X]
\end{aligned}
$$

Unbiasedness Proof

$$
\begin{aligned}
E[\hat{\beta}_1 - \beta_1|X] &= E\left[\sum_{i=1}^{n} W_i u_i|X\right] \\
&= \sum_{i=1}^{n} E[W_i u_i|X] \\
&= \sum_{i=1} W_i E[u_i|X] \\
&= \sum_{i=1} W_i 0
\end{aligned}
$$

Unbiasedness Proof

$$
\begin{aligned}
E[\hat{\beta}_1 - \beta_1|X] &= E\left[\sum_{i=1}^{n} W_i u_i|X\right] \\
&= \sum_{i=1}^{n} E[W_i u_i|X] \\
&= \sum_{i=1} W_i E[u_i|X] \\
&= \sum_{i=1} W_i 0 \\
&= 0
\end{aligned}
$$

Unbiasedness Proof

$$\begin{aligned} E[\hat{\beta}_1 - \beta_1|X] &= E\left[\sum_{i=1}^{n} W_i u_i|X\right] \\ &= \sum_{i=1}^{n} E[W_i u_i|X] \\ &= \sum_{i=1} W_i E[u_i|X] \\ &= \sum_{i=1} W_i 0 \\ &= 0 \end{aligned}$$

Using iterated expectations we can show that it is also unconditionally biased $E[\hat{\beta}_1] = E[E[\hat{\beta}_1|X]] = E[\beta_1] = \beta_1$.

Consistency

Consistency

- Recall the estimation error,

$$\hat{\beta}_1 = \beta_1 + \sum_{i=1}^{n} W_i u_i$$

Consistency

- Recall the estimation error,

$$\hat{\beta}_1 = \beta_1 + \sum_{i=1}^{n} W_i u_i$$

- Under iid sampling we have

$$\sum_{i=1}^{n} W_i u_i = \frac{\sum_{i=1}^{n}(X_i - \overline{X})u_i}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \xrightarrow{p} \frac{\text{Cov}[X_i, u_i]}{V[X_i]}$$

Consistency

- Recall the estimation error,

$$\hat{\beta}_1 = \beta_1 + \sum_{i=1}^{n} W_i u_i$$

- Under iid sampling we have

$$\sum_{i=1}^{n} W_i u_i = \frac{\sum_{i=1}^{n}(X_i - \overline{X})u_i}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \xrightarrow{p} \frac{\text{Cov}[X_i, u_i]}{V[X_i]}$$

- Under A4 (zero conditional mean error) we have the slightly weaker property $Cov[X_i, u_i] = 0$ so as long as $V[X] > 0$, then we have,

$$\hat{\beta}_1 \xrightarrow{p} \beta_1$$

Where are we?

Where are we?

- Now we know that, under Assumptions 1-4, we know that

$$\widehat{\beta}_1 \sim ?(\beta_1, ?)$$

Where are we?

- Now we know that, under Assumptions 1-4, we know that

$$\widehat{\beta}_1 \sim ?(\beta_1, ?)$$

- That is we know that the sampling distribution is centered on the true population slope, but we don't know the population variance.

Sampling variance of estimated slope

Sampling variance of estimated slope

- In order to derive the sampling variance of the OLS estimator,

Sampling variance of estimated slope

- In order to derive the sampling variance of the OLS estimator,

1. Linearity

Sampling variance of estimated slope

- In order to derive the sampling variance of the OLS estimator,

1. Linearity
2. Random (iid) sample

Sampling variance of estimated slope

- In order to derive the sampling variance of the OLS estimator,

1. Linearity
2. Random (iid) sample
3. Variation in X_i

Sampling variance of estimated slope

- In order to derive the sampling variance of the OLS estimator,

1. Linearity
2. Random (iid) sample
3. Variation in X_i
4. Zero conditional mean of the errors

Sampling variance of estimated slope

- In order to derive the sampling variance of the OLS estimator,

1. Linearity
2. Random (iid) sample
3. Variation in X_i
4. Zero conditional mean of the errors
5. Homoskedasticity

Variance of OLS Estimators

Variance of OLS Estimators

How can we derive $\text{Var}[\hat{\beta}_0]$ and $\text{Var}[\hat{\beta}_1]$? Let's make the following additional assumption:

Variance of OLS Estimators

How can we derive Var$[\hat{\beta}_0]$ and Var$[\hat{\beta}_1]$? Let's make the following additional assumption:

Assumption (V. Homoskedasticity)

The conditional variance of the error term is constant and does not vary as a function of the explanatory variable:

$$Var[u|X] = \sigma_u^2$$

Variance of OLS Estimators

How can we derive $\text{Var}[\hat{\beta}_0]$ and $\text{Var}[\hat{\beta}_1]$? Let's make the following additional assumption:

Assumption (V. Homoskedasticity)

The conditional variance of the error term is constant and does not vary as a function of the explanatory variable:

$$Var[u|X] = \sigma_u^2$$

- This implies $\text{Var}[u] = \sigma_u^2$
  $\rightarrow$ all errors have an identical error variance ($\sigma_{u_i}^2 = \sigma_u^2$ for all i)

Variance of OLS Estimators

How can we derive $\text{Var}[\hat{\beta}_0]$ and $\text{Var}[\hat{\beta}_1]$? Let's make the following additional assumption:

Assumption (V. Homoskedasticity)

The conditional variance of the error term is constant and does not vary as a function of the explanatory variable:

$$Var[u|X] = \sigma_u^2$$

- This implies $\text{Var}[u] = \sigma_u^2$
  $\rightarrow$ all errors have an identical error variance ($\sigma_{u_i}^2 = \sigma_u^2$ for all i)
- Taken together, Assumptions I–V imply:

$$\begin{aligned} E[Y|X] &= \beta_0 + \beta_1 X \\ \text{Var}[Y|X] &= \sigma_u^2 \end{aligned}$$

Variance of OLS Estimators

How can we derive $\text{Var}[\hat{\beta}_0]$ and $\text{Var}[\hat{\beta}_1]$? Let's make the following additional assumption:

Assumption (V. Homoskedasticity)

The conditional variance of the error term is constant and does not vary as a function of the explanatory variable:

$$Var[u|X] = \sigma_u^2$$

- This implies $\text{Var}[u] = \sigma_u^2$
  $\rightarrow$ all errors have an identical error variance ($\sigma_{u_i}^2 = \sigma_u^2$ for all i)
- Taken together, Assumptions I–V imply:

$$\begin{aligned} E[Y|X] &= \beta_0 + \beta_1 X \\ \text{Var}[Y|X] &= \sigma_u^2 \end{aligned}$$

- Violation: $\text{Var}[u|X = x_1] \neq \text{Var}[u|X = x_2]$ called heteroskedasticity.

Variance of OLS Estimators

How can we derive $\text{Var}[\hat{\beta}_0]$ and $\text{Var}[\hat{\beta}_1]$? Let's make the following additional assumption:

Assumption (V. Homoskedasticity)

The conditional variance of the error term is constant and does not vary as a function of the explanatory variable:

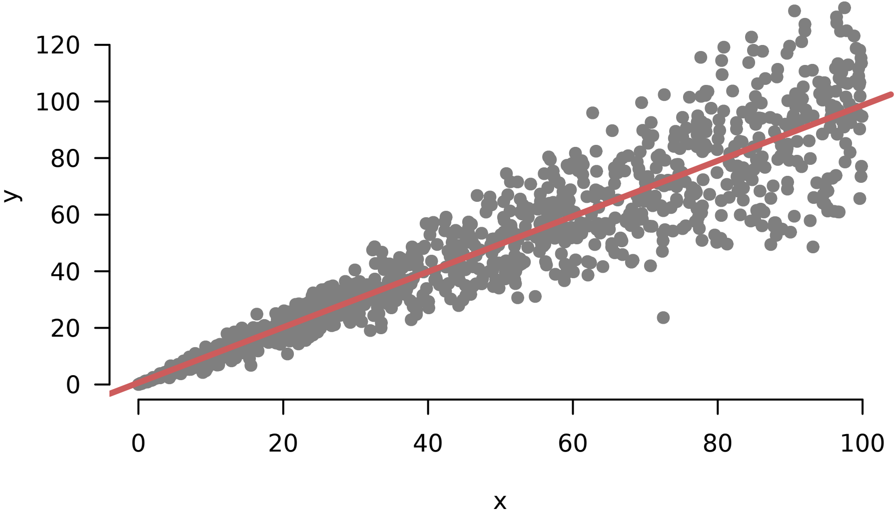
$$Var[u|X] = \sigma_u^2$$

- This implies $\text{Var}[u] = \sigma_u^2$
  $\rightarrow$ all errors have an identical error variance ($\sigma_{u_i}^2 = \sigma_u^2$ for all i)
- Taken together, Assumptions I–V imply:

$$\begin{aligned} E[Y|X] &= \beta_0 + \beta_1 X \\ \text{Var}[Y|X] &= \sigma_u^2 \end{aligned}$$

- Violation: $\text{Var}[u|X = x_1] \neq \text{Var}[u|X = x_2]$ called heteroskedasticity.
- Assumptions I–V are collectively known as the Gauss-Markov assumptions

Heteroskedasticity



Deriving the sampling variance

$$V[\widehat{\beta}_1|X] = ??$$

Deriving the sampling variance

$$V[\widehat{\beta}_1|X] = ??$$

$$V[\widehat{\beta}_1|X] = V\left[\sum_{i=1}^{n} W_i u_i \middle| X\right]$$

Deriving the sampling variance

$$V[\widehat{\beta}_1|X] = ??$$

$$\begin{aligned} V[\widehat{\beta}_1|X] &= V\left[\sum_{i=1}^{n} W_i u_i \middle| X\right] \\ &= \sum_{i=1}^{n} W_i^2 V\left[u_i|X\right] & \text{(A2: iid)} \end{aligned}$$

Deriving the sampling variance

$$V[\widehat{\beta}_1|X] = ??$$

$$\begin{aligned}
V[\widehat{\beta}_1|X] &= V\left[\sum_{i=1}^{n} W_i u_i \middle| X\right] \\
&= \sum_{i=1}^{n} W_i^2 V\left[u_i|X\right] && \text{(A2: iid)} \\
&= \sum_{i=1}^{n} W_i^2 \sigma_u^2 && \text{(A5: homoskedastic)}
\end{aligned}$$

Deriving the sampling variance

$$V[\widehat{\beta}_1|X] = ??$$

$$\begin{aligned}
V[\widehat{\beta}_1|X] &= V\left[\sum_{i=1}^{n} W_i u_i \middle| X\right] \\
&= \sum_{i=1}^{n} W_i^2 V\left[u_i|X\right] && \text{(A2: iid)} \\
&= \sum_{i=1}^{n} W_i^2 \sigma_u^2 && \text{(A5: homoskedastic)} \\
&= \sigma_u^2 \sum_{i=1}^{n} \left(\frac{(X_i - \overline{X})}{\sum_{i'=1}^{n}(X_{i'} - \overline{X})^2}\right)^2
\end{aligned}$$

Deriving the sampling variance

$$V[\widehat{\beta}_1|X] = ??$$

$$\begin{aligned}
V[\widehat{\beta}_1|X] &= V\left[\sum_{i=1}^{n} W_i u_i \middle| X\right] \\
&= \sum_{i=1}^{n} W_i^2 V\left[u_i|X\right] && \text{(A2: iid)} \\
&= \sum_{i=1}^{n} W_i^2 \sigma_u^2 && \text{(A5: homoskedastic)} \\
&= \sigma_u^2 \sum_{i=1}^{n} \left(\frac{(X_i - \overline{X})}{\sum_{i'=1}^{n}(X_{i'} - \overline{X})^2}\right)^2 \\
&= \frac{\sigma_u^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}
\end{aligned}$$

Variance of OLS Estimators

Theorem (Variance of OLS Estimators)

Given OLS Assumptions I–V (Gauss-Markov Assumptions):

$$V[\hat{\beta}_1 \mid X] = \frac{\sigma_u^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$V[\hat{\beta}_0 \mid X] = \sigma_u^2 \left\{ \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right\}$$

where $V[u \mid X] = \sigma_u^2$ *(the error variance).*

Understanding the sampling variance

$$V[\widehat{\beta}_1|X_1, \ldots, X_n] = \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \overline{X})^2}$$

- What drives the sampling variability of the OLS estimator?

Understanding the sampling variance

$$V[\widehat{\beta}_1 | X_1, \ldots, X_n] = \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \overline{X})^2}$$

- What drives the sampling variability of the OLS estimator?
  - The higher the variance of $Y_i | X_i$, the higher the sampling variance

Understanding the sampling variance

$$V[\widehat{\beta}_1|X_1,\ldots,X_n] = \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \overline{X})^2}$$

- What drives the sampling variability of the OLS estimator?
  - The higher the variance of $Y_i|X_i$, the higher the sampling variance
  - The lower the variance of X_i, the higher the sampling variance

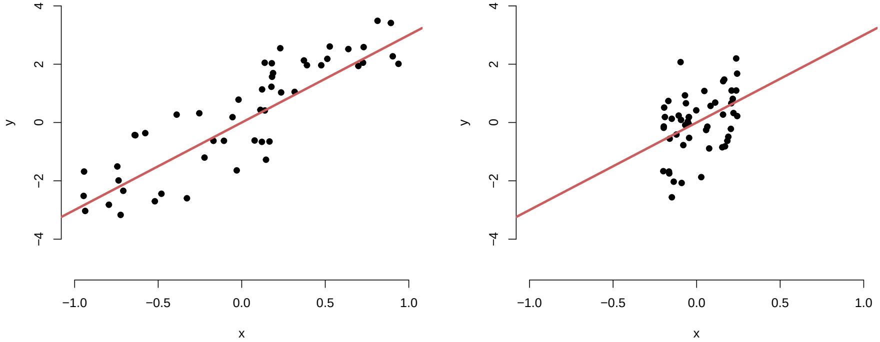
Understanding the sampling variance

$$V[\widehat{\beta}_1|X_1,\ldots,X_n] = \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \overline{X})^2}$$

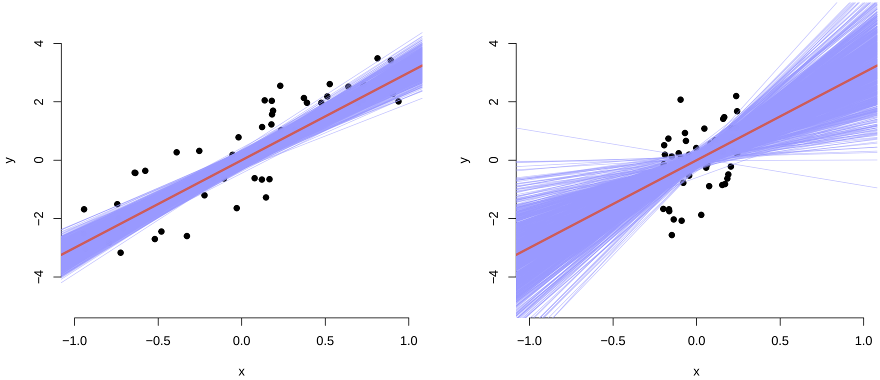
- What drives the sampling variability of the OLS estimator?
    - The higher the variance of $Y_i|X_i$, the higher the sampling variance
    - The lower the variance of X_i, the higher the sampling variance
    - As we increase n, the denominator gets large, while the numerator is fixed and so the sampling variance shrinks to 0.

Variance in X Reduces Standard Errors

Variance in X Reduces Standard Errors



Variance in X Reduces Standard Errors



Estimating the Variance of OLS Estimators

How can we estimate the unobserved error variance $Var[u] = \sigma_u^2$?

Estimating the Variance of OLS Estimators

How can we estimate the unobserved error variance $Var[u] = \sigma_u^2$?
We can derive an estimator based on the residuals:

$$\hat{u}_i \ = \ y_i - \hat{y}_i \ = \ y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

Recall: The errors u_i are NOT the same as the residuals $\hat{u}_i$.

Estimating the Variance of OLS Estimators

How can we estimate the unobserved error variance $Var[u] = \sigma_u^2$?
We can derive an estimator based on the residuals:

$$\hat{u}_i \ = \ y_i - \hat{y}_i \ = \ y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

Recall: The errors u_i are NOT the same as the residuals $\hat{u}_i$.

Intuitively, the scatter of the residuals around the fitted regression line should reflect the unseen scatter about the true population regression line.

We can measure scatter with the mean squared deviation:

$$MSD(\hat{u}) \equiv \frac{1}{n}\sum_{i=1}^{n}(\hat{u}_i - \bar{\hat{u}})^2 =$$

Estimating the Variance of OLS Estimators

How can we estimate the unobserved error variance $Var[u] = \sigma_u^2$?
We can derive an estimator based on the residuals:

$$\hat{u}_i \; = \; y_i - \hat{y}_i \; = \; y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

Recall: The errors u_i are NOT the same as the residuals $\hat{u}_i$.

Intuitively, the scatter of the residuals around the fitted regression line should reflect the unseen scatter about the true population regression line.

We can measure scatter with the mean squared deviation:

$$MSD(\hat{u}) \equiv \frac{1}{n}\sum_{i=1}^{n}(\hat{u}_i - \bar{\hat{u}})^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2$$

Estimating the Variance of OLS Estimators

- By construction, the regression line is closer since it is drawn to fit the sample we observe

Estimating the Variance of OLS Estimators

- By construction, the regression line is closer since it is drawn to fit the sample we observe
- Specifically, the regression line is drawn so as to minimize the sum of the squares of the distances between it and the observations

Estimating the Variance of OLS Estimators

- By construction, the regression line is closer since it is drawn to fit the sample we observe
- Specifically, the regression line is drawn so as to minimize the sum of the squares of the distances between it and the observations
- So the spread of the residuals $MSD(\hat{u})$ will slightly *underestimate* the error variance $\text{Var}[u] = \sigma_u^2$ on average

Estimating the Variance of OLS Estimators

- By construction, the regression line is closer since it is drawn to fit the sample we observe
- Specifically, the regression line is drawn so as to minimize the sum of the squares of the distances between it and the observations
- So the spread of the residuals $MSD(\hat{u})$ will slightly *underestimate* the error variance $\text{Var}[u] = \sigma_u^2$ on average
- In fact, we can show that with a single regressor X we have:

$$E[MSD(\hat{u})] = \frac{n-2}{n}\sigma_u^2 \text{ (degrees of freedom adjustment)}$$

Estimating the Variance of OLS Estimators

- By construction, the regression line is closer since it is drawn to fit the sample we observe
- Specifically, the regression line is drawn so as to minimize the sum of the squares of the distances between it and the observations
- So the spread of the residuals $MSD(\hat{u})$ will slightly *underestimate* the error variance $\text{Var}[u] = \sigma_u^2$ on average
- In fact, we can show that with a single regressor X we have:

$$E[MSD(\hat{u})] = \frac{n-2}{n}\sigma_u^2 \text{ (degrees of freedom adjustment)}$$

- Thus, an unbiased estimator for the error variance is:

$$\hat{\sigma}_u^2 = \frac{n}{n-2} MSD(\hat{u}) = \frac{n}{n-2}\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i = \frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2$$

We plug this estimate into the variance estimators for $\hat{\beta}_0$ and $\hat{\beta}_1$.

Where are we?

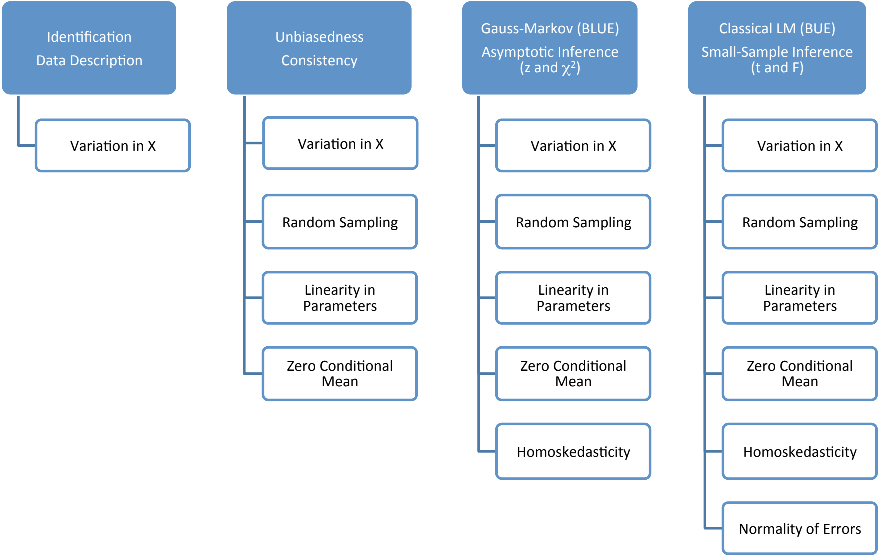
- Under Assumptions 1-5, we know that

$$\widehat{\beta}_1 \sim ?\left(\beta_1, \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right)$$

- Now we know the mean and sampling variance of the sampling distribution.
- How does this compare to other estimators for the population slope?

Where are we?

Where are we?


Identification
Data Description
Variation in X
Unbiasedness
Consistency
Variation in X
Random Sampling
Linearity in Parameters
Zero Conditional Mean
Gauss-Markov (BLUE)
Asymptotic Inference (z and χ^2)
Variation in X
Random Sampling
Linearity in Parameters
Zero Conditional Mean
Homoskedasticity
Classical LM (BUE)
Small-Sample Inference (t and F)
Variation in X
Random Sampling
Linearity in Parameters
Zero Conditional Mean
Homoskedasticity
Normality of Errors

OLS is BLUE :(

OLS is BLUE :(

Theorem (Gauss-Markov)

Given OLS Assumptions I–V, the OLS estimator is BLUE, i.e. the

1. *Best: Lowest variance in class*
2. *Linear: Among Linear estimators*
3. *Unbiased: Among Linear Unbiased estimators*
4. *Estimator.*

OLS is BLUE :(

Theorem (Gauss-Markov)

Given OLS Assumptions I–V, the OLS estimator is BLUE, i.e. the

1. *Best: Lowest variance in class*
2. *Linear: Among Linear estimators*
3. *Unbiased: Among Linear Unbiased estimators*
4. *Estimator.*

- A linear estimator is one that can be written as $\hat{\beta} = \mathbf{W}y$

OLS is BLUE :(

Theorem (Gauss-Markov)

Given OLS Assumptions I–V, the OLS estimator is BLUE, i.e. the

1. *Best: Lowest variance in class*
2. *Linear: Among Linear estimators*
3. *Unbiased: Among Linear Unbiased estimators*
4. *Estimator.*

- A linear estimator is one that can be written as $\hat{\beta} = \mathbf{W}y$
- Assumptions 1-5 are called the "Gauss Markov Assumptions"

OLS is BLUE :(

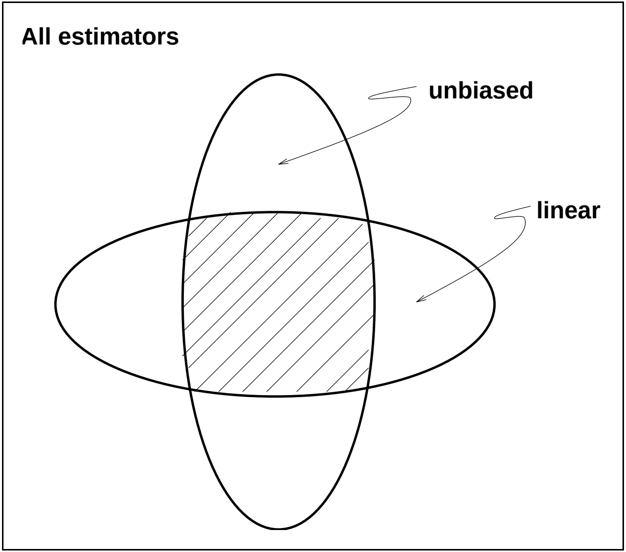
Theorem (Gauss-Markov)

Given OLS Assumptions I–V, the OLS estimator is BLUE, i.e. the

1. *Best: Lowest variance in class*
2. *Linear: Among Linear estimators*
3. *Unbiased: Among Linear Unbiased estimators*
4. *Estimator.*

- A linear estimator is one that can be written as $\hat{\beta} = \mathbf{W}y$
- Assumptions 1-5 are called the "Gauss Markov Assumptions"
- Result fails to hold when the assumptions are violated!

Gauss-Markov Theorem



Where are we?

- Under Assumptions 1-5, we know that

$$\widehat{\beta}_1 \sim ?\left(\beta_1, \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right)$$

Where are we?

- Under Assumptions 1-5, we know that

$$\widehat{\beta}_1 \sim ?\left(\beta_1, \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right)$$

- And we know that $\frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \overline{X})^2}$ is the lowest variance of any linear estimator of β_1

Where are we?

- Under Assumptions 1-5, we know that

$$\widehat{\beta}_1 \sim ? \left(\beta_1, \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right)$$

- And we know that $\frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \overline{X})^2}$ is the lowest variance of any linear estimator of β_1
- What about the last question mark? What's the form of the distribution?

Large-sample distribution of OLS estimators

- Remember that the OLS estimator is the sum of independent r.v.'s:

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

Large-sample distribution of OLS estimators

- Remember that the OLS estimator is the sum of independent r.v.'s:

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

- Mantra of the Central Limit Theorem:

Large-sample distribution of OLS estimators

- Remember that the OLS estimator is the sum of independent r.v.'s:

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

- Mantra of the Central Limit Theorem:

Large-sample distribution of OLS estimators

- Remember that the OLS estimator is the sum of independent r.v.'s:

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

- Mantra of the Central Limit Theorem:
  "the sums and means of random variables tend to be Normally distributed in large samples."

Large-sample distribution of OLS estimators

- Remember that the OLS estimator is the sum of independent r.v.'s:

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

- Mantra of the Central Limit Theorem:
  "the sums and means of random variables tend to be Normally distributed in large samples."

- True here as well, so we know that in large samples:

$$\frac{\widehat{\beta}_1 - \beta_1}{SE[\widehat{\beta}_1]} \sim N(0, 1)$$

Large-sample distribution of OLS estimators

- Remember that the OLS estimator is the sum of independent r.v.'s:

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

- Mantra of the Central Limit Theorem:
  "the sums and means of random variables tend to be Normally distributed in large samples."

- True here as well, so we know that in large samples:

$$\frac{\widehat{\beta}_1 - \beta_1}{SE[\widehat{\beta}_1]} \sim N(0,1)$$

- Can also replace SE with an estimate:

$$\frac{\widehat{\beta}_1 - \beta_1}{\widehat{SE}[\widehat{\beta}_1]} \sim N(0,1)$$

Where are we?

Under Assumptions 1-5 and in large samples, we know that

$$\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right)$$

Where are we?

Under Assumptions 1-5 and in large samples, we know that

$$\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right)$$

Sampling distribution in small samples

- What if we have a small sample? What can we do then?

Sampling distribution in small samples

- What if we have a small sample? What can we do then?
- Can't get something for nothing, but we can make progress if we make another assumption:

Sampling distribution in small samples

- What if we have a small sample? What can we do then?
- Can't get something for nothing, but we can make progress if we make another assumption:

1. Linearity
2. Random (iid) sample
3. Variation in X_i
4. Zero conditional mean of the errors
5. Homoskedasticity

Sampling distribution in small samples

- What if we have a small sample? What can we do then?
- Can't get something for nothing, but we can make progress if we make another assumption:

1. Linearity
2. Random (iid) sample
3. Variation in X_i
4. Zero conditional mean of the errors
5. Homoskedasticity
6. Errors are conditionally Normal

OLS Assumptions VI

OLS Assumptions VI

Assumption (VI. Normality)

The population error term is independent of the explanatory variable, $u \perp\!\!\!\perp X$, and is normally distributed with mean zero and variance σ_u^2:

$$u \sim N(0, \sigma_u^2), \quad \textit{which implies} \quad Y|X \sim N(\beta_0 + \beta_1 X, \sigma_u^2)$$

Note: This also implies homoskedasticity and zero conditional mean.

OLS Assumptions VI

Assumption (VI. Normality)

The population error term is independent of the explanatory variable, $u \perp\!\!\!\perp X$, and is normally distributed with mean zero and variance σ_u^2:

$$u \sim N(0, \sigma_u^2), \quad \textit{which implies} \quad Y|X \sim N(\beta_0 + \beta_1 X, \sigma_u^2)$$

Note: This also implies homoskedasticity and zero conditional mean.

- Together Assumptions I–VI are the classical linear model (CLM) assumptions.

OLS Assumptions VI

Assumption (VI. Normality)

The population error term is independent of the explanatory variable, $u \perp\!\!\!\perp X$, and is normally distributed with mean zero and variance σ_u^2:

$$u \sim N(0, \sigma_u^2), \quad \textit{which implies} \quad Y|X \sim N(\beta_0 + \beta_1 X, \sigma_u^2)$$

Note: This also implies homoskedasticity and zero conditional mean.

- Together Assumptions I–VI are the classical linear model (CLM) assumptions.
- The CLM assumptions imply that OLS is BUE (i.e. minimum variance among all linear or non-linear unbiased estimators)

OLS Assumptions VI

Assumption (VI. Normality)

The population error term is independent of the explanatory variable, $u \perp\!\!\!\perp X$, and is normally distributed with mean zero and variance σ_u^2:

$$u \sim N(0, \sigma_u^2), \quad \textit{which implies} \quad Y|X \sim N(\beta_0 + \beta_1 X, \sigma_u^2)$$

Note: This also implies homoskedasticity and zero conditional mean.

- Together Assumptions I–VI are the classical linear model (CLM) assumptions.
- The CLM assumptions imply that OLS is BUE (i.e. minimum variance among all linear or non-linear unbiased estimators)
- Non-normality of the errors is a serious concern in small samples. We can *partially* check this assumption by looking at the residuals (more in coming weeks)

OLS Assumptions VI

Assumption (VI. Normality)

The population error term is independent of the explanatory variable, $u \perp\!\!\!\perp X$, and is normally distributed with mean zero and variance σ_u^2:

$$u \sim N(0, \sigma_u^2), \quad \textit{which implies} \quad Y|X \sim N(\beta_0 + \beta_1 X, \sigma_u^2)$$

Note: This also implies homoskedasticity and zero conditional mean.

- Together Assumptions I–VI are the classical linear model (CLM) assumptions.
- The CLM assumptions imply that OLS is BUE (i.e. minimum variance among all linear or non-linear unbiased estimators)
- Non-normality of the errors is a serious concern in small samples. We can *partially* check this assumption by looking at the residuals (more in coming weeks)
- Variable transformations can help to come closer to normality

OLS Assumptions VI

Assumption (VI. Normality)

The population error term is independent of the explanatory variable, $u \perp\!\!\!\perp X$, and is normally distributed with mean zero and variance σ_u^2:

$$u \sim N(0, \sigma_u^2), \quad \textit{which implies} \quad Y|X \sim N(\beta_0 + \beta_1 X, \sigma_u^2)$$

Note: This also implies homoskedasticity and zero conditional mean.

- Together Assumptions I–VI are the classical linear model (CLM) assumptions.
- The CLM assumptions imply that OLS is BUE (i.e. minimum variance among all linear or non-linear unbiased estimators)
- Non-normality of the errors is a serious concern in small samples. We can *partially* check this assumption by looking at the residuals (more in coming weeks)
- Variable transformations can help to come closer to normality
- Reminder: we don't need normality assumption in large samples

Sampling distribution of OLS slope

- If we have Y_i given X_i is distributed $N(\beta_0 + \beta_1 X_i, \sigma_u^2)$, then we have the following at any sample size:

$$\frac{\widehat{\beta}_1 - \beta_1}{SE[\widehat{\beta}_1]} \sim N(0, 1)$$

Sampling distribution of OLS slope

- If we have Y_i given X_i is distributed $N(\beta_0 + \beta_1 X_i, \sigma_u^2)$, then we have the following at any sample size:

$$\frac{\widehat{\beta}_1 - \beta_1}{SE[\widehat{\beta}_1]} \sim N(0,1)$$

- Furthermore, if we replace the true standard error with the estimated standard error, then we get the following:

$$\frac{\widehat{\beta}_1 - \beta_1}{\widehat{SE}[\widehat{\beta}_1]} \sim t_{n-2}$$

Sampling distribution of OLS slope

- If we have Y_i given X_i is distributed $N(\beta_0 + \beta_1 X_i, \sigma_u^2)$, then we have the following at any sample size:

$$\frac{\widehat{\beta}_1 - \beta_1}{SE[\widehat{\beta}_1]} \sim N(0,1)$$

- Furthermore, if we replace the true standard error with the estimated standard error, then we get the following:

$$\frac{\widehat{\beta}_1 - \beta_1}{\widehat{SE}[\widehat{\beta}_1]} \sim t_{n-2}$$

- The standardized coefficient follows a t distribution $n-2$ degrees of freedom. We take off an extra degree of freedom because we had to estimate one more parameter than just the sample mean.

Sampling distribution of OLS slope

- If we have Y_i given X_i is distributed $N(\beta_0 + \beta_1 X_i, \sigma_u^2)$, then we have the following at any sample size:

$$\frac{\widehat{\beta}_1 - \beta_1}{SE[\widehat{\beta}_1]} \sim N(0,1)$$

- Furthermore, if we replace the true standard error with the estimated standard error, then we get the following:

$$\frac{\widehat{\beta}_1 - \beta_1}{\widehat{SE}[\widehat{\beta}_1]} \sim t_{n-2}$$

- The standardized coefficient follows a t distribution $n-2$ degrees of freedom. We take off an extra degree of freedom because we had to estimate one more parameter than just the sample mean.
- All of this depends on Normal errors!

Where are we?

Where are we?

- Under Assumptions 1-5 and in large samples, we know that

$$\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right)$$

Where are we?

- Under Assumptions 1-5 and in large samples, we know that

$$\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right)$$

- Under Assumptions 1-6 and in any sample, we know that

$$\frac{\widehat{\beta}_1 - \beta_1}{\widehat{SE}[\widehat{\beta}_1]} \sim t_{n-2}$$

Hierarchy of OLS Assumptions

Hierarchy of OLS Assumptions

Identification
Data Description
- Variation in X

Unbiasedness
Consistency
- Variation in X
- Random Sampling
- Linearity in Parameters
- Zero Conditional Mean

Gauss-Markov (BLUE)
Asymptotic Inference (z and χ^2)
- Variation in X
- Random Sampling
- Linearity in Parameters
- Zero Conditional Mean
- Homoskedasticity

Classical LM (BUE)
Small-Sample Inference (t and F)
- Variation in X
- Random Sampling
- Linearity in Parameters
- Zero Conditional Mean
- Homoskedasticity
- Normality of Errors

Regression as parametric modeling

Let's summarize the parametric view we have taken thus far.

Regression as parametric modeling

Let's summarize the parametric view we have taken thus far.

- Gauss-Markov assumptions:
  - (A1) linearity, (A2) i.i.d. sample, (A3) variation in X, (A4) zero conditional mean error, (A5) homoskedasticity.

Regression as parametric modeling

Let's summarize the parametric view we have taken thus far.

- Gauss-Markov assumptions:
  - (A1) linearity, (A2) i.i.d. sample, (A3) variation in X, (A4) zero conditional mean error, (A5) homoskedasticity.
  - basically, assume the model is right

Regression as parametric modeling

Let's summarize the parametric view we have taken thus far.

- Gauss-Markov assumptions:
  - (A1) linearity, (A2) i.i.d. sample, (A3) variation in X, (A4) zero conditional mean error, (A5) homoskedasticity.
  - basically, assume the model is right
- $\rightsquigarrow$ OLS is BLUE, plus (A6) normality of the errors and we get small sample SEs and BUE.

Regression as parametric modeling

Let's summarize the parametric view we have taken thus far.

- Gauss-Markov assumptions:
  - (A1) linearity, (A2) i.i.d. sample, (A3) variation in X, (A4) zero conditional mean error, (A5) homoskedasticity.
  - basically, assume the model is right
- $\rightsquigarrow$ OLS is BLUE, plus (A6) normality of the errors and we get small sample SEs and BUE.
- What is the basic approach here?

Regression as parametric modeling

Let's summarize the parametric view we have taken thus far.

- Gauss-Markov assumptions:
  - (A1) linearity, (A2) i.i.d. sample, (A3) variation in X, (A4) zero conditional mean error, (A5) homoskedasticity.
  - basically, assume the model is right
- $\rightsquigarrow$ OLS is BLUE, plus (A6) normality of the errors and we get small sample SEs and BUE.
- What is the basic approach here?
  - A1 defines a linear model for the outcome:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

Regression as parametric modeling

Let's summarize the parametric view we have taken thus far.

- Gauss-Markov assumptions:
  - (A1) linearity, (A2) i.i.d. sample, (A3) variation in X, (A4) zero conditional mean error, (A5) homoskedasticity.
  - basically, assume the model is right
- $\rightsquigarrow$ OLS is BLUE, plus (A6) normality of the errors and we get small sample SEs and BUE.
- What is the basic approach here?
  - A1 defines a linear model for the outcome:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

  - A2 and A4 let us write the CEF as function of X_i alone.

$$E[Y_i|X_i] = \mu_i = \beta_0 + \beta_1 X_i$$

Regression as parametric modeling

Let's summarize the parametric view we have taken thus far.

- Gauss-Markov assumptions:
  - (A1) linearity, (A2) i.i.d. sample, (A3) variation in X, (A4) zero conditional mean error, (A5) homoskedasticity.
  - basically, assume the model is right
- $\rightsquigarrow$ OLS is BLUE, plus (A6) normality of the errors and we get small sample SEs and BUE.
- What is the basic approach here?
  - A1 defines a linear model for the outcome:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

  - A2 and A4 let us write the CEF as function of X_i alone.

$$E[Y_i|X_i] = \mu_i = \beta_0 + \beta_1 X_i$$

  - A5-6, define a probabilistic model for the conditional distribution:

$$Y_i|X_i \sim \mathcal{N}(\mu_i, \sigma^2)$$

Regression as parametric modeling

Let's summarize the parametric view we have taken thus far.

- Gauss-Markov assumptions:
  - (A1) linearity, (A2) i.i.d. sample, (A3) variation in X, (A4) zero conditional mean error, (A5) homoskedasticity.
  - basically, assume the model is right
- $\rightsquigarrow$ OLS is BLUE, plus (A6) normality of the errors and we get small sample SEs and BUE.
- What is the basic approach here?
  - A1 defines a linear model for the outcome:

    $$Y_i = \beta_0 + \beta_1 X_i + u_i$$

  - A2 and A4 let us write the CEF as function of X_i alone.

    $$E[Y_i|X_i] = \mu_i = \beta_0 + \beta_1 X_i$$

  - A5-6, define a probabilistic model for the conditional distribution:

    $$Y_i|X_i \sim \mathcal{N}(\mu_i, \sigma^2)$$

  - A3 covers the edge-case that the β can't be estimated due to 0 variance in X.

Agnostic views on regression

- These assumptions assume we know a lot about how Y is 'generated'.

Agnostic views on regression

- These assumptions assume we know a lot about how Y is ‘generated’.
- Justifications for using OLS (like BLUE/BUE) often invoke these assumptions which are unlikely to hold exactly.

Agnostic views on regression

- These assumptions assume we know a lot about how Y is 'generated'.
- Justifications for using OLS (like BLUE/BUE) often invoke these assumptions which are unlikely to hold exactly.
- Alternative: take an agnostic view on regression.

Agnostic views on regression

- These assumptions assume we know a lot about how Y is 'generated'.
- Justifications for using OLS (like BLUE/BUE) often invoke these assumptions which are unlikely to hold exactly.
- Alternative: take an agnostic view on regression.
  - use OLS without believing these assumptions.

Agnostic views on regression

- These assumptions assume we know a lot about how Y is 'generated'.
- Justifications for using OLS (like BLUE/BUE) often invoke these assumptions which are unlikely to hold exactly.
- Alternative: take an agnostic view on regression.
  - use OLS without believing these assumptions.
  - lean on two things: A2 i.i.d. sample, asymptotics (large-sample properties)

Agnostic views on regression

- These assumptions assume we know a lot about how Y is 'generated'.
- Justifications for using OLS (like BLUE/BUE) often invoke these assumptions which are unlikely to hold exactly.
- Alternative: take an agnostic view on regression.
  - use OLS without believing these assumptions.
  - lean on two things: A2 i.i.d. sample, asymptotics (large-sample properties)
- Lose the distributional assumptions and focus on approximating the best linear predictor.

Agnostic views on regression

- These assumptions assume we know a lot about how Y is 'generated'.
- Justifications for using OLS (like BLUE/BUE) often invoke these assumptions which are unlikely to hold exactly.
- Alternative: take an agnostic view on regression.
  - use OLS without believing these assumptions.
  - lean on two things: A2 i.i.d. sample, asymptotics (large-sample properties)
- Lose the distributional assumptions and focus on approximating the best linear predictor.
- If the true CEF happens to be linear, the best linear predictor is it.

Agnostic views on regression

- These assumptions assume we know a lot about how Y is 'generated'.
- Justifications for using OLS (like BLUE/BUE) often invoke these assumptions which are unlikely to hold exactly.
- Alternative: take an agnostic view on regression.
  - use OLS without believing these assumptions.
  - lean on two things: A2 i.i.d. sample, asymptotics (large-sample properties)
- Lose the distributional assumptions and focus on approximating the best linear predictor.
- If the true CEF happens to be linear, the best linear predictor is it.
- Again, we will come back to this in a couple of weeks.

Where are we?

Where are we?

- Under Assumptions 1-5 and in large samples, we know that

$$\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right)$$

- Under Assumptions 1-6 and in any sample, we know that

$$\frac{\widehat{\beta}_1 - \beta_1}{\widehat{SE}[\widehat{\beta}_1]} \sim t_{n-2}$$

Null and alternative hypotheses review

- Null: $H_0 : \beta_1 = 0$

Null and alternative hypotheses review

- Null: $H_0 : \beta_1 = 0$
  - The null is the straw man we want to knock down.

Null and alternative hypotheses review

- Null: $H_0 : \beta_1 = 0$
  - The null is the straw man we want to knock down.
  - With regression, almost always null of no relationship

Null and alternative hypotheses review

- Null: $H_0 : \beta_1 = 0$
  - The null is the straw man we want to knock down.
  - With regression, almost always null of no relationship
- Alternative: $H_a : \beta_1 \neq 0$

Null and alternative hypotheses review

- Null: $H_0 : \beta_1 = 0$
  - The null is the straw man we want to knock down.
  - With regression, almost always null of no relationship
- Alternative: $H_a : \beta_1 \neq 0$
  - Claim we want to test

Null and alternative hypotheses review

- Null: $H_0 : \beta_1 = 0$
  - The null is the straw man we want to knock down.
  - With regression, almost always null of no relationship
- Alternative: $H_a : \beta_1 \neq 0$
  - Claim we want to test
  - Could do one-sided test, but you shouldn't

Null and alternative hypotheses review

- Null: $H_0 : \beta_1 = 0$
  - The null is the straw man we want to knock down.
  - With regression, almost always null of no relationship
- Alternative: $H_a : \beta_1 \neq 0$
  - Claim we want to test
  - Could do one-sided test, but you shouldn't
- Notice these are statements about the population parameters, not the OLS estimates.

Test statistic

- Under the null of $H_0 : \beta_1 = c$, we can use the following familiar test statistic:

$$T = \frac{\widehat{\beta}_1 - c}{\widehat{SE}[\widehat{\beta}_1]}$$

Test statistic

- Under the null of $H_0 : \beta_1 = c$, we can use the following familiar test statistic:
$$T = \frac{\widehat{\beta}_1 - c}{\widehat{SE}[\widehat{\beta}_1]}$$
- Under the null hypothesis:

Test statistic

- Under the null of $H_0 : \beta_1 = c$, we can use the following familiar test statistic:
$$T = \frac{\widehat{\beta}_1 - c}{\widehat{SE}[\widehat{\beta}_1]}$$
- Under the null hypothesis:
  - large samples: $T \sim \mathcal{N}(0, 1)$

Test statistic

- Under the null of $H_0 : \beta_1 = c$, we can use the following familiar test statistic:
$$T = \frac{\widehat{\beta}_1 - c}{\widehat{SE}[\widehat{\beta}_1]}$$
- Under the null hypothesis:
  - large samples: $T \sim \mathcal{N}(0, 1)$
  - any size sample with normal errors: $T \sim t_{n-2}$

Test statistic

- Under the null of $H_0 : \beta_1 = c$, we can use the following familiar test statistic:

$$T = \frac{\widehat{\beta}_1 - c}{\widehat{SE}[\widehat{\beta}_1]}$$

- Under the null hypothesis:
  - large samples: $T \sim \mathcal{N}(0, 1)$
  - any size sample with normal errors: $T \sim t_{n-2}$
  - conservative to use t_{n-2} anyways since t_{n-2} is approximately normal in large samples.

Test statistic

- Under the null of $H_0 : \beta_1 = c$, we can use the following familiar test statistic:
$$T = \frac{\widehat{\beta}_1 - c}{\widehat{SE}[\widehat{\beta}_1]}$$
- Under the null hypothesis:
  - large samples: $T \sim \mathcal{N}(0, 1)$
  - any size sample with normal errors: $T \sim t_{n-2}$
  - conservative to use t_{n-2} anyways since t_{n-2} is approximately normal in large samples.
- Thus, under the null, we know the distribution of T and can use that to formulate a rejection region and calculate p-values.

Test statistic

- Under the null of $H_0 : \beta_1 = c$, we can use the following familiar test statistic:
$$T = \frac{\widehat{\beta}_1 - c}{\widehat{SE}[\widehat{\beta}_1]}$$
- Under the null hypothesis:
  - large samples: $T \sim \mathcal{N}(0, 1)$
  - any size sample with normal errors: $T \sim t_{n-2}$
  - conservative to use t_{n-2} anyways since t_{n-2} is approximately normal in large samples.
- Thus, under the null, we know the distribution of T and can use that to formulate a rejection region and calculate p-values.
- By default, R shows you the test statistic for $\beta_1 = 0$ and uses the t distribution.

Rejection region

- Choose a level of the test, α, and find rejection regions that correspond to that value under the null distribution:

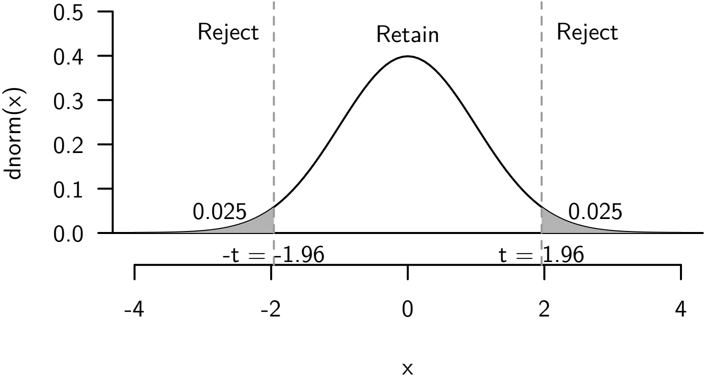
$$\mathbb{P}(-t_{\alpha/2,n-2} < T < t_{\alpha/2,n-2}) = 1 - \alpha$$

Rejection region

- Choose a level of the test, α, and find rejection regions that correspond to that value under the null distribution:

$$\mathbb{P}(-t_{\alpha/2,n-2} < T < t_{\alpha/2,n-2}) = 1 - \alpha$$

- This is exactly the same as with sample means and sample differences in means, except that the degrees of freedom on the t distribution have changed.



p-value

- The interpretation of the p-value is the same: the probability of seeing a test statistic at least this extreme if the null hypothesis were true

p-value

- The interpretation of the p-value is the same: the probability of seeing a test statistic at least this extreme if the null hypothesis were true
- Mathematically:

$$\mathbb{P}\left(\left|\frac{\widehat{\beta}_1 - c}{\widehat{SE}[\widehat{\beta}_1]}\right| \geq |T_{obs}|\right)$$

p-value

- The interpretation of the p-value is the same: the probability of seeing a test statistic at least this extreme if the null hypothesis were true
- Mathematically:

$$\mathbb{P}\left(\left|\frac{\widehat{\beta}_1 - c}{\widehat{SE}[\widehat{\beta}_1]}\right| \geq |T_{obs}|\right)$$

- If the p-value is less than α we would reject the null at the α level.

Confidence intervals

- Very similar to the approach with sample means. By the sampling distribution of the OLS estimator, we know that we can find t-values such that:

$$\mathbb{P}\left(-t_{\alpha/2,n-2} \leq \frac{\widehat{\beta}_1 - \beta_1}{\widehat{SE}[\widehat{\beta}_1]} \leq t_{\alpha/2,n-2}\right) = 1 - \alpha$$

Confidence intervals

- Very similar to the approach with sample means. By the sampling distribution of the OLS estimator, we know that we can find t-values such that:

$$\mathbb{P}\Big(- t_{\alpha/2,n-2} \leq \frac{\widehat{\beta}_1 - \beta_1}{\widehat{SE}[\widehat{\beta}_1]} \leq t_{\alpha/2,n-2}\Big) = 1 - \alpha$$

- If we rearrange this as before, we can get an expression for confidence intervals:

$$\mathbb{P}\Big(\widehat{\beta}_1 - t_{\alpha/2,n-2}\widehat{SE}[\widehat{\beta}_1] \leq \beta_1 \leq \widehat{\beta}_1 + t_{\alpha/2,n-2}\widehat{SE}[\widehat{\beta}_1]\Big) = 1 - \alpha$$

Confidence intervals

- Very similar to the approach with sample means. By the sampling distribution of the OLS estimator, we know that we can find t-values such that:

$$\mathbb{P}\left(- t_{\alpha/2,n-2} \leq \frac{\widehat{\beta}_1 - \beta_1}{\widehat{SE}[\widehat{\beta}_1]} \leq t_{\alpha/2,n-2} \right) = 1 - \alpha$$

- If we rearrange this as before, we can get an expression for confidence intervals:

$$\mathbb{P}\left(\widehat{\beta}_1 - t_{\alpha/2,n-2}\widehat{SE}[\widehat{\beta}_1] \leq \beta_1 \leq \widehat{\beta}_1 + t_{\alpha/2,n-2}\widehat{SE}[\widehat{\beta}_1] \right) = 1 - \alpha$$

- Thus, we can write the confidence intervals as:

$$\widehat{\beta}_1 \pm t_{\alpha/2,n-2}\widehat{SE}[\widehat{\beta}_1]$$

Confidence intervals

- Very similar to the approach with sample means. By the sampling distribution of the OLS estimator, we know that we can find t-values such that:

$$\mathbb{P}\left(-t_{\alpha/2,n-2} \leq \frac{\widehat{\beta}_1 - \beta_1}{\widehat{SE}[\widehat{\beta}_1]} \leq t_{\alpha/2,n-2} \right) = 1 - \alpha$$

- If we rearrange this as before, we can get an expression for confidence intervals:

$$\mathbb{P}\left(\widehat{\beta}_1 - t_{\alpha/2,n-2}\widehat{SE}[\widehat{\beta}_1] \leq \beta_1 \leq \widehat{\beta}_1 + t_{\alpha/2,n-2}\widehat{SE}[\widehat{\beta}_1] \right) = 1 - \alpha$$

- Thus, we can write the confidence intervals as:

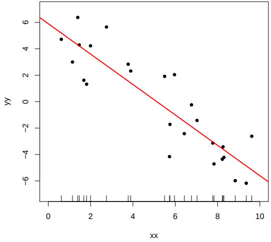
$$\widehat{\beta}_1 \pm t_{\alpha/2,n-2}\widehat{SE}[\widehat{\beta}_1]$$

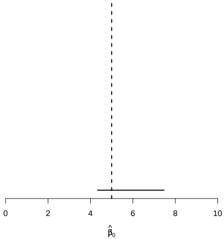
- We can derive these for the intercept as well:

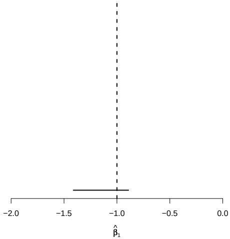
$$\widehat{\beta}_0 \pm t_{\alpha/2,n-2}\widehat{SE}[\widehat{\beta}_0]$$

CIs Simulation Example

Returning to our simulation example we can simulate the sampling distributions of the 95 % confidence interval estimates for $\widehat{\beta}_1$ and $\widehat{\beta}_0$

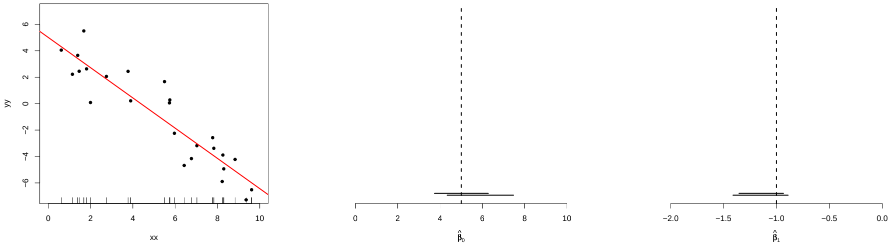




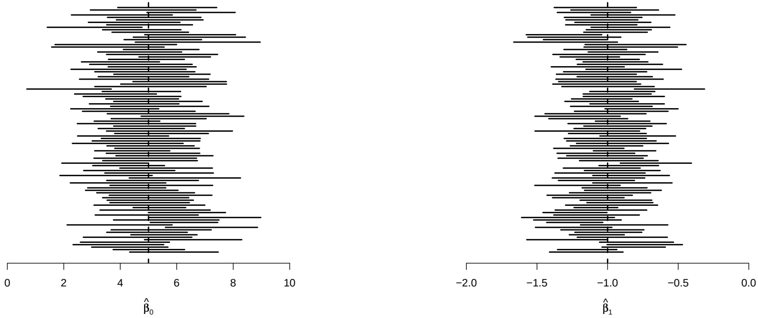


CIs Simulation Example

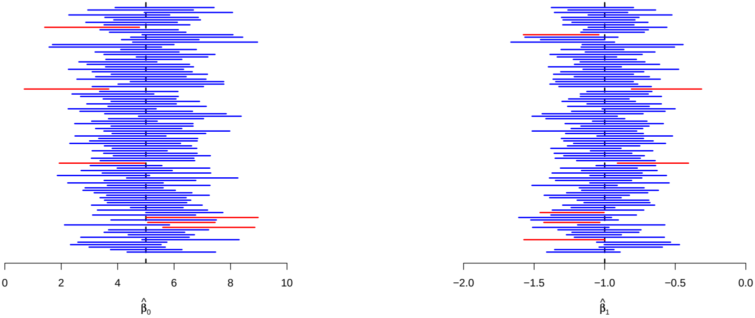
Returning to our simulation example we can simulate the sampling distributions of the 95 % confidence interval estimates for $\widehat{\beta}_1$ and $\widehat{\beta}_0$



CIs Simulation Example



CIs Simulation Example



Prediction error

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

Prediction error

- 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 error

- 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:

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

Prediction error

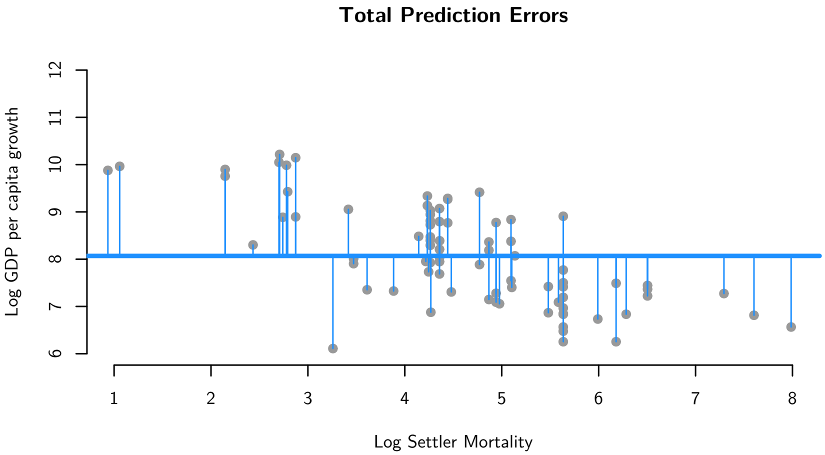
- How do we judge how well a line fits the data?
- One way is to find out how much better we do at predicting Y once we include X into the regression model.
- Prediction errors without X: best prediction is the mean, so our squared errors, or the total sum of squares (SS_{tot}) would be:

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

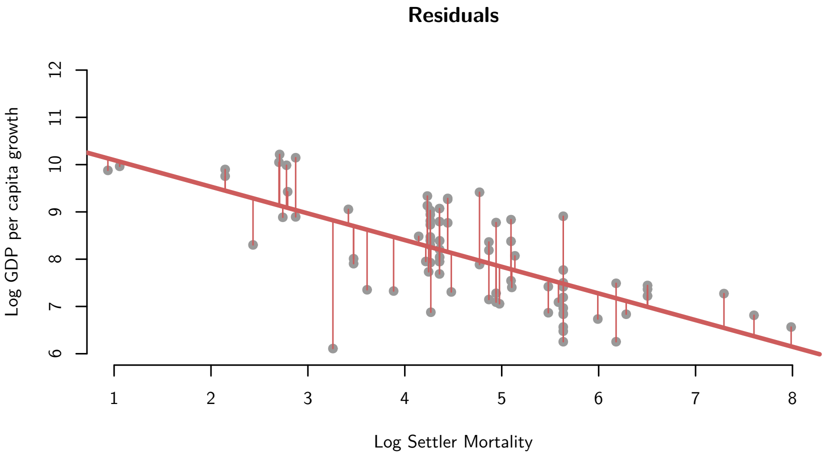
- Once we have estimated our model, we have new prediction errors, which are just the sum of the squared residuals or SS_{res}:

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

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

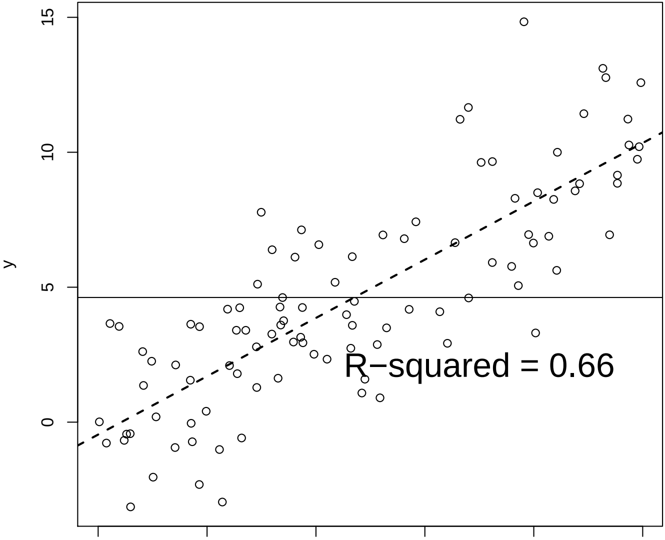
R-square

- 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:

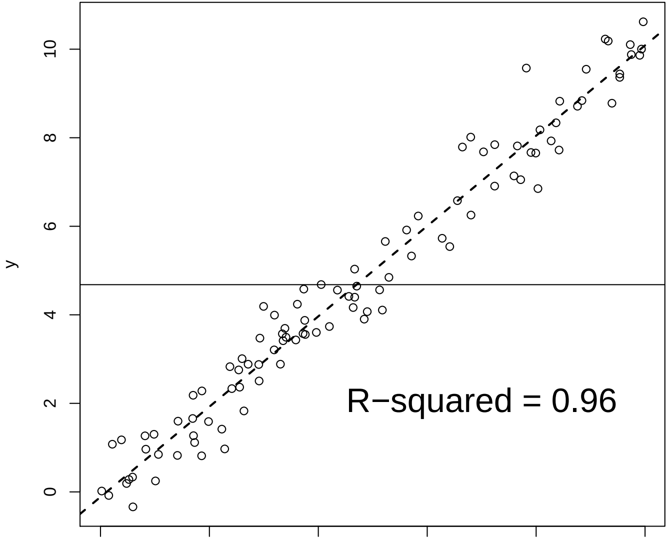
$$R^2 = \frac{SS_{tot} - SS_{res}}{SS_{tot}} = 1 - \frac{SS_{res}}{SS_{tot}}$$

- 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

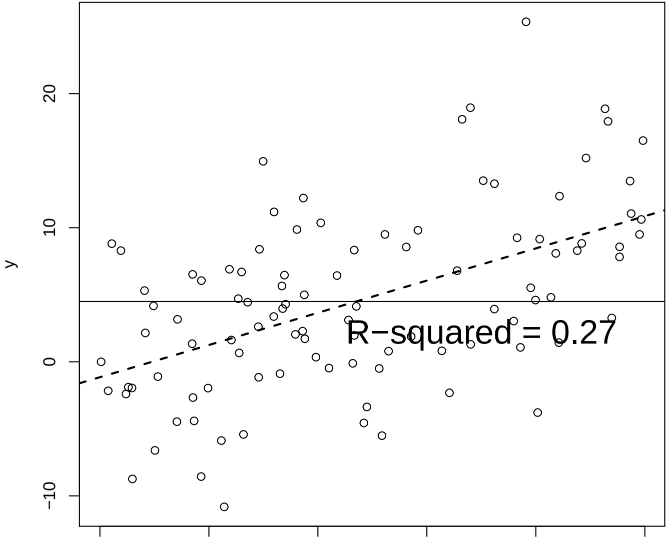
Is R-squared useful?



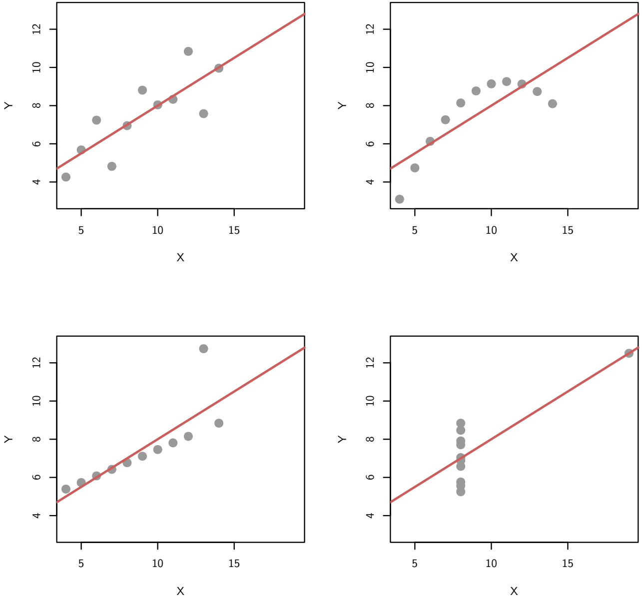
Is R-squared useful?



Is R-squared useful?

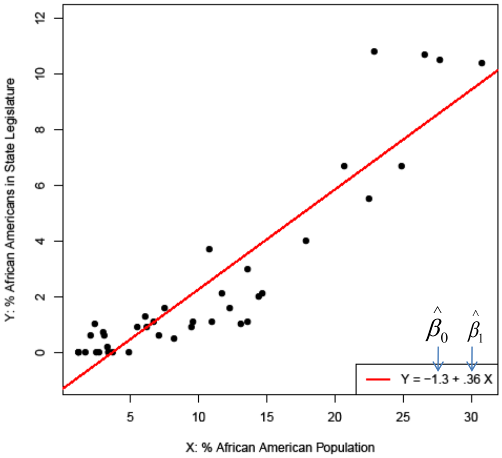


Is R-squared useful?



Interpreting a Regression

Let's have a quick chat about interpretation.



State Legislators and African American Population

Interpretations of increasing quality:

```
> summary(lm(beo ~ bpop, data = D))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.31489    0.32775  -4.012 0.000264 ***
bpop         0.35848    0.02519  14.232  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.317 on 39 degrees of freedom
Multiple R-squared: 0.8385,Adjusted R-squared: 0.8344
F-statistic: 202.6 on 1 and 39 DF,  p-value: < 2.2e-16
```

"African American population is statistically significant ($p < 0.001$)"

(no effect size or direction)

State Legislators and African American Population

Interpretations of increasing quality:

```
> summary(lm(beo ~ bpop, data = D))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.31489    0.32775  -4.012 0.000264 ***
bpop         0.35848    0.02519  14.232  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.317 on 39 degrees of freedom
Multiple R-squared: 0.8385,Adjusted R-squared: 0.8344
F-statistic: 202.6 on 1 and 39 DF,  p-value: < 2.2e-16
```

"Percent African American legislators increases with African American population ($p < 0.001$)"

(direction, but no effect size)

State Legislators and African American Population

Interpretations of increasing quality:

```
> summary(lm(beo ~ bpop, data = D))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.31489    0.32775  -4.012 0.000264 ***
bpop         0.35848    0.02519  14.232  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.317 on 39 degrees of freedom
Multiple R-squared: 0.8385,Adjusted R-squared: 0.8344
F-statistic: 202.6 on 1 and 39 DF,  p-value: < 2.2e-16
```

"A one percentage point increase in the African American population causes a 0.35 percentage point increase in the fraction of African American state legislators ($p < 0.001$)."

(unwarranted causal language)

State Legislators and African American Population

Interpretations of increasing quality:

```
> summary(lm(beo ~ bpop, data = D))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.31489    0.32775  -4.012 0.000264 ***
bpop         0.35848    0.02519  14.232  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.317 on 39 degrees of freedom
Multiple R-squared: 0.8385,Adjusted R-squared: 0.8344
F-statistic: 202.6 on 1 and 39 DF,  p-value: < 2.2e-16
```

"A one percentage point increase in the African American population is associated with a 0.35 percentage point increase in the fraction of African American state legislators ($p < 0.001$)."

(hints at causality)

State Legislators and African American Population

Interpretations of increasing quality:

```
> summary(lm(beo ~ bpop, data = D))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.31489    0.32775  -4.012 0.000264 ***
bpop         0.35848    0.02519  14.232  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.317 on 39 degrees of freedom
Multiple R-squared: 0.8385,Adjusted R-squared: 0.8344
F-statistic: 202.6 on 1 and 39 DF,  p-value: < 2.2e-16
```

"In states where an additional percentage point of the population is African American, we observe on average 0.35 percentage point more African American state legislators ($p < 0.001$)."

(p value doesn't help people with uncertainty)

State Legislators and African American Population

Interpretations of increasing quality:

```
> summary(lm(beo ~ bpop, data = D))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.31489    0.32775  -4.012 0.000264 ***
bpop         0.35848    0.02519  14.232  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.317 on 39 degrees of freedom
Multiple R-squared: 0.8385,Adjusted R-squared: 0.8344
F-statistic: 202.6 on 1 and 39 DF,  p-value: < 2.2e-16
```

"In states where an additional percentage point of the population is African American, we observe on average 0.35 percentage points more African American state legislators (between .35 and .40 with 95% confidence)."

(still not perfect, the best will be subject matter specific. is fairly clear it is non-causal, gives uncertainty.)

Ground Rules: Interpretation of the Slope

Ground Rules: Interpretation of the Slope

I almost didn't include the last example in the slides. It is hard to give ground rules that cover all cases. Regressions are a part of marshaling evidence in an argument which makes them naturally specific to context.

Ground Rules: Interpretation of the Slope

I almost didn't include the last example in the slides. It is hard to give ground rules that cover all cases. Regressions are a part of marshaling evidence in an argument which makes them naturally specific to context.

1. Give a short, but precise interpretation of the association using interpretable language and units

Ground Rules: Interpretation of the Slope

I almost didn't include the last example in the slides. It is hard to give ground rules that cover all cases. Regressions are a part of marshaling evidence in an argument which makes them naturally specific to context.

1. Give a short, but precise interpretation of the association using interpretable language and units
2. If the association has a causal interpretation explain why, otherwise do not imply a causal interpretation.

Ground Rules: Interpretation of the Slope

I almost didn't include the last example in the slides. It is hard to give ground rules that cover all cases. Regressions are a part of marshaling evidence in an argument which makes them naturally specific to context.

1. Give a short, but precise interpretation of the association using interpretable language and units
2. If the association has a causal interpretation explain why, otherwise do not imply a causal interpretation.
3. Provide a meaningful sense of uncertainty

Ground Rules: Interpretation of the Slope

I almost didn't include the last example in the slides. It is hard to give ground rules that cover all cases. Regressions are a part of marshaling evidence in an argument which makes them naturally specific to context.

1. Give a short, but precise interpretation of the association using interpretable language and units
2. If the association has a causal interpretation explain why, otherwise do not imply a causal interpretation.
3. Provide a meaningful sense of uncertainty
4. Indicate the practical significance of the finding for your argument.

Goal Check: Understand `lm()` Output

```
Call:
lm(formula = sr ~ pop15, data = LifeCycleSavings)

Residuals:
   Min     1Q Median     3Q    Max
-8.637 -2.374  0.349  2.022 11.155

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.49660    2.27972   7.675 6.85e-10 ***
pop15       -0.22302    0.06291  -3.545 0.000887 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.03 on 48 degrees of freedom
Multiple R-squared:  0.2075,Adjusted R-squared:  0.191
F-statistic: 12.57 on 1 and 48 DF,  p-value: 0.0008866
```

This Week in Review

- CEFs!
- OLS!
- Classical regression assumptions!
- Inference!

This Week in Review

- CEFs!
- OLS!
- Classical regression assumptions!
- Inference!

Going Deeper:

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

This Week in Review

- CEFs!
- OLS!
- Classical regression assumptions!
- Inference!

Going Deeper:

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

Next week: Linear Regression with Two Variables!