Week 8: What Can Go Wrong and How To Fix It, Diagnostics and Solutions

Brandon Stewart[1]

Princeton

October 31–November 2, 2022

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

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

- Last Week
  - multiple regression

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

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

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

- Last Week
  - multiple regression
- This Week
  - peeking under the hood of the linear models
  - non-linearity
  - unusual and influential data
  - clustered standard errors
- Next Week
  - frameworks for causal inference

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

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

Topics We Will Cover

Topics We Will Cover

- Non-linearity
- Non-Normal Errors
- Extreme Values
- Clustering

Topics We Will Cover

- Non-linearity
- Non-Normal Errors
- Extreme Values
- Clustering

Regrettably we won't have time to cover many topics that are important like missing data, sensitivity analysis, etc.

1. Non-Linearity
   - Polynomials
   - LOESS
   - Log Transformations
2. Non-Normality
   - Hat Matrix
   - Standardized and Studentized Residuals
   - QQ Plots
3. Extreme Values
   - Outliers
   - Leverage Points
   - Influence Points
4. Clustering

1. Non-Linearity
   - Polynomials
   - LOESS
   - Log Transformations
2. Non-Normality
   - Hat Matrix
   - Standardized and Studentized Residuals
   - QQ Plots
3. Extreme Values
   - Outliers
   - Leverage Points
   - Influence Points
4. Clustering

Polynomial Terms

Polynomial Terms

- Polynomial terms are a special case of the continuous variable interactions.

Polynomial Terms

- Polynomial terms are a special case of the continuous variable interactions.
- For example, when $X_1 = X_2$ in the interaction model, we get a quadratic:

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + u \\
Y &= \beta_0 + (\beta_1 + \beta_2) X_1 + \beta_3 X_1 X_1 + u \\
Y &= \beta_0 + \tilde{\beta}_1 X_1 + \tilde{\beta}_2 X_1^2 + u
\end{aligned}
$$

- This is called a second order polynomial in X_1

Polynomial Terms

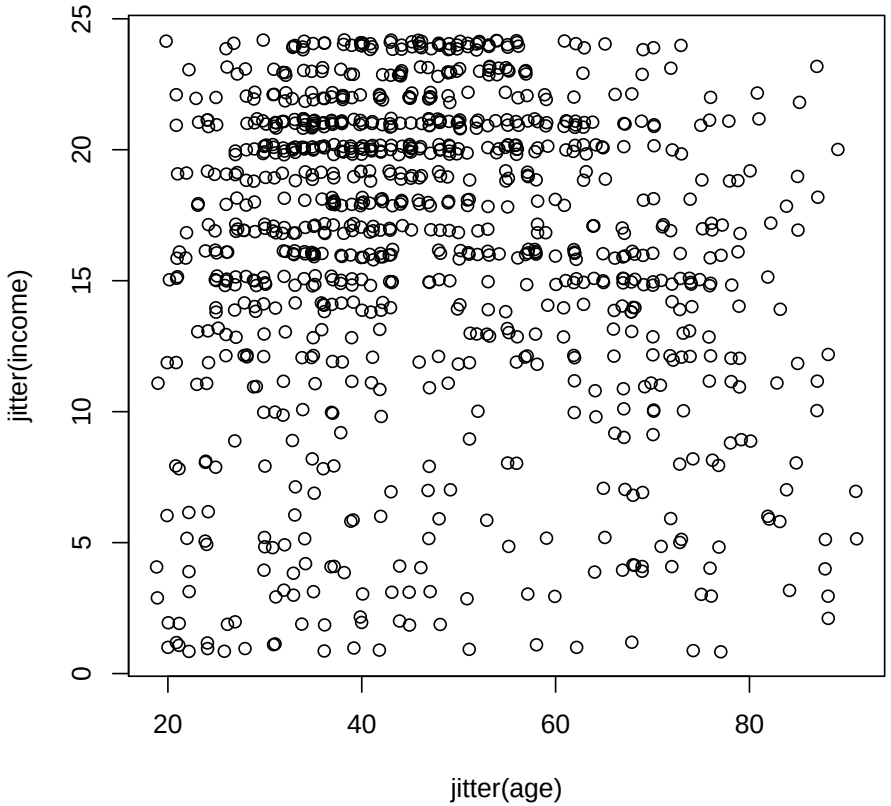
- Polynomial terms are a special case of the continuous variable interactions.
- For example, when $X_1 = X_2$ in the interaction model, we get a quadratic:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + u$$
$$Y = \beta_0 + (\beta_1 + \beta_2) X_1 + \beta_3 X_1 X_1 + u$$
$$Y = \beta_0 + \tilde{\beta}_1 X_1 + \tilde{\beta}_2 X_1^2 + u$$

- This is called a second order polynomial in X_1
- A third order polynomial is given by:
  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_1^3 + u$

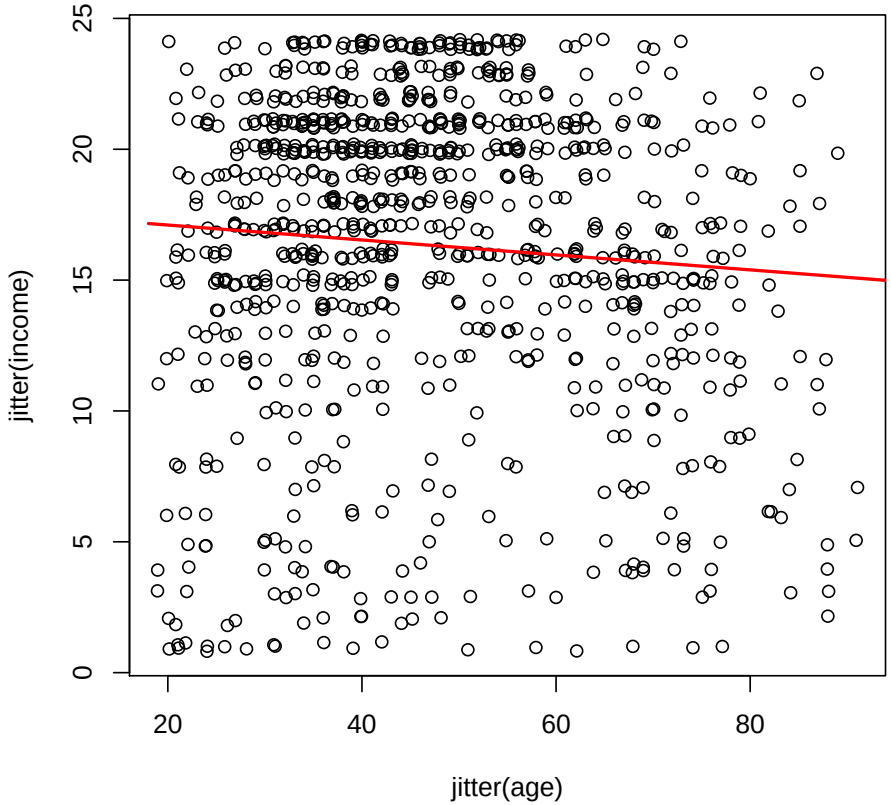
Polynomial Example: Income and Age

- Let's look at data from the U.S. and examine the relationship between Y: income and X: age



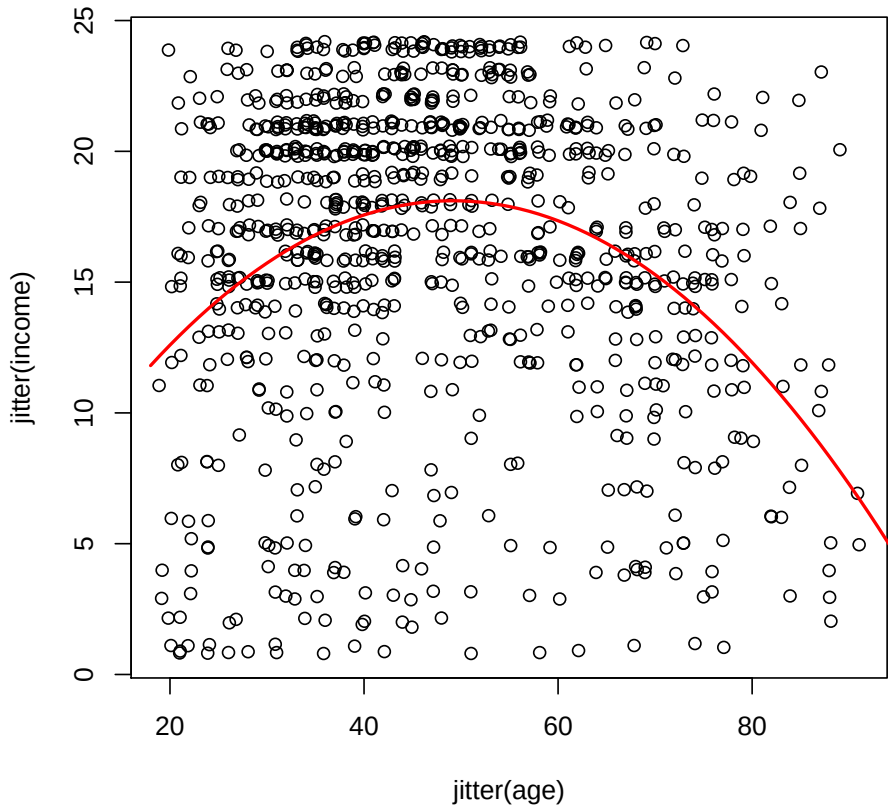
Polynomial Example: Income and Age

- Let's look at data from the U.S. and examine the relationship between Y: income and X: age
- We see that a simple linear specification does not fit the data very well:
  $Y = \beta_0 + \beta_1 X_1 + u$



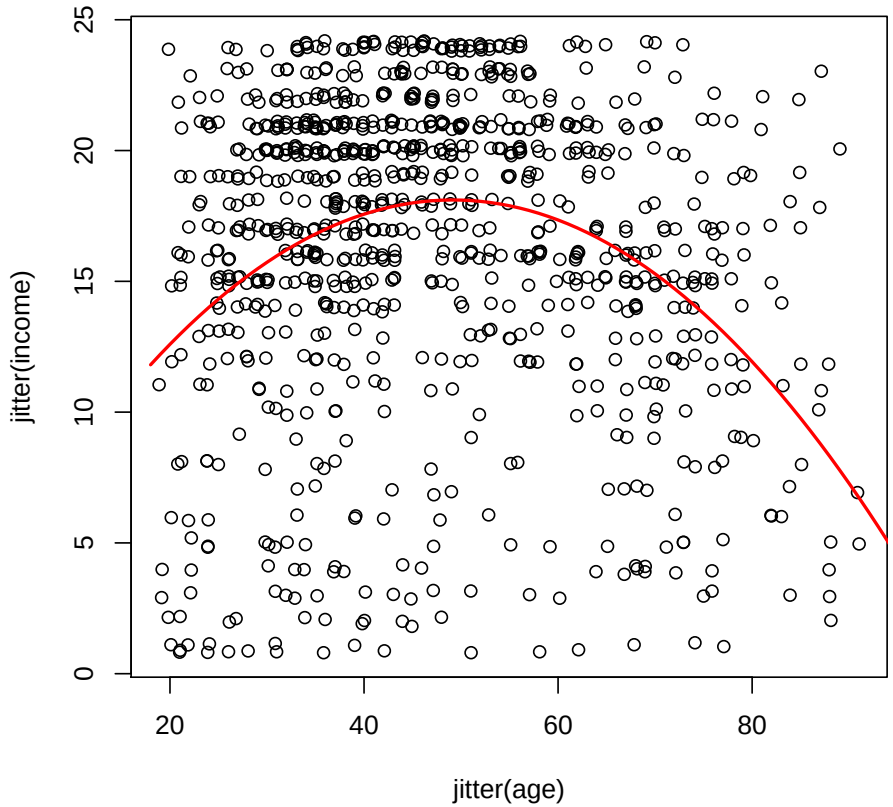
Polynomial Example: Income and Age

- Let's look at data from the U.S. and examine the relationship between Y: income and X: age
- We see that a simple linear specification does not fit the data very well:
  $Y = \beta_0 + \beta_1 X_1 + u$
- A second order polynomial in age fits the data a lot better:
  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + u$



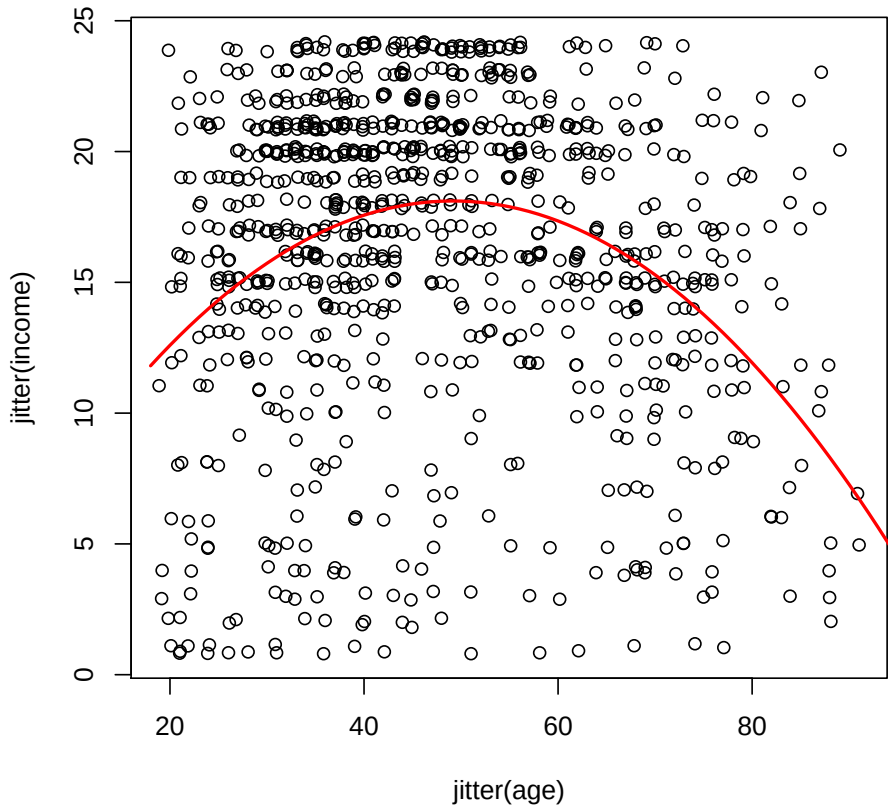
Polynomial Example: Income and Age

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + u$



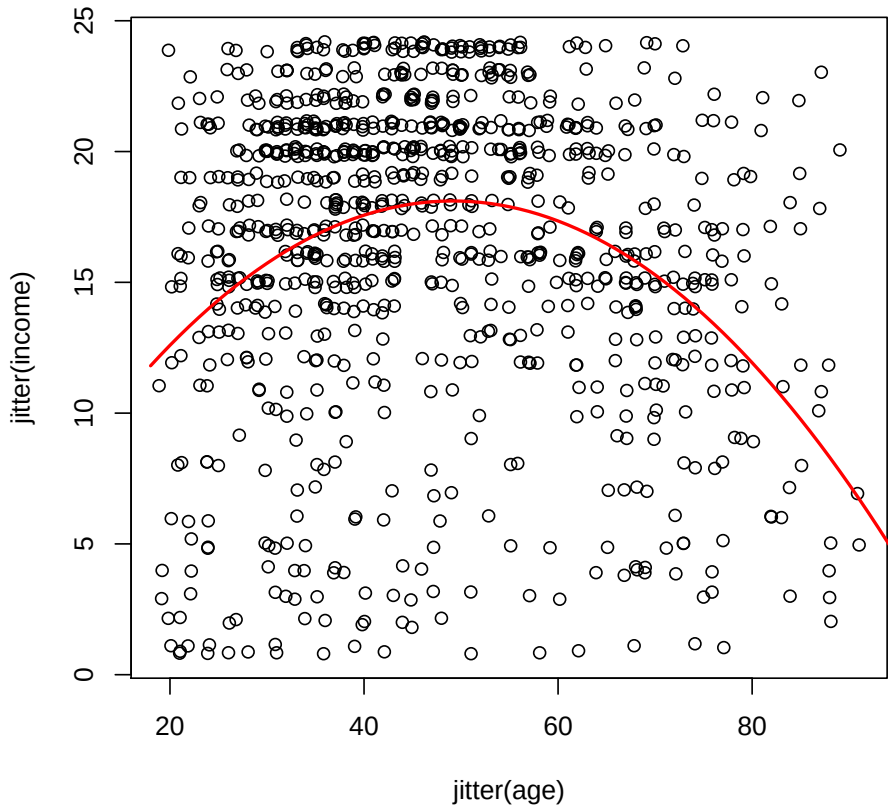
Polynomial Example: Income and Age

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + u$
- Is β_1 the marginal effect of age on income?



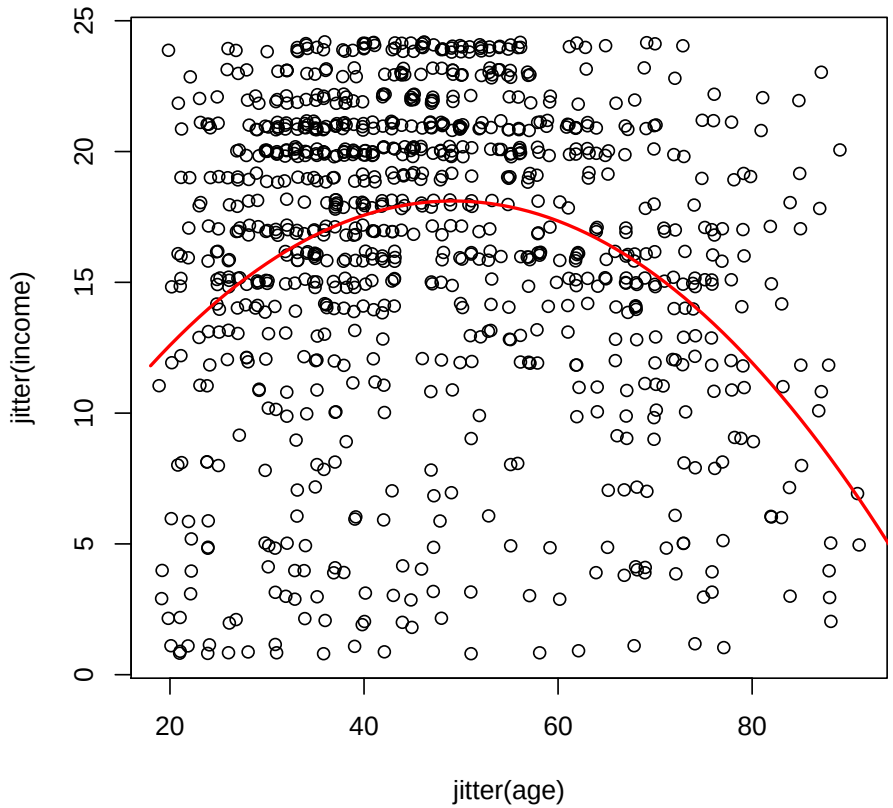
Polynomial Example: Income and Age

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + u$
- Is β_1 the marginal effect of age on income?
- No! The marginal effect of age depends on the level of age:
  $\frac{\partial Y}{\partial X_1} = \widehat{\beta}_1 + 2\,\widehat{\beta}_2\, X_1$
  Here the effect of age changes monotonically from positive to negative with income.



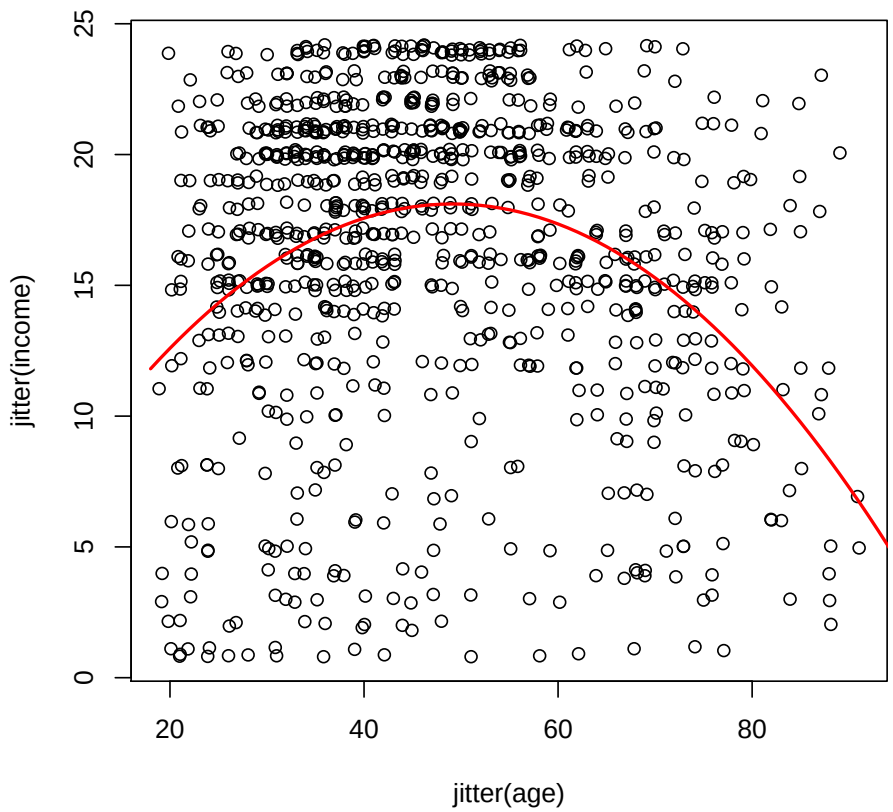
Polynomial Example: Income and Age

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + u$
- Is β_1 the marginal effect of age on income?
- No! The marginal effect of age depends on the level of age:
  $\frac{\partial Y}{\partial X_1} = \widehat{\beta}_1 + 2\,\widehat{\beta}_2\,X_1$
  Here the effect of age changes monotonically from positive to negative with income.
- If $\beta_2 > 0$ we get a U-shape, and if $\beta_2 < 0$ we get an inverted U-shape.

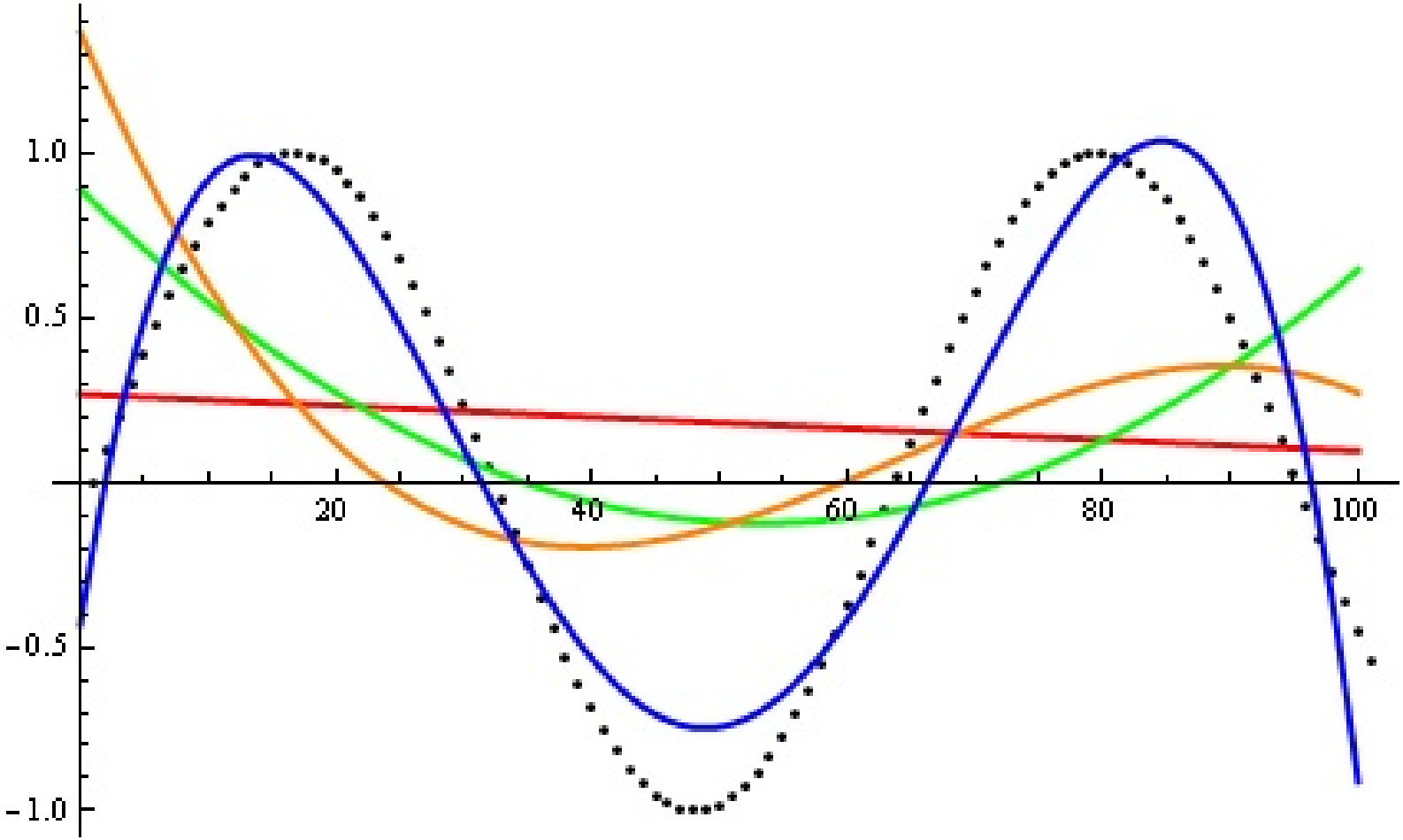


Polynomial Example: Income and Age

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + u$
- Is β_1 the marginal effect of age on income?
- No! The marginal effect of age depends on the level of age:
  $\frac{\partial Y}{\partial X_1} = \widehat{\beta}_1 + 2\,\widehat{\beta}_2\, X_1$
  Here the effect of age changes monotonically from positive to negative with income.
- If $\beta_2 > 0$ we get a U-shape, and if $\beta_2 < 0$ we get an inverted U-shape.
- Maximum/Minimum occurs at $|\frac{\beta_1}{2\beta_2}|$. Here turning point is at $X_1 = 50$.



Higher Order Polynomials



Approximating data generated with a sine function. Red line is a first degree polynomial, green line is second degree, orange line is third degree and blue is fourth degree

Complex Parameter Interpretation

We can mix and match these model specifications

Complex Parameter Interpretation

We can mix and match these model specifications

- Interactions with higher order terms

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 X_i^2 + \beta_4 X_i * Z_i + \beta_5 X_i^2 * Z_i + u$$

Complex Parameter Interpretation

We can mix and match these model specifications

- Interactions with higher order terms

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 X_i^2 + \beta_4 X_i * Z_i + \beta_5 X_i^2 * Z_i + u$$

- World is your oyster!

Complex Parameter Interpretation

We can mix and match these model specifications

- Interactions with higher order terms

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 X_i^2 + \beta_4 X_i * Z_i + \beta_5 X_i^2 * Z_i + u$$

- World is your oyster!
- But interpretation gets difficult.

Complex Parameter Interpretation

We can mix and match these model specifications

- Interactions with higher order terms

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 X_i^2 + \beta_4 X_i * Z_i + \beta_5 X_i^2 * Z_i + u$$

- World is your oyster!
- But interpretation gets difficult.
- We can take partial derivatives as a way to get a sense of the shape of the estimated CEF.

Complex Parameter Interpretation

We can mix and match these model specifications

- Interactions with higher order terms

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 X_i^2 + \beta_4 X_i * Z_i + \beta_5 X_i^2 * Z_i + u$$

- World is your oyster!
- But interpretation gets difficult.
- We can take partial derivatives as a way to get a sense of the shape of the estimated CEF.
- An easier approach can be to make predictions and then average over the data

$$\frac{1}{n} \sum_{i}^{n} \widehat{E}[Y|X = x_i + 1, Z = z_i] - \widehat{E}[Y|X = x_i, Z = z_i]$$

Complex Parameter Interpretation

We can mix and match these model specifications

- Interactions with higher order terms

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 X_i^2 + \beta_4 X_i * Z_i + \beta_5 X_i^2 * Z_i + u$$

- World is your oyster!
- But interpretation gets difficult.
- We can take partial derivatives as a way to get a sense of the shape of the estimated CEF.
- An easier approach can be to make predictions and then average over the data

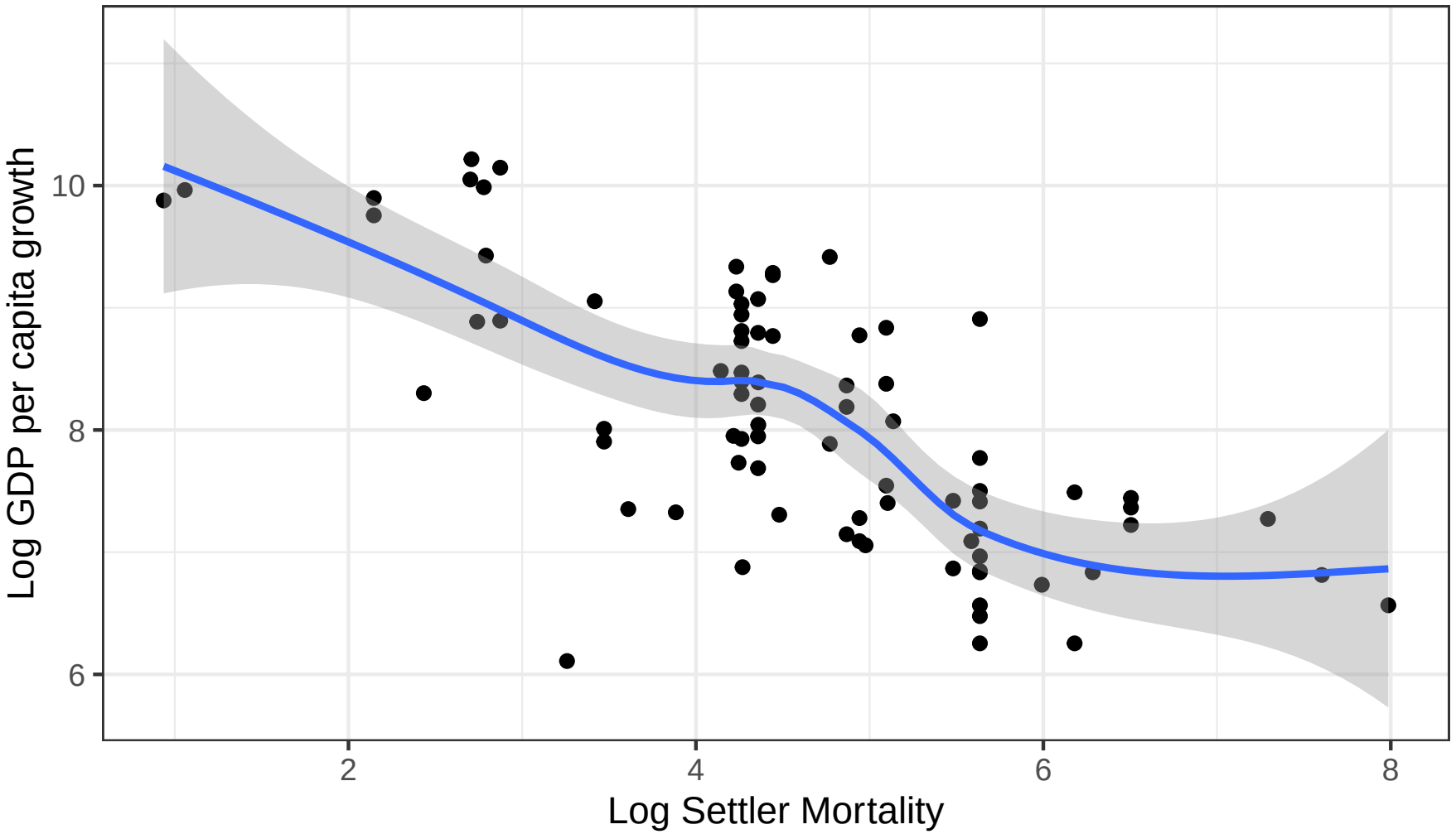
$$\frac{1}{n} \sum_{i}^{n} \widehat{E}[Y|X = x_i + 1, Z = z_i] - \widehat{E}[Y|X = x_i, Z = z_i]$$

- Getting uncertainty on this quantity is a great use case for the bootstrap!

1. Non-Linearity
   - Polynomials
   - LOESS
   - Log Transformations
2. Non-Normality
   - Hat Matrix
   - Standardized and Studentized Residuals
   - QQ Plots
3. Extreme Values
   - Outliers
   - Leverage Points
   - Influence Points
4. Clustering

1. Non-Linearity
   - Polynomials
   - LOESS
   - Log Transformations
2. Non-Normality
   - Hat Matrix
   - Standardized and Studentized Residuals
   - QQ Plots
3. Extreme Values
   - Outliers
   - Leverage Points
   - Influence Points
4. Clustering

So what is ggplot2 doing?



LOESS

LOESS

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

LOESS

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

LOESS

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

LOESS

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

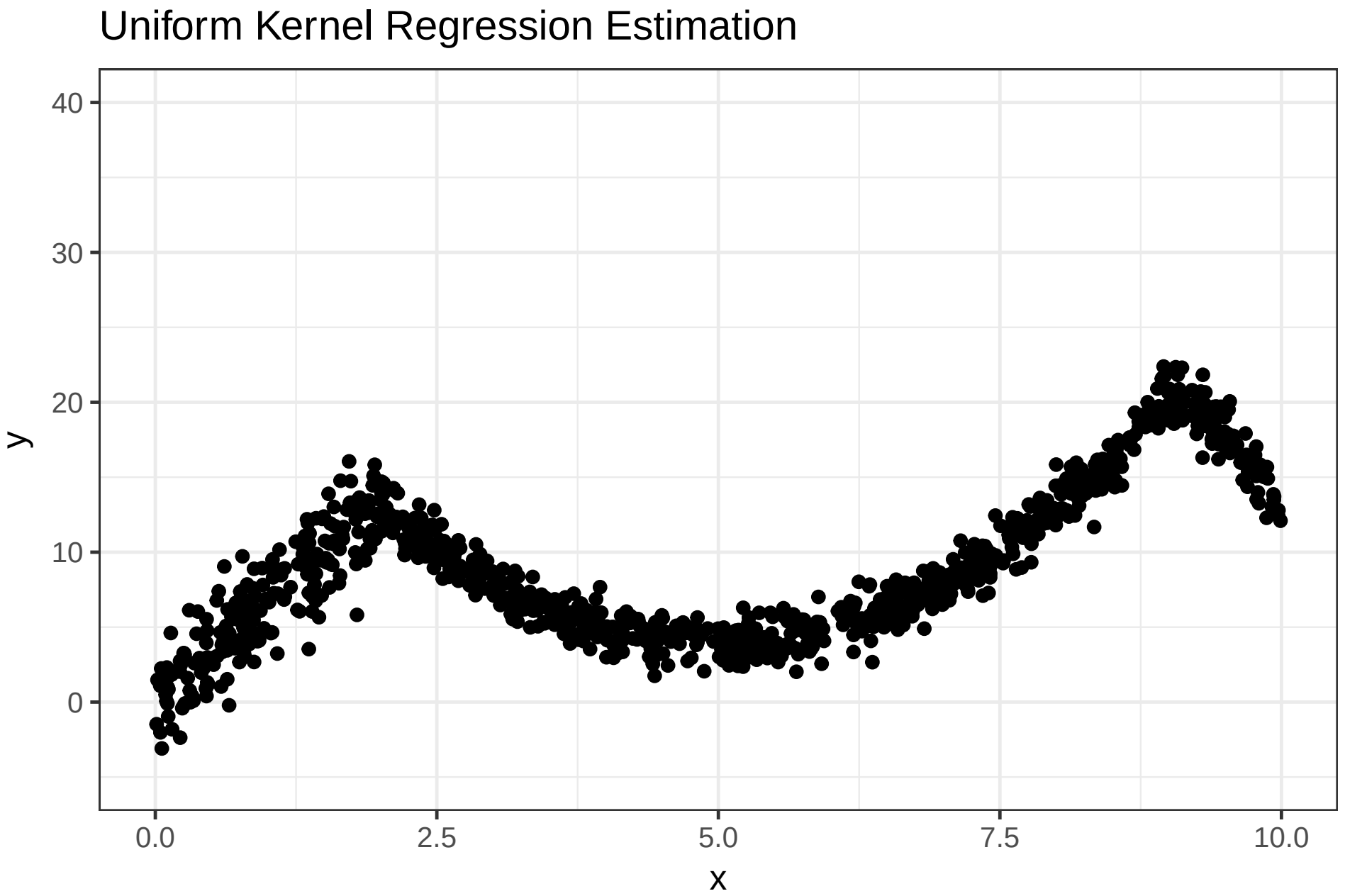
LOESS

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

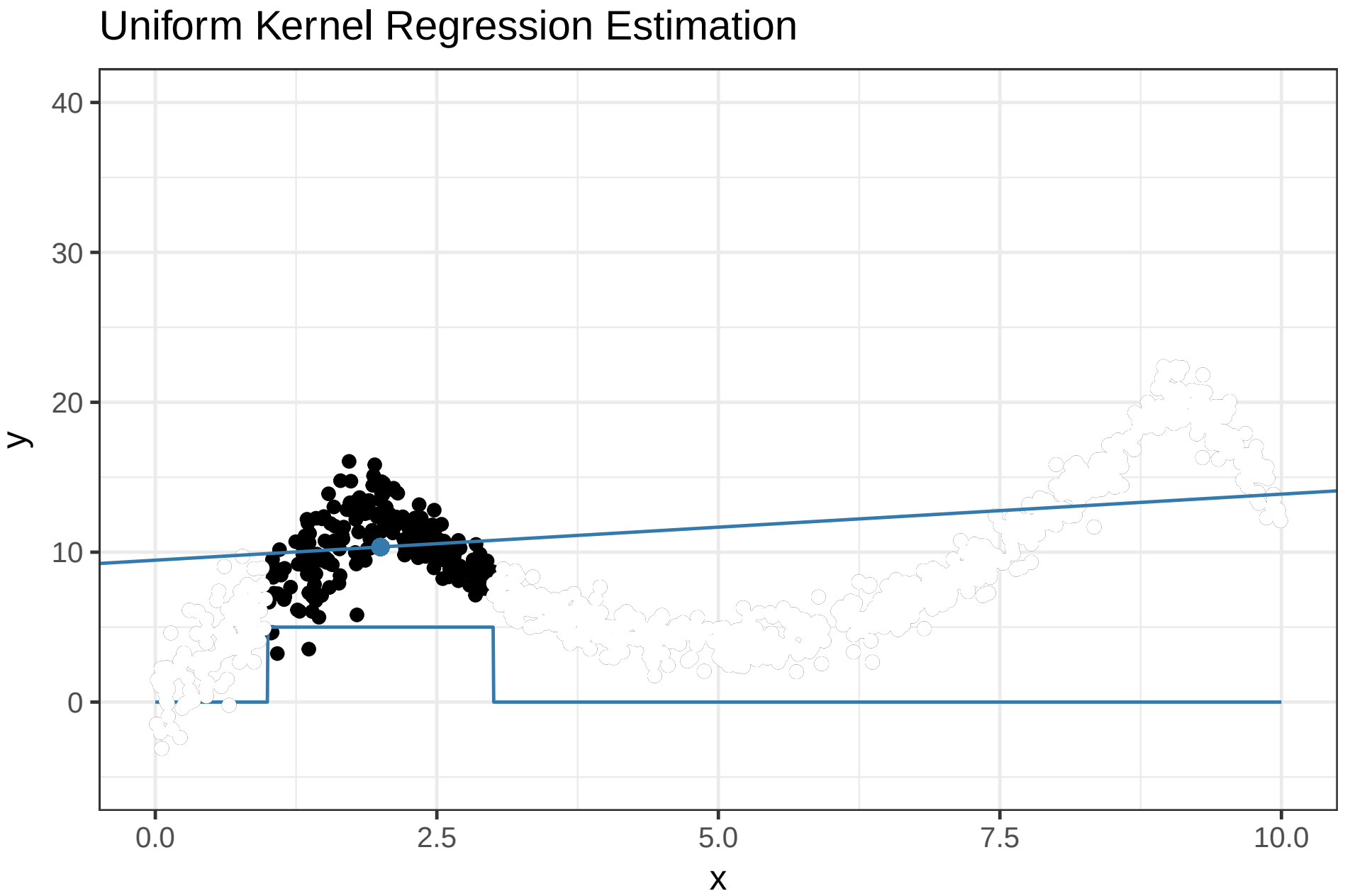
LOESS

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

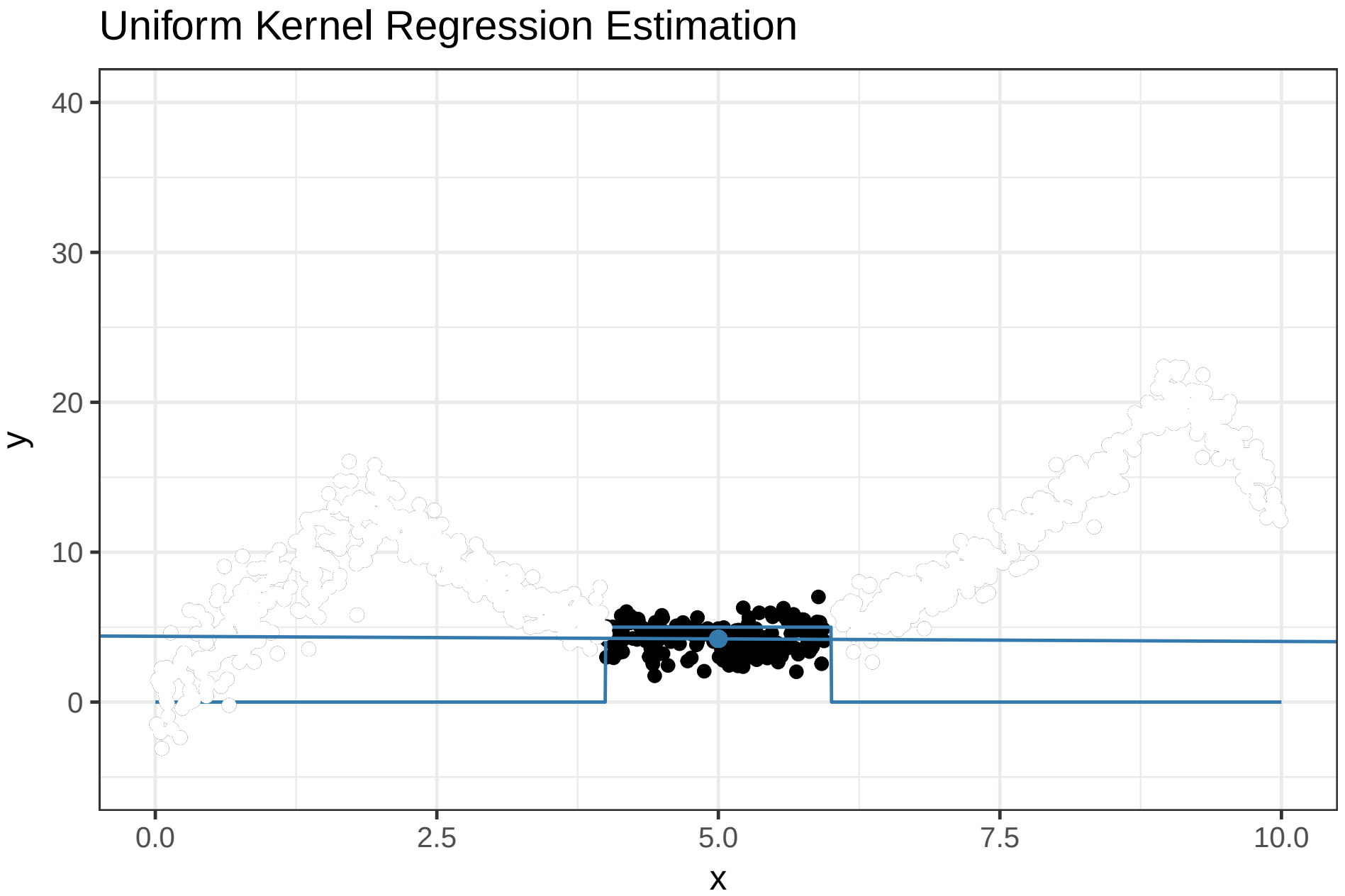
LOESS Example



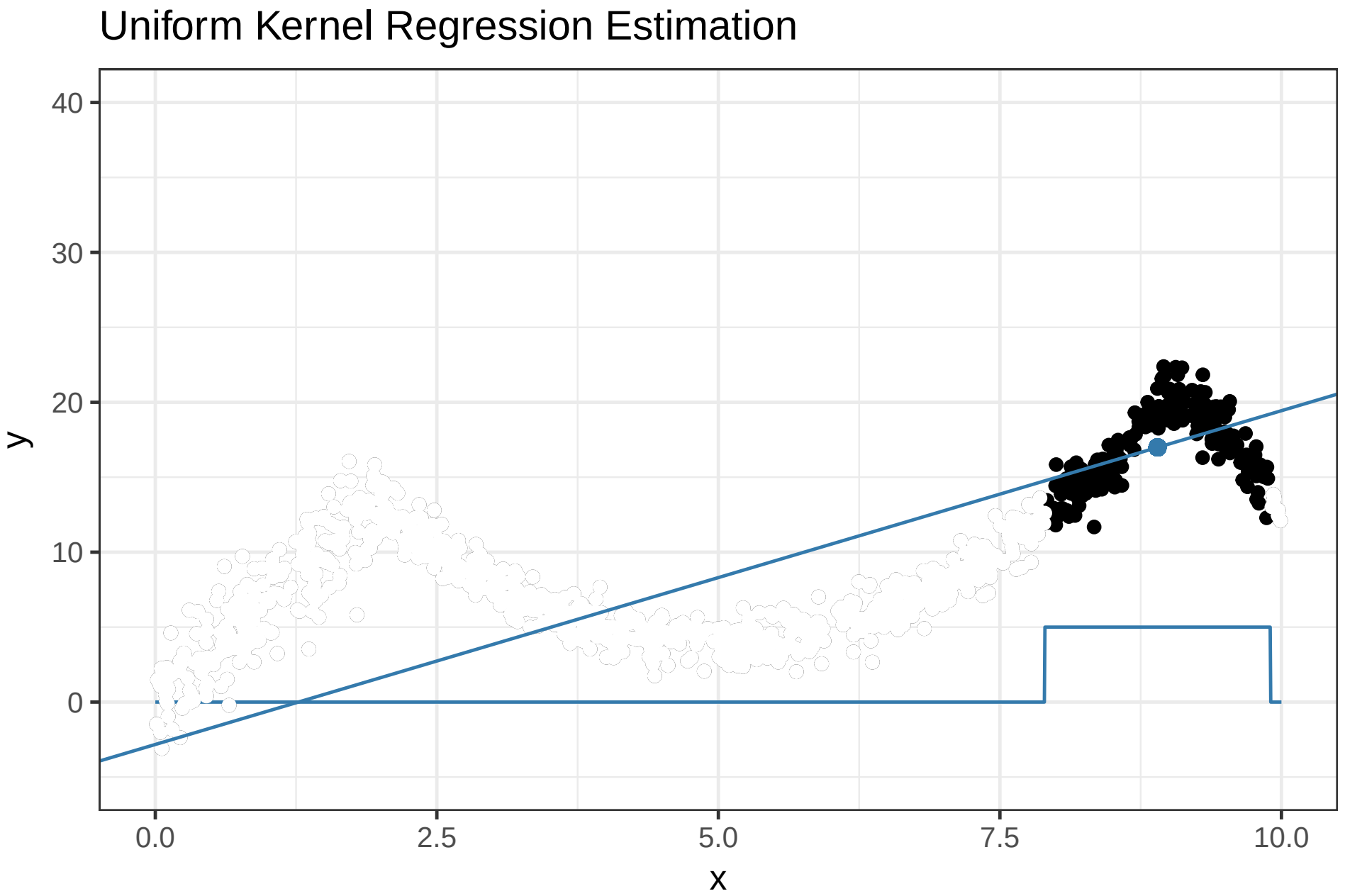
LOESS Example



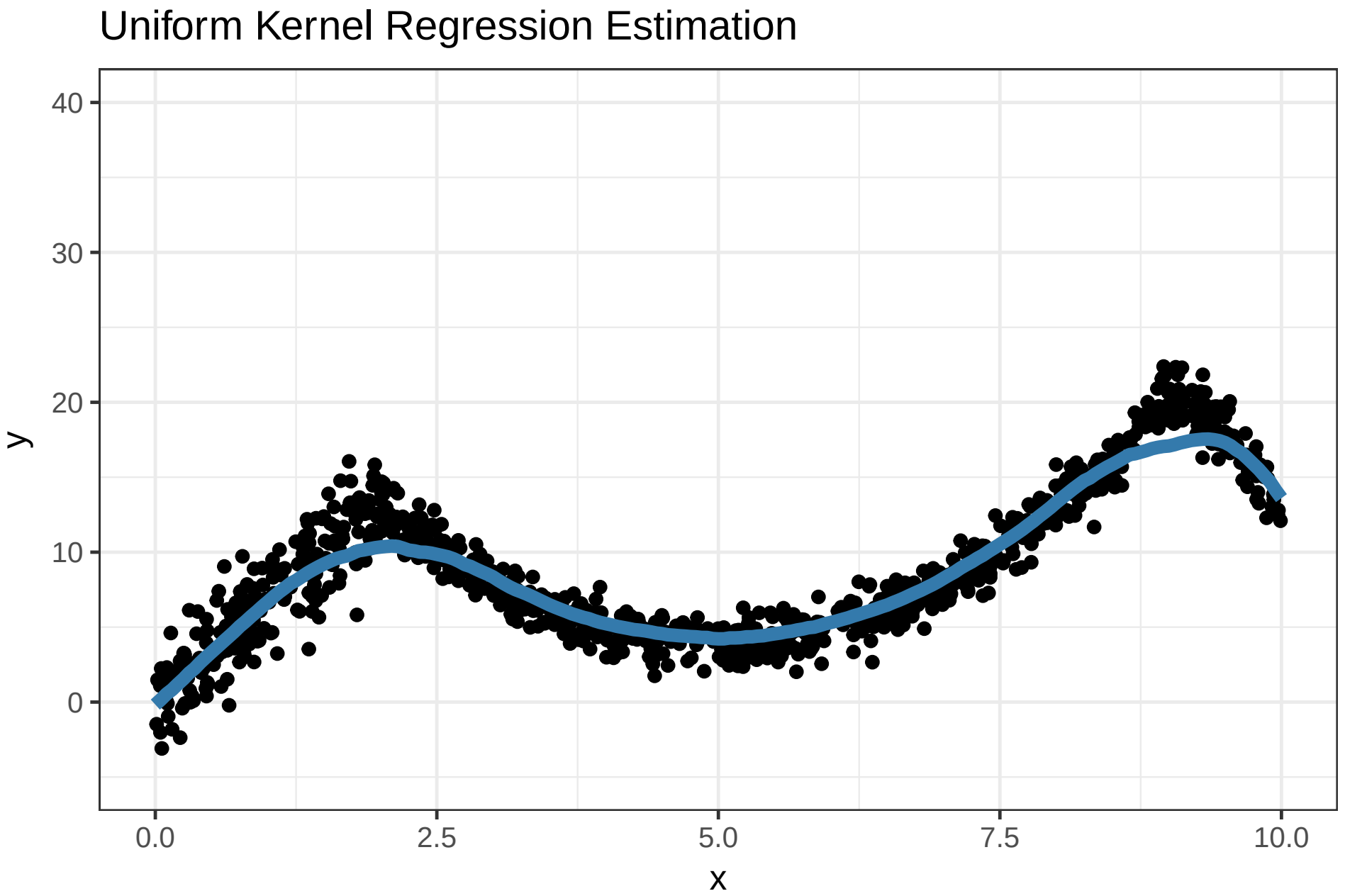
LOESS Example



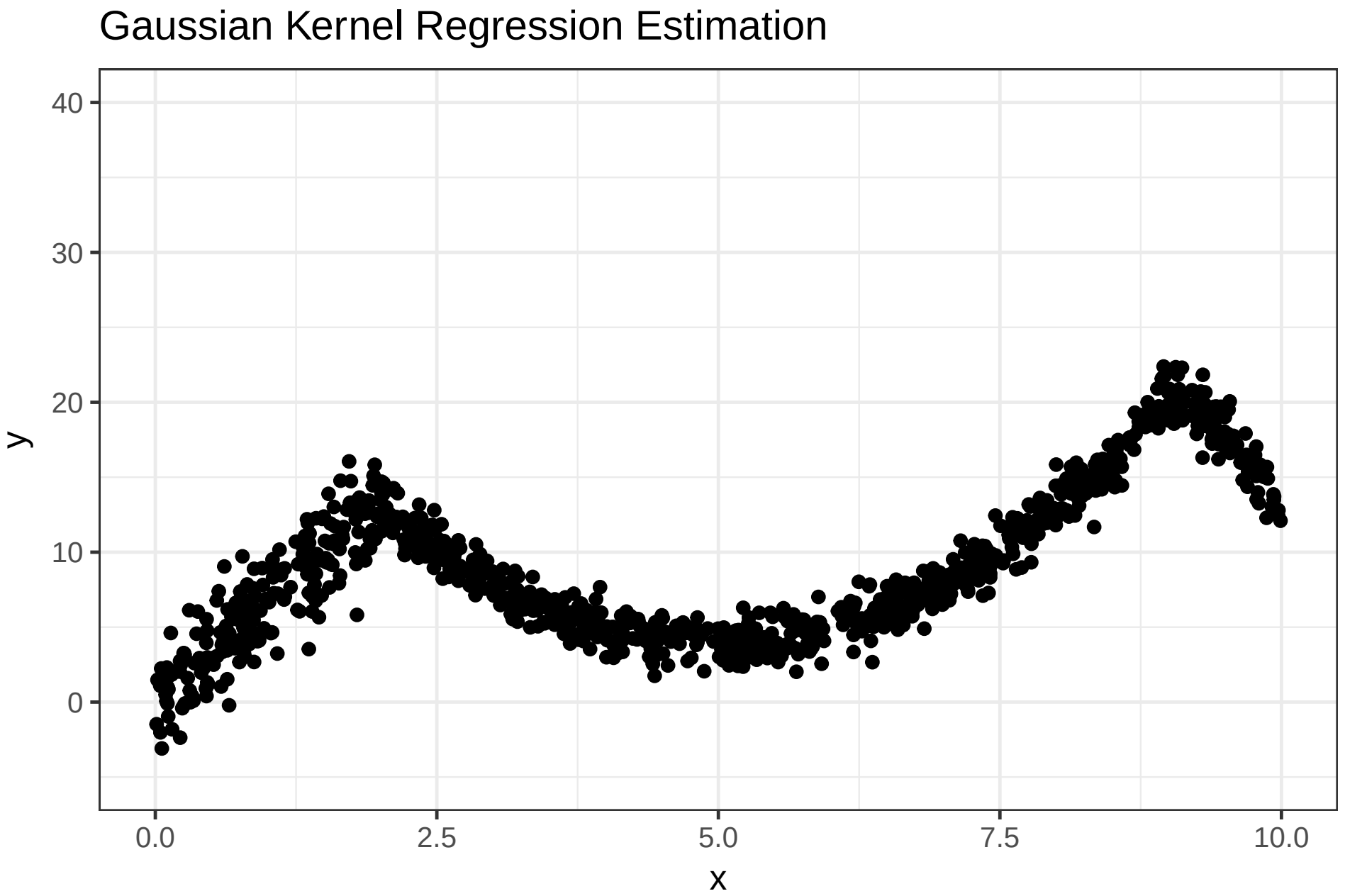
LOESS Example



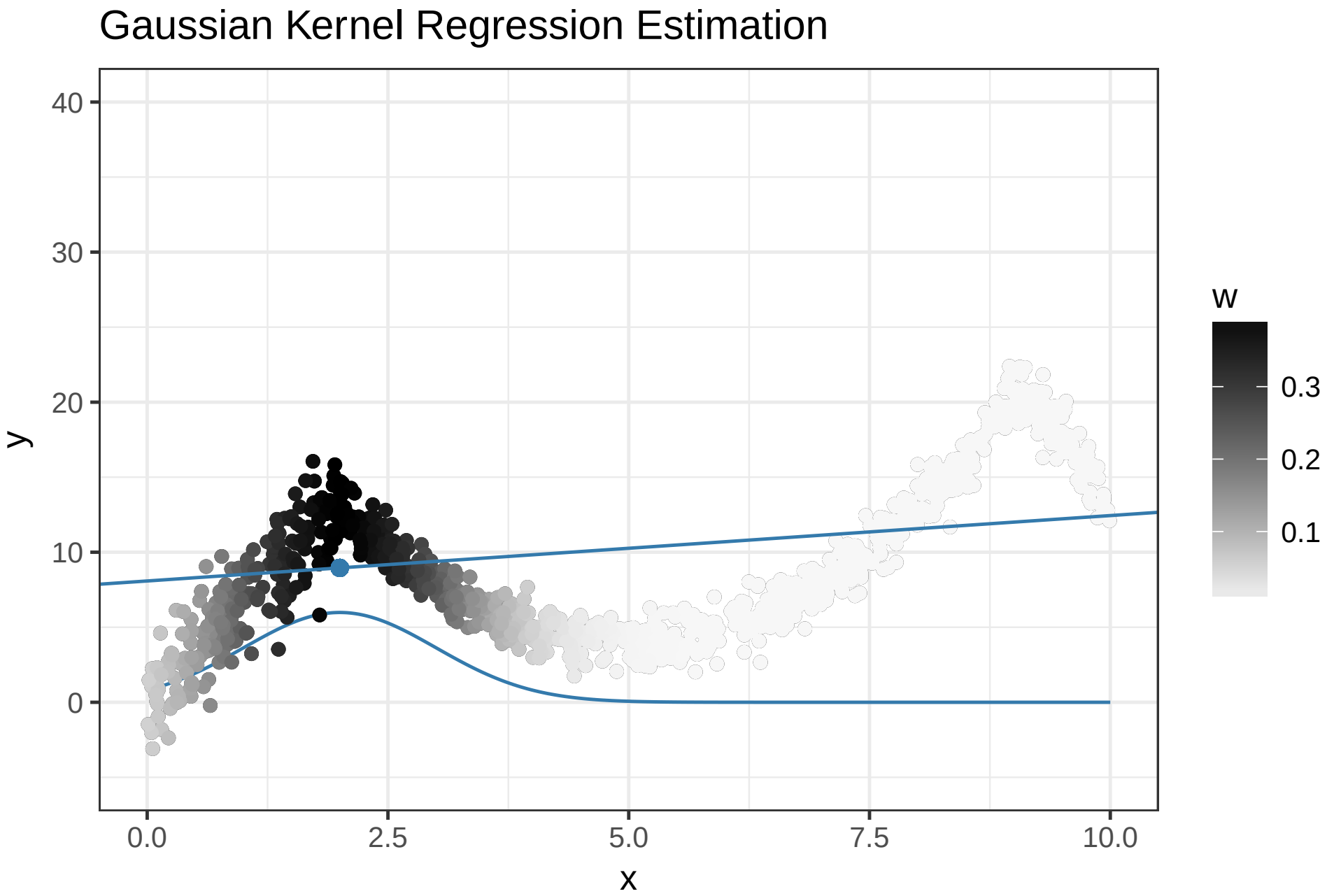
LOESS Example



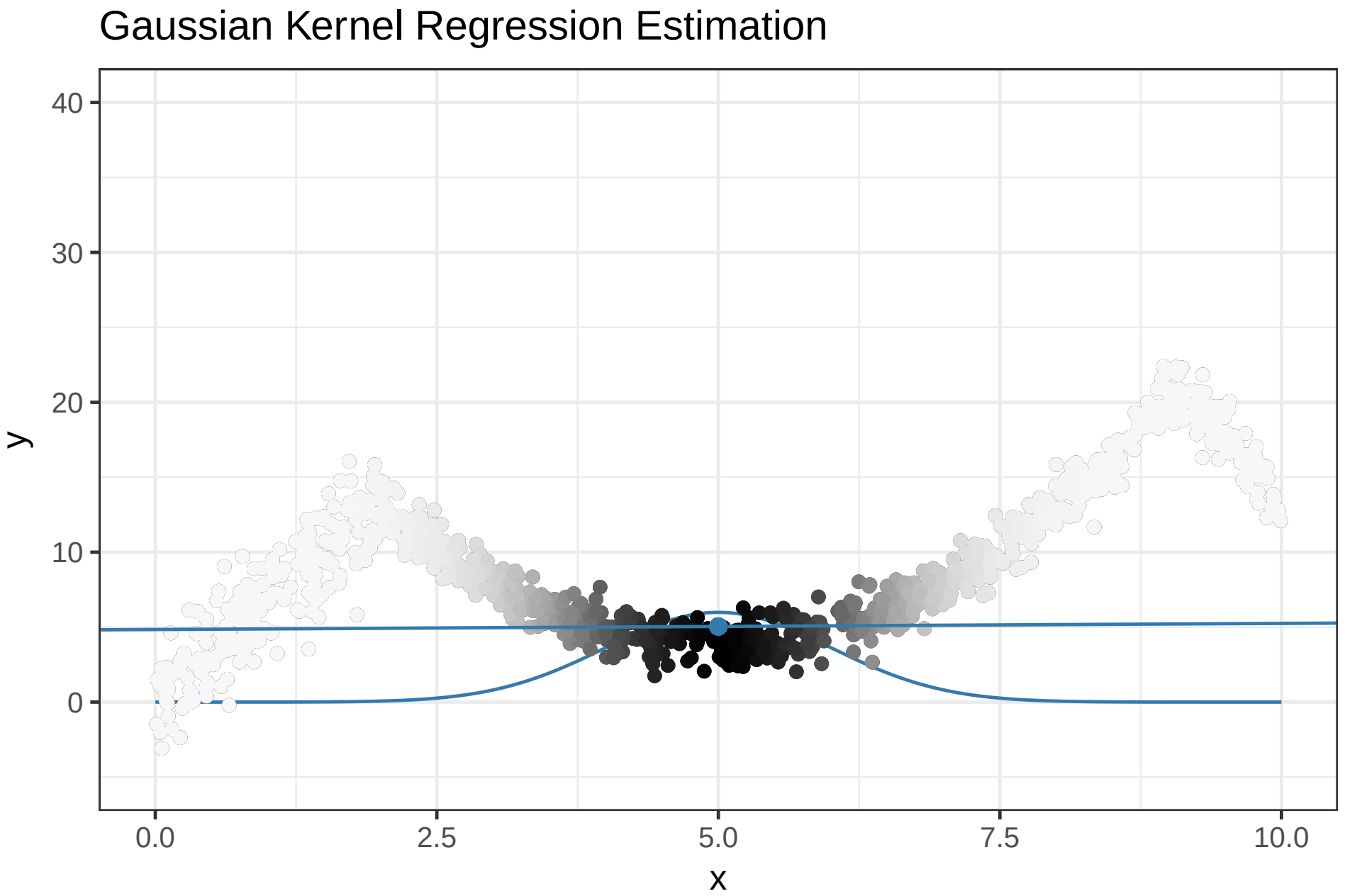
LOESS Example



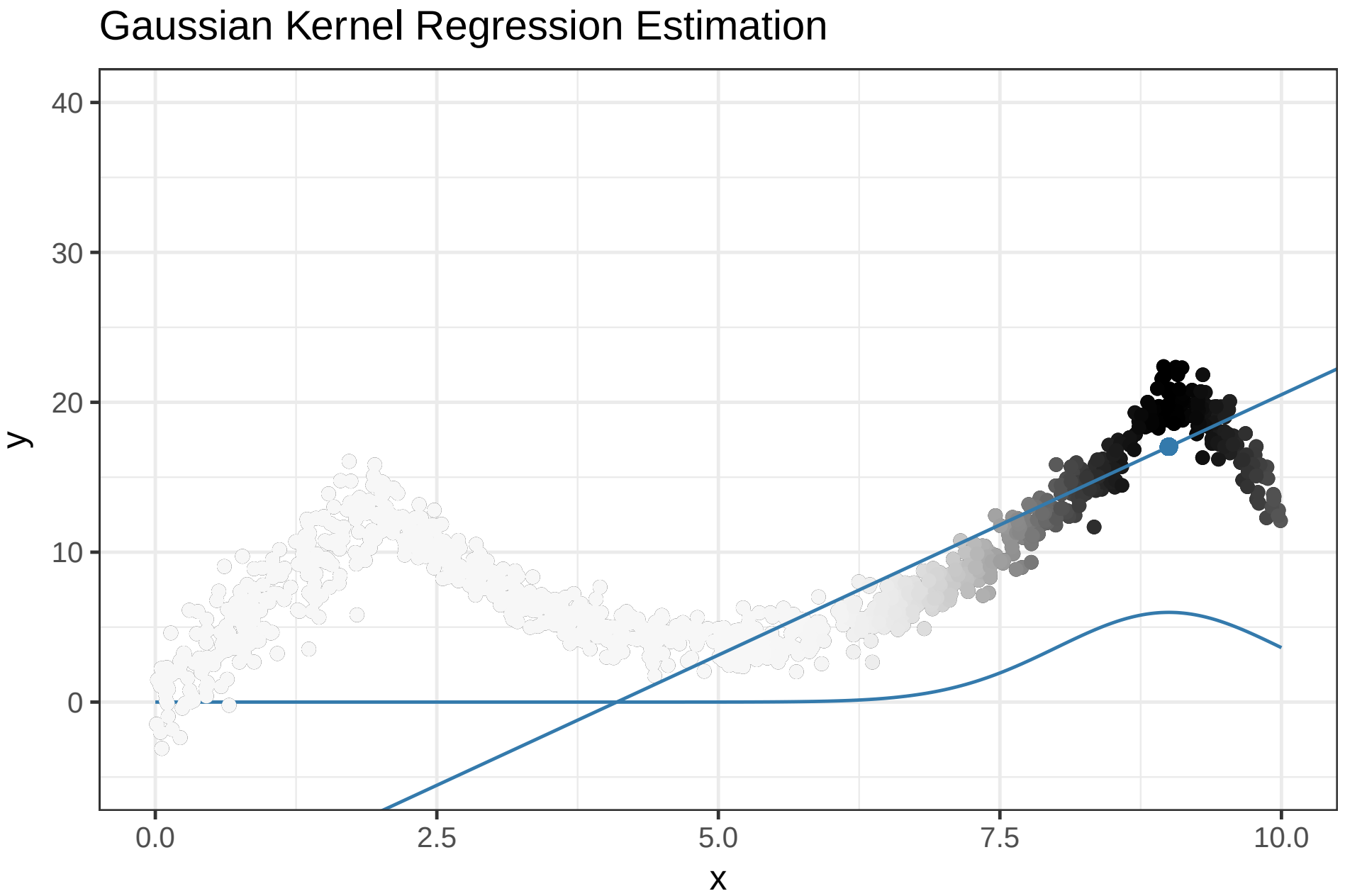
LOESS Example



LOESS Example



LOESS Example



Non-linear CEFs

Non-linear CEFs

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

Non-linear CEFs

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

Non-linear CEFs

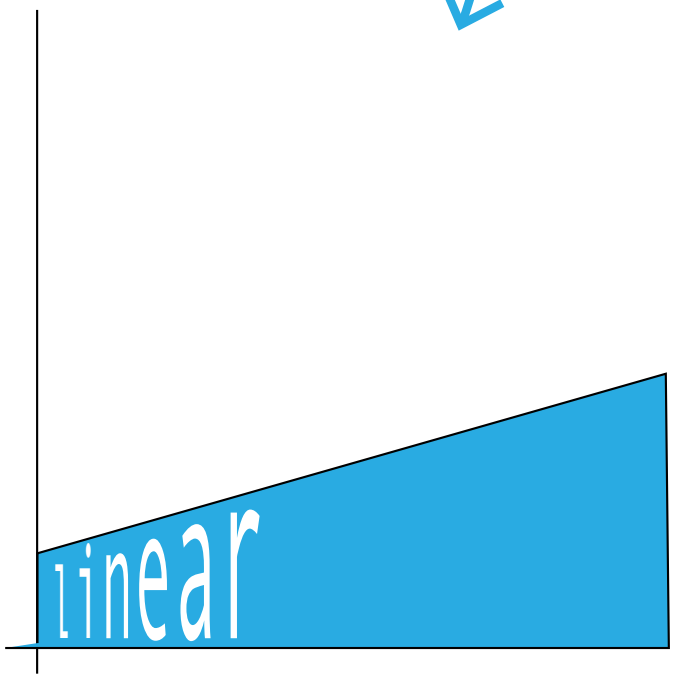
- When we say that CEFs are linear with regression, we mean linear in parameters but by including transformations of our variables we can make non-linear shapes of pre-specified functional forms.
- Many of these non-linear transformations are made by creating multiple variables out of a single X and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

Non-linear CEFs

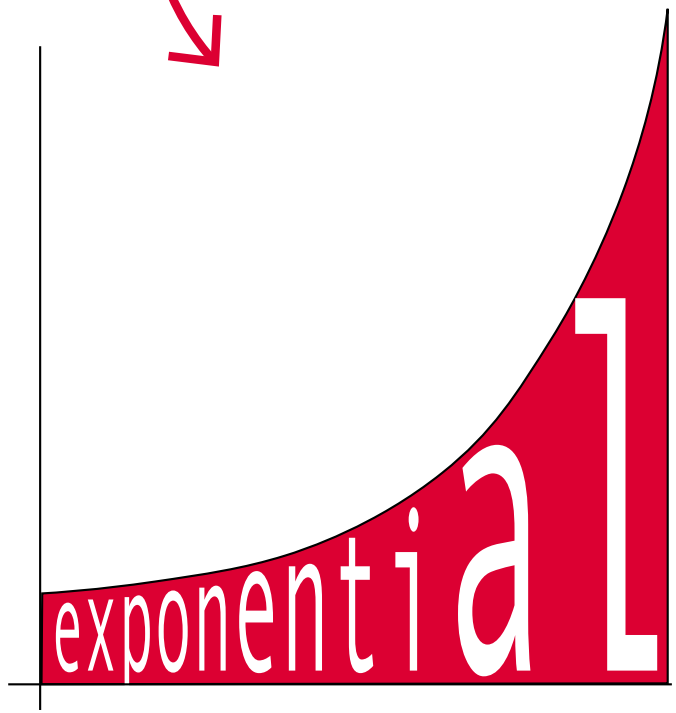
- When we say that CEFs are linear with regression, we mean linear in parameters but by including transformations of our variables we can make non-linear shapes of pre-specified functional forms.
- Many of these non-linear transformations are made by creating multiple variables out of a single X and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

Why does everyone keep logging stuff??

Logs linearize exponential growth.



grows by a fixed amount.



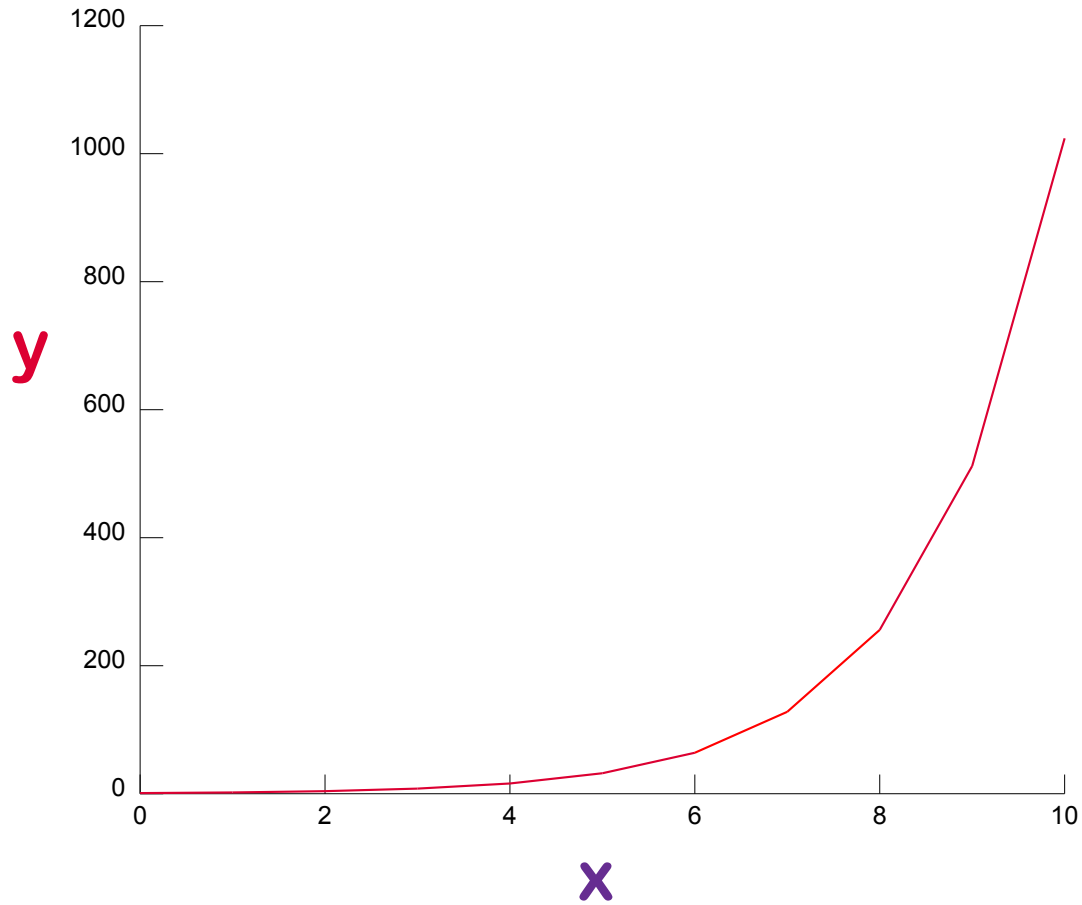
grows by a fixed percent.

How? Let's look.

First, here's a graph showing exponential growth.

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

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

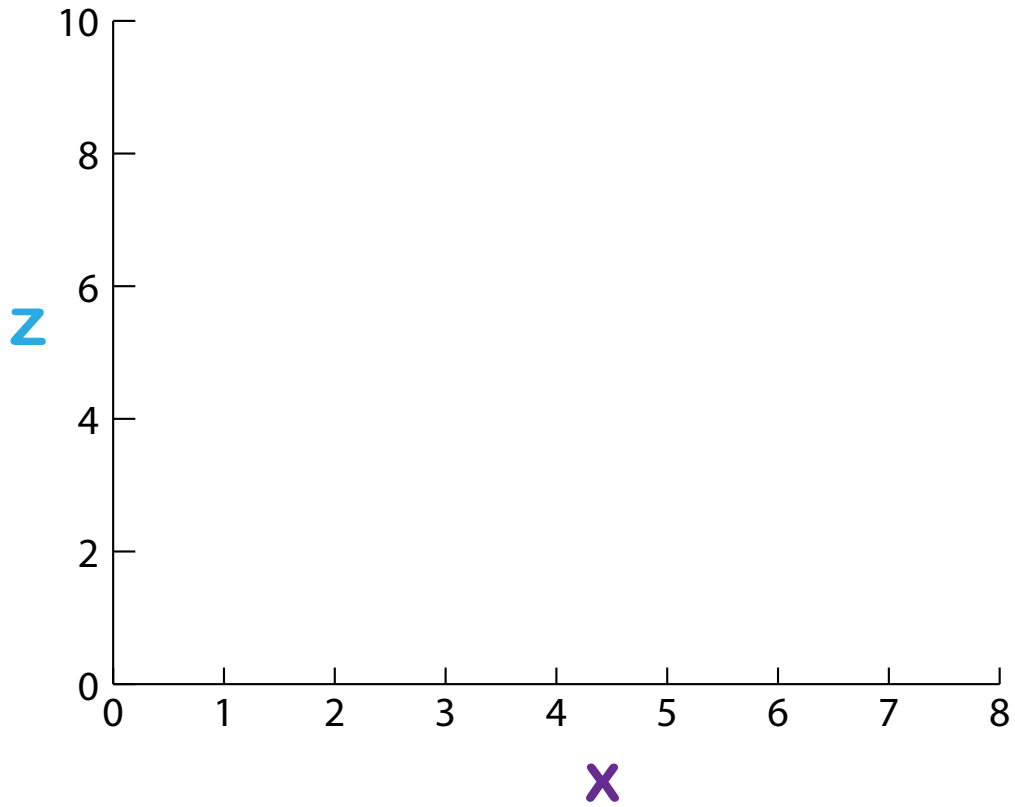


What happens when we take the log of y?

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

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

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

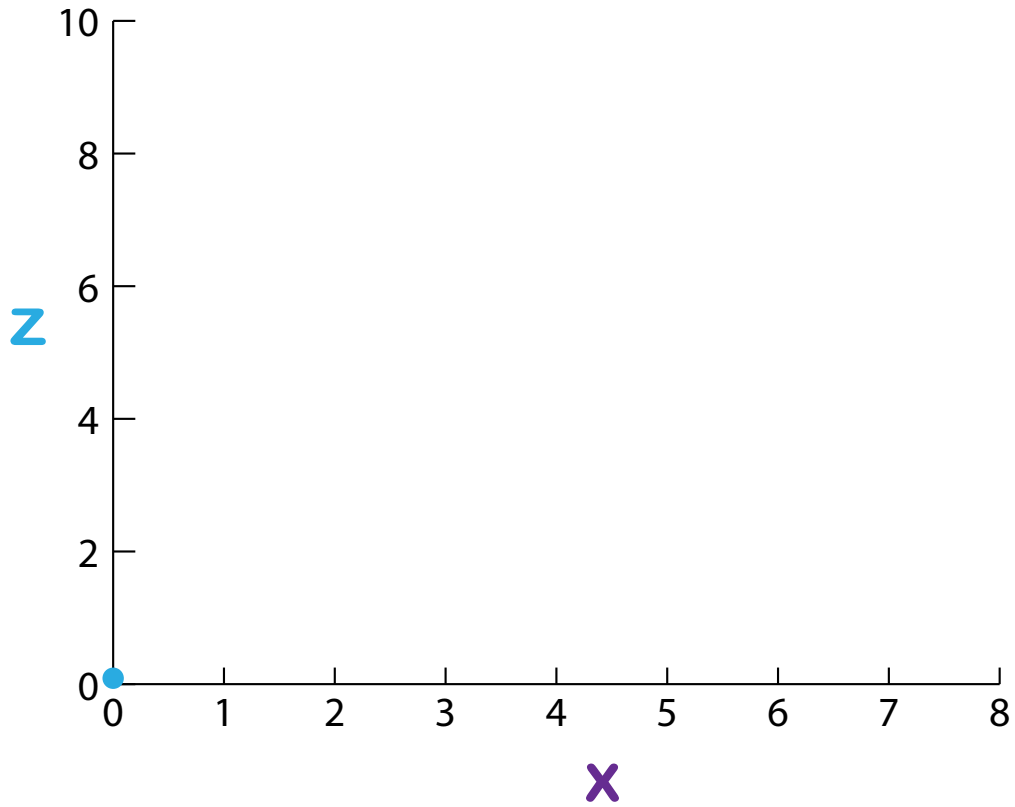


What happens when we take the log of y?

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

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

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

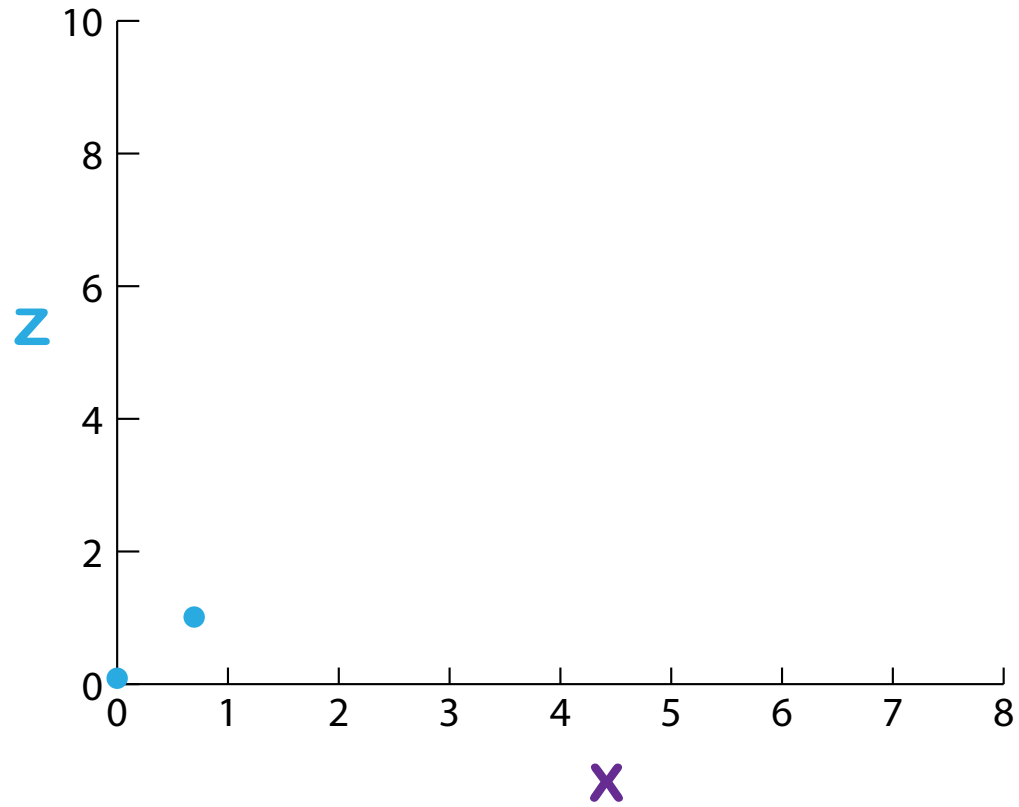


What happens when we take the log of y?

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

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

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

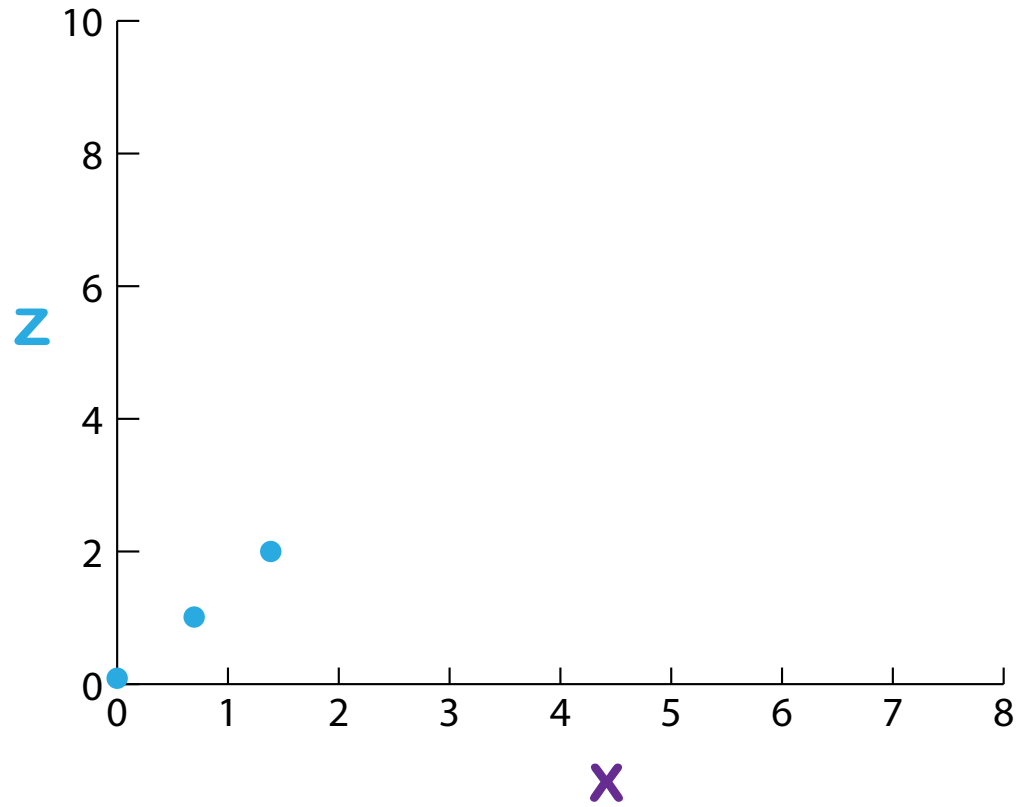


What happens when we take the log of y?

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

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

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

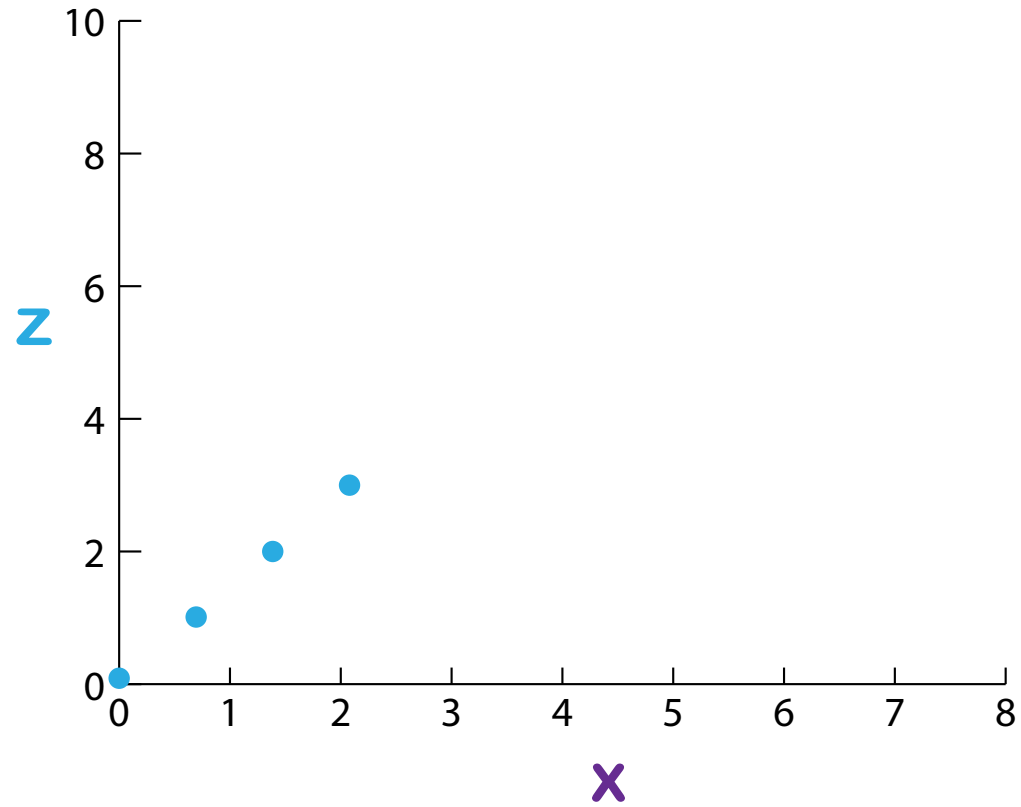


What happens when we take the log of y?

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

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

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

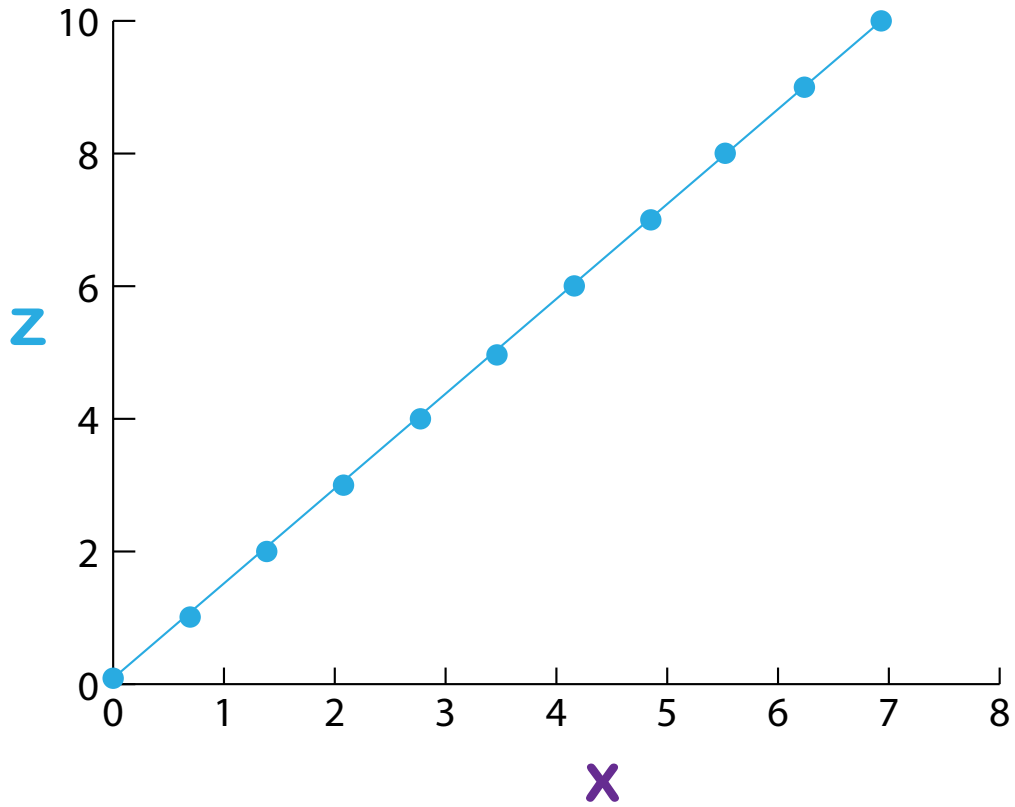


What happens when we take the log of y?

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

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

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



Interpretation

Interpretation

The log transformation changes the interpretation of β_1:

Interpretation

The log transformation changes the interpretation of β_1:

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

Interpretation

The log transformation changes the interpretation of β_1:

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

Interpretation

The log transformation changes the interpretation of β_1:

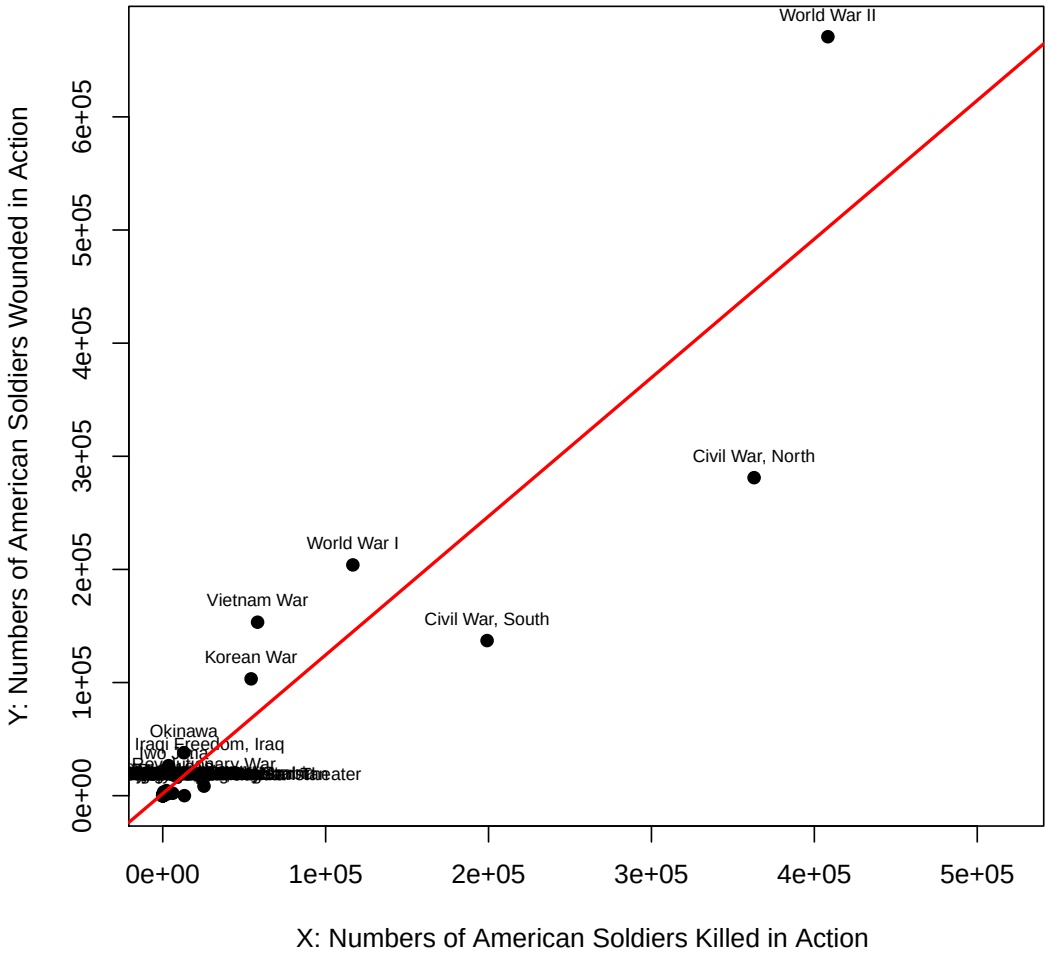
- Regress $\log(Y)$ on $X \longrightarrow \beta_1$ approximates percent increase in our prediction of Y associated with one unit increase in X.
- Regress Y on $\log(X) \longrightarrow \beta_1$ approximates increase in Y associated with a percent increase in X.
- Note that these approximations work only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

Interpretation

The log transformation changes the interpretation of β_1:

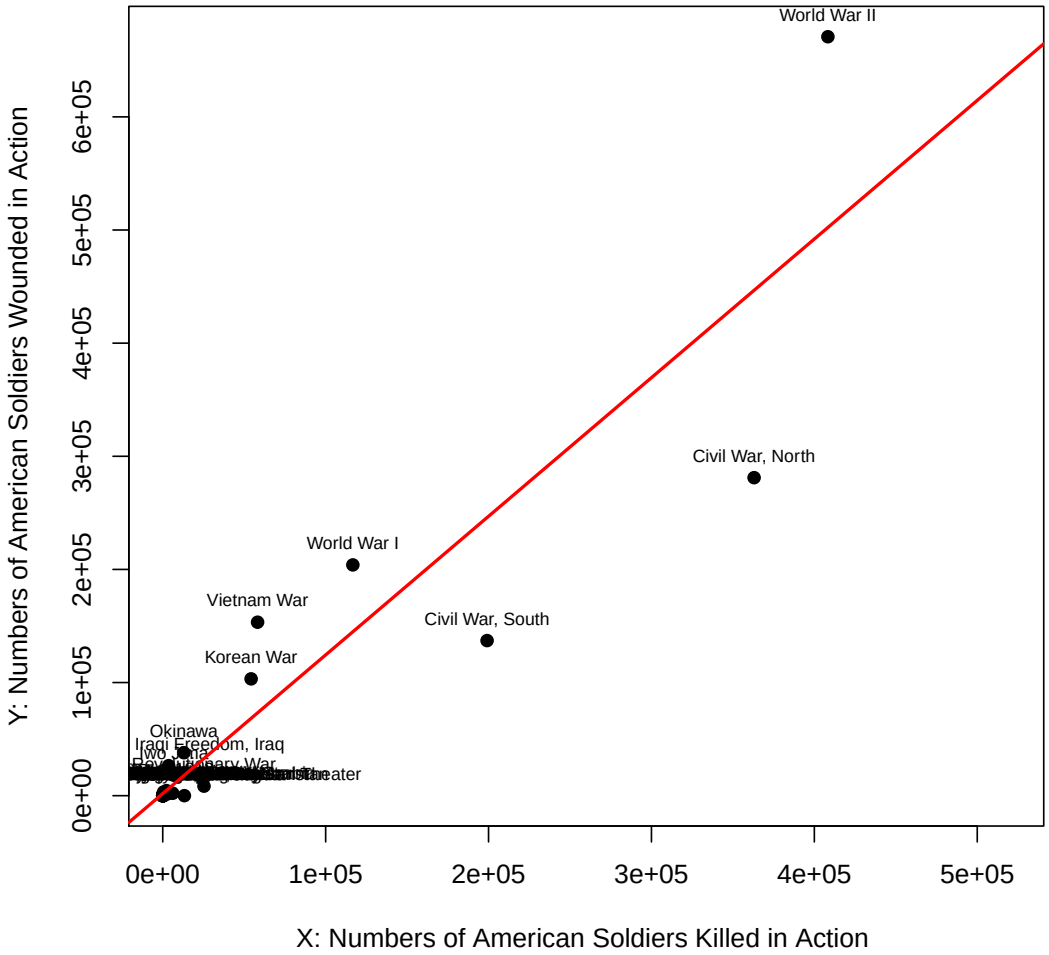
- Regress $\log(Y)$ on $X \longrightarrow \beta_1$ approximates percent increase in our prediction of Y associated with one unit increase in X.
- Regress Y on $\log(X) \longrightarrow \beta_1$ approximates increase in Y associated with a percent increase in X.
- Note that these approximations work only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

Example from the American War Library



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

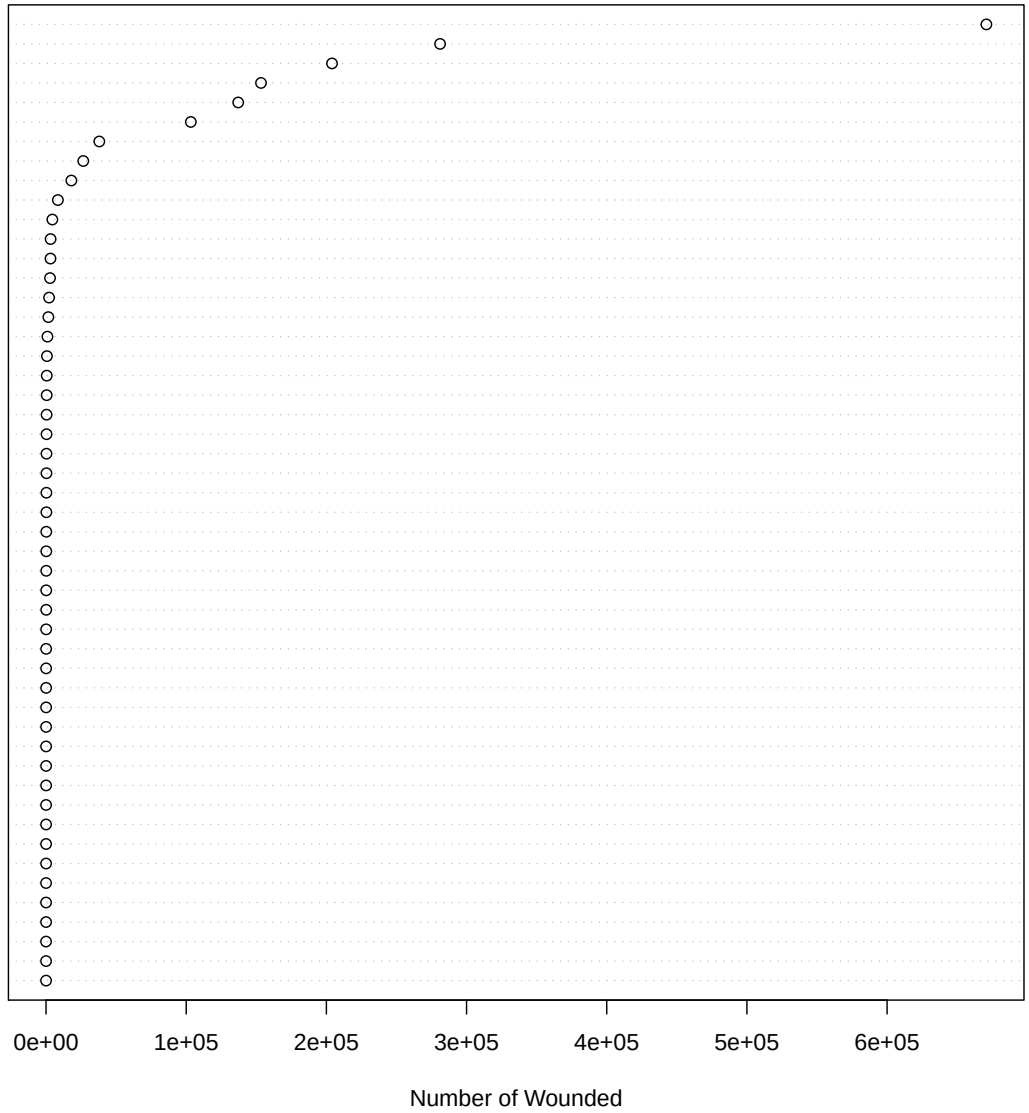
Example from the American War Library



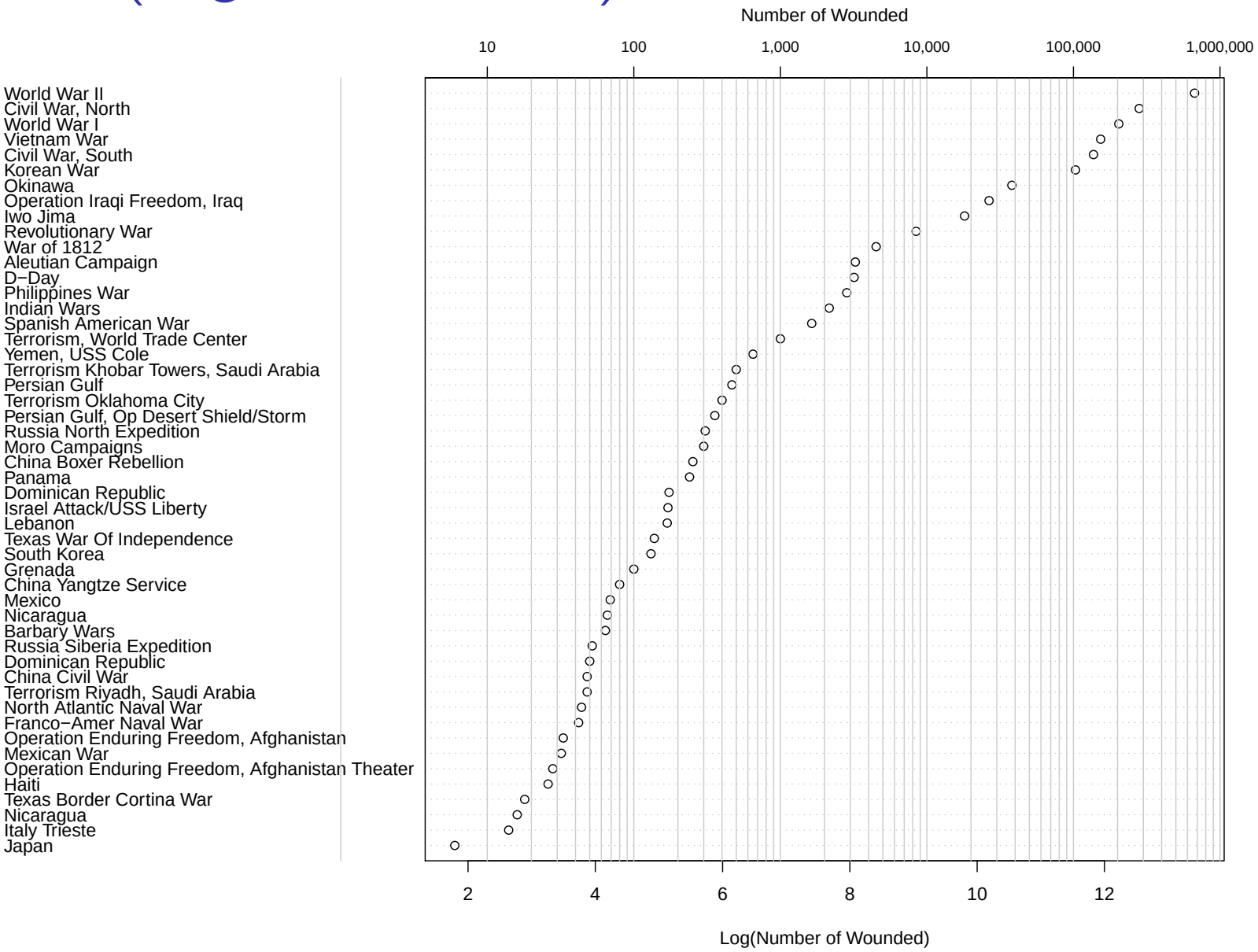
$\hat{\beta}_1 = 1.23 \longrightarrow$ One additional soldier killed predicts 1.23 additional soldiers wounded

Wounded (Scale in Levels)

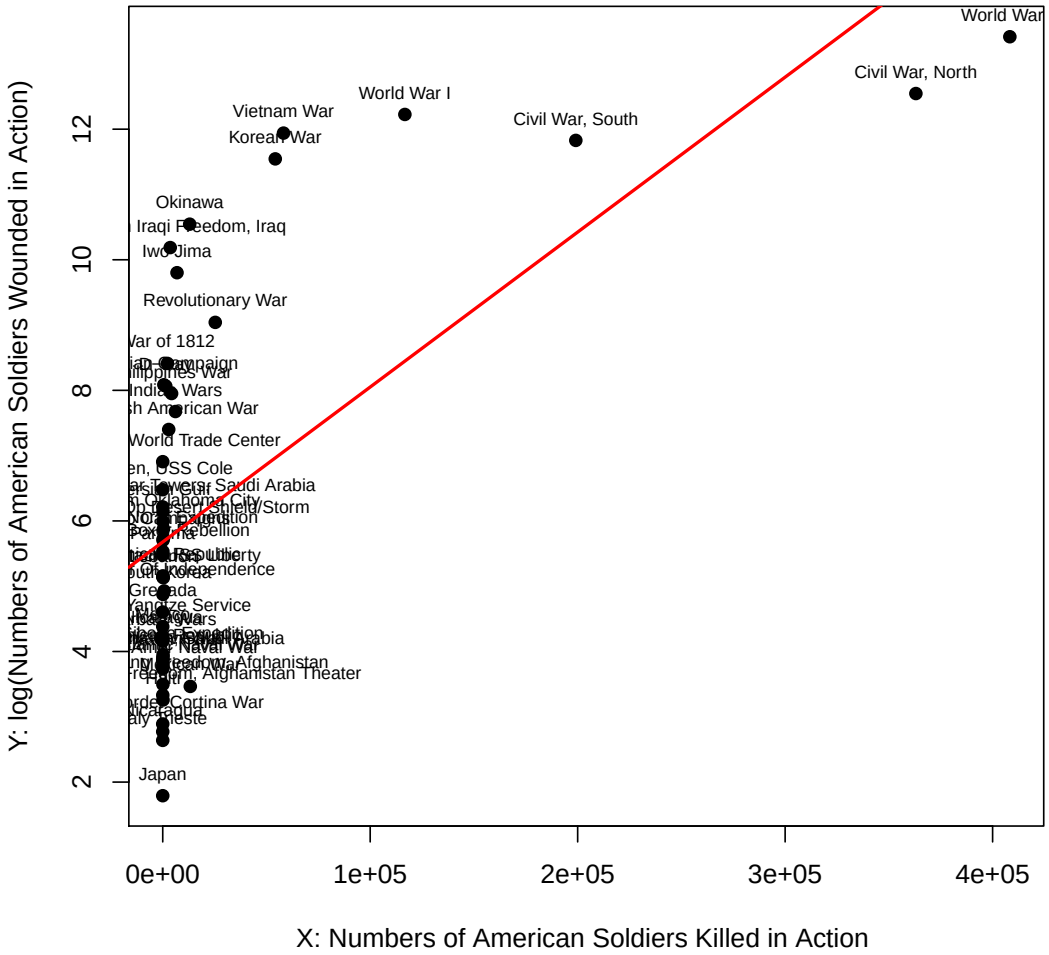
World War II
Civil War, North
World War I
Vietnam War
Civil War, South
Korean War
Okinawa
Operation Iraqi Freedom, Iraq
Iwo Jima
Revolutionary War
War of 1812
Aleutian Campaign
D−Day
Philippines War
Indian Wars
Spanish American War
Terrorism, World Trade Center
Yemen, USS Cole
Terrorism Khobar Towers, Saudi Arabia
Persian Gulf
Terrorism Oklahoma City
Persian Gulf, Op Desert Shield/Storm
Russia North Expedition
Moro Campaigns
China Boxer Rebellion
Panama
Dominican Republic
Israel Attack/USS Liberty
Lebanon
Texas War Of Independence
South Korea
Grenada
China Yangtze Service
Mexico
Nicaragua
Barbary Wars
Russia Siberia Expedition
Dominican Republic
China Civil War
Terrorism Riyadh, Saudi Arabia
North Atlantic Naval War
Franco−Amer Naval War
Operation Enduring Freedom, Afghanistan
Mexican War
Operation Enduring Freedom, Afghanistan Theater
Haiti
Texas Border Cortina War
Nicaragua
Italy Trieste
Japan



Wounded (Logarithmic Scale)

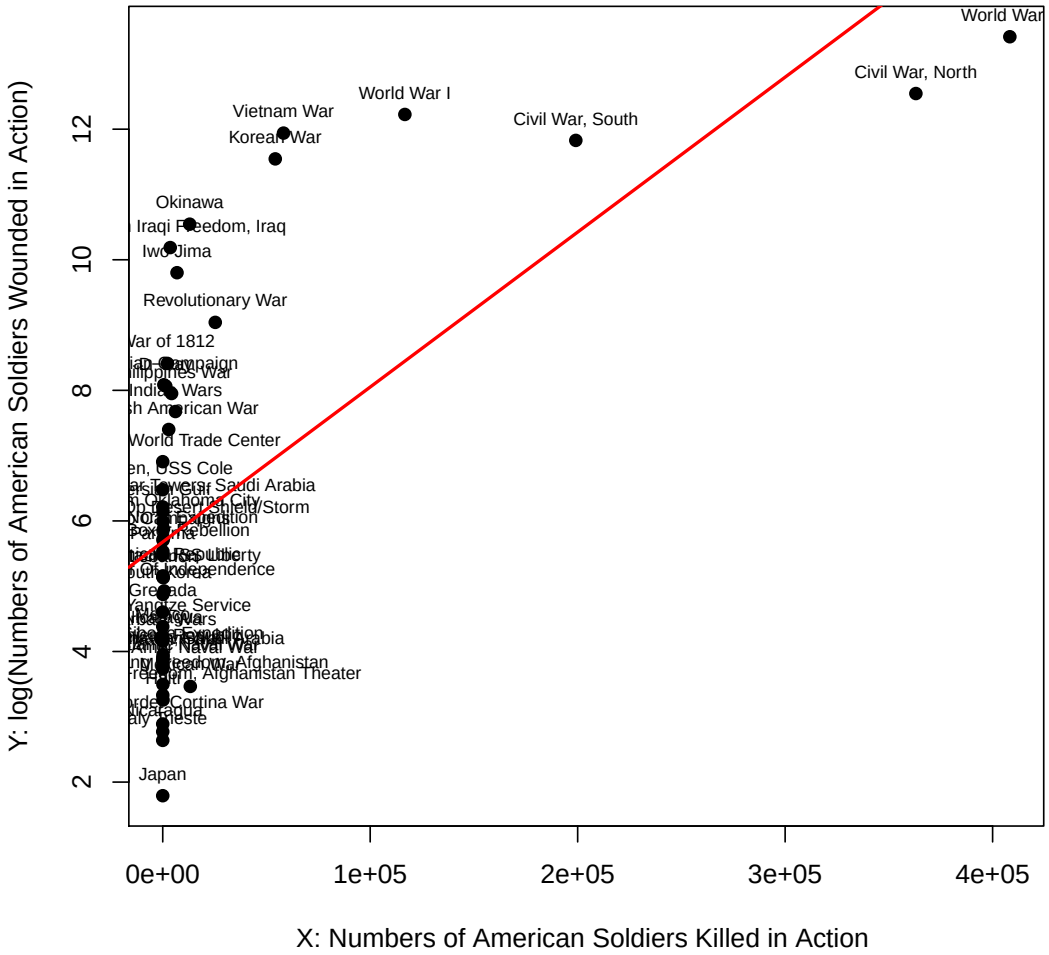


Regression: Log-Level



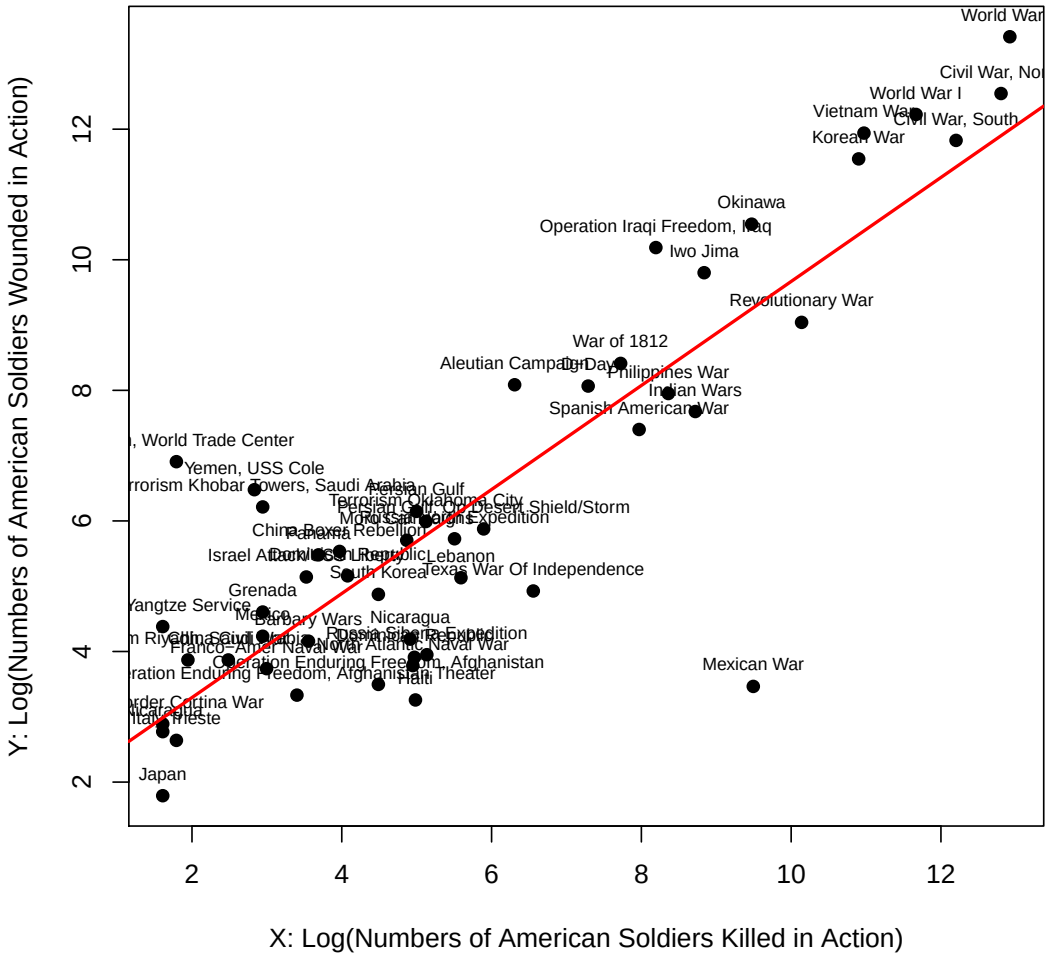
$\hat{\beta}_1 = 0.0000237 \longrightarrow$

Regression: Log-Level



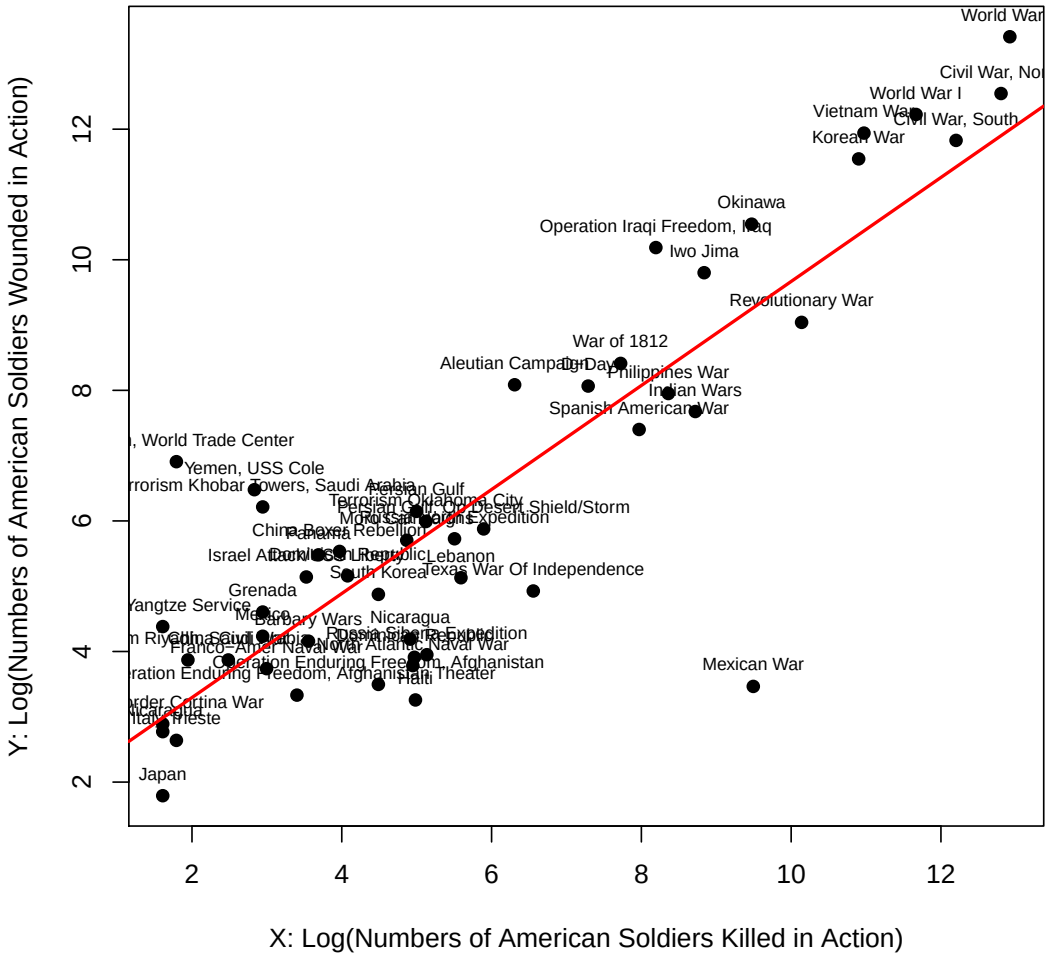
$\hat{\beta}_1 = 0.0000237 \longrightarrow$ One additional soldier killed predicts 0.0023 percent increase in the number of soldiers wounded

Regression: Log-Log



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

Regression: Log-Log



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

Four Most Commonly Used Models

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

Why Does This Approximation Work?

Why Does This Approximation Work?

A useful thing to know is that for small x,

$$\log(1+x) \approx x$$
$$\exp(x) \approx 1+x$$

Why Does This Approximation Work?

A useful thing to know is that for small x,

$$\log(1+x) \approx x$$
$$\exp(x) \approx 1+x$$

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

Why Does This Approximation Work?

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

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

Why Does This Approximation Work?

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

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

We can rewrite this as

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

Why Does This Approximation Work?

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

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

We can rewrite this as

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

Taking natural logs

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

Why Does This Approximation Work?

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

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

We can rewrite this as

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

Taking natural logs

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

Applying our approximation and multiplying by 100 we find,

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

Be Careful: Log-Level with binary X

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

Be Careful: Log-Level with binary X

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

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

Be Careful: Log-Level with binary X

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

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

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

Be Careful: Log-Level with binary X

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

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

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

Be Careful: Log-Level with binary X

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

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

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

Be Careful: Log-Level with binary X

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

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

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

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

Be Careful: Log-Level with binary X

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

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

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

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

Be Careful: Log-Level with binary X

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

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

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

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

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

Interpreting a Logged Outcome

Interpreting a Logged Outcome

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

Interpreting a Logged Outcome

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

Interpreting a Logged Outcome

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

Interpreting a Logged Outcome

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

Interpreting a Logged Outcome

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

Geometric Mean

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

Geometric Mean

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

Geometric Mean

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

Geometric Mean

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

Geometric Mean

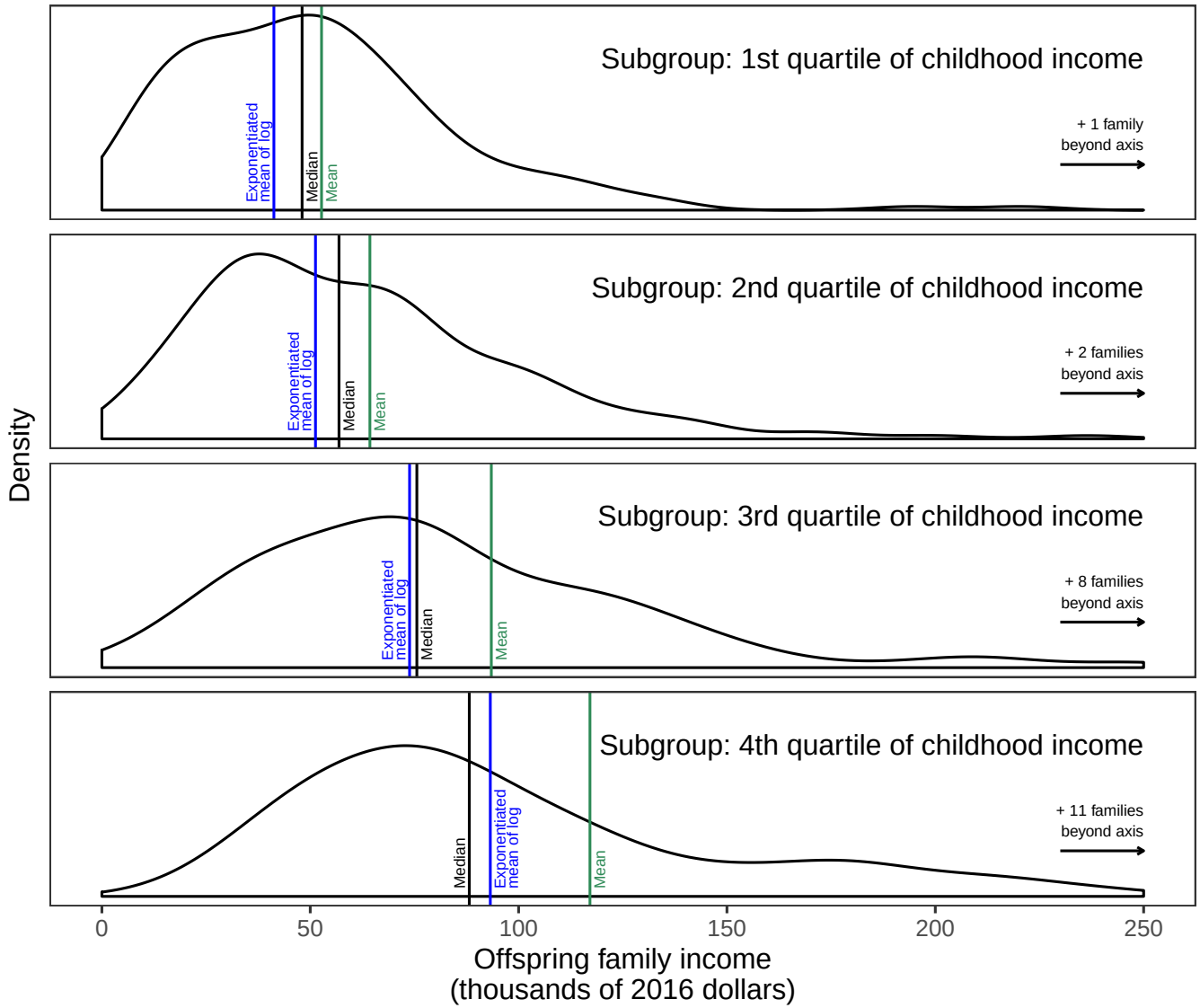
$$
\begin{aligned}
\exp(E(\log(Y))) &= \exp\left(\frac{1}{N}\sum_{i=1}^{N}\log(Y_i)\right) \\
&= \exp\left(\frac{1}{N}\log(\prod_{i=1}^{N}Y_i)\right) \\
&= \exp\left(\log\left(\left(\prod_{i=1}^{N}Y_i\right)^{\frac{1}{N}}\right)\right) \\
&= \left(\prod_{i=1}^{N}Y_i\right)^{\frac{1}{N}} \\
&= \text{Geometric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

Geometric Mean

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

The geometric mean is a robust measure of central tendency.

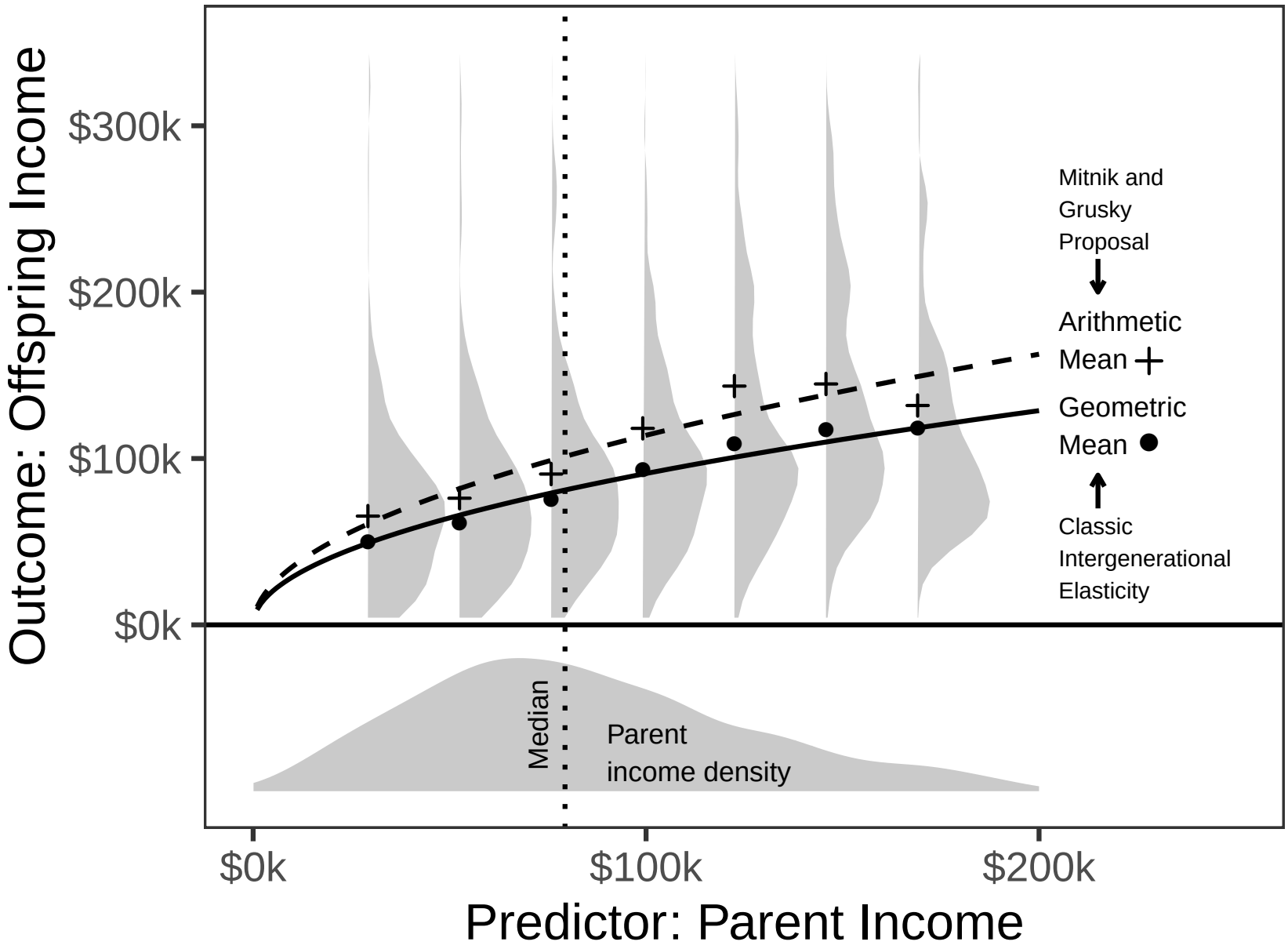
Geometric Mean is Closer to the Median Than the Mean



(Lundberg and Stewart, 2020)

Lundberg and Stewart, 2020

Lundberg and Stewart, 2020



1. Non-Linearity
   - Polynomials
   - LOESS
   - Log Transformations
2. Non-Normality
   - Hat Matrix
   - Standardized and Studentized Residuals
   - QQ Plots
3. Extreme Values
   - Outliers
   - Leverage Points
   - Influence Points
4. Clustering

1. Non-Linearity
   - Polynomials
   - LOESS
   - Log Transformations
2. Non-Normality
   - Hat Matrix
   - Standardized and Studentized Residuals
   - QQ Plots
3. Extreme Values
   - Outliers
   - Leverage Points
   - Influence Points
4. Clustering

Review of the Normality assumption

- In matrix notation:

$$\mathbf{u}|\mathbf{X} \sim \mathcal{N}(0, \sigma_u^2 \mathbf{I})$$

Review of the Normality assumption

- In matrix notation:

$$\mathbf{u}|\mathbf{X} \sim \mathcal{N}(0, \sigma_u^2 \mathbf{I})$$

- Equivalent to:

$$u_i|\mathbf{x}_i' \sim \mathcal{N}(0, \sigma_u^2)$$

Review of the Normality assumption

- In matrix notation:
$$\mathbf{u}|\mathbf{X} \sim \mathcal{N}(0, \sigma_u^2 \mathbf{I})$$
- Equivalent to:
$$u_i|\mathbf{x}_i' \sim \mathcal{N}(0, \sigma_u^2)$$
- Fix $\mathbf{x}_i'$ and the distribution of errors are Normal.

Consequences of non-Normal Errors?

- In small samples:

Consequences of non-Normal Errors?

- In small samples:
  - Sampling distribution of $\widehat{\beta}$ will not be Normal

Consequences of non-Normal Errors?

- In small samples:
  - Sampling distribution of $\widehat{\beta}$ will not be Normal
  - Test statistics will not have t or F distributions

Consequences of non-Normal Errors?

- In small samples:
  - Sampling distribution of $\widehat{\boldsymbol{\beta}}$ will not be Normal
  - Test statistics will not have t or F distributions
  - Probability of Type I error will not be α

Consequences of non-Normal Errors?

- In small samples:
  - Sampling distribution of $\widehat{\boldsymbol{\beta}}$ will not be Normal
  - Test statistics will not have t or F distributions
  - Probability of Type I error will not be α
  - $1 - \alpha$ confidence interval will not have $1 - \alpha$ coverage

Consequences of non-Normal Errors?

- In small samples:
  - Sampling distribution of $\widehat{\boldsymbol{\beta}}$ will not be Normal
  - Test statistics will not have t or F distributions
  - Probability of Type I error will not be α
  - $1 - \alpha$ confidence interval will not have $1 - \alpha$ coverage
- In large samples:

Consequences of non-Normal Errors?

- In small samples:
  - Sampling distribution of $\widehat{\boldsymbol{\beta}}$ will not be Normal
  - Test statistics will not have t or F distributions
  - Probability of Type I error will not be α
  - $1 - \alpha$ confidence interval will not have $1 - \alpha$ coverage
- In large samples:
  - Sampling distribution of $\widehat{\boldsymbol{\beta}} \approx$ Normal by the CLT

Consequences of non-Normal Errors?

- In small samples:
  - Sampling distribution of $\widehat{\boldsymbol{\beta}}$ will not be Normal
  - Test statistics will not have t or F distributions
  - Probability of Type I error will not be α
  - $1 - \alpha$ confidence interval will not have $1 - \alpha$ coverage
- In large samples:
  - Sampling distribution of $\widehat{\boldsymbol{\beta}} \approx$ Normal by the CLT
  - Test statistics will be $\approx t$ or F by the CLT

Consequences of non-Normal Errors?

- In small samples:
  - Sampling distribution of $\widehat{\boldsymbol{\beta}}$ will not be Normal
  - Test statistics will not have t or F distributions
  - Probability of Type I error will not be α
  - $1 - \alpha$ confidence interval will not have $1 - \alpha$ coverage
- In large samples:
  - Sampling distribution of $\widehat{\boldsymbol{\beta}} \approx$ Normal by the CLT
  - Test statistics will be $\approx t$ or F by the CLT
  - Probability of Type I error $\approx \alpha$

Consequences of non-Normal Errors?

- In small samples:
  - Sampling distribution of $\widehat{\boldsymbol{\beta}}$ will not be Normal
  - Test statistics will not have t or F distributions
  - Probability of Type I error will not be α
  - $1 - \alpha$ confidence interval will not have $1 - \alpha$ coverage
- In large samples:
  - Sampling distribution of $\widehat{\boldsymbol{\beta}} \approx$ Normal by the CLT
  - Test statistics will be $\approx t$ or F by the CLT
  - Probability of Type I error $\approx \alpha$
  - $1 - \alpha$ confidence interval will have $\approx 1 - \alpha$ coverage

Consequences of non-Normal Errors?

- In small samples:
  - Sampling distribution of $\widehat{\boldsymbol{\beta}}$ will not be Normal
  - Test statistics will not have t or F distributions
  - Probability of Type I error will not be α
  - $1 - \alpha$ confidence interval will not have $1 - \alpha$ coverage
- In large samples:
  - Sampling distribution of $\widehat{\boldsymbol{\beta}} \approx$ Normal by the CLT
  - Test statistics will be $\approx t$ or F by the CLT
  - Probability of Type I error $\approx \alpha$
  - $1 - \alpha$ confidence interval will have $\approx 1 - \alpha$ coverage
- The sample size (n) needed for approximation to hold depends on how far the errors are from Normal.

Consequences of non-Normal Errors?

- In small samples:
  - Sampling distribution of $\widehat{\boldsymbol{\beta}}$ will not be Normal
  - Test statistics will not have t or F distributions
  - Probability of Type I error will not be α
  - $1 - \alpha$ confidence interval will not have $1 - \alpha$ coverage
- In large samples:
  - Sampling distribution of $\widehat{\boldsymbol{\beta}} \approx$ Normal by the CLT
  - Test statistics will be $\approx t$ or F by the CLT
  - Probability of Type I error $\approx \alpha$
  - $1 - \alpha$ confidence interval will have $\approx 1 - \alpha$ coverage
- The sample size (n) needed for approximation to hold depends on how far the errors are from Normal.

Consequences of non-Normal Errors?

- In small samples:
  - Sampling distribution of $\widehat{\boldsymbol{\beta}}$ will not be Normal
  - Test statistics will not have t or F distributions
  - Probability of Type I error will not be α
  - $1-\alpha$ confidence interval will not have $1-\alpha$ coverage
- In large samples:
  - Sampling distribution of $\widehat{\boldsymbol{\beta}} \approx$ Normal by the CLT
  - Test statistics will be $\approx t$ or F by the CLT
  - Probability of Type I error $\approx \alpha$
  - $1-\alpha$ confidence interval will have $\approx 1-\alpha$ coverage
- The sample size (n) needed for approximation to hold depends on how far the errors are from Normal.
- Reasonable question: if we have enough data are non-normal errors even a problem?

Clarifying a Point of Confusion: Marginal versus Conditional

- Be careful with this assumption: distribution of the error, not the distribution of the outcome is the key assumption

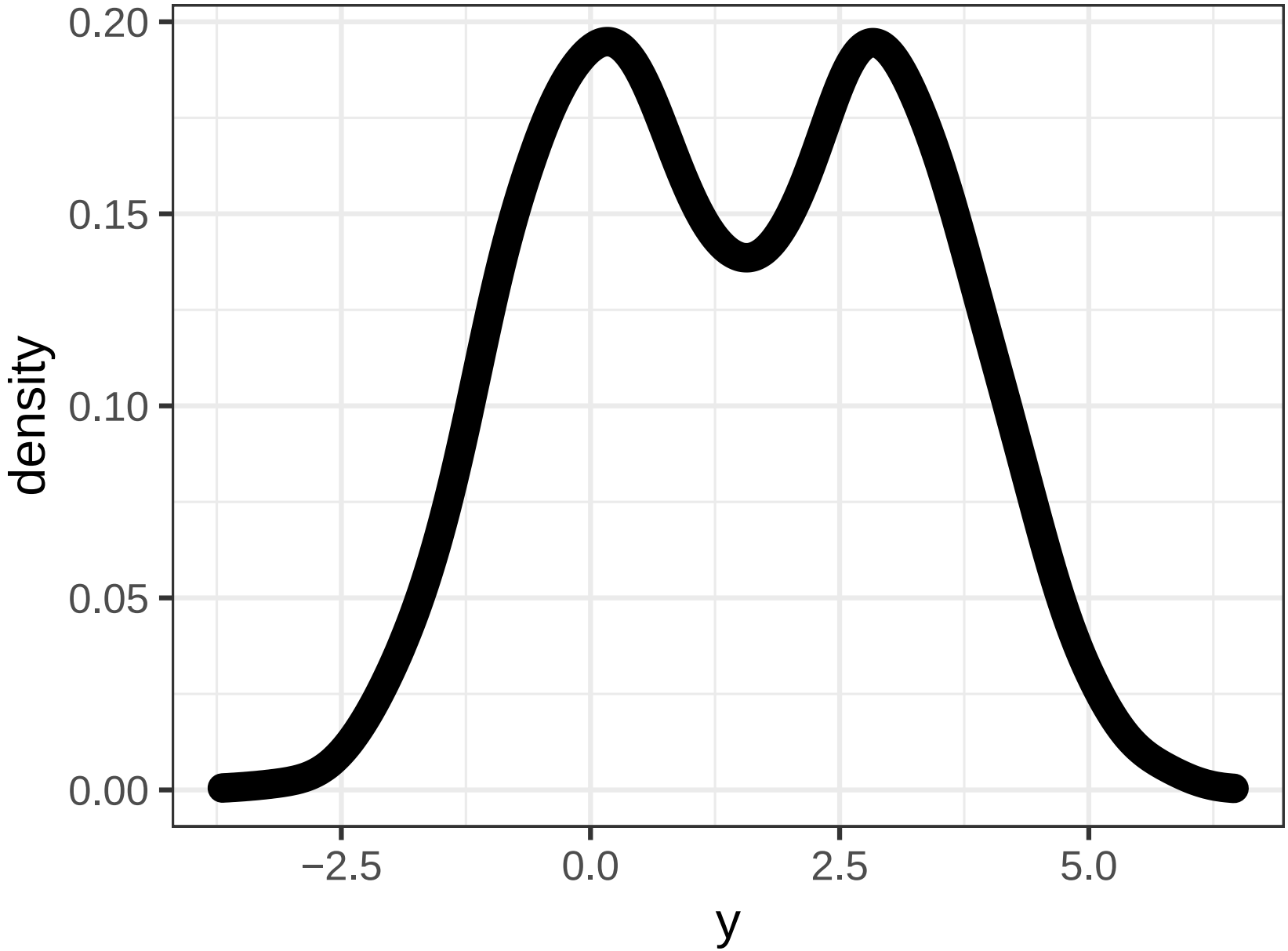
Clarifying a Point of Confusion: Marginal versus Conditional

- Be careful with this assumption: distribution of the error, not the distribution of the outcome is the key assumption
- The marginal distribution of y can be non-Normal even if the conditional distribution is Normal!

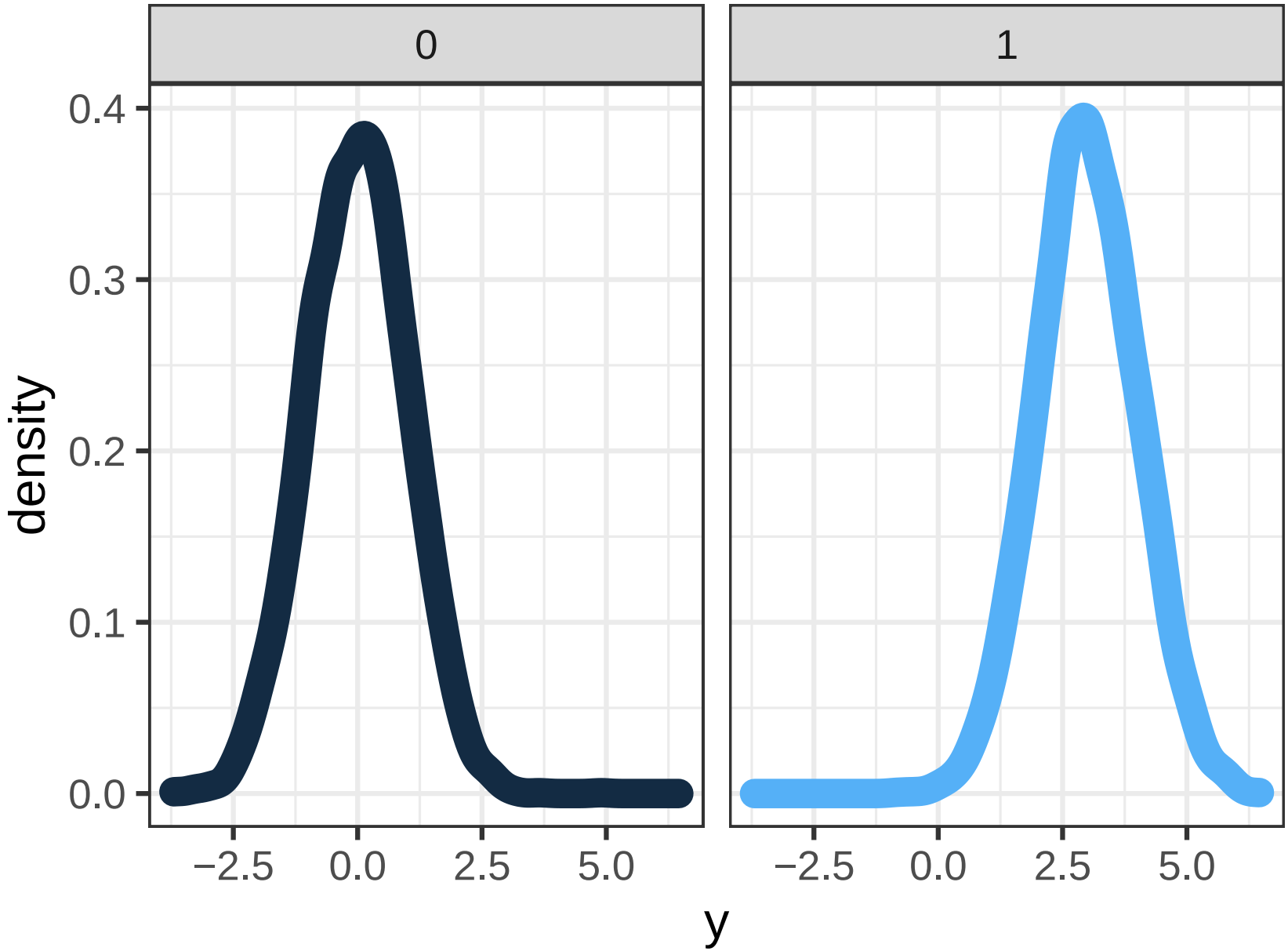
Clarifying a Point of Confusion: Marginal versus Conditional

- Be careful with this assumption: distribution of the error, not the distribution of the outcome is the key assumption
- The marginal distribution of y can be non-Normal even if the conditional distribution is Normal!
- The plausibility depends on the X chosen by the researcher.

Example: Is this a Violation?



Example: Is this a Violation?



How to Diagnose?

- Assumption is about unobserved errors $\mathbf{u} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$

How to Diagnose?

- Assumption is about unobserved errors $\mathbf{u} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$
- We can only observe residuals, $\widehat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$

How to Diagnose?

- Assumption is about unobserved errors $\mathbf{u} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$
- We can only observe residuals, $\widehat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$
- If distribution of residuals $\approx$ distribution of errors, we could check residuals as a proxy for the errors.

How to Diagnose?

- Assumption is about unobserved errors $\mathbf{u} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$
- We can only observe residuals, $\widehat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$
- If distribution of residuals $\approx$ distribution of errors, we could check residuals as a proxy for the errors.
- Unfortunately, this is not true—the distribution of the residuals is more complicated

How to Diagnose?

- Assumption is about unobserved errors $\mathbf{u} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$
- We can only observe residuals, $\widehat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$
- If distribution of residuals $\approx$ distribution of errors, we could check residuals as a proxy for the errors.
- Unfortunately, this is not true—the distribution of the residuals is more complicated

How to Diagnose?

- Assumption is about unobserved errors $\mathbf{u} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$
- We can only observe residuals, $\widehat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$
- If distribution of residuals $\approx$ distribution of errors, we could check residuals as a proxy for the errors.
- Unfortunately, this is not true—the distribution of the residuals is more complicated

Solution: Carefully investigate the residuals numerically and graphically.

How to Diagnose?

- Assumption is about unobserved errors $\mathbf{u} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$
- We can only observe residuals, $\widehat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$
- If distribution of residuals $\approx$ distribution of errors, we could check residuals as a proxy for the errors.
- Unfortunately, this is not true—the distribution of the residuals is more complicated

Solution: Carefully investigate the residuals numerically and graphically.

To understand the relationship between residuals and errors, we need to derive the distribution of the residuals (which we will do over the next few slides).

Defining the Hat Matrix

Defining the Hat Matrix

- We want to figure out how the distribution of errors $\mathbf{u}$ relates to the distribution of residuals $\widehat{\mathbf{u}}$.

Defining the Hat Matrix

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

Defining the Hat Matrix

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

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

Defining the Hat Matrix

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

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

Defining the Hat Matrix

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

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

Defining the Hat Matrix

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

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

Defining the Hat Matrix

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

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

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

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

Defining the Hat Matrix

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

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

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

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

it has a few nice properties:

Defining the Hat Matrix

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

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

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

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

it has a few nice properties:

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

Defining the Hat Matrix

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

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

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

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

it has a few nice properties:

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

Relating the Residuals to the Errors

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

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

Relating the Residuals to the Errors

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

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

Relating the Residuals to the Errors

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

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

Relating the Residuals to the Errors

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

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

Relating the Residuals to the Errors

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

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

Relating the Residuals to the Errors

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

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

Relating the Residuals to the Errors

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

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

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

Relating the Residuals to the Errors

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

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

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

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

Relating the Residuals to the Errors

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

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

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

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

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

Characterizing the Distribution of the Residuals

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

Characterizing the Distribution of the Residuals

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

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

Characterizing the Distribution of the Residuals

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

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

Characterizing the Distribution of the Residuals

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

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

Characterizing the Distribution of the Residuals

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

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

Characterizing the Distribution of the Residuals

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

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

Characterizing the Distribution of the Residuals

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

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

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

Properties of the Distribution of Residuals

Properties of the Distribution of Residuals

Notice in contrast to the unobserved errors, the estimated residuals have some different properties. They

Properties of the Distribution of Residuals

Notice in contrast to the unobserved errors, the estimated residuals have some different properties. They

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

Properties of the Distribution of Residuals

Notice in contrast to the unobserved errors, the estimated residuals have some different properties. They

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

Properties of the Distribution of Residuals

Notice in contrast to the unobserved errors, the estimated residuals have some different properties. They

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

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

Properties of the Distribution of Residuals

Notice in contrast to the unobserved errors, the estimated residuals have some different properties. They

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

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

Properties of the Distribution of Residuals

Notice in contrast to the unobserved errors, the estimated residuals have some different properties. They

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

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

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

Standardized Residuals

Standardized Residuals

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

Standardized Residuals

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

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

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

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

Standardized Residuals

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

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

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

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

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

Studentized Residuals

Studentized Residuals

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

Studentized Residuals

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

- estimate error variance without residual i:

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

Studentized Residuals

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

- estimate error variance without residual i:

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

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

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

Studentized Residuals

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

- estimate error variance without residual i:

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

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

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

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

Studentized Residuals

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

- estimate error variance without residual i:

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

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

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

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

Example: Buchanan votes in Florida, 2000

Example: Buchanan votes in Florida, 2000

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

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

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

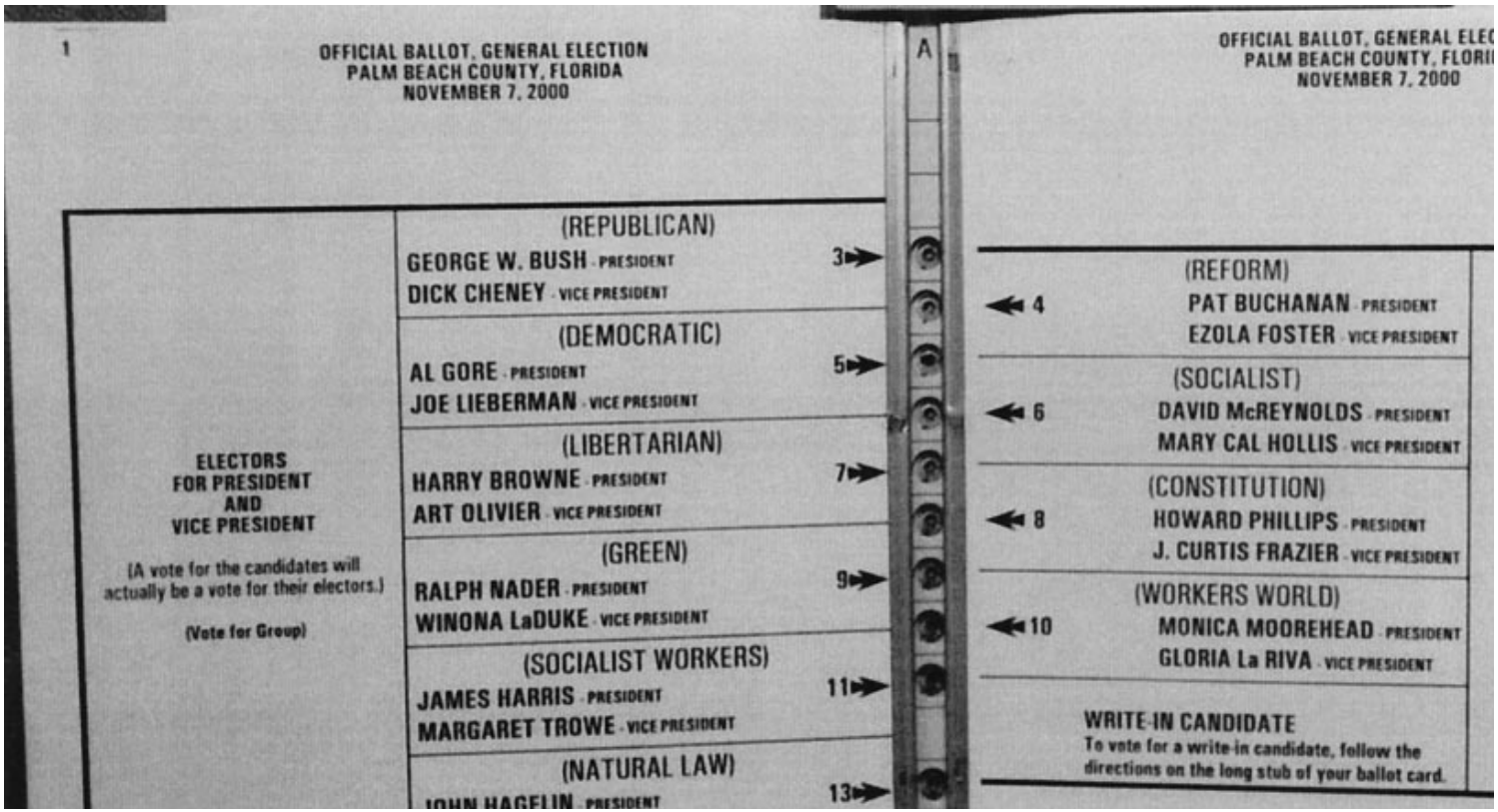
JONATHAN N. WAND *Cornell University*
KENNETH W. SHOTTS *Northwestern University*
JASJEET S. SEKHON *Harvard University*
WALTER R. MEBANE, JR. *Cornell University*
MICHAEL C. HERRON *Northwestern University*
HENRY E. BRADY *University of California, Berkeley*

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

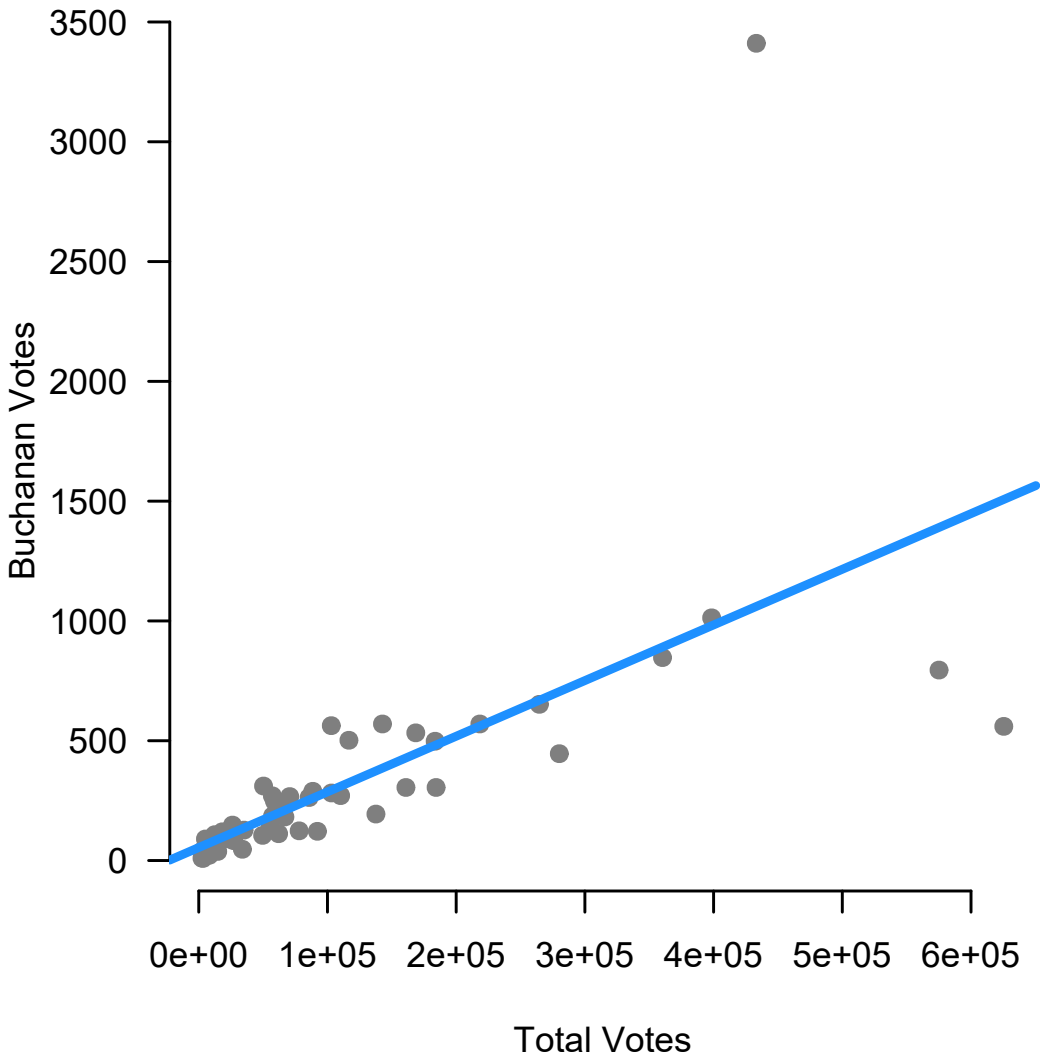
Example: Buchanan votes in Florida, 2000

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

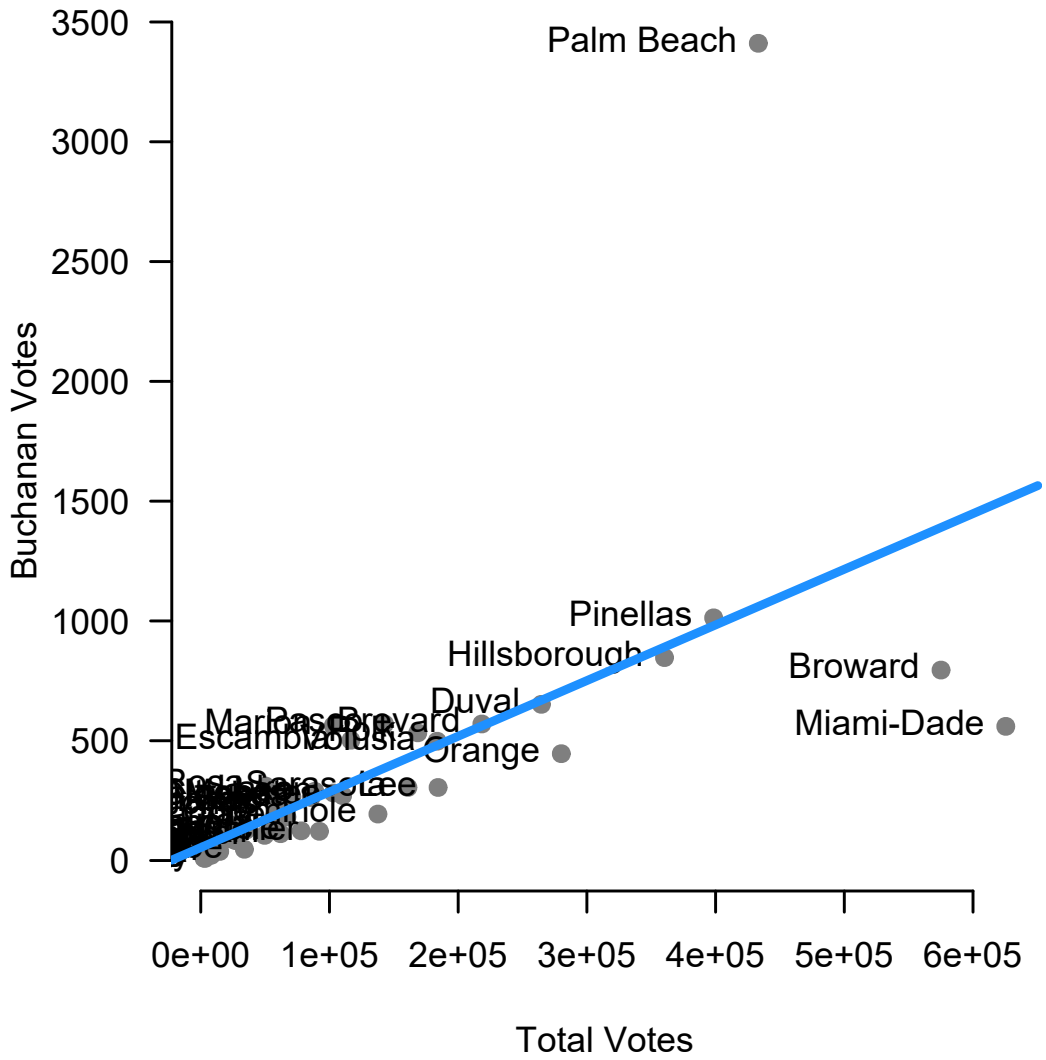
FIGURE 1. The Palm Beach County Bufferfly Ballot



Example: Buchanan votes in Florida, 2000



Example: Buchanan votes in Florida, 2000



Example: Buchanan Votes in Florida

Example: Buchanan Votes in Florida

- Now that our studentized residuals follow a known standard distribution, we can proceed with diagnostic analysis for the nonnormal errors.

Example: Buchanan Votes in Florida

- Now that our studentized residuals follow a known standard distribution, we can proceed with diagnostic analysis for the nonnormal errors.
- We examine data from the 2000 presidential election in Florida used in Wand et al. (2001).

Example: Buchanan Votes in Florida

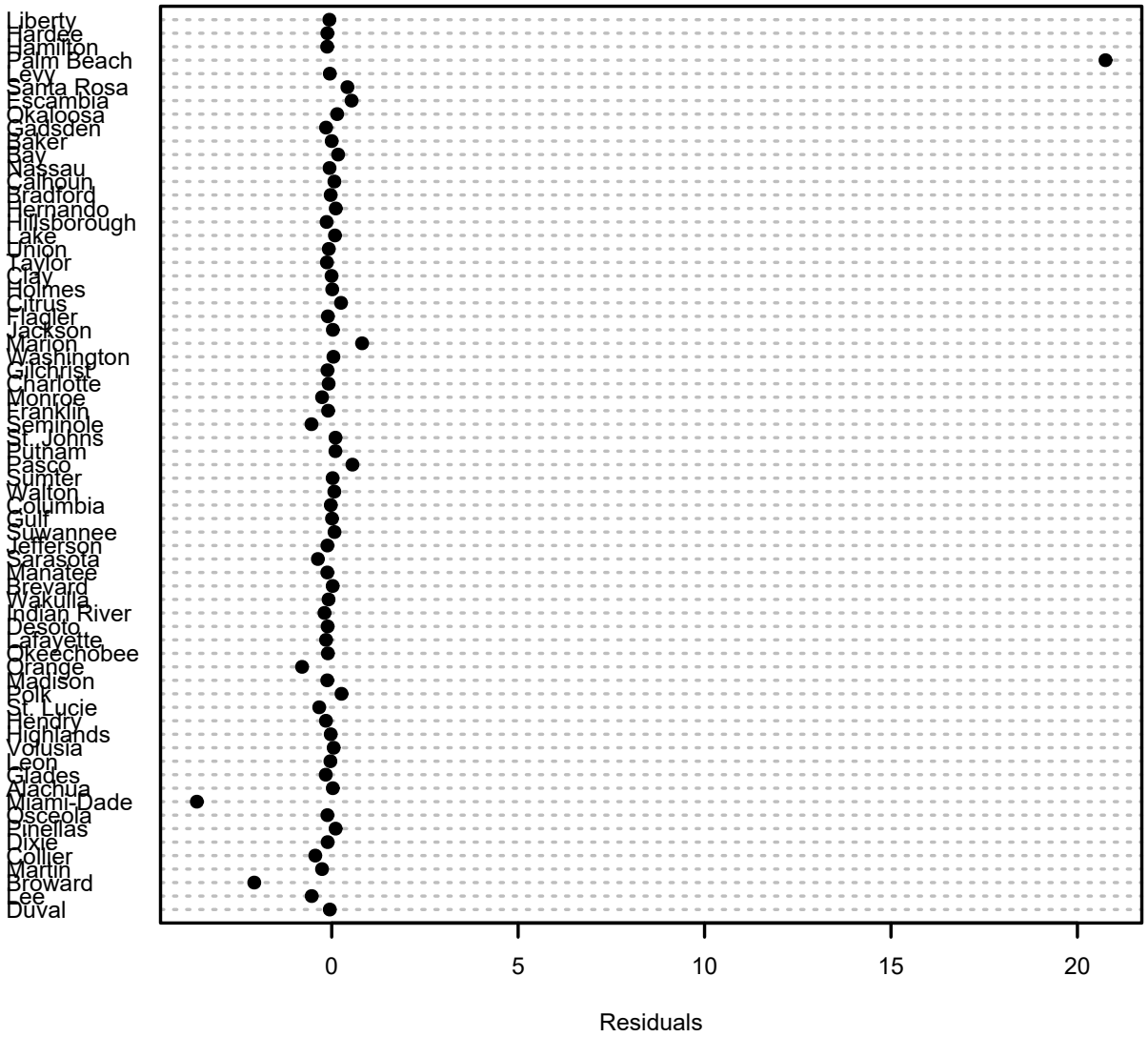
- Now that our studentized residuals follow a known standard distribution, we can proceed with diagnostic analysis for the nonnormal errors.
- We examine data from the 2000 presidential election in Florida used in Wand et al. (2001).
- Our analysis takes place at the county level and we will regress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

Buchanan Votes and Total Votes

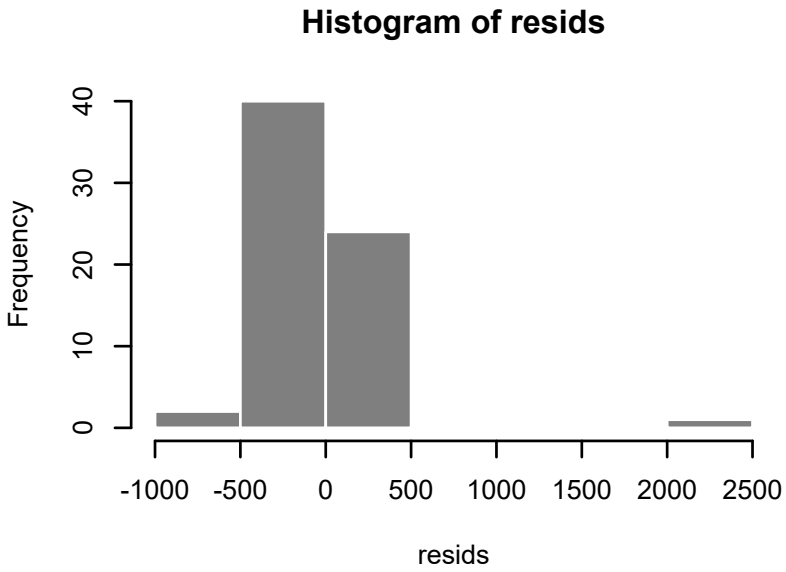
R Code

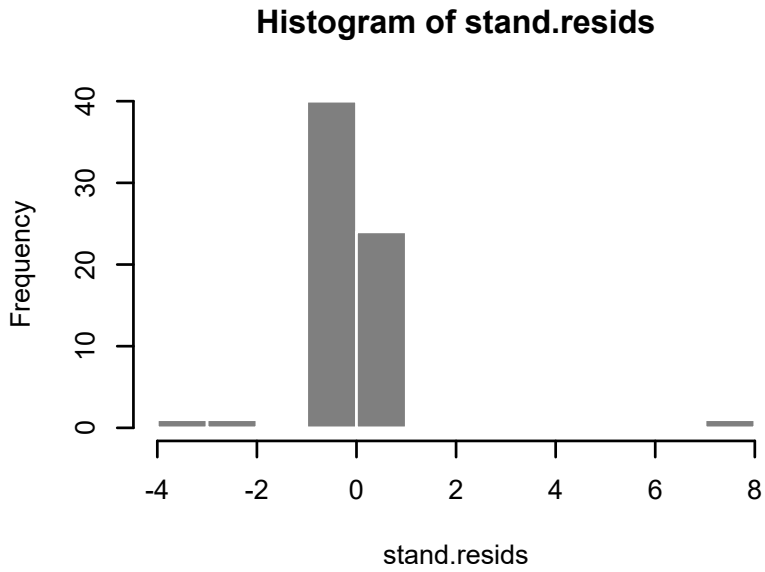
```
> mod1 <- lm(buchanan00~TotalVotes00,data=dta)
> summary(mod1)
Residuals:
    Min      1Q  Median      3Q     Max
-947.05  -41.74  -19.47   20.20 2350.54

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

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

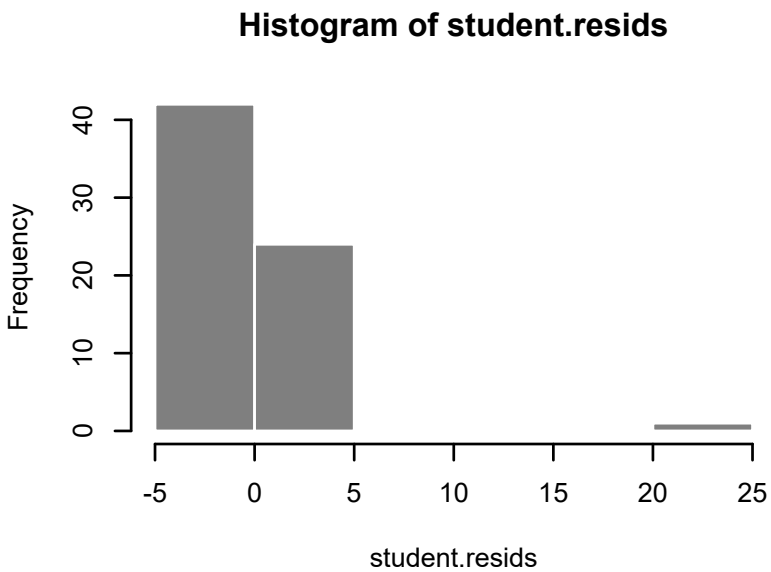
Plotting the residuals



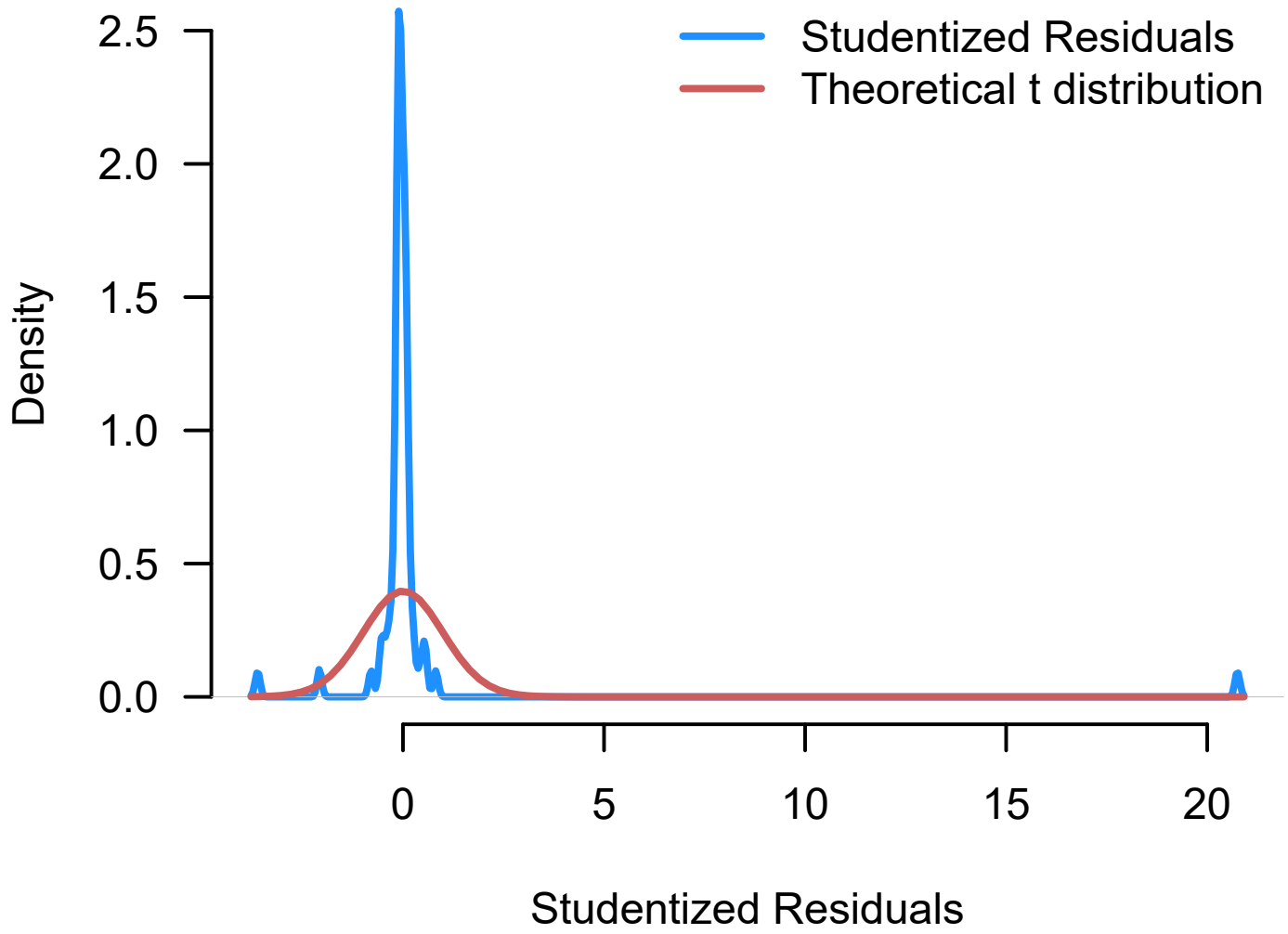
Plotting the residuals







Plotting the residuals



Quantile-Quantile plots

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

Quantile-Quantile plots

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

Quantile-Quantile plots

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

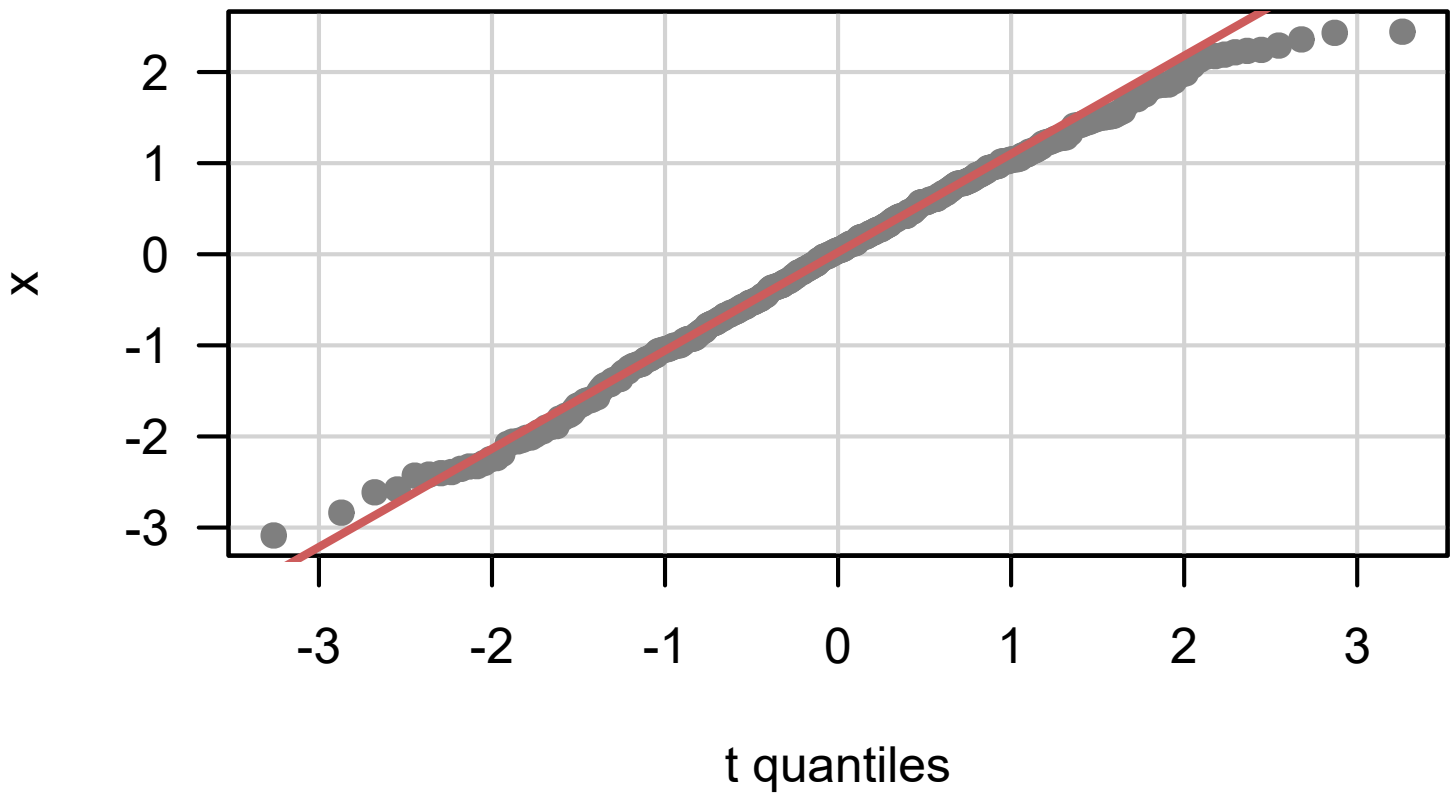
Quantile-Quantile plots

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

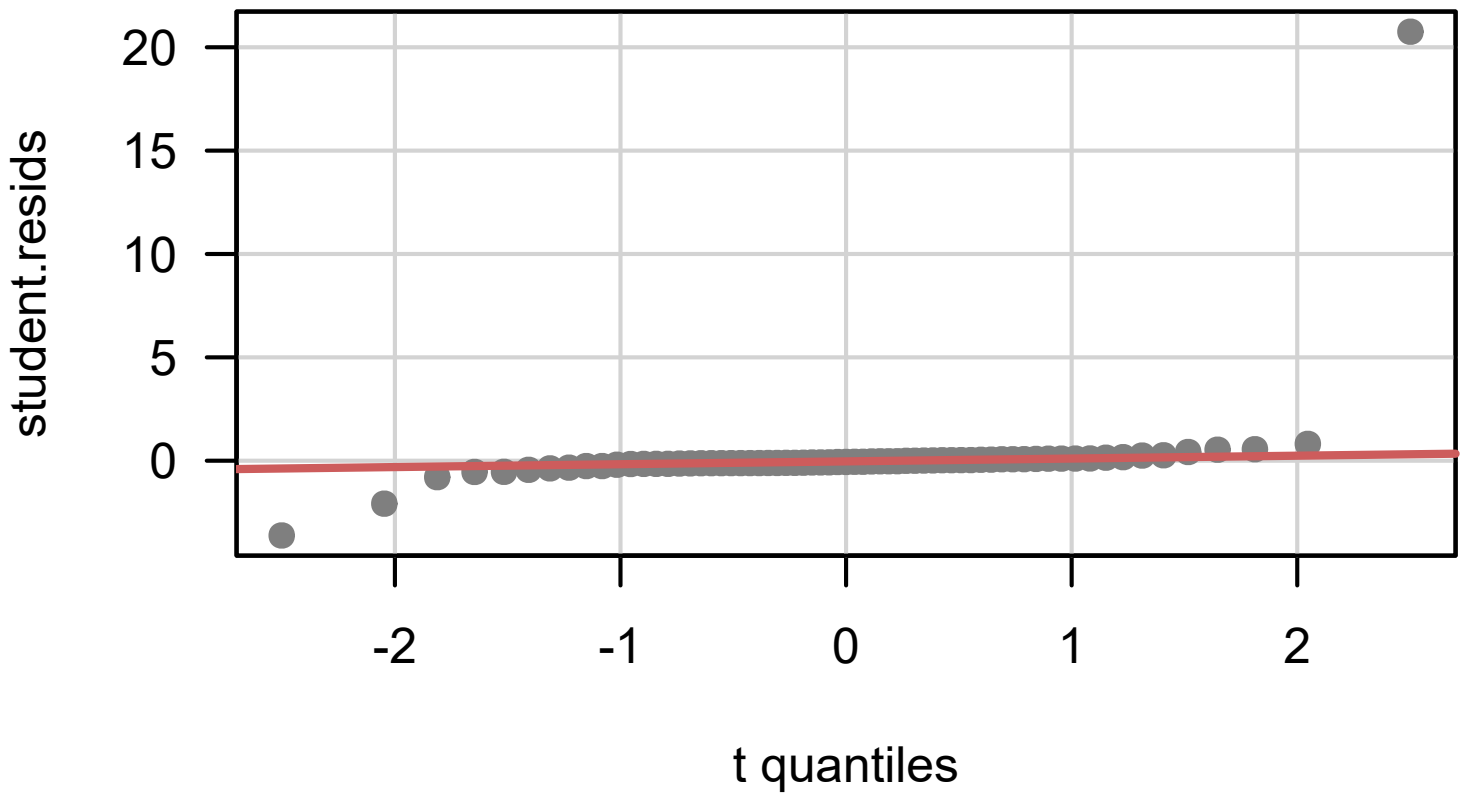
Quantile-Quantile plots

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

Good QQ-plot



Buchanan QQ-plot



How Can we Deal with non-Normal Errors?

How Can we Deal with non-Normal Errors?

- Drop or change problematic observations (could be a bad idea unless you have some reason to believe the data are wrong or corrupted)

How Can we Deal with non-Normal Errors?

- Drop or change problematic observations (could be a bad idea unless you have some reason to believe the data are wrong or corrupted)
- Add variables to $\mathbf{X}$ (remember that the errors are defined in terms of explanatory variables)

How Can we Deal with non-Normal Errors?

- Drop or change problematic observations (could be a bad idea unless you have some reason to believe the data are wrong or corrupted)
- Add variables to $\mathbf{X}$ (remember that the errors are defined in terms of explanatory variables)
- Use transformations (this may work, but a transformation affects all the assumptions of the model)

How Can we Deal with non-Normal Errors?

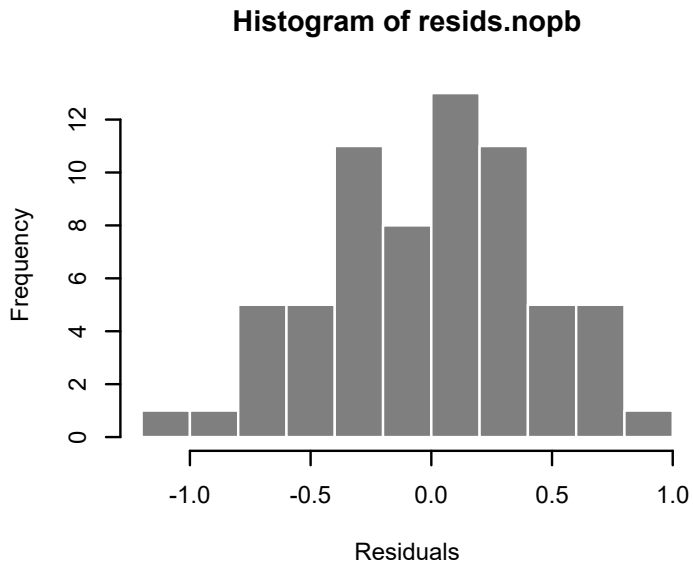
- Drop or change problematic observations (could be a bad idea unless you have some reason to believe the data are wrong or corrupted)
- Add variables to $\mathbf{X}$ (remember that the errors are defined in terms of explanatory variables)
- Use transformations (this may work, but a transformation affects all the assumptions of the model)
- Use estimators other than OLS that are robust to nonnormality (skipped this year- but see materials from prior years)

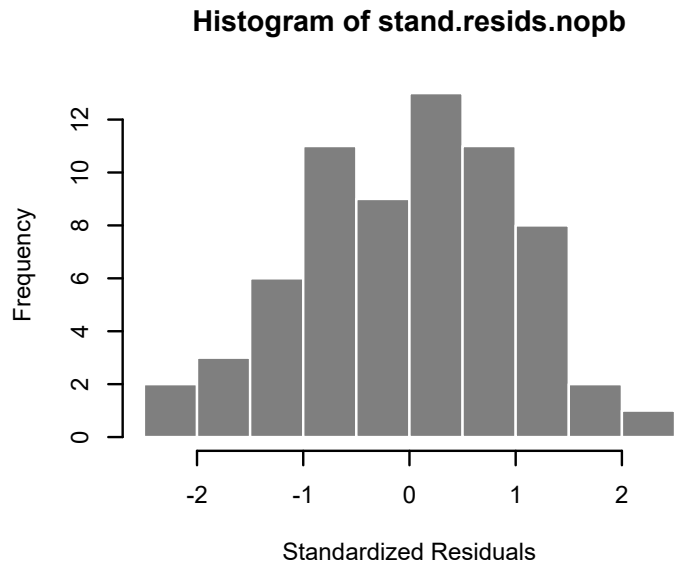
Buchanan Revisited

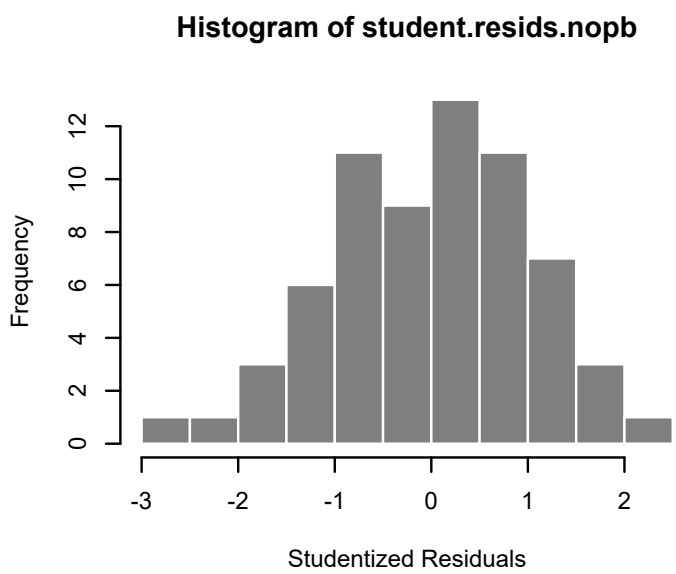
Let's delete Palm Beach and also use log transformations for both variables

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

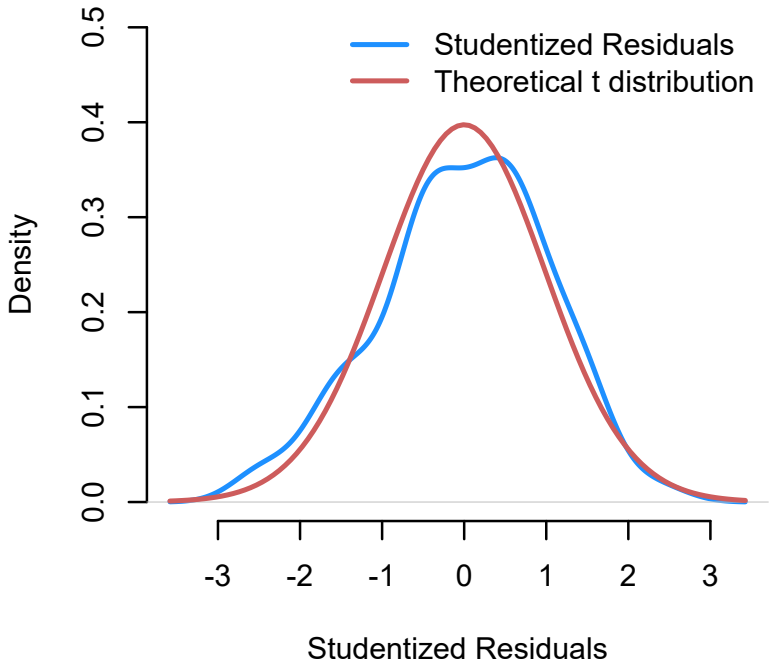
Buchanan Revisited

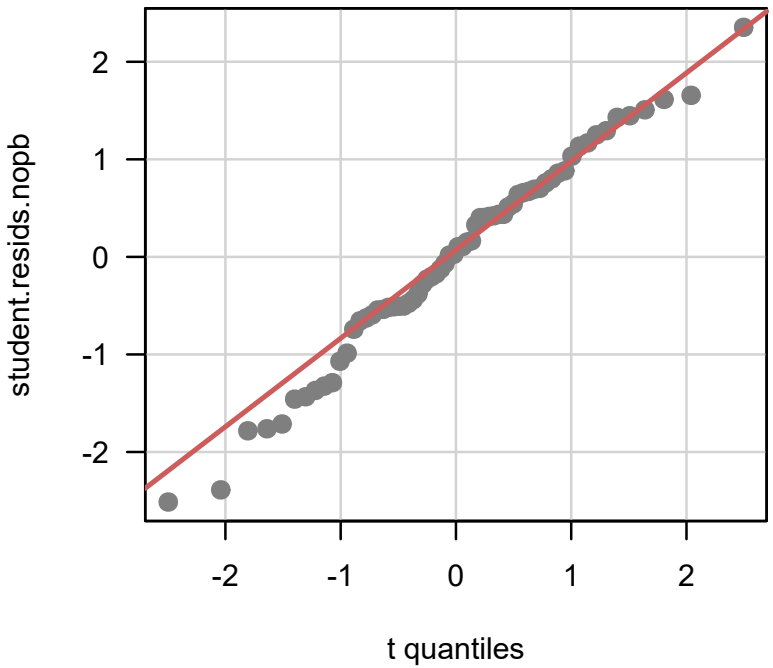






Buchanan Revisited





1. Non-Linearity
   - Polynomials
   - LOESS
   - Log Transformations
2. Non-Normality
   - Hat Matrix
   - Standardized and Studentized Residuals
   - QQ Plots
3. Extreme Values
   - Outliers
   - Leverage Points
   - Influence Points
4. Clustering

1. Non-Linearity
   - Polynomials
   - LOESS
   - Log Transformations
2. Non-Normality
   - Hat Matrix
   - Standardized and Studentized Residuals
   - QQ Plots
3. Extreme Values
   - Outliers
   - Leverage Points
   - Influence Points
4. Clustering

The Trouble with Norway

The Trouble with Norway

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

The Trouble with Norway

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

The Trouble with Norway

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

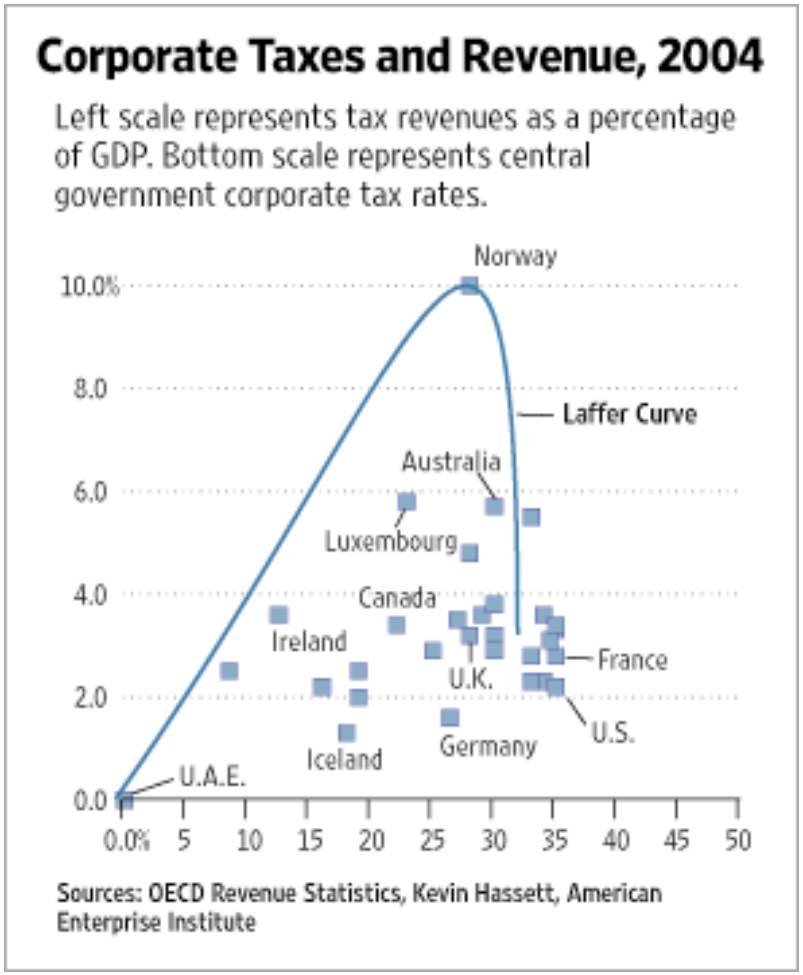
The Trouble with Norway

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

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

Creative Curve Fitting with Norway

Creative Curve Fitting with Norway



The Most Important Lesson: Check Your Data

The Most Important Lesson: Check Your Data

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

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

Draper and Smith (1981, p. 418)

The Most Important Lesson: Check Your Data

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

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

Draper and Smith (1981, p. 418)

Carefully Examine the Data First!!

1. Examine summary statistics: `summary(data)`
2. Scatterplot matrix for densities and bivariate relationships:
   E.g. `scatterplotMatrix(data)` from `car` library.
3. Further conditional plots for multivariate data:
   E.g. `ggplot2`

Three Types of Extreme Values

Three Types of Extreme Values

1. Outlier: extreme in the y direction

Three Types of Extreme Values

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

Three Types of Extreme Values

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

Three Types of Extreme Values

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

- Not all of these are problematic

Three Types of Extreme Values

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

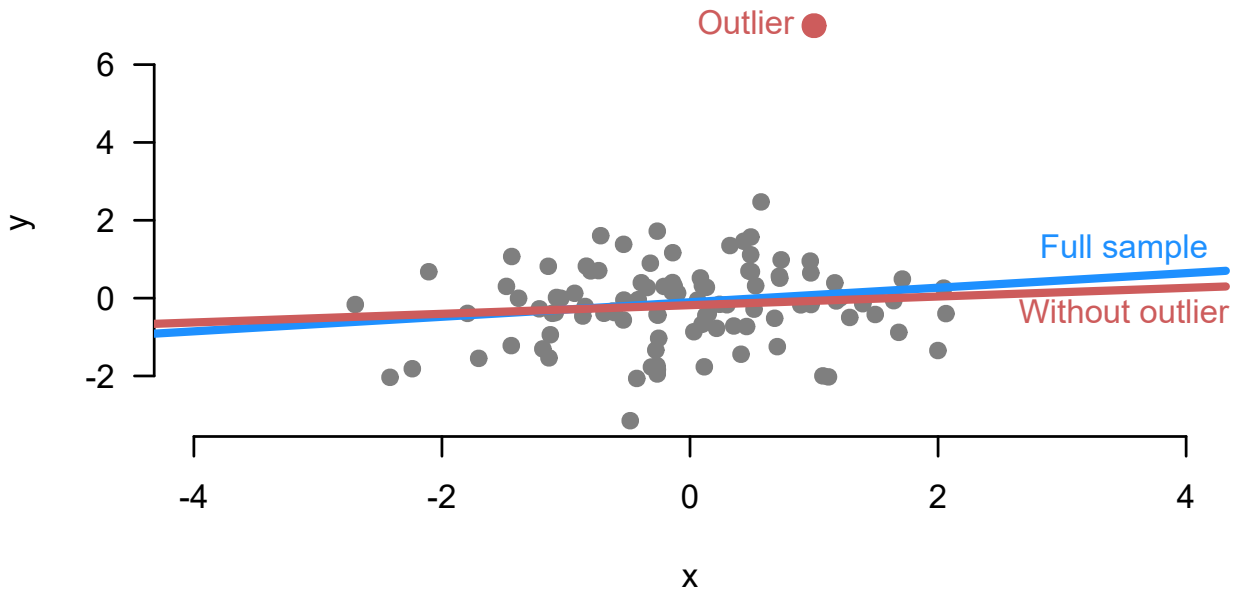
- Not all of these are problematic
- If the data are truly "contaminated" (come from a different distribution), can cause inefficiency and possibly bias

Three Types of Extreme Values

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

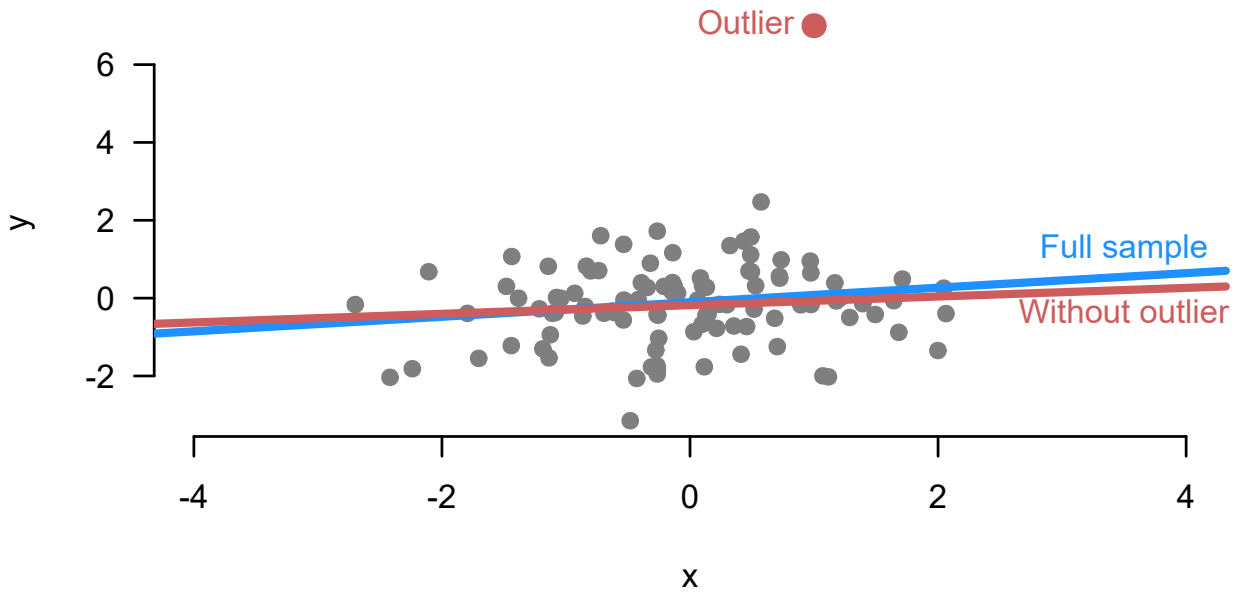
- Not all of these are problematic
- If the data are truly "contaminated" (come from a different distribution), can cause inefficiency and possibly bias
- Can be a violation of iid (not identically distributed)

Outlier Definition



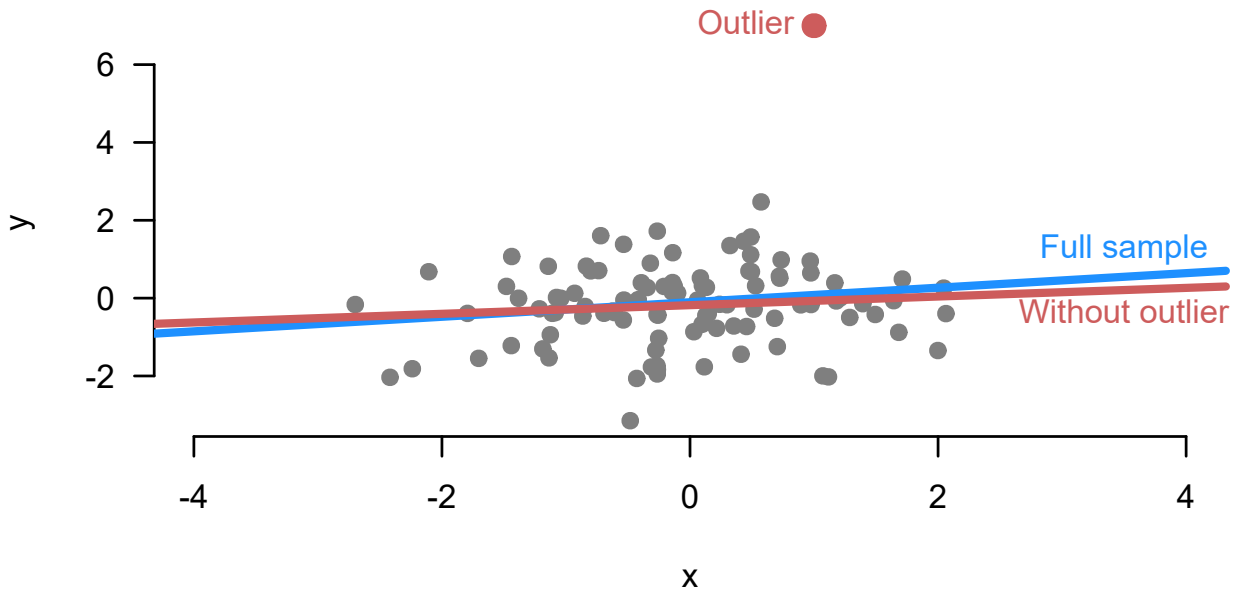
- An **outlier** is a data point with very large regression errors, u_i

Outlier Definition



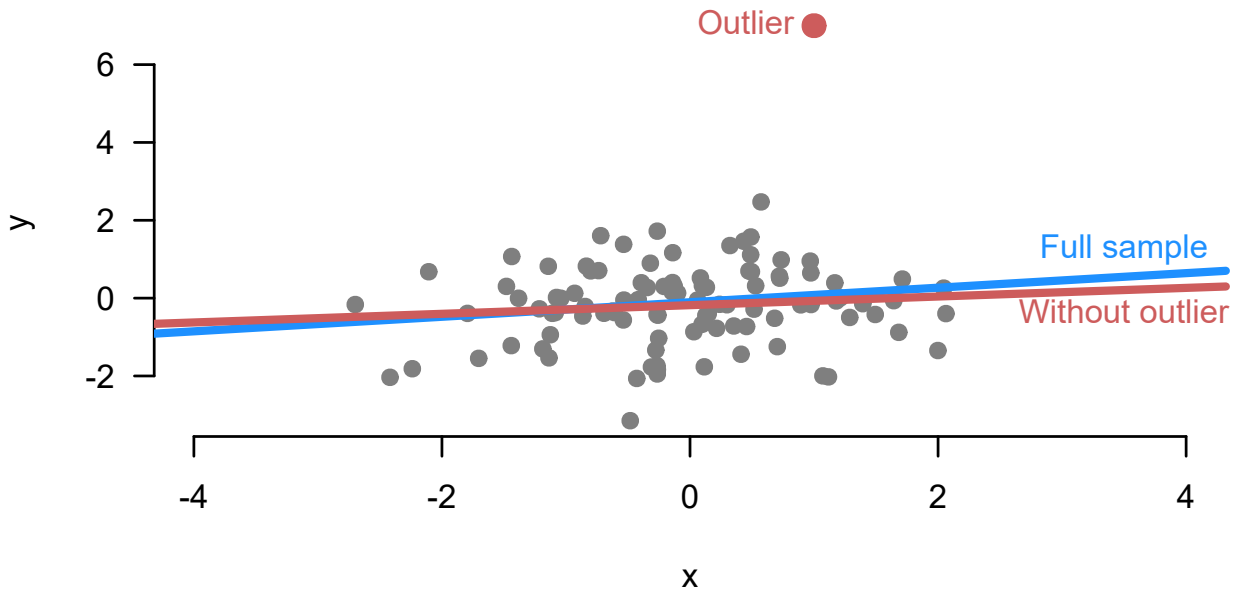
- An **outlier** is a data point with very large regression errors, u_i
- Very distant from the rest of the data in the y-dimension

Outlier Definition



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

Outlier Definition



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

Detecting Outliers

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

Detecting Outliers

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

Detecting Outliers

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

Detecting Outliers

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

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

Detecting Outliers

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

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

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

Cutoff Rules for Outliers

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

Cutoff Rules for Outliers

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

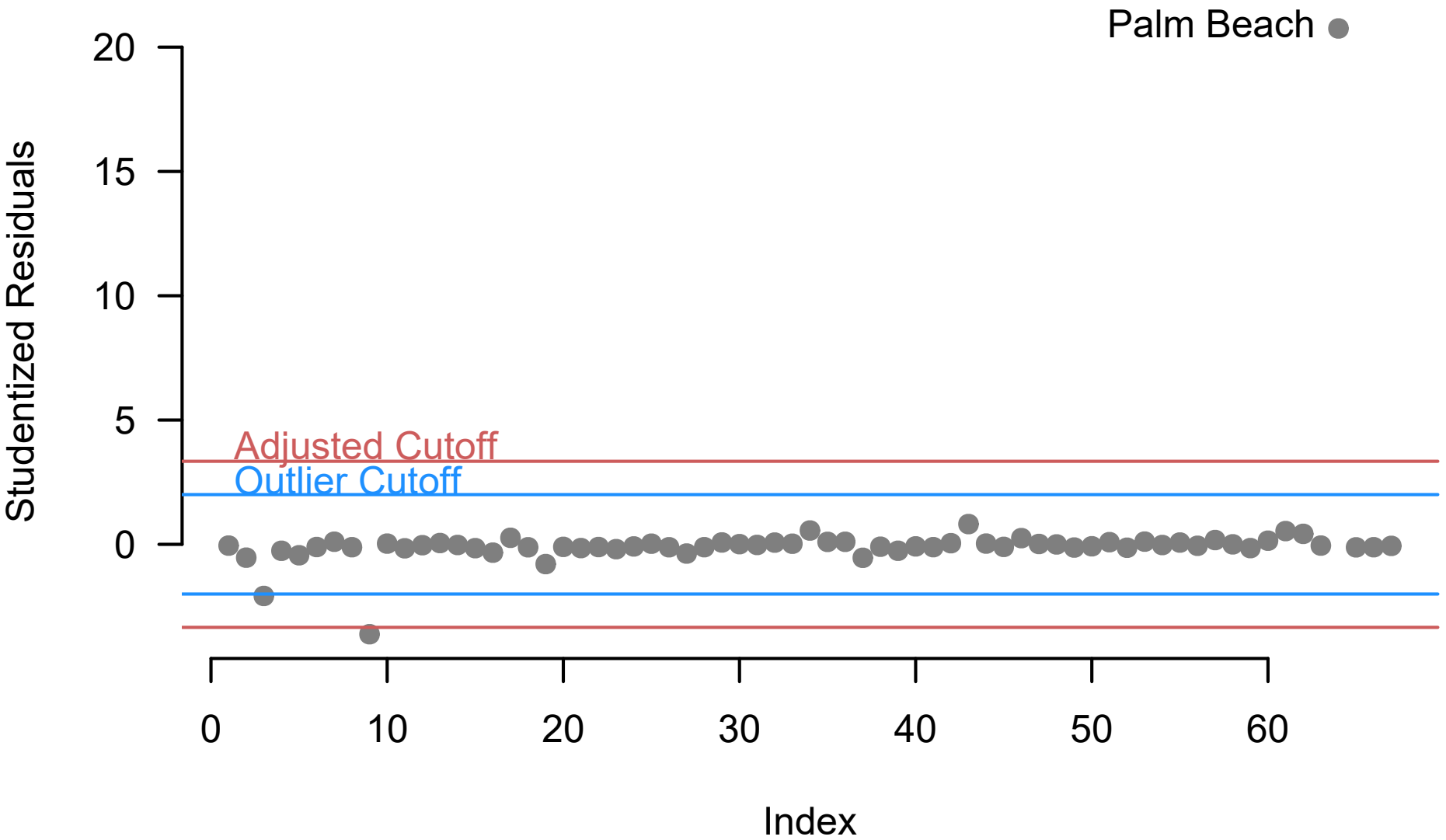
Cutoff Rules for Outliers

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

Cutoff Rules for Outliers

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

Buchanan outliers



What to do about outliers

- Is the data corrupted?

What to do about outliers

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

What to do about outliers

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

What to do about outliers

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

What to do about outliers

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

What to do about outliers

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

What to do about outliers

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

What to do about outliers

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

A Cautionary Tale: The "Discovery" of the Ozone Hole

A Cautionary Tale: The "Discovery" of the Ozone Hole

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

A Cautionary Tale: The "Discovery" of the Ozone Hole

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

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

A Cautionary Tale: The "Discovery" of the Ozone Hole

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

A Cautionary Tale: The "Discovery" of the Ozone Hole

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

A Sociological Cautionary Tale

Comment on Herring, ASR, April 2009

AMERICAN SOCIOLOGICAL ASSOCIATION

American Sociological Review
2017, Vol. 82(4) 857–867

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

Does Diversity Pay? A Replication of Herring (2009)

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

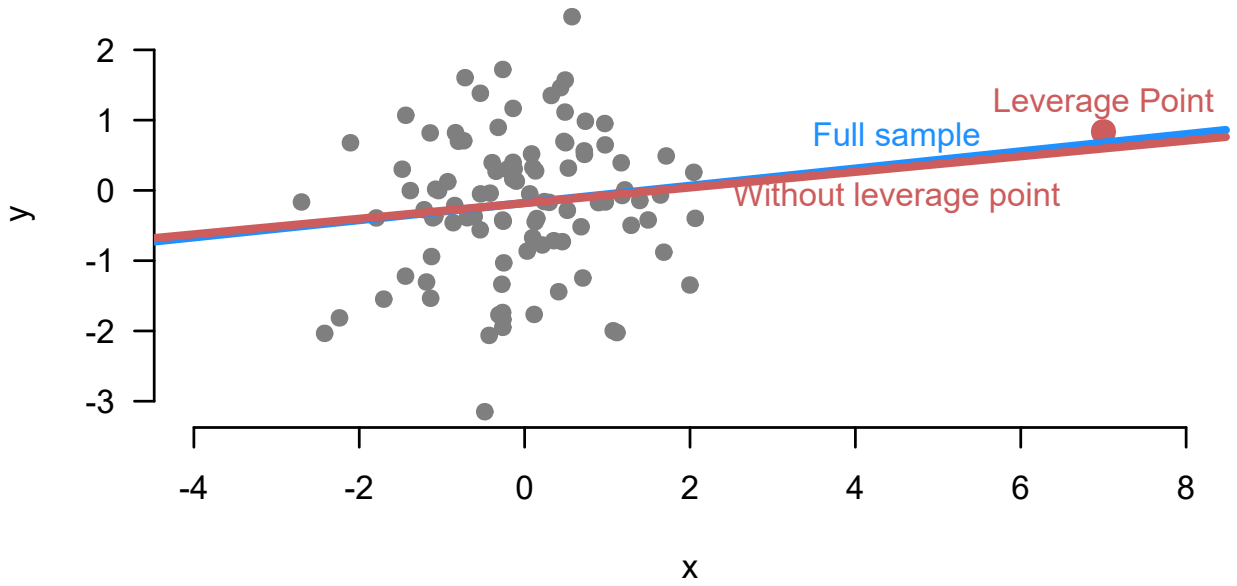
Abstract

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

A Sociological Cautionary Tale

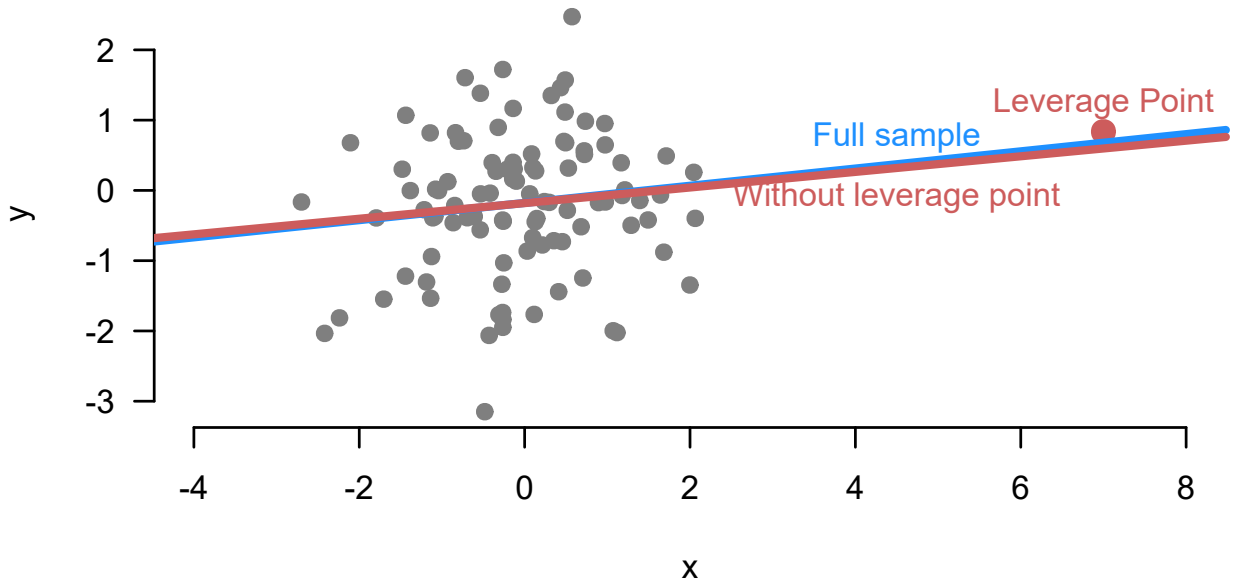
In our correspondence with Herring, he did not offer a definitive explanation for these discrepancies, but indicated that he may have treated all codes other than "not applicable" (–999) as substantive codes. Given (1) the large difference between his sample size and the number of valid observations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

Leverage Point Definition



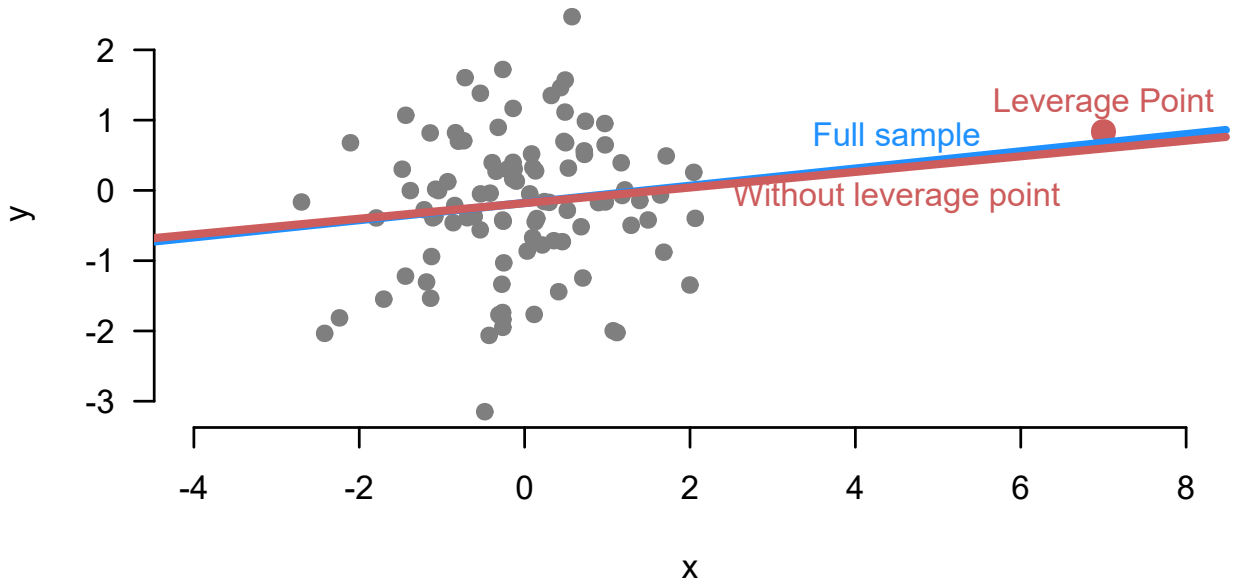
- Values that are extreme in the x direction

Leverage Point Definition



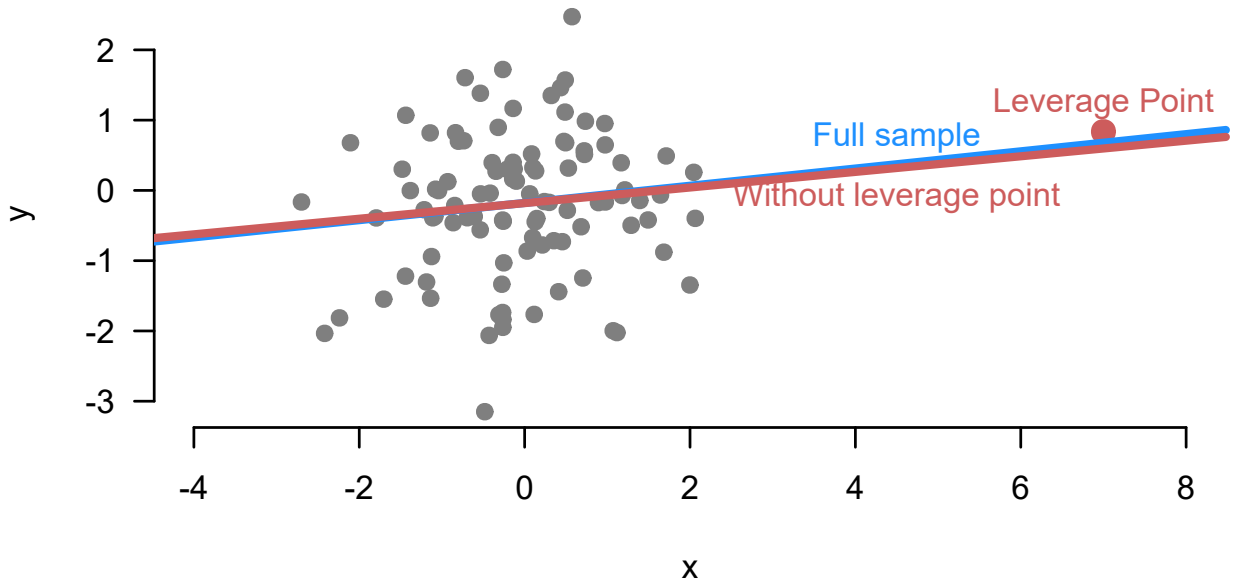
- Values that are extreme in the x direction
- That is, values far from the center of the covariate distribution

Leverage Point Definition



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

Leverage Point Definition



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

Leverage Points: Hat values

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

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

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

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

Leverage Points: Hat values

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

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

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

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

Therefore,

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

Leverage Points: Hat values

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

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

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

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

Therefore,

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

Leverage Points: Hat values

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

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

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

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

Therefore,

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

Leverage Points: Hat values

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

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

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

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

Therefore,

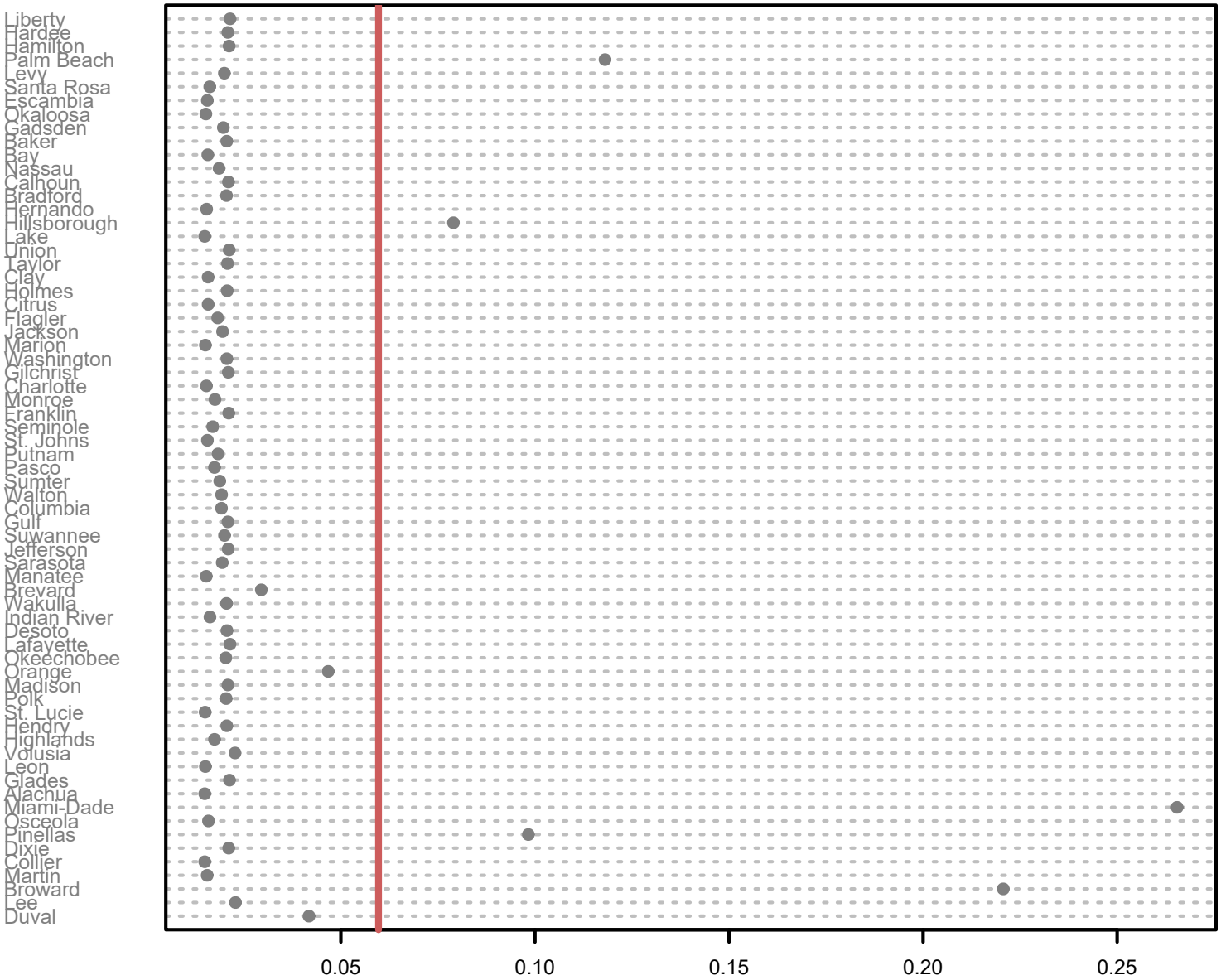
- h_{ij} dictates how important y_j is for the fitted value $\hat{y}_i$ (regardless of the actual value of y_j, since **H** depends only on **X**)
- The diagonal entries $h_{ii} = \sum_{j=1}^{n} h_{ij}^2$, so they summarize how important y_i is for all the fitted values. We call them the hat values or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

Appendix: Facts about Hat Values

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

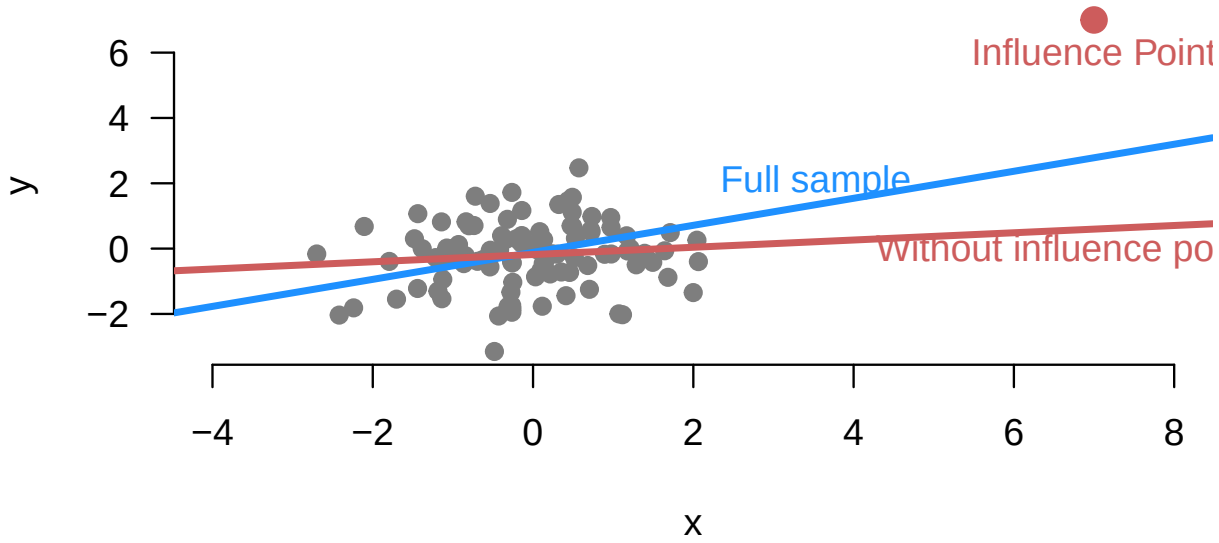
$$h_i = \frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum_{j=1}^{n}(X_j - \overline{X})^2}$$

Buchanan hats

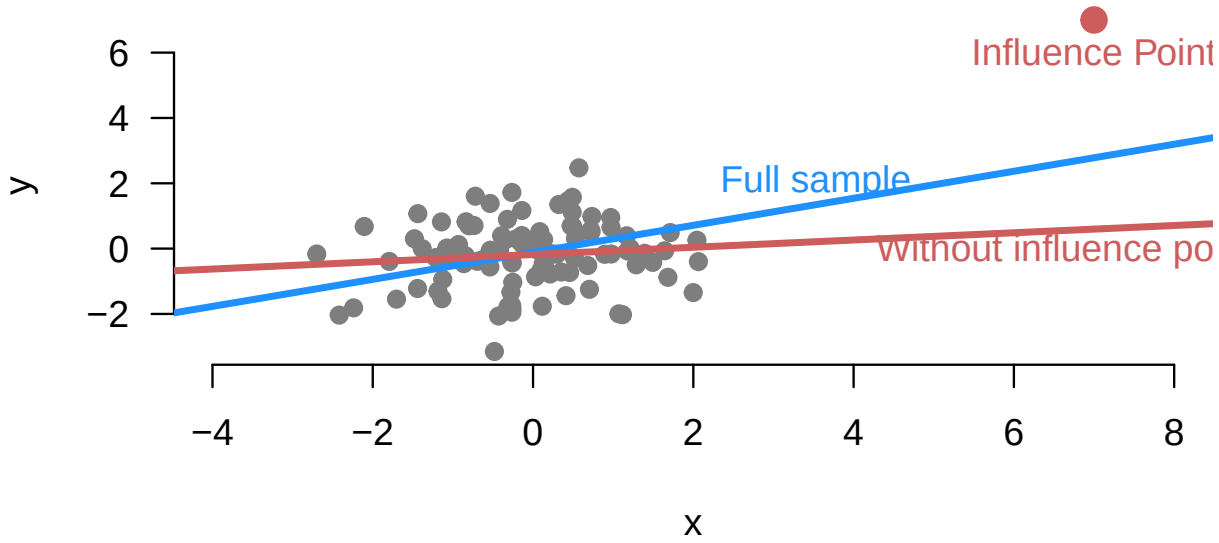


Influence points

Influence points

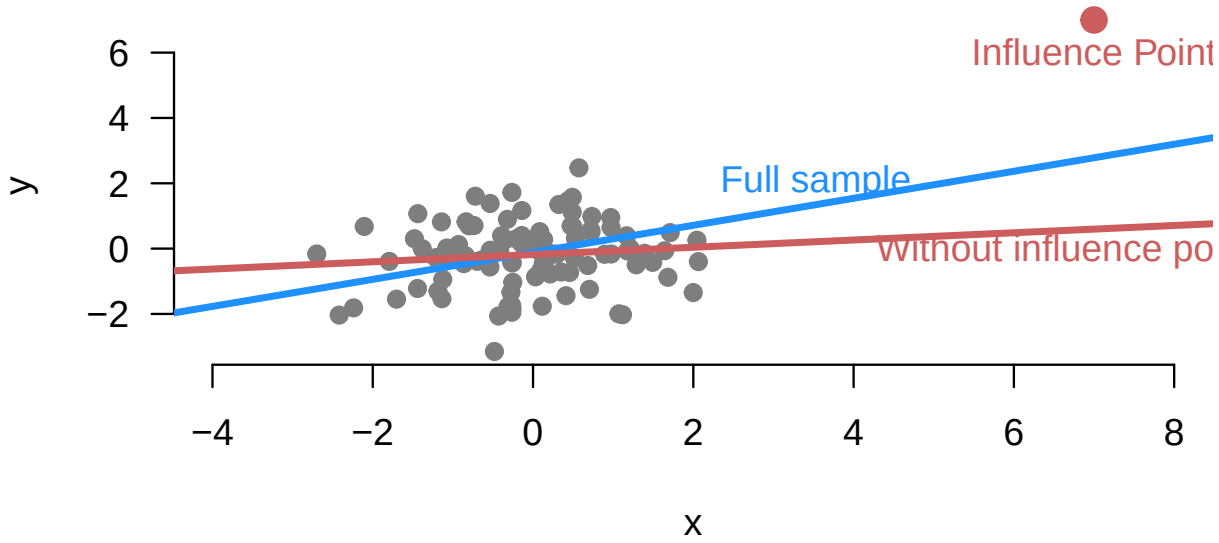


Influence points



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

Influence points



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

Detecting Influence Points/Bad Leverage Points

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

Detecting Influence Points/Bad Leverage Points

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

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

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

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

Detecting Influence Points/Bad Leverage Points

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

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

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

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

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

Standardized Influence

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

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

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

Standardized Influence

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

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

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

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

Standardized Influence

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

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

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

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

Standardized Influence

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

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

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

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

Standardized Influence

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

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

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

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

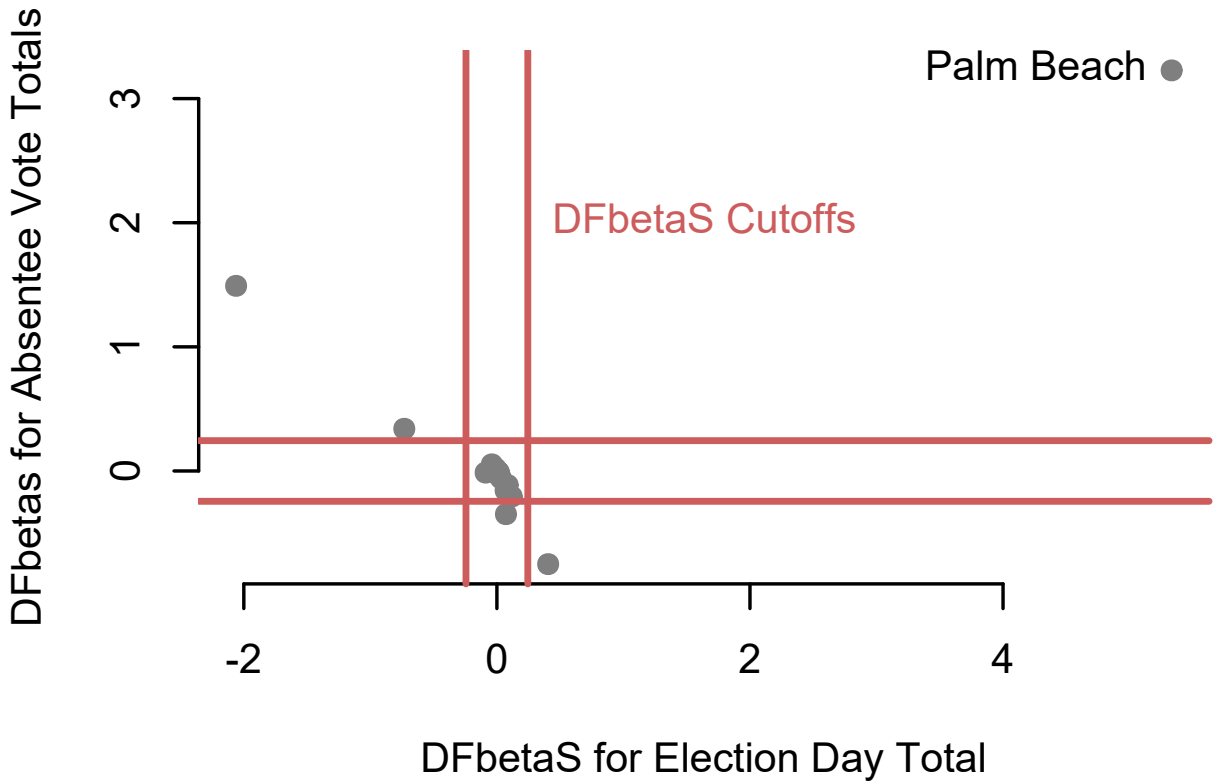
Buchanan influence

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

Buchanan influence

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

Buchanan influence



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

Summarizing Influence across All Coefficients

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

Summarizing Influence across All Coefficients

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

Summarizing Influence across All Coefficients

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

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

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

Summarizing Influence across All Coefficients

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

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

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

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

Summarizing Influence across All Coefficients

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

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

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

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

Summarizing Influence across All Coefficients

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

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

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

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

Summarizing Influence across All Coefficients

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

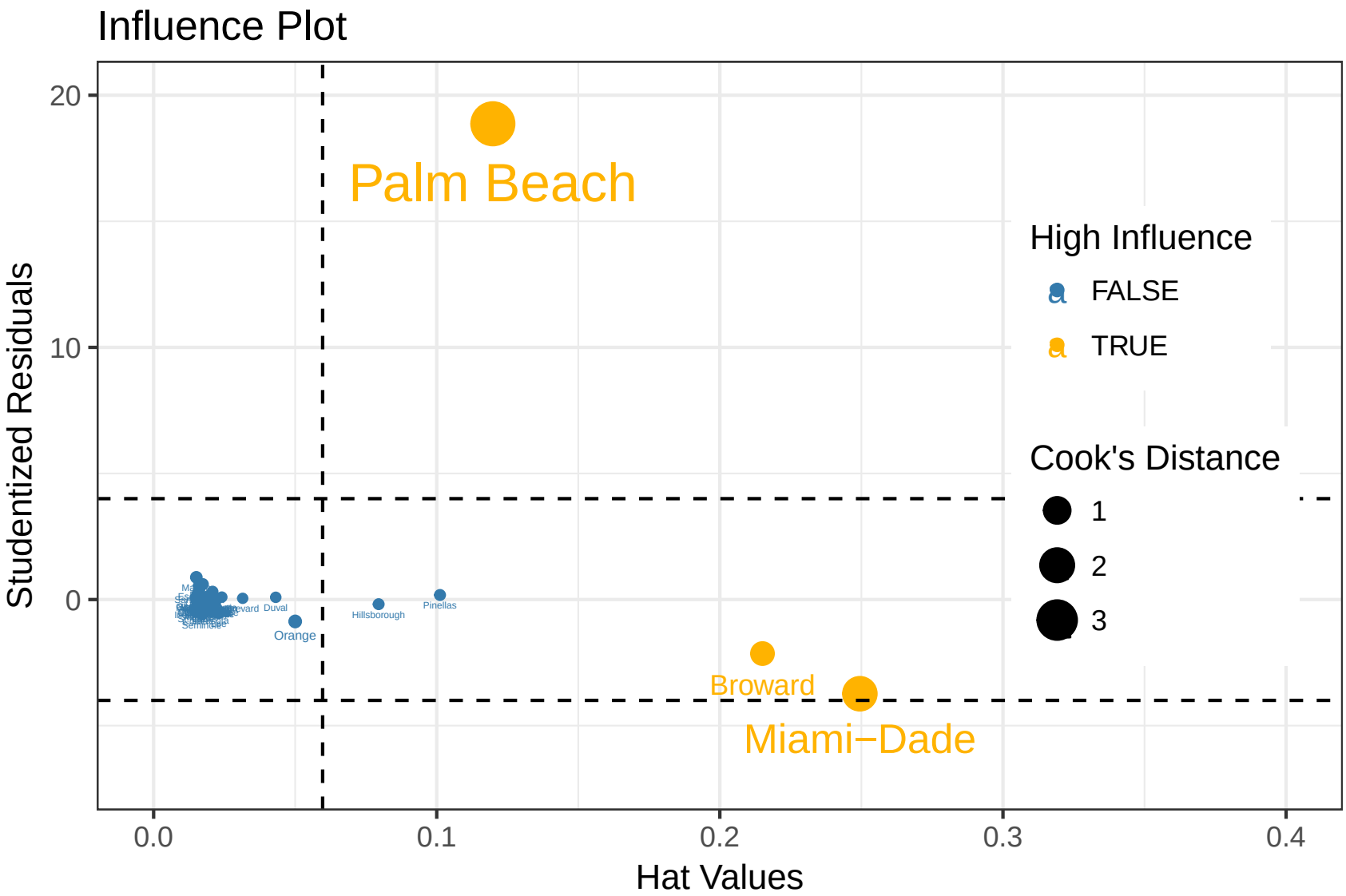
$$D_i = \underbrace{\frac{\hat{u}_i'^2}{k+1}}_{\text{outlyingness}} \times \underbrace{\frac{h_i}{1-h_i}}_{\text{leverage}}$$

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

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

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

Influence Plot Buchanan



Courtesy of Erin Hartman

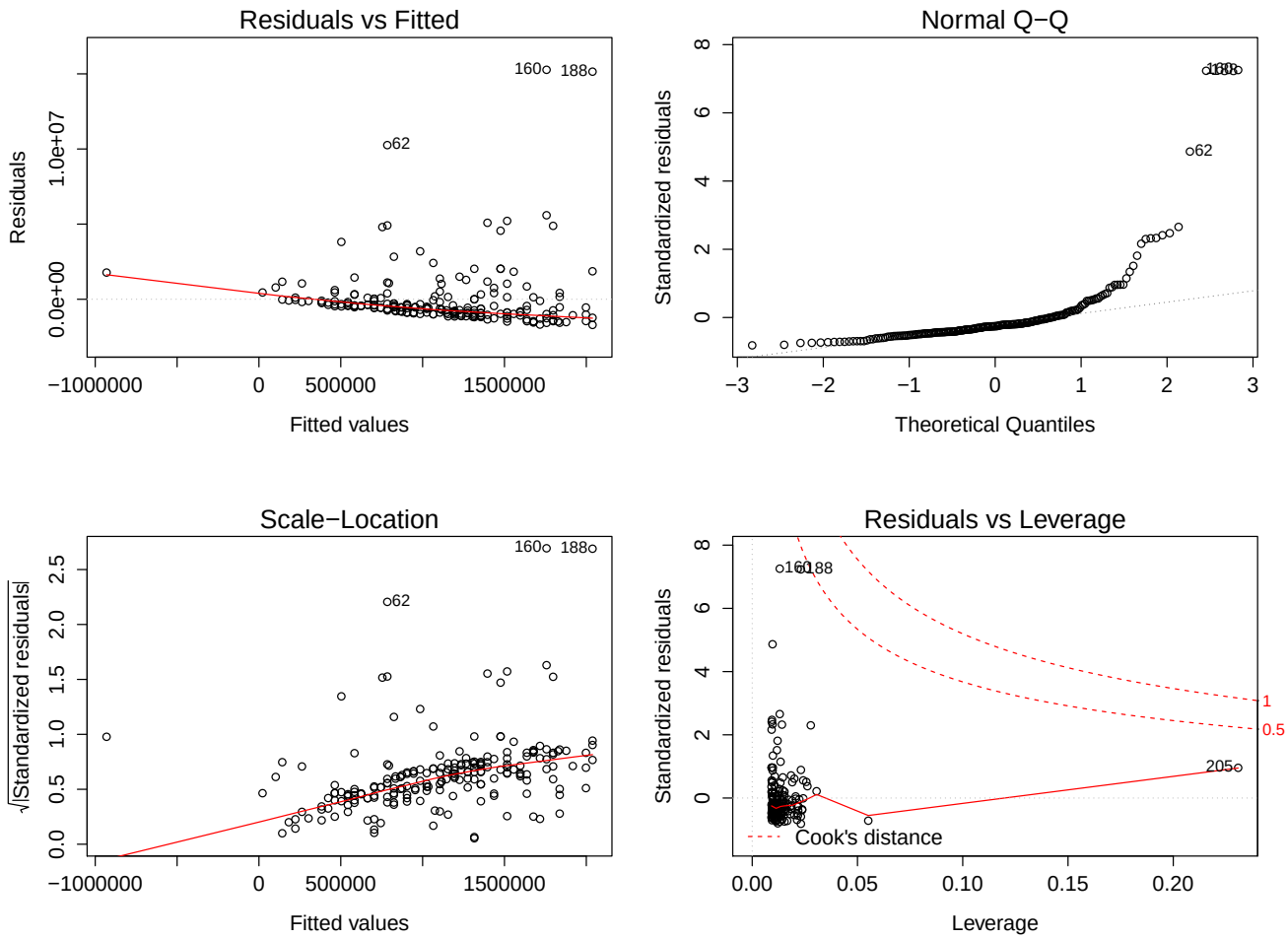
Code for Influence Plot

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

A Quick Function for Standard Diagnostic Plots

R Code

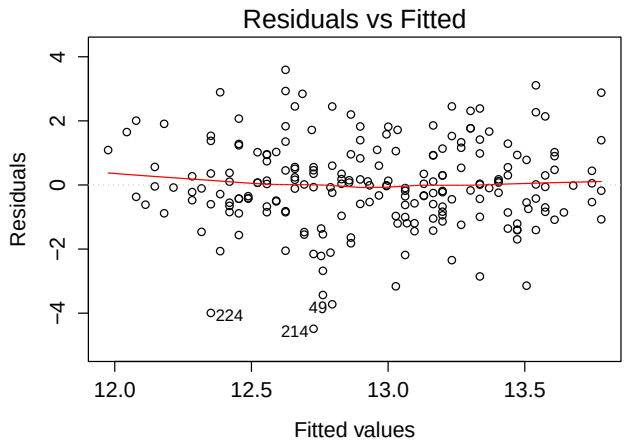
```
> par(mfrow=c(2,2))
> plot(mod1)
```

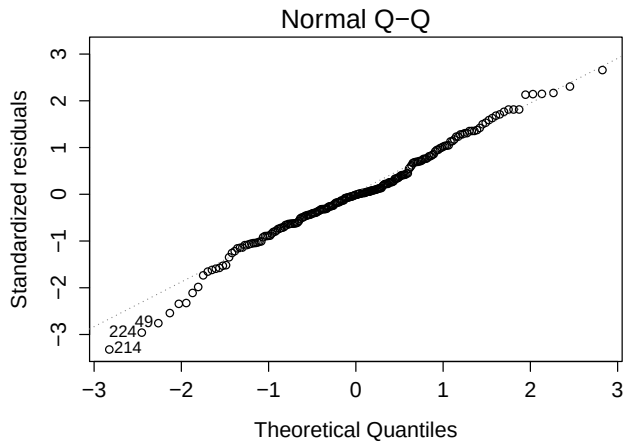


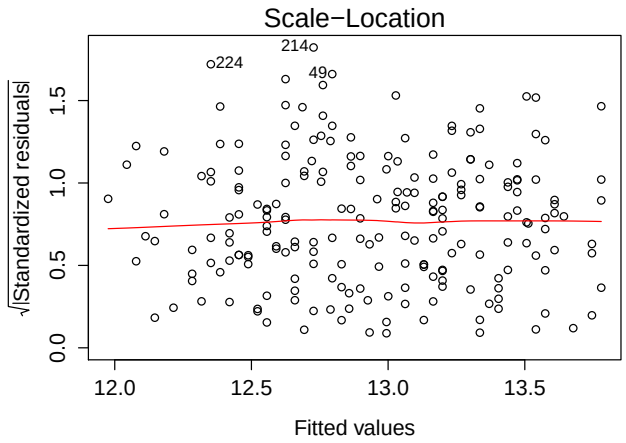
The Improved Model

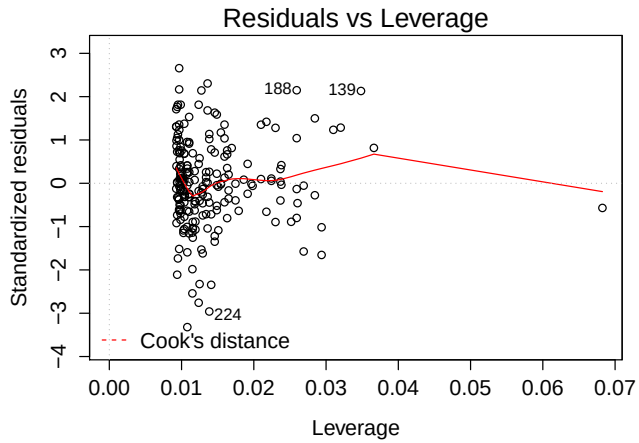
R Code

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









1. Non-Linearity
   - Polynomials
   - LOESS
   - Log Transformations
2. Non-Normality
   - Hat Matrix
   - Standardized and Studentized Residuals
   - QQ Plots
3. Extreme Values
   - Outliers
   - Leverage Points
   - Influence Points
4. Clustering

1. Non-Linearity
   - Polynomials
   - LOESS
   - Log Transformations
2. Non-Normality
   - Hat Matrix
   - Standardized and Studentized Residuals
   - QQ Plots
3. Extreme Values
   - Outliers
   - Leverage Points
   - Influence Points
4. Clustering

Clustered Dependence: Intuition

Clustered Dependence: Intuition

- Think back to the Gerber, Green, and Larimer (2008) social pressure mailer example.

Clustered Dependence: Intuition

- Think back to the Gerber, Green, and Larimer (2008) social pressure mailer example.
- Their design: randomly sample households and randomly assign them to different treatment conditions.

Clustered Dependence: Intuition

- Think back to the Gerber, Green, and Larimer (2008) social pressure mailer example.
- Their design: randomly sample households and randomly assign them to different treatment conditions.
- But the measurement of turnout is at the individual level and we have one unit per person.

Clustered Dependence: Intuition

- Think back to the Gerber, Green, and Larimer (2008) social pressure mailer example.
- Their design: randomly sample households and randomly assign them to different treatment conditions.
- But the measurement of turnout is at the individual level and we have one unit per person.
- Violation of iid/random sampling called clustering or clustered dependence.

Clustered Dependence: Intuition

- Think back to the Gerber, Green, and Larimer (2008) social pressure mailer example.
- Their design: randomly sample households and randomly assign them to different treatment conditions.
- But the measurement of turnout is at the individual level and we have one unit per person.
- Violation of iid/random sampling called clustering or clustered dependence.
- Loosely speaking, you are tricking the regression into thinking it has more information than it does.

Sampling Variance under Cluster Dependence

Sampling Variance under Cluster Dependence

- Recall our general sandwich expression:

$$V[\hat{\beta}|\mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\Sigma\mathbf{X}\,(\mathbf{X}'\mathbf{X})^{-1}$$

Sampling Variance under Cluster Dependence

- Recall our general sandwich expression:

$$V[\hat{\beta}|\mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\Sigma\mathbf{X}\,(\mathbf{X}'\mathbf{X})^{-1}$$

- We need to adjust to address our new sampling strategy.

Sampling Variance under Cluster Dependence

- Recall our general sandwich expression:

$$V[\hat{\beta}|\mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\Sigma\mathbf{X}\,(\mathbf{X}'\mathbf{X})^{-1}$$

- We need to adjust to address our new sampling strategy.
- We assume that respondents are independent across clusters, but dependent within clusters. Thus, we have

$$\text{Var}[\boldsymbol{u}_g|\mathbf{X}_g] = \boldsymbol{\Sigma}_g$$

Sampling Variance under Cluster Dependence

- Recall our general sandwich expression:

$$V[\hat{\beta}|\mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\Sigma\mathbf{X}\,(\mathbf{X}'\mathbf{X})^{-1}$$

- We need to adjust to address our new sampling strategy.
- We assume that respondents are independent across clusters, but dependent within clusters. Thus, we have

$$\text{Var}[\boldsymbol{u}_g|\mathbf{X}_g] = \Sigma_g$$

- This leads to a block diagonal expression for Σ overall

$$V[\boldsymbol{u}] = \Sigma = \left[\begin{array}{c|c|c|c} \Sigma_1 & 0 & \dots & 0 \\ \hline \mathbf{0} & \Sigma_2 & \dots & \mathbf{0} \\ \hline & & \ddots & \\ \hline \mathbf{0} & \mathbf{0} & \dots & \Sigma_M \end{array}\right]$$

Sampling Variance under Cluster Dependence

- Recall our general sandwich expression:

$$V[\hat{\beta}|\mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\Sigma\mathbf{X}\,(\mathbf{X}'\mathbf{X})^{-1}$$

- We need to adjust to address our new sampling strategy.
- We assume that respondents are independent across clusters, but dependent within clusters. Thus, we have

$$\text{Var}[\boldsymbol{u}_g|\mathbf{X}_g] = \Sigma_g$$

- This leads to a block diagonal expression for Σ overall

$$V[\boldsymbol{u}] = \Sigma = \left[\begin{array}{c|c|c|c} \Sigma_1 & 0 & \dots & 0 \\ \hline \mathbf{0} & \Sigma_2 & \dots & \mathbf{0} \\ \hline & & \ddots & \\ \hline \mathbf{0} & \mathbf{0} & \dots & \Sigma_M \end{array}\right]$$

- Under this clustered dependence, we can write this as:

$$V[\hat{\beta}|\mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{g=1}^{m}\mathbf{X}_g'\Sigma_g\mathbf{X}_g\right)(\mathbf{X}'\mathbf{X})^{-1}$$

Cluster-Robust Variance Estimation

Cluster-Robust Variance Estimation

- The independence is now across clusters, so we will need asymptotics in the number of clusters for CLT.

Cluster-Robust Variance Estimation

- The independence is now across clusters, so we will need asymptotics in the number of clusters for CLT.
- We can use a plugin estimator for the sampling variance on the previous slide with a small sample correction (many variants available, here we give the CR1 variant Stata uses):

$$\hat{V}_{CR1}[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \frac{m}{(m-1)}\frac{n-1}{n-k}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\sum_{g=1}^{m}\mathbf{X}_g'\hat{\boldsymbol{u}}_g\hat{\boldsymbol{u}}_g{}'\mathbf{X}_g\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

- Another excellent strategy is to use the block bootstrap (resample clusters instead of units).

Notes on Cluster-Robust Standard Errors

Notes on Cluster-Robust Standard Errors

- CRSE do not change our estimates $\widehat{\beta}$, cannot fix bias.

Notes on Cluster-Robust Standard Errors

- CRSE do not change our estimates $\widehat{\beta}$, cannot fix bias.
- Consistency of the CRSE are in the number of groups, not the number of individuals.

Notes on Cluster-Robust Standard Errors

- CRSE do not change our estimates $\widehat{\beta}$, cannot fix bias.
- Consistency of the CRSE are in the number of groups, not the number of individuals.
- Works well in the following settings:
  - sampling complete blocks of units where sampled clusters are a small fraction of the population.

Notes on Cluster-Robust Standard Errors

- CRSE do not change our estimates $\widehat{\beta}$, cannot fix bias.
- Consistency of the CRSE are in the number of groups, not the number of individuals.
- Works well in the following settings:
  - sampling complete blocks of units where sampled clusters are a small fraction of the population.
  - sampling a small fraction of units from all or nearly all clusters in the population.

Notes on Cluster-Robust Standard Errors

- CRSE do not change our estimates $\widehat{\beta}$, cannot fix bias.
- Consistency of the CRSE are in the number of groups, not the number of individuals.
- Works well in the following settings:
  - sampling complete blocks of units where sampled clusters are a small fraction of the population.
  - sampling a small fraction of units from all or nearly all clusters in the population.
  - assignment of key variable is constant within the cluster.

Notes on Cluster-Robust Standard Errors

- CRSE do not change our estimates $\widehat{\beta}$, cannot fix bias.
- Consistency of the CRSE are in the number of groups, not the number of individuals.
- Works well in the following settings:
  - sampling complete blocks of units where sampled clusters are a small fraction of the population.
  - sampling a small fraction of units from all or nearly all clusters in the population.
  - assignment of key variable is constant within the cluster.
- Just because something is a group, doesn't mean you need to cluster standard errors, key is to think about the sampling strategy.

Notes on Cluster-Robust Standard Errors

- CRSE do not change our estimates $\widehat{\beta}$, cannot fix bias.
- Consistency of the CRSE are in the number of groups, not the number of individuals.
- Works well in the following settings:
  - sampling complete blocks of units where sampled clusters are a small fraction of the population.
  - sampling a small fraction of units from all or nearly all clusters in the population.
  - assignment of key variable is constant within the cluster.
- Just because something is a group, doesn't mean you need to cluster standard errors, key is to think about the sampling strategy.
- Abadie, Athey, Imbens and Wooldridge (2022) "When Should You Adjust Standard Errors for Clustering" *QJE* provides guidance and approaches for other cases.

Notes on Cluster-Robust Standard Errors

- CRSE do not change our estimates $\widehat{\beta}$, cannot fix bias.
- Consistency of the CRSE are in the number of groups, not the number of individuals.
- Works well in the following settings:
  - sampling complete blocks of units where sampled clusters are a small fraction of the population.
  - sampling a small fraction of units from all or nearly all clusters in the population.
  - assignment of key variable is constant within the cluster.
- Just because something is a group, doesn't mean you need to cluster standard errors, key is to think about the sampling strategy.
- Abadie, Athey, Imbens and Wooldridge (2022) "When Should You Adjust Standard Errors for Clustering" *QJE* provides guidance and approaches for other cases.
- NB: this broad strategy of sandwich estimators works for other kinds of settings (time-series etc.). See `sandwich` package in `R`.

This Week in Review

This Week in Review

- Approaches to modeling non-linearity in the CEF.

This Week in Review

- Approaches to modeling non-linearity in the CEF.
- Diagnostics for the normality assumption in the classical linear model (including general tools for looking at residuals)

This Week in Review

- Approaches to modeling non-linearity in the CEF.
- Diagnostics for the normality assumption in the classical linear model (including general tools for looking at residuals)
- Analysis of extreme values: outliers, leverage points, and influence points.

This Week in Review

- Approaches to modeling non-linearity in the CEF.
- Diagnostics for the normality assumption in the classical linear model (including general tools for looking at residuals)
- Analysis of extreme values: outliers, leverage points, and influence points.
- Cluster-robust standard errors for addressing clustering in our sampling strategy.

This Week in Review

- Approaches to modeling non-linearity in the CEF.
- Diagnostics for the normality assumption in the classical linear model (including general tools for looking at residuals)
- Analysis of extreme values: outliers, leverage points, and influence points.
- Cluster-robust standard errors for addressing clustering in our sampling strategy.

Next week: Frameworks for Causal Inference!