FLS 6441 - Methods III: Explanation and Causation

Week 1 - Review of Regression

Jonathan Phillips

March 2020

Course Objectives

1. Change how you think about quantitative methods, *explaining* politics, and not just describing it

Course Objectives

1. Change how you think about quantitative methods, *explaining* politics, and not just describing it
2. Understand the 'toolkit' of causal methods used in top journals

Course Objectives

1. Change how you think about quantitative methods, *explaining* politics, and not just describing it
2. Understand the 'toolkit' of causal methods used in top journals
3. Apply those methods to your own research questions

Course Objectives

1. Change how you think about quantitative methods, *explaining* politics, and not just describing it
2. Understand the 'toolkit' of causal methods used in top journals
3. Apply those methods to your own research questions

Course Website

Course Topics

1. Review of Regression (5th March)

Course Topics

1. Review of Regression (5th March)
2. A Framework for Explanation (12th March)

Course Topics

1. Review of Regression (5th March)
2. A Framework for Explanation (12th March)
3. Field Experiments (19th March)
4. Survey and Lab Experiments (26th March)
5. Randomized Natural Experiments (2nd April, Semana Santa)

Course Topics

1. Review of Regression (5th March)
2. A Framework for Explanation (12th March)
3. Field Experiments (19th March)
4. Survey and Lab Experiments (26th March)
5. Randomized Natural Experiments (2nd April, Semana Santa)
6. Instrumental Variables (16th April)
7. Discontinuities (23rd April)

Course Topics

1. Review of Regression (5th March)
2. A Framework for Explanation (12th March)
3. Field Experiments (19th March)
4. Survey and Lab Experiments (26th March)
5. Randomized Natural Experiments (2nd April, Semana Santa)
6. Instrumental Variables (16th April)
7. Discontinuities (23rd April)
8. Difference-in-Differences (30th April)
9. Controlling for Confounding (7th May)
10. Matching (14th May)
11. Comparative Cases and Process Tracing (21st May)

Course Topics

1. Review of Regression (5th March)
2. A Framework for Explanation (12th March)
3. Field Experiments (19th March)
4. Survey and Lab Experiments (26th March)
5. Randomized Natural Experiments (2nd April, Semana Santa)
6. Instrumental Variables (16th April)
7. Discontinuities (23rd April)
8. Difference-in-Differences (30th April)
9. Controlling for Confounding (7th May)
10. Matching (14th May)
11. Comparative Cases and Process Tracing (21st May)
12. Generalizability, Reproducibility and Mechanisms (28th May)

Course Schedule

- Wednesday 18h - Submit Replication Task

Course Schedule

- Wednesday 18h - Submit Replication Task
- Thursday 14h-16h - Room 105

Course Schedule

- Wednesday 18h - Submit Replication Task
- Thursday 14h-16h - Room 105
- Thursday 16.15-18.00 - Lab 122

Evaluation

- Replication Tasks - 40%

Evaluation

- Replication Tasks - 40%
  - 8 best grades out of 10 tasks

Evaluation

- Replication Tasks - 40%
  - 8 best grades out of 10 tasks
- Short Research Paper - 40%

Evaluation

- Replication Tasks - 40%
  - 8 best grades out of 10 tasks
- Short Research Paper - 40%
- Participation - 20%

Short Research Paper

- Quality > Quantity

Short Research Paper

- Quality > Quantity
- Max 15 pages, English or Portuguese

Short Research Paper

- Quality > Quantity
- Max 15 pages, English or Portuguese
- Submit paper and code by email to me by 24th July 2020

Short Research Paper

- Quality > Quantity
- Max 15 pages, English or Portuguese
- Submit paper and code by email to me by 24th July 2020
- Use at least one of the methods studied in class

Short Research Paper

- Quality > Quantity
- Max 15 pages, English or Portuguese
- Submit paper and code by email to me by 24th July 2020
- Use at least one of the methods studied in class
- *Tip:* Pick a simple *causal* question and dataset

If you get Lost:

1. Don't panic! Everyone needs to see this content 3 or 4 times to 'get' it

If you get Lost:

1. Don't panic! Everyone needs to see this content 3 or 4 times to 'get' it
2. Simplify your thoughts - all the methods are doing *less* than you think they are

If you get Lost:

1. Don't panic! Everyone needs to see this content 3 or 4 times to 'get' it
2. Simplify your thoughts - all the methods are doing *less* than you think they are
3. Re-read the slides and core readings

If you get Lost:

1. Don't panic! Everyone needs to see this content 3 or 4 times to 'get' it
2. Simplify your thoughts - all the methods are doing *less* than you think they are
3. Re-read the slides and core readings
4. Search online

If you get Lost:

1. Don't panic! Everyone needs to see this content 3 or 4 times to 'get' it
2. Simplify your thoughts - all the methods are doing *less* than you think they are
3. Re-read the slides and core readings
4. Search online
5. Ask your friends - they can explain better than me

If you get Lost:

1. Don’t panic! Everyone needs to see this content 3 or 4 times to ’get’ it
2. Simplify your thoughts - all the methods are doing *less* than you think they are
3. Re-read the slides and core readings
4. Search online
5. Ask your friends - they can explain better than me
6. Ask me

Today's Objectives

1. What Does Regression Actually Do?
2. Guide to 'Smart' Regression
3. What Does Regression NOT Do?

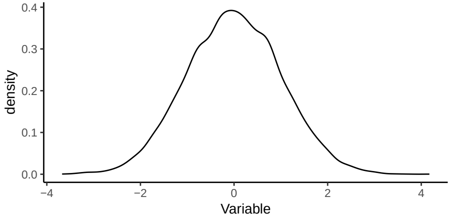
Section 1

What Does Regression Actually Do?

Data

- We work with variables, which VARY!

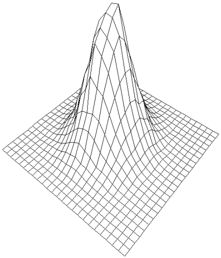
Variable
0.30
-0.67
0.39
0.03
-1.26
1.26
-1.44
0.16
0.50
0.01



Data

- We work with variables, which VARY!

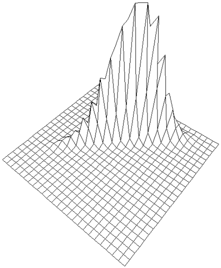
Variable_1	Variable_2
-0.44	0.63
0.06	0.68
-0.21	-0.02
0.44	-0.25
1.29	0.46
-0.38	-0.81
-1.04	0.24
-0.16	1.84
1.29	0.06
-0.10	-0.18



Data

- We work with variables, which VARY!

Variable_1	Variable_2
-0.86	-0.69
-0.35	-0.44
1.27	0.42
-0.35	-0.22
-0.43	-0.56
0.05	-0.04
0.69	0.70
1.27	1.07
0.22	-0.00
-0.28	-0.13



What Does Regression Actually Do?

1. Regression as Least Squares
2. Regression as Conditional Expectation
3. Regression as (Partial) Correlation

1. Regression as Least Squares

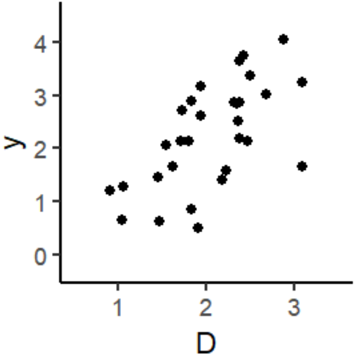
- Regression identifies the line through the data that minimizes the sum of squared vertical distances

1. Regression as Least Squares

- Regression identifies the line through the data that minimizes the sum of squared vertical distances
- $y_i = \alpha + \beta D_i + \epsilon_i$

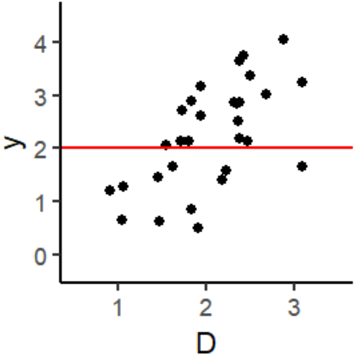
1. Regression as Least Squares

- Regression identifies the line through the data that minimizes the sum of squared vertical distances
- $y_i = \alpha + \beta D_i + \epsilon_i$



1. Regression as Least Squares

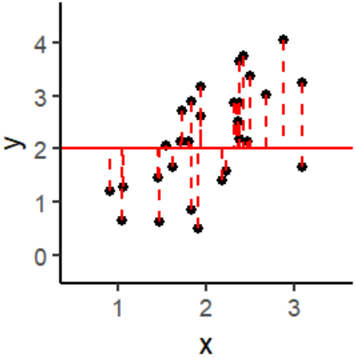
- Regression identifies the line through the data that minimizes the sum of squared vertical distances
- $y_i = \alpha + \beta D_i + \epsilon_i$



1. Regression as Least Squares

- Regression identifies the line through the data that minimizes the sum of squared vertical distances
- $y_i = \alpha + \beta D_i + \epsilon_i$

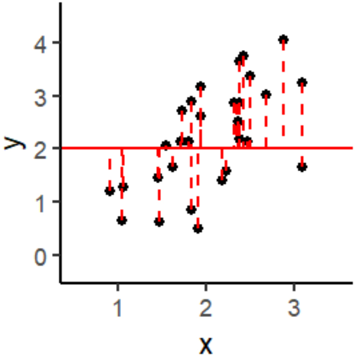
Slope = 0



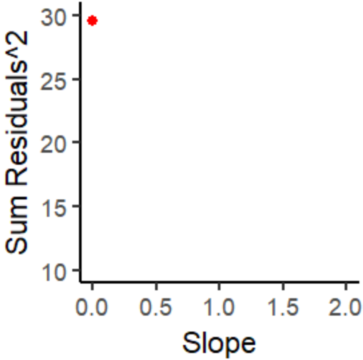
1. Regression as Least Squares

- Regression identifies the line through the data that minimizes the sum of squared vertical distances
- $y_i = \alpha + \beta D_i + \epsilon_i$

Slope = 0



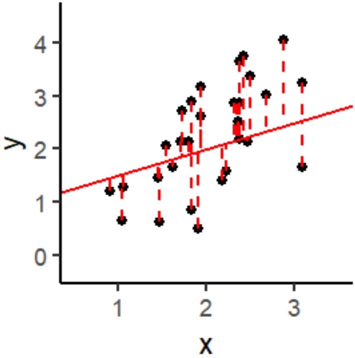
Sum of Residuals^2 = 29.6



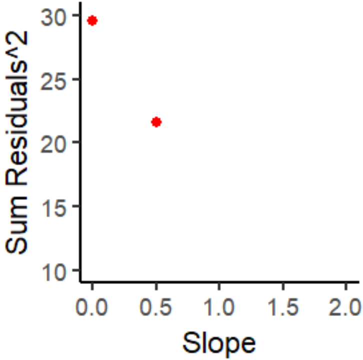
1. Regression as Least Squares

- Regression identifies the line through the data that minimizes the sum of squared vertical distances
- $y_i = \alpha + \beta D_i + \epsilon_i$

Slope = 0.5



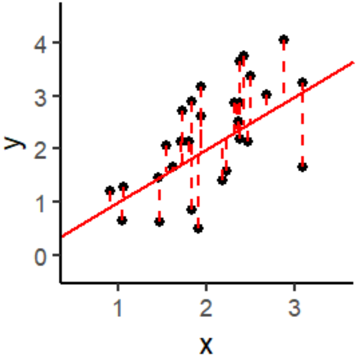
Sum of Residuals^2 = 21.6



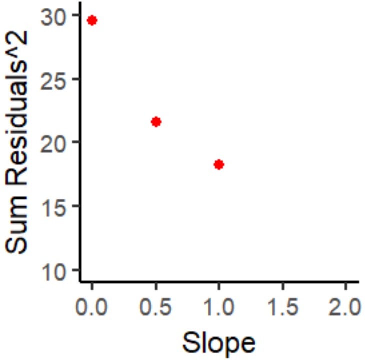
1. Regression as Least Squares

- Regression identifies the line through the data that minimizes the sum of squared vertical distances
- $y_i = \alpha + \beta D_i + \epsilon_i$

Slope = 1



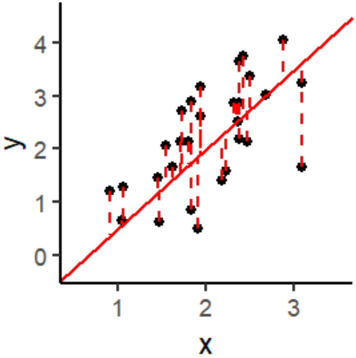
Sum of Residuals^2 = 18.3



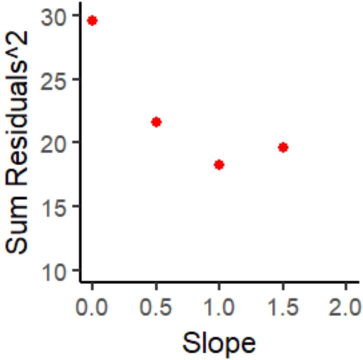
1. Regression as Least Squares

- Regression identifies the line through the data that minimizes the sum of squared vertical distances
- $y_i = \alpha + \beta D_i + \epsilon_i$

Slope = 1.5



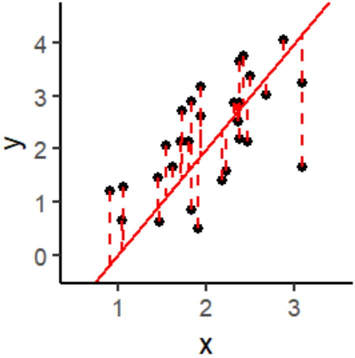
Sum of Residuals^2 = 19.6



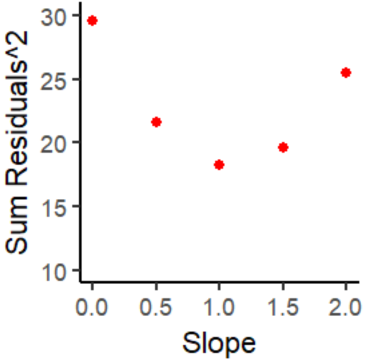
1. Regression as Least Squares

- Regression identifies the line through the data that minimizes the sum of squared vertical distances
- $y_i = \alpha + \beta D_i + \epsilon_i$

Slope = 2



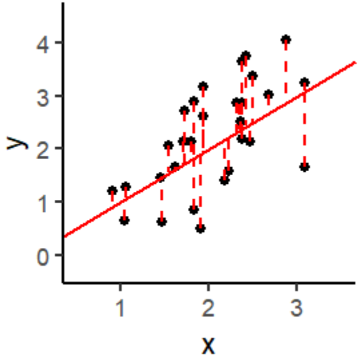
Sum of Residuals^2 = 25.5



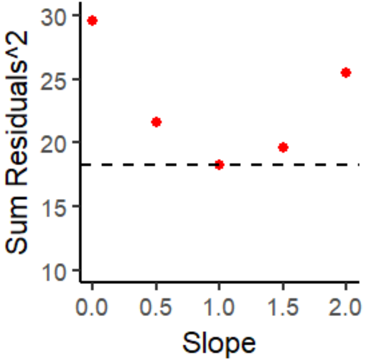
1. Regression as Least Squares

- Regression identifies the line through the data that minimizes the sum of squared vertical distances
- $y_i = \alpha + \beta D_i + \epsilon_i$

Slope = 1



Sum of Residuals^2 = 18.3

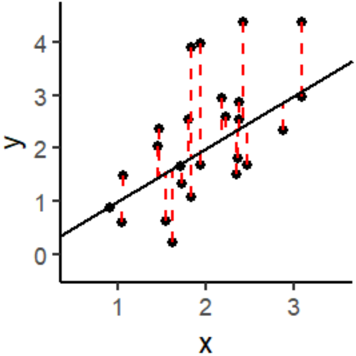


1. Regression as Least Squares

- If we add pure *noise* to y, our estimate of β is unchanged

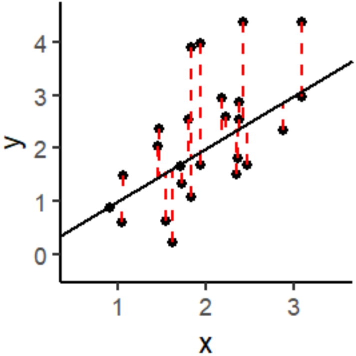
1. Regression as Least Squares
   - If we add pure *noise* to y, our estimate of β is unchanged
     - The residual error increases
   - $y_i = \alpha + \beta D_i + \epsilon_i$

Slope = 1

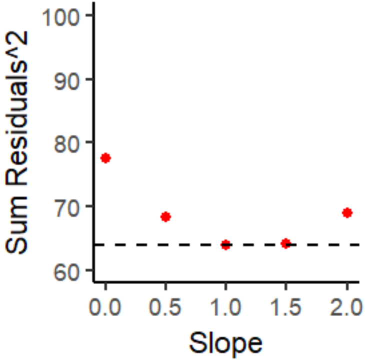


1. Regression as Least Squares
   - If we add pure *noise* to y, our estimate of β is unchanged
     - The residual error increases
   - $y_i = \alpha + \beta D_i + \epsilon_i$

Slope = 1



Sum of Residuals^2 = 63.9

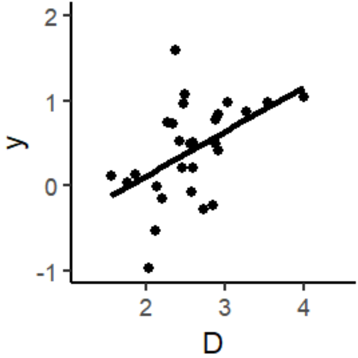


1. Regression as Least Squares

- Dummy control variables *remove variation* associated with specific levels or categories
  - The same as Fixed Effects

1. Regression as Least Squares

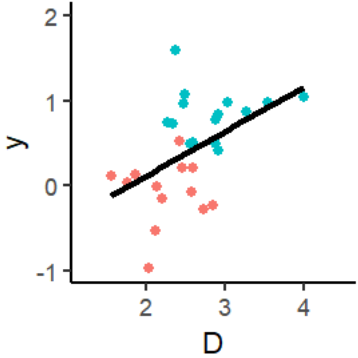
- Dummy control variables *remove variation* associated with specific levels or categories
  - The same as Fixed Effects
- $y_{ij} = \alpha + \beta_1 D_{ij} + \epsilon_i$



Ignoring the dummy control variable, the slope coefficient is 1

1. Regression as Least Squares

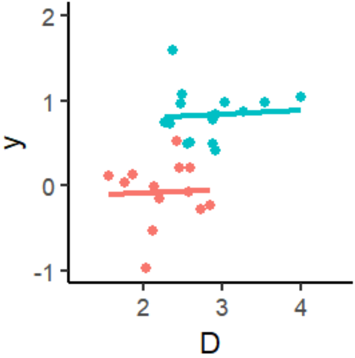
- Dummy control variables *remove variation* associated with specific levels or categories
  - The same for fixed effects
- $y_{ij} = \alpha + \beta_1 D_{ij} + \epsilon_i$



But the data points really represent two very different groups, blues and reds

1. Regression as Least Squares

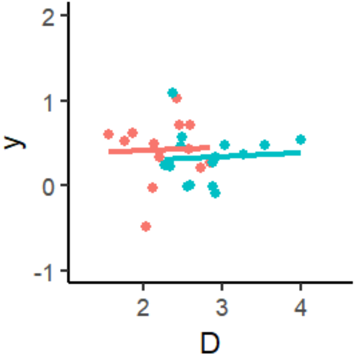
- Dummy control variables *remove variation* associated with specific levels or categories
  - The same for fixed effects
- $y_{ij} = \alpha + \beta_1 D_{ij} + \beta_2 X_j + \epsilon_i$



What if we ran the regression for each group *separately*?

1. Regression as Least Squares

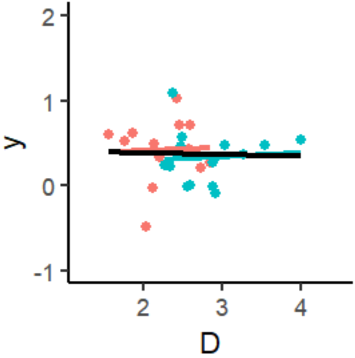
- Dummy control variables *remove variation* associated with specific levels or categories
  - The same for fixed effects
- $y_{ij} = \alpha + \beta_1 D_{ij} + \beta_2 X_j + \epsilon_i$



Dummy control variables *remove* the average Y differences between blues and reds

1. Regression as Least Squares

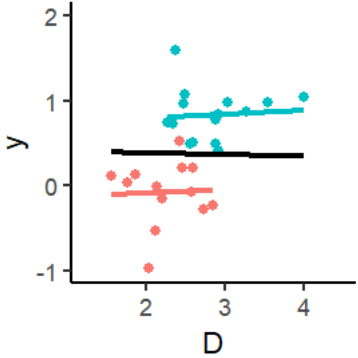
- Dummy control variables *remove variation* associated with specific levels or categories
  - The same for fixed effects
- $y_{ij} = \alpha + \beta_1 D_{ij} + \beta_2 X_j + \epsilon_i$



The new regression line for the full data now has a slope of zero

1. Regression as Least Squares

- Dummy control variables *remove variation* associated with specific levels or categories
  - The same for fixed effects
- $y_{ij} = \alpha + \beta_1 D_{ij} + \beta_2 X_j + \epsilon_i$



Equivalently, dummy control variables restrict comparisons to **within the same group**:

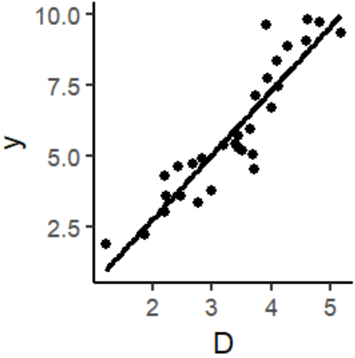
1. How much does D affect Y within the blue group? 0
2. How much does D affect Y within the red group? 0
3. What's the average of (1) and (2) (weighted by the number of units in each group)? 0

1. Regression as Least Squares

- Continuous control variables *remove variation* based on how much the control explains y

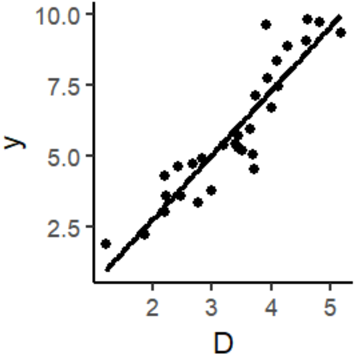
1. Regression as Least Squares

- Continuous control variables *remove variation* based on how much the control explains y
- $y_i = \alpha + \beta_1 D_i + \epsilon_i$



1. Regression as Least Squares

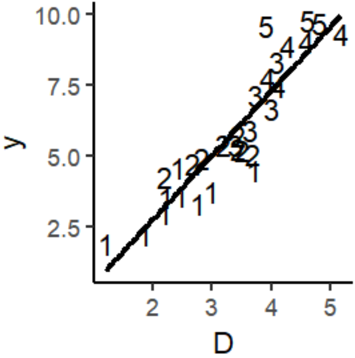
- Continuous control variables *remove variation* based on how much the control explains y
- $y_i = \alpha + \beta_1 D_i + \epsilon_i$



The coefficient β_1 is 2.267
Real effect = 1

1. Regression as Least Squares

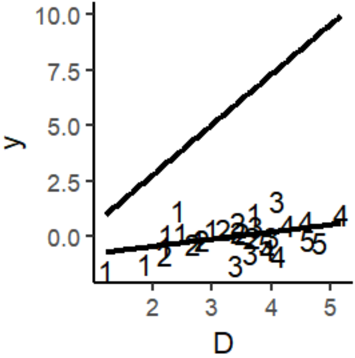
- Continuous control variables *remove variation* based on how much the control explains y
- $y_i = \alpha + \beta_1 D_i + \epsilon_i$



The coefficient β_1 is 2.267
Real effect = 1

1. Regression as Least Squares

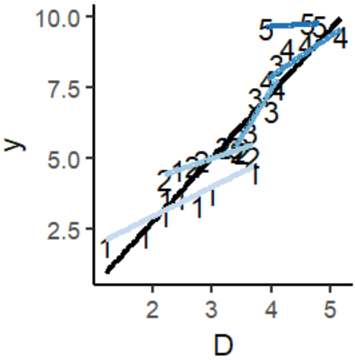
- Continuous control variables *remove variation* based on how much the control explains y
- $y_i = \alpha + \beta_1 D_i + \beta_2 X_i + \epsilon_i$



The coefficient β_1 is 1.024
Real effect = 1

1. Regression as Least Squares

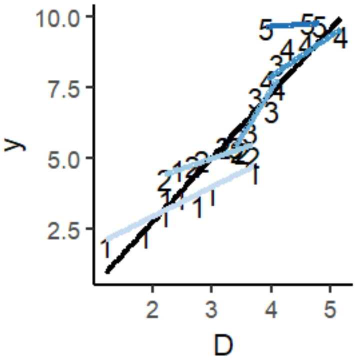
- Continuous control variables *remove variation* based on how much the control explains y
- $y_i = \alpha + \beta_1 D_i + \beta_2 X_i + \epsilon_i$



- Equivalently, we subset the data to each value of X, and find each slope

1. Regression as Least Squares

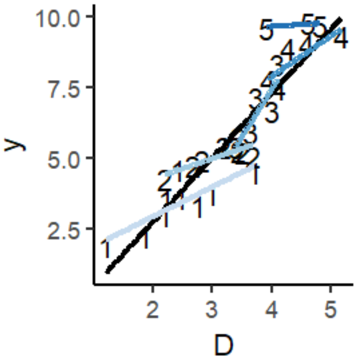
- Continuous control variables *remove variation* based on how much the control explains y
- $y_i = \alpha + \beta_1 D_i + \beta_2 X_i + \epsilon_i$



- Equivalently, we subset the data to each value of X, and find each slope
- Then average these slopes, $\beta_1 \sim 1$

1. Regression as Least Squares

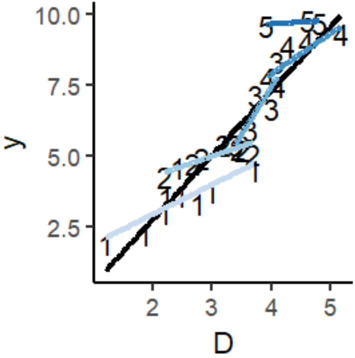
- Continuous control variables *remove variation* based on how much the control explains y
- $y_i = \alpha + \beta_1 D_i + \beta_2 X_i + \epsilon_i$



- Equivalently, we subset the data to each value of X, and find each slope
- Then average these slopes, $\beta_1 \sim 1$
- Impossible with truly continuous variables

1. Regression as Least Squares

- Continuous control variables *remove variation* based on how much the control explains y
- $y_i = \alpha + \beta_1 D_i + \beta_2 X_i + \epsilon_i$



- Equivalently, we subset the data to each value of X, and find each slope
- Then average these slopes, $\beta_1 \sim 1$
- Impossible with truly continuous variables
- So regression uses linearity to fill in the gaps

2. Regression as Conditional Expectation

- Regression is also a **Conditional Expectation Function**

2. Regression as Conditional Expectation

- Regression is also a **Conditional Expectation Function**
- **Conditional on D,** What is our expectation (mean value) of y?

$$y_i = \alpha + \beta_1 D_i + \epsilon_i$$

2. Regression as Conditional Expectation

- Regression is also a **Conditional Expectation Function**
- **Conditional on D,** What is our expectation (mean value) of y?

$$y_i = \alpha + \beta_1 D_i + \epsilon_i$$

$$E(y) = \alpha + \beta_1 D$$

2. Regression as Conditional Expectation

- **Conditional on a specific value of D,** what is our expectation (mean value) of y?

$$y_i = \alpha + \beta_1 D_i + \epsilon_i$$

2. Regression as Conditional Expectation

- **Conditional on a specific value of D,** what is our expectation (mean value) of y?

$$y_i = \alpha + \beta_1 D_i + \epsilon_i$$
$$Attitude_i = \alpha + \beta_1 Income_i + N(0, \sigma^2)$$

2. Regression as Conditional Expectation

- **Conditional on a specific value of D,** what is our expectation (mean value) of *y*?

$$y_i = \alpha + \beta_1 D_i + \epsilon_i$$

$$Attitude_i = \alpha + \beta_1 Income_i + N(0, \sigma^2)$$

$$Attitude_i = 2.235 - 0.000818 * Income_i + N(0, 2.38)$$

2. Regression as Conditional Expectation

- **Conditional on a specific value of D,** what is our expectation (mean value) of y?

$$y_i = \alpha + \beta_1 D_i + \epsilon_i$$

$$Attitude_i = \alpha + \beta_1 Income_i + N(0, \sigma^2)$$

$$Attitude_i = 2.235 - 0.000818 * Income_i + N(0, 2.38)$$

$$(Attitude_i | Income_i = 3000) = 2.235 - 0.000818 * 3000 + N(0, 2.38)$$

2. Regression as Conditional Expectation

- **Conditional on a specific value of D,** what is our expectation (mean value) of y?

$$y_i = \alpha + \beta_1 D_i + \epsilon_i$$
$$Attitude_i = \alpha + \beta_1 Income_i + N(0, \sigma^2)$$
$$Attitude_i = 2.235 - 0.000818 * Income_i + N(0, 2.38)$$
$$(Attitude_i | Income_i = 3000) = 2.235 - 0.000818 * 3000 + N(0, 2.38)$$
$$(Attitude_i | Income_i = 3000) = -0.22 + N(0, 2.38)$$

2. Regression as Conditional Expectation

- **Conditional on a specific value of D,** what is our expectation (mean value) of y?

$$y_i = \alpha + \beta_1 D_i + \epsilon_i$$

$$Attitude_i = \alpha + \beta_1 Income_i + N(0, \sigma^2)$$

$$Attitude_i = 2.235 - 0.000818 * Income_i + N(0, 2.38)$$

$$(Attitude_i | Income_i = 3000) = 2.235 - 0.000818 * 3000 + N(0, 2.38)$$

$$(Attitude_i | Income_i = 3000) = -0.22 + N(0, 2.38)$$

$$E(Attitude | Income = 3000) = -0.22$$

2. Regression as Conditional Expectation

- **Conditional on a specific value of D,** what is our expectation (mean value) of y?

$$y_i = \alpha + \beta_1 D_i + \epsilon_i$$

$$Attitude_i = \alpha + \beta_1 Income_i + N(0, \sigma^2)$$

$$Attitude_i = 2.235 - 0.000818 * Income_i + N(0, 2.38)$$

2. Regression as Conditional Expectation

- **Conditional on a specific value of D,** what is our expectation (mean value) of y?

$$y_i = \alpha + \beta_1 D_i + \epsilon_i$$

$$Attitude_i = \alpha + \beta_1 Income_i + N(0, \sigma^2)$$

$$Attitude_i = 2.235 - 0.000818 * Income_i + N(0, 2.38)$$

- $E(Attitude|Income)$
  - When income is 3000, the average attitude is -0.22

2. Regression as Conditional Expectation

- **Conditional on a specific value of D,** what is our expectation (mean value) of y?

$$y_i = \alpha + \beta_1 D_i + \epsilon_i$$

$$Attitude_i = \alpha + \beta_1 Income_i + N(0, \sigma^2)$$

$$Attitude_i = 2.235 - 0.000818 * Income_i + N(0, 2.38)$$

- $E(Attitude|Income)$
  - When income is 3000, the average attitude is -0.22
  - When income is 6000, the average attitude is -2.67

2. Regression as Conditional Expectation

- **Conditional on a specific value of D,** what is our expectation (mean value) of y?

$$y_i = \alpha + \beta_1 D_i + \epsilon_i$$

$$Attitude_i = \alpha + \beta_1 Income_i + N(0, \sigma^2)$$

$$Attitude_i = 2.235 - 0.000818 * Income_i + N(0, 2.38)$$

- $E(Attitude|Income)$
  - When income is 3000, the average attitude is -0.22
  - When income is 6000, the average attitude is -2.67
  - When income is -1000, the average attitude is 3.05

2. Regression as Conditional Expectation

- **Conditional on a specific value of D,** what is our expectation (mean value) of y?

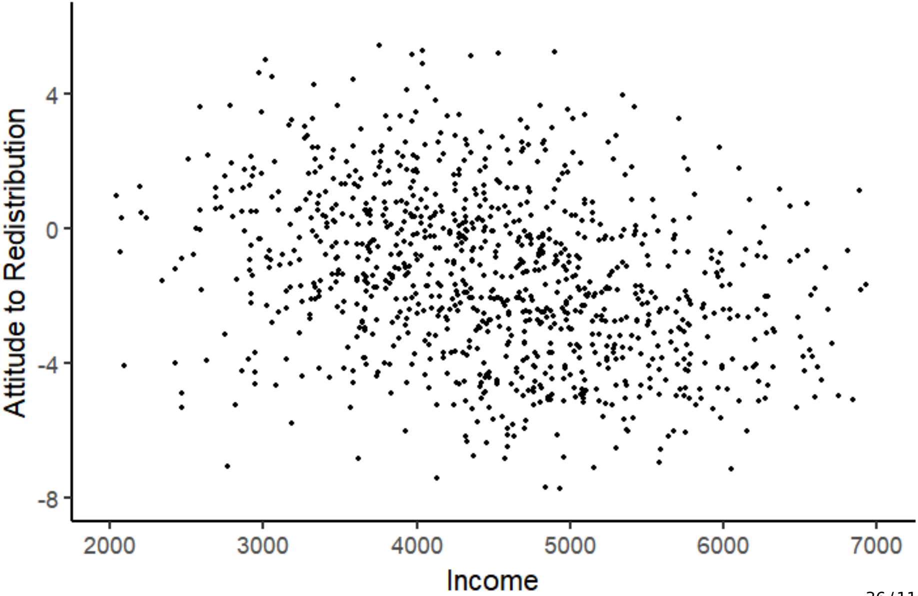
$$y_i = \alpha + \beta_1 D_i + \epsilon_i$$

$$Attitude_i = \alpha + \beta_1 Income_i + N(0, \sigma^2)$$

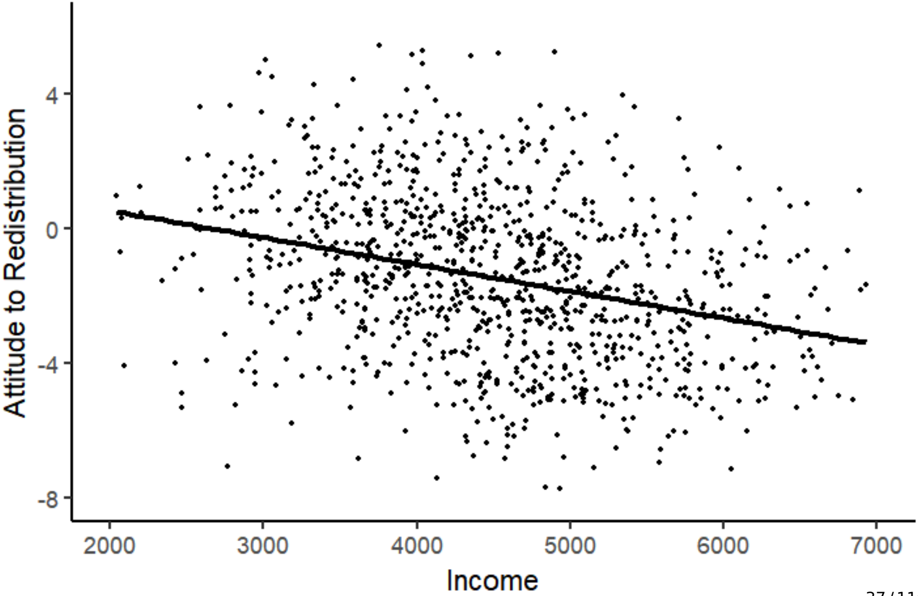
$$Attitude_i = 2.235 - 0.000818 * Income_i + N(0, 2.38)$$

- $E(Attitude|Income)$
  - When income is 3000, the average attitude is -0.22
  - When income is 6000, the average attitude is -2.67
  - When income is -1000, the average attitude is 3.05
- $E(Attitude|$Income, Age, Gender, Municipality)

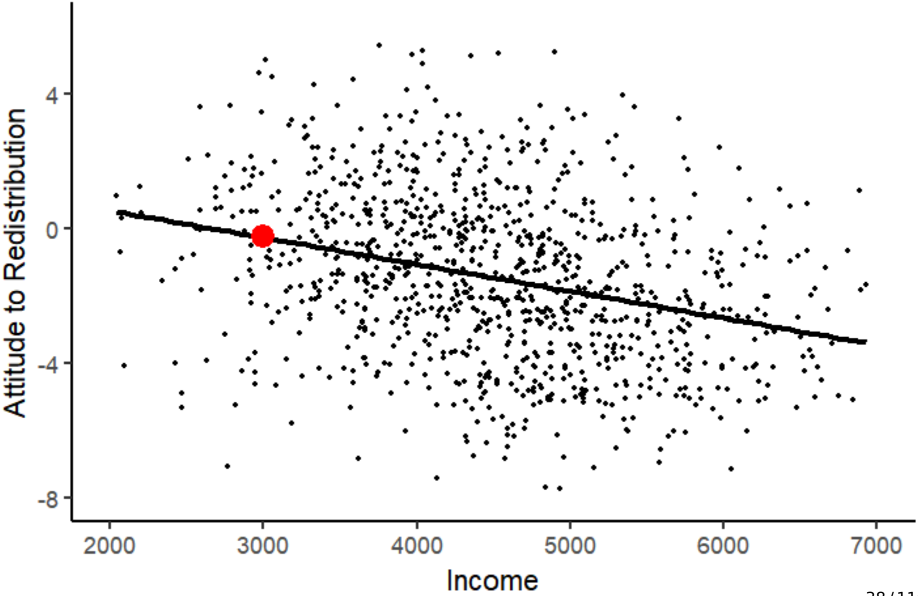
2. Regression as Conditional Expectation



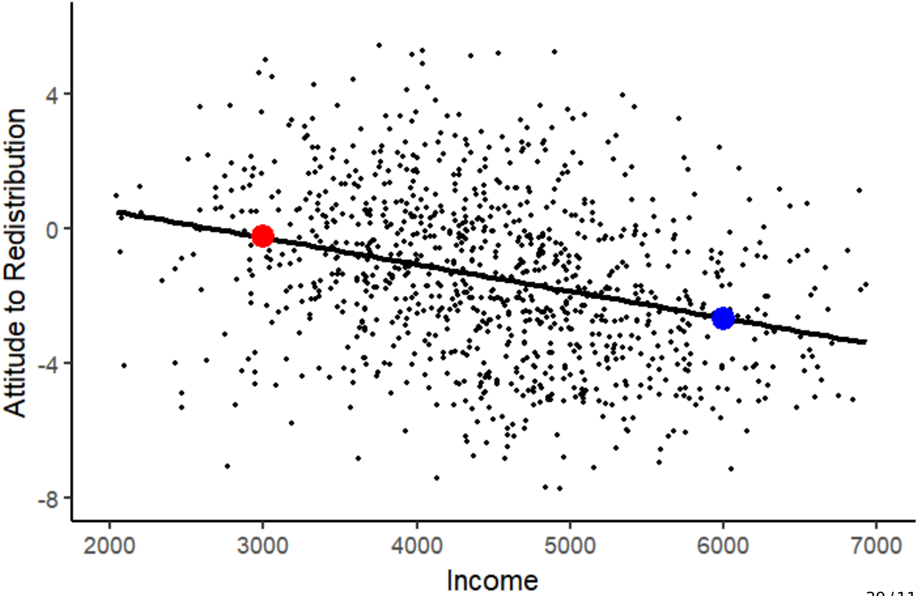
2. Regression as Conditional Expectation



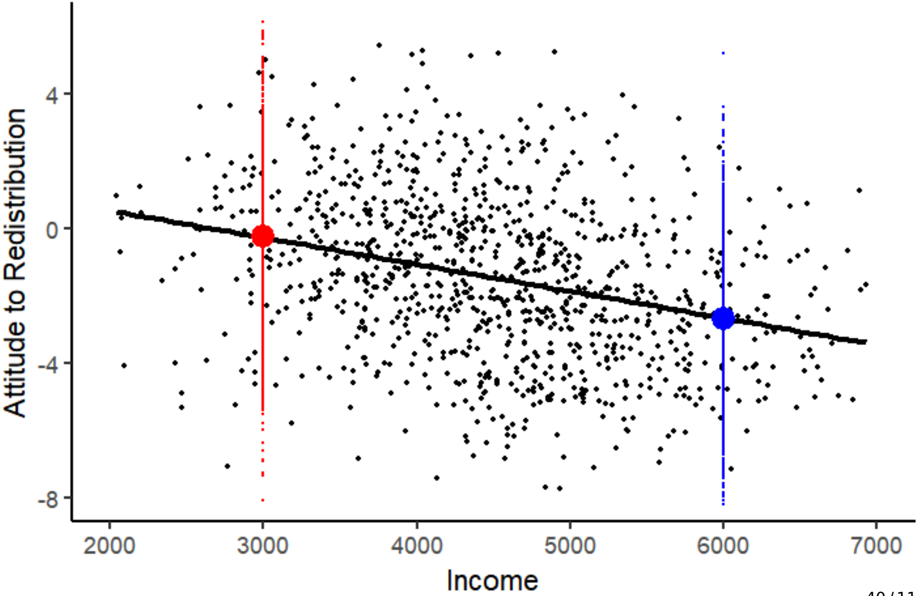
2. Regression as Conditional Expectation



2. Regression as Conditional Expectation



2. Regression as Conditional Expectation



2. Regression as Conditional Expectation

- How do we work out the conditional expectation? We estimate the βs

2. Regression as Conditional Expectation

- How do we work out the conditional expectation? We estimate the βs
- But we **NEVER** know the exact value of β

2. Regression as Conditional Expectation

- How do we work out the conditional expectation? We estimate the βs
- But we **NEVER** know the exact value of β
- Regression **estimates a distribution** for each β

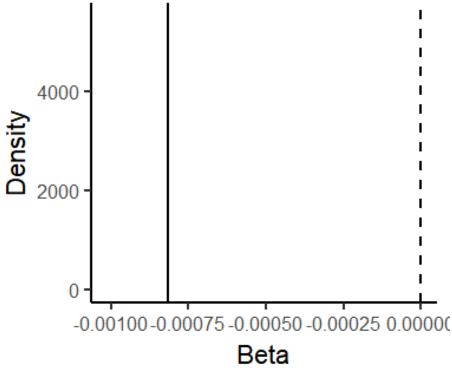
2. Regression as Conditional Expectation

- How do we work out the conditional expectation? We estimate the βs
- But we **NEVER** know the exact value of β
- Regression **estimates a distribution** for each β
  - That's why every β comes with a standard error

2. Regression as Conditional Expectation

- How do we work out the conditional expectation? We estimate the βs
- But we **NEVER** know the exact value of β
- Regression **estimates a distribution** for each β
  - That's why every β comes with a standard error

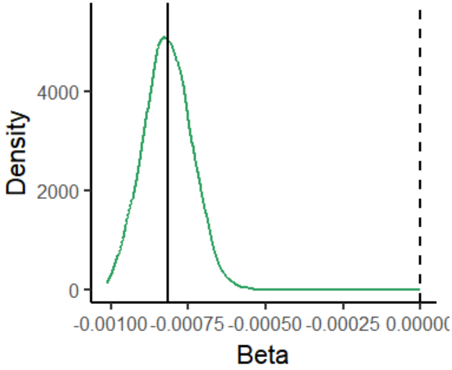
	Dependent variable:
	redist
income	-0.000818^{***} (0.000078)
Constant	2.234719^{***} (0.361477)
Observations	1,000
Note:	*p<0.1; **p<0.05; ***p<0.01



2. Regression as Conditional Expectation

- How do we work out the conditional expectation? We estimate the βs
- But we **NEVER** know the exact value of β
- Regression **estimates a distribution** for each β
  - That's why every β comes with a standard error

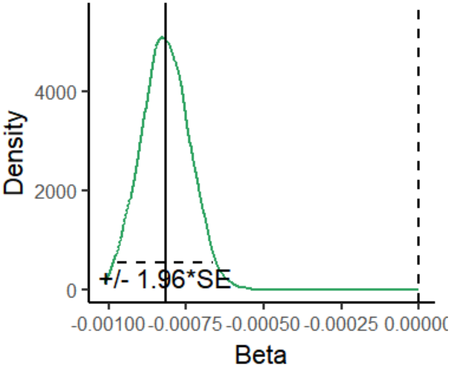
	Dependent variable:
	redist
income	-0.000818^{***} (0.000078)
Constant	2.234719^{***} (0.361477)
Observations	1,000
Note:	*p<0.1; **p<0.05; ***p<0.01



2. Regression as Conditional Expectation

- How do we work out the conditional expectation? We estimate the βs
- But we **NEVER** know the exact value of β
- Regression **estimates a distribution** for each β
  - That's why every β comes with a standard error

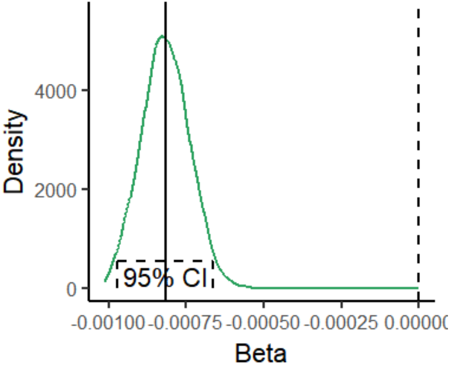
	Dependent variable:
	redist
income	-0.000818^{***} (0.000078)
Constant	2.234719^{***} (0.361477)
Observations	1,000
Note:	*p<0.1; **p<0.05; ***p<0.01



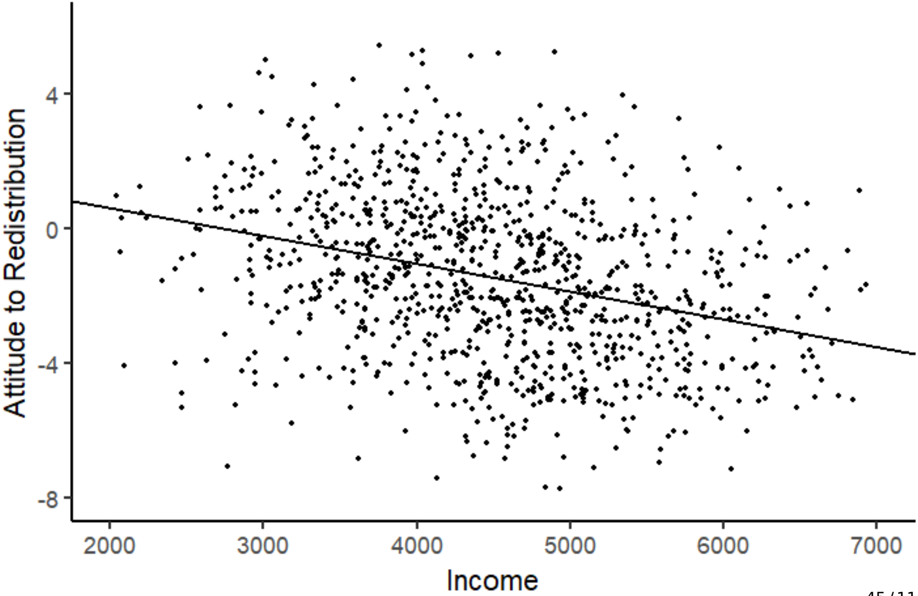
2. Regression as Conditional Expectation

- How do we work out the conditional expectation? We estimate the βs
- But we **NEVER** know the exact value of β
- Regression **estimates a distribution** for each β
  - That's why every β comes with a standard error

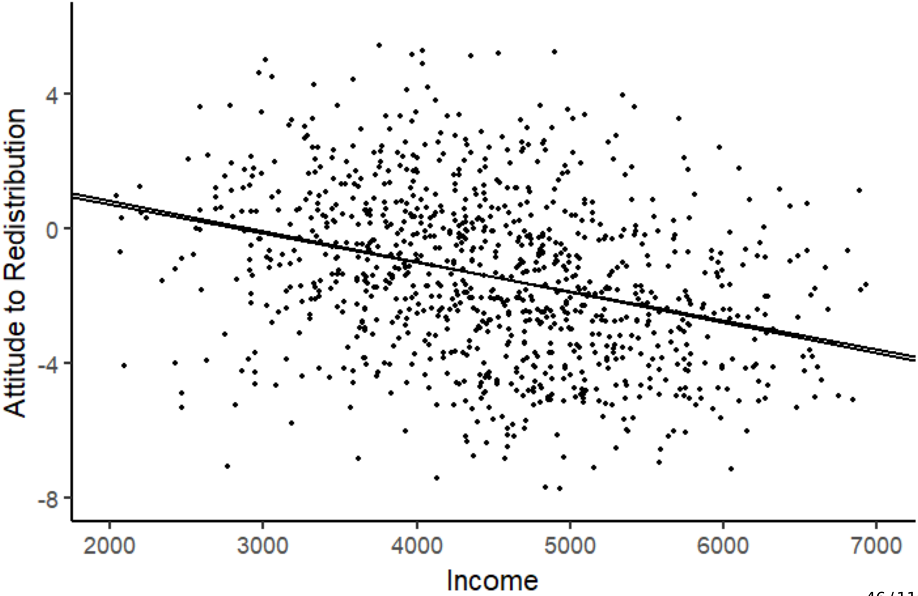
	Dependent variable:
	redist
income	-0.000818^{***} (0.000078)
Constant	2.234719^{***} (0.361477)
Observations	1,000
Note:	*p<0.1; **p<0.05; ***p<0.01



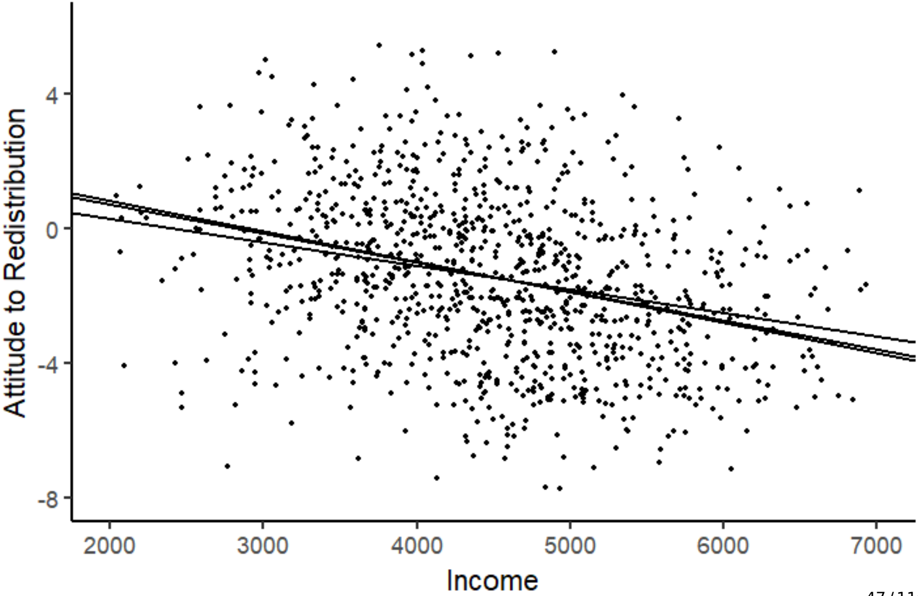
2. Regression as Conditional Expectation



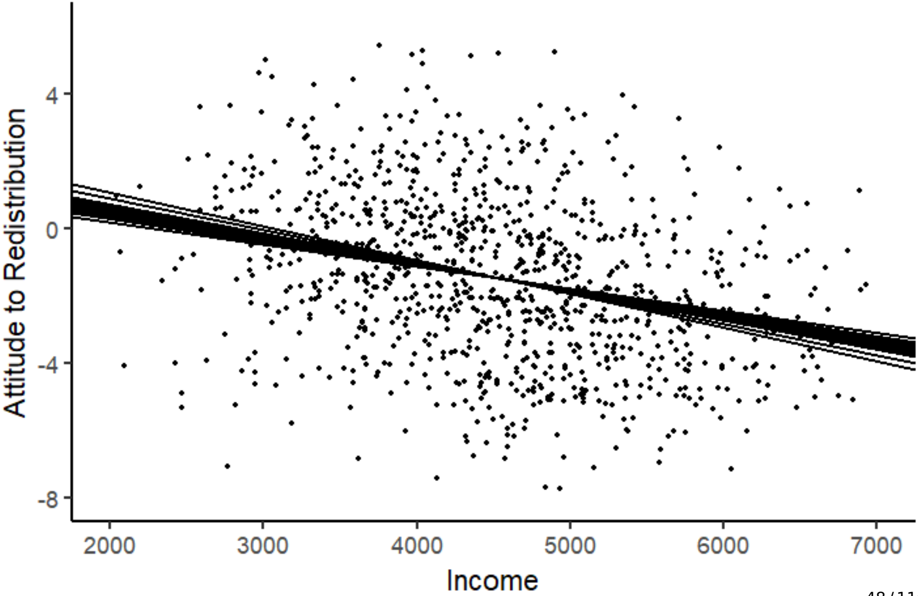
2. Regression as Conditional Expectation



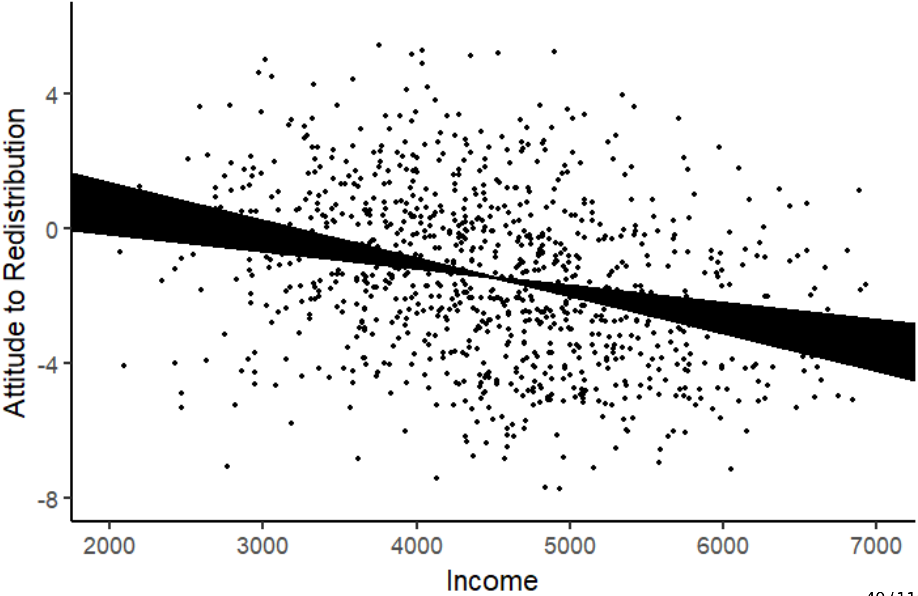
2. Regression as Conditional Expectation



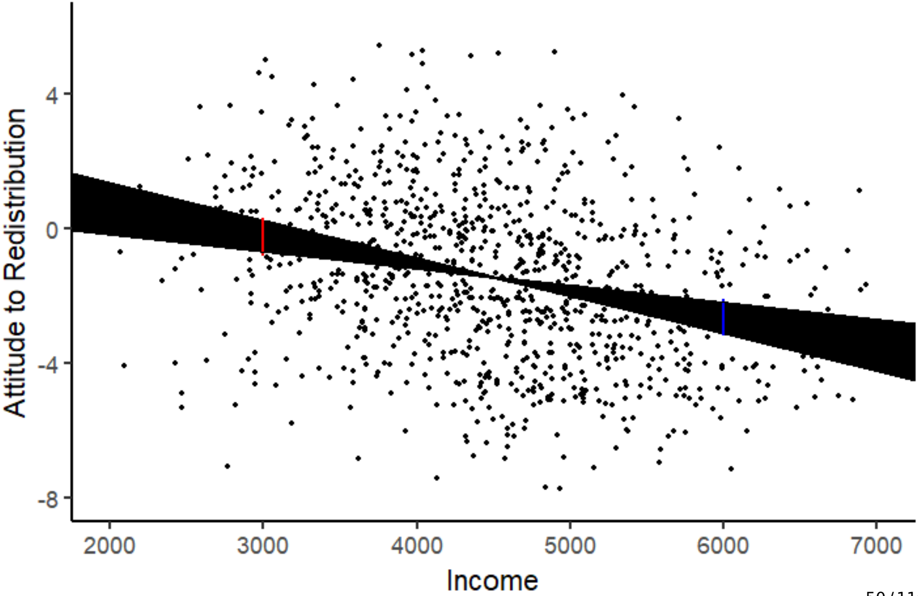
2. Regression as Conditional Expectation



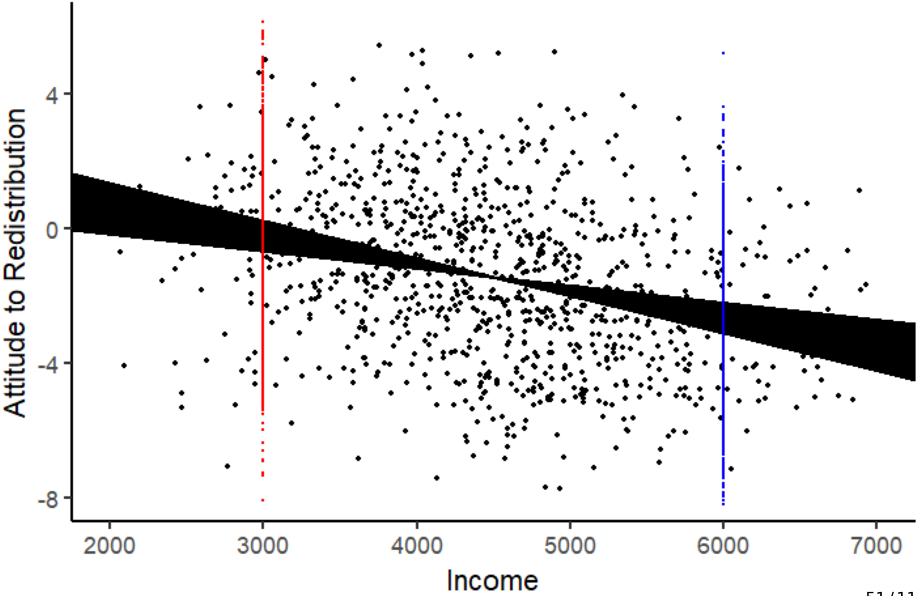
2. Regression as Conditional Expectation



2. Regression as Conditional Expectation



2. Regression as Conditional Expectation



3. Regression as (Partial) Correlation

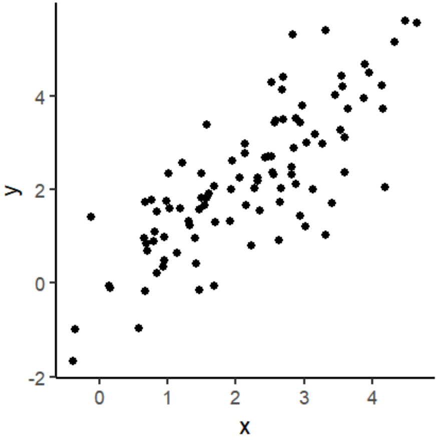
- Regression with two variables is very similar to calculating correlation:

3. Regression as (Partial) Correlation

- Regression with two variables is very similar to calculating correlation:
- $\hat{\beta} = cor(x, y) * \frac{\sigma_Y}{\sigma_X}$

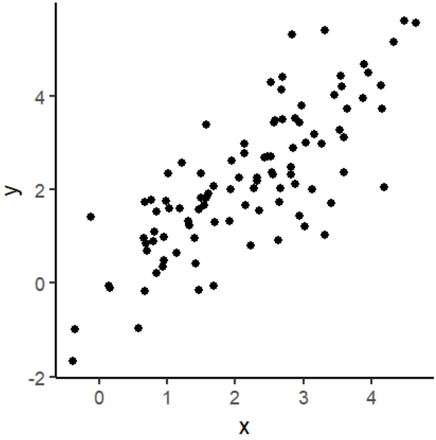
3. Regression as (Partial) Correlation

- Regression with two variables is very similar to calculating correlation:
- $\hat{\beta} = cor(x, y) * \frac{\sigma_Y}{\sigma_X}$



3. Regression as (Partial) Correlation

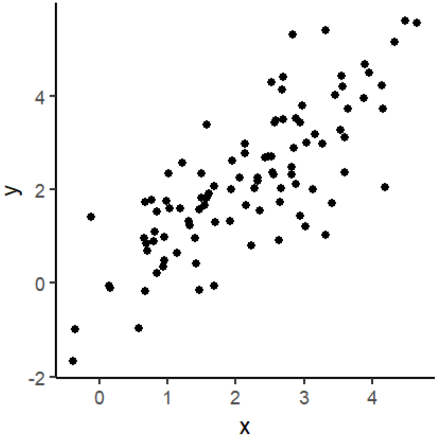
- Regression with two variables is very similar to calculating correlation:
- $\hat{\beta} = cor(x, y) * \frac{\sigma_Y}{\sigma_X}$



- Correlation is 0.781

3. Regression as (Partial) Correlation

- Regression with two variables is very similar to calculating correlation:
- $\hat{\beta} = cor(x, y) * \frac{\sigma_Y}{\sigma_X}$

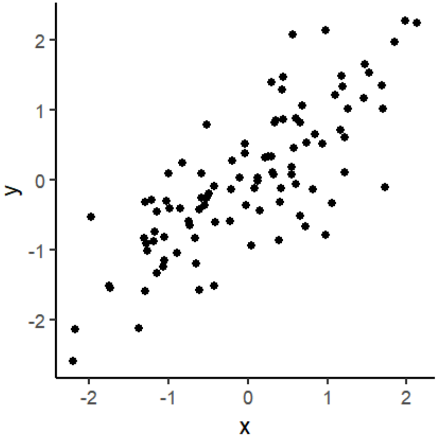


- Correlation is 0.781
- Regression Results:

	term	estimate
1	(Intercept)	0.006
2	x	1.008

3. Regression as (Partial) Correlation

- Regression with two variables is very similar to calculating correlation:
- $\hat{\beta} = cor(x, y) * \frac{\sigma_Y}{\sigma_X}$



- Correlation is 0.781
- It's **identical** if we standardize both variables first ($\frac{(x_i - \bar{x})}{\sigma_x}$)
- Standardized Regression Results:

	term	estimate
1	(Intercept)	0.000
2	x	0.781

3. Regression as (Partial) Correlation

- Regression with **multiple** variables is very similar to calculating **partial** correlation

3. Regression as (Partial) Correlation

- Regression with **multiple** variables is very similar to calculating **partial** correlation
- $y_i = \alpha + \beta_1 x_1 + \beta_2 x_2 + \epsilon_i$

3. Regression as (Partial) Correlation

- Regression with **multiple** variables is very similar to calculating **partial** correlation
- $y_i = \alpha + \beta_1 x_1 + \beta_2 x_2 + \epsilon_i$
- Just a small difference in the denominator (how we standardize the measure)

3. Regression as (Partial) Correlation

- Regression with **multiple** variables is very similar to calculating **partial** correlation
- $y_i = \alpha + \beta_1 x_1 + \beta_2 x_2 + \epsilon_i$
- Just a small difference in the denominator (how we standardize the measure)

$$\beta_{x_1} = \frac{r_{yx_1} - r_{yx_2} r_{x_1 x_2}}{1 - r^2_{x_1 x_2}}$$

$$r_{yx_1|x_2} = \frac{r_{yx_1} - r_{yx_2} r_{x_1 x_2}}{\sqrt{(1 - r^2_{yx_2})(1 - r^2_{x_1 x_2})}}$$

- **There is no magic in regression, it's just 'extra' correlation**

Section 2

Guide to 'Smart' Regression

Regression Guide

1. We will use regression throughout this course

Regression Guide

1. We will use regression throughout this course
2. But in a very **precise** way for each methodology

Regression Guide

1. We will use regression throughout this course
2. But in a very **precise** way for each methodology
3. There are fundamental best practices that apply to all the methodologies

Regression Guide

1. **Choose Variables and Measures:** To test a specific hypothesis

Regression Guide

1. **Choose Variables and Measures:** To test a specific hypothesis
2. **Choose the Data:** Throw out data we cannot learn from!

Regression Guide

1. **Choose Variables and Measures:** To test a specific hypothesis
2. **Choose the Data:** Throw out data we cannot learn from!
3. **Choose a Model/Link Function:** To match the data type of your outcome variable

Regression Guide

1. **Choose Variables and Measures:** To test a specific hypothesis
2. **Choose the Data:** Throw out data we cannot learn from!
3. **Choose a Model/Link Function:** To match the data type of your outcome variable
4. **Choose Covariates:** To make specific comparisons

Regression Guide

1. **Choose Variables and Measures:** To test a specific hypothesis
2. **Choose the Data:** Throw out data we cannot learn from!
3. **Choose a Model/Link Function:** To match the data type of your outcome variable
4. **Choose Covariates:** To make specific comparisons
5. **Choose Fixed Effects:** To focus on comparisons at a specific level

Regression Guide

1. **Choose Variables and Measures:** To test a specific hypothesis
2. **Choose the Data:** Throw out data we cannot learn from!
3. **Choose a Model/Link Function:** To match the data type of your outcome variable
4. **Choose Covariates:** To make specific comparisons
5. **Choose Fixed Effects:** To focus on comparisons at a specific level
6. **Choose Error Structure:** To match known dependencies/clustering in the data or sampling

Regression Guide

1. **Choose Variables and Measures:** To test a specific hypothesis
2. **Choose the Data:** Throw out data we cannot learn from!
3. **Choose a Model/Link Function:** To match the data type of your outcome variable
4. **Choose Covariates:** To make specific comparisons
5. **Choose Fixed Effects:** To focus on comparisons at a specific level
6. **Choose Error Structure:** To match known dependencies/clustering in the data or sampling
7. **Interpret the Coefficients:** To match the type/scale of the explanatory variable, outcome variable and model

Regression Guide

1. **Choose Variables and Measures:** To test a specific hypothesis
2. **Choose the Data:** Throw out data we cannot learn from!
3. **Choose a Model/Link Function:** To match the data type of your outcome variable
4. **Choose Covariates:** To make specific comparisons
5. **Choose Fixed Effects:** To focus on comparisons at a specific level
6. **Choose Error Structure:** To match known dependencies/clustering in the data or sampling
7. **Interpret the Coefficients:** To match the type/scale of the explanatory variable, outcome variable and model
8. **Predict Meaningful Comparisons:** To communicate your findings

1. Variables and Measures

- For the research question “Does income affect attitudes to redistribution?”

1. Variables and Measures

- For the research question “Does income affect attitudes to redistribution?”
- What measure of income should we use?

1. Variables and Measures

- For the research question "Does income affect attitudes to redistribution?"
- What measure of income should we use?
  - Pre-tax, post-tax, after government benefits?
- It depends on the theory we are testing

2. Data Sample

- For the research question "Does income affect attitudes to redistribution?"

2. Data Sample

- For the research question “Does income affect attitudes to redistribution?”
- We are conducting a within-country analysis

2. Data Sample

- For the research question “Does income affect attitudes to redistribution?”
- We are conducting a within-country analysis
- But everyone in our data from Qatar earns exactly $1m - no variation in income!

2. Data Sample

- For the research question “Does income affect attitudes to redistribution?”
- We are conducting a within-country analysis
- But everyone in our data from Qatar earns exactly \$1m - no variation in income!
- We may as well throw the Qatar data away

3. Regression Models

The Regression Model reflects the data type of the outcome variable:

- Continuous -> Ordinary Least Squares
  - "Pick a precise number that reflects your attitude to redistribution"

3. Regression Models

The Regression Model reflects the data type of the outcome variable:

- Continuous -> Ordinary Least Squares
  - “Pick a precise number that reflects your attitude to redistribution”
- Binary -> Logit
  - “Do you support redistribution, yes or no?”

3. Regression Models

The Regression Model reflects the data type of the outcome variable:

- Continuous -> Ordinary Least Squares
  - "Pick a precise number that reflects your attitude to redistribution"
- Binary -> Logit
  - "Do you support redistribution, yes or no?"
- Unordered categories -> Multinomial logit
  - "Do you think redistribution is a western, oriental or african concept?"

3. Regression Models

The Regression Model reflects the data type of the outcome variable:

- Continuous -> Ordinary Least Squares
  - "Pick a precise number that reflects your attitude to redistribution"
- Binary -> Logit
  - "Do you support redistribution, yes or no?"
- Unordered categories -> Multinomial logit
  - "Do you think redistribution is a western, oriental or african concept?"
- Ordered categories -> Ordered logit
  - "Do you want a lot more, more, the same, less, or a lot less redistribution?"

3. Regression Models

The Regression Model reflects the data type of the outcome variable:

- Continuous -> Ordinary Least Squares
  - "Pick a precise number that reflects your attitude to redistribution"
- Binary -> Logit
  - "Do you support redistribution, yes or no?"
- Unordered categories -> Multinomial logit
  - "Do you think redistribution is a western, oriental or african concept?"
- Ordered categories -> Ordered logit
  - "Do you want a lot more, more, the same, less, or a lot less redistribution?"
- Count -> Poisson
  - "In the past year, how many times have you complained about redistribution?"

4. Covariates

- Which covariates should we include?

4. Covariates

- Which covariates should we include?
- Which comparisons do we want to make?

4. Covariates

- Which covariates should we include?
- Which comparisons do we want to make?
- Control for gender if we want to compare men with men, women with women

4. Covariates

- Which covariates should we include?
- Which comparisons do we want to make?
- Control for gender if we want to compare men with men, women with women
- Only include controls where there is theory or evidence that this variable could be an **omitted variable**

4. Covariates

- Which covariates should we include?
- Which comparisons do we want to make?
- Control for gender if we want to compare men with men, women with women
- Only include controls where there is theory or evidence that this variable could be an **omitted variable**
- Controlling for post-treatment variables can make your estimate *worse*

5. Fixed Effects

- Data are usually hierarchical: countries, states, municipalities, neighbourhoods, families, individuals

5. Fixed Effects

- Data are usually hierarchical: countries, states, municipalities, neighbourhoods, families, individuals
- A fixed effect for countries means we only compare people within the same country

5. Fixed Effects

- Data are usually hierarchical: countries, states, municipalities, neighbourhoods, families, individuals
- A fixed effect for countries means we only compare people within the same country
- Removing *ALL* the variation between countries
  - If rich *countries* have stronger attitudes to redistribution, we control for this
  - So we can ask whether richer *people* have stronger attitudes

5. Fixed Effects

- Data are usually hierarchical: countries, states, municipalities, neighbourhoods, families, individuals
- A fixed effect for countries means we only compare people within the same country
- Removing *ALL* the variation between countries
  - If rich *countries* have stronger attitudes to redistribution, we control for this
  - So we can ask whether richer *people* have stronger attitudes
- Our question becomes: How do variations within income in the same country affect attitudes to redistribution?

6. Errors Structure

- An assumption of regression analysis is that the errors are independent

6. Errors Structure

- An assumption of regression analysis is that the errors are independent
  - Knowing the value of one error tells you *nothing* about the value of the next error

6. Errors Structure

- An assumption of regression analysis is that the errors are independent
  - Knowing the value of one error tells you *nothing* about the value of the next error
- But attitudes to redistribution are probably very similar to everyone you live with, even after controlling for income etc.

6. Errors Structure

- An assumption of regression analysis is that the errors are independent
  - Knowing the value of one error tells you *nothing* about the value of the next error
- But attitudes to redistribution are probably very similar to everyone you live with, even after controlling for income etc.
- Due to 'unobservable' variables (conversations over dinner...)

6. Errors Structure

- An assumption of regression analysis is that the errors are independent
  - Knowing the value of one error tells you *nothing* about the value of the next error
- But attitudes to redistribution are probably very similar to everyone you live with, even after controlling for income etc.
- Due to 'unobservable' variables (conversations over dinner...)
- So we don't really have 2 observations, we have closer to 1 'independent' observation

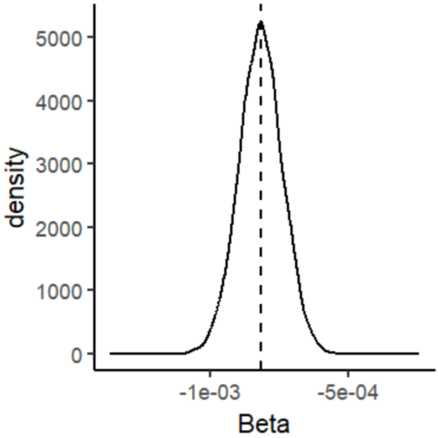
6. Errors Structure

- An assumption of regression analysis is that the errors are independent
  - Knowing the value of one error tells you *nothing* about the value of the next error
- But attitudes to redistribution are probably very similar to everyone you live with, even after controlling for income etc.
- Due to 'unobservable' variables (conversations over dinner...)
- So we don't really have 2 observations, we have closer to 1 'independent' observation
- So the standard errors for our β's are *over-confident* (too small)

6. Errors Structure

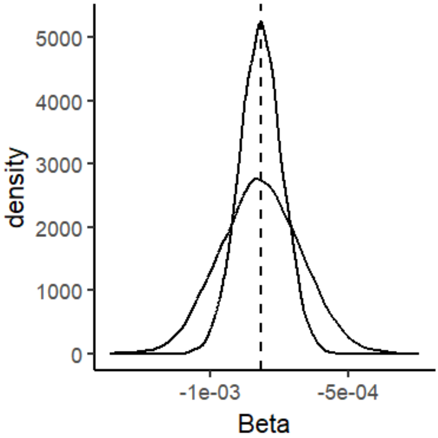
- An assumption of regression analysis is that the errors are independent
  - Knowing the value of one error tells you *nothing* about the value of the next error
- But attitudes to redistribution are probably very similar to everyone you live with, even after controlling for income etc.
- Due to 'unobservable' variables (conversations over dinner...)
- So we don't really have 2 observations, we have closer to 1 'independent' observation
- So the standard errors for our β's are *over-confident* (too small)
- We need to adjust for these dependencies with clustered standard errors
  - Created by the underlying structure of the data
  - Or by our data sampling process

6. Errors Structure



- The distribution of our estimated betas suggests we're pretty confident β is close to -0.0008175

6. Errors Structure



- With clustered SEs, the wider distribution of our betas suggests we're *less* confident β is close to -0.0008175

7. Interpreting Regression Results

- Difficult! It depends on:
  1. The scale of the explanatory variable

7. Interpreting Regression Results

- Difficult! It depends on:
  1. The scale of the explanatory variable
  2. The scale of the outcome

7. Interpreting Regression Results

- Difficult! It depends on:
  1. The scale of the explanatory variable
  2. The scale of the outcome
  3. The regression model we used

7. Interpreting Regression Results

- Difficult! It depends on:
  1. The scale of the explanatory variable
  2. The scale of the outcome
  3. The regression model we used
  4. The presence of any interaction

7. Interpreting Regression Results

- Difficult! It depends on:
  1. The scale of the explanatory variable
  2. The scale of the outcome
  3. The regression model we used
  4. The presence of any interaction
- Basic OLS: $y_i = \alpha + \beta D_i + \epsilon$

7. Interpreting Regression Results

- Difficult! It depends on:
    1. The scale of the explanatory variable
    2. The scale of the outcome
    3. The regression model we used
    4. The presence of any interaction
- Basic OLS: $y_i = \alpha + \beta D_i + \epsilon$
    - A 1 [unit of D] change in the explanatory variable is associated with a β [unit of y] change in the outcome, holding other variables constant

7. Interpreting Regression Results

- Difficult! It depends on:
  1. The scale of the explanatory variable
  2. The scale of the outcome
  3. The regression model we used
  4. The presence of any interaction
- Basic OLS **with log outcome**: $log(y_i) = \alpha + \beta D_i + \epsilon$
  - A 1 [unit of D] change in the explanatory variable is associated with a $100 * (e^{\beta} - 1)\%$ change in the outcome, holding other variables constant

7. Interpreting Regression Results

- Difficult! It depends on:
  1. The scale of the explanatory variable
  2. The scale of the outcome
  3. The regression model we used
  4. The presence of any interaction
- Basic OLS **with log treatment**: $y_i = \alpha + \beta log(D_i) + \epsilon$
  - A 1% change in the explanatory variable is associated with a $\beta * log(\frac{101}{100})$ change in the outcome, holding other variables constant

7. Interpreting Regression Results

- Difficult! It depends on:
  1. The scale of the explanatory variable
  2. The scale of the outcome
  3. The regression model we used
  4. The presence of any interaction
- **Logit:** $log(\frac{Pr(y_i=1)}{Pr(y_i=0)}) = \alpha + \beta D_i + \epsilon$
  - A 1 [unit of D] change in the explanatory variable is associated with a β change in the log-odds of $y_i = 1$, holding other variables constant

7. Interpreting Regression Results

- Difficult! It depends on:
  1. The scale of the explanatory variable
  2. The scale of the outcome
  3. The regression model we used
  4. The presence of any interaction
- **Logit:** $log(\frac{Pr(y_i=1)}{Pr(y_i=0)}) = \alpha + \beta D_i + \epsilon$
  - A 1 [unit of D] change in the explanatory variable is associated with a $100 * (e^{\beta} - 1)$% change in the odds (relative probability, $\frac{p}{1-p}$) of $y_i = 1$, holding other variables constant

7. Interpreting Regression Results

- Difficult! It depends on:
  1. The scale of the explanatory variable
  2. The scale of the outcome
  3. The regression model we used
  4. The presence of any interaction
- **Multinomial:** $log(\frac{Pr(y_i=C)}{Pr(y_i=B)}) = \alpha + \beta D_i + \epsilon$
  - A 1 [unit of D] change in the explanatory variable is associated with a $100 * (e^{\beta_C} - 1)$% change in the odds (relative probability, $\frac{p}{1-p}$) of moving from the baseline category B to the outcome category C, holding other variables constant

7. Interpreting Regression Results

- Difficult! It depends on:
  1. The scale of the explanatory variable
  2. The scale of the outcome
  3. The regression model we used
  4. The presence of any interaction
- **Ordered Multinomial:** $log(\frac{Pr(y_i=C)}{Pr(y_i=C-1)}) = \alpha + \beta D_i + \epsilon$
  - A 1 [unit of D] change in the explanatory variable is associated with a $100 * (e^{\beta} - 1)$% change in the odds (relative probability, $\frac{p}{1-p}$) of moving up one unit on the outcome scale, holding other variables constant

7. Interpreting Regression Results

- Difficult! It depends on:
  1. The scale of the explanatory variable
  2. The scale of the outcome
  3. The regression model we used
  4. The presence of any interaction
- **OLS with Interaction:** $y_i = \alpha + \beta_1 D_i + \beta_2 X_i + \beta_3 D_i * X_i + \epsilon$
  - $$\frac{\partial y}{\partial D} = \beta_1 + \beta_3 X$$
  - β_1 is the effect of D when $X = 0$: *May not make sense!*
  - Insert values for X and see how the marginal effect changes

7. Interpreting Regression Results

OLS with Interaction:

$$Redist_i = \alpha + \beta_1 Gender_i + \beta_2 Income_i$$
$$+ \beta_3 Gender_i * Income_i + \epsilon_i$$

7. Interpreting Regression Results

OLS with Interaction:

$$Redist_i = \alpha + \beta_1 Gender_i + \beta_2 Income_i$$
$$+ \beta_3 Gender_i * Income_i + \epsilon_i$$

$$\frac{\partial Redist}{\partial Gender} = \beta_1 + \beta_3 * Income$$

7. Interpreting Regression Results

OLS with Interaction:

$$Redist_i = \alpha + \beta_1 Gender_i + \beta_2 Income_i + \beta_3 Gender_i * Income_i + \epsilon_i$$

$$\frac{\partial Redist}{\partial Gender} = \beta_1 + \beta_3 * Income$$

$$\frac{\partial Redist}{\partial Income} = \beta_2 + \beta_3 * Gender$$

7. Interpreting Regression Results

OLS with Interaction:

$$Redist_i = \alpha + \beta_1 Gender_i + \beta_2 Income_i + \beta_3 Gender_i * Income_i + \epsilon_i$$

$$\frac{\partial Redist}{\partial Gender} = \beta_1 + \beta_3 * Income$$

$$\frac{\partial Redist}{\partial Income} = \beta_2 + \beta_3 * Gender$$

	Dependent variable:
	redist
gender1	-2.942614^{***} (0.700510)
income	0.079980^{***} (0.000110)
gender1:income	0.000986^{***} (0.000152)
Constant	0.112903 (0.454926)
Observations	1,000
Note:	*p<0.1; **p<0.05; ***p<0.01

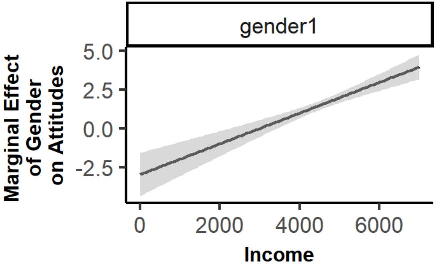
7. Interpreting Regression Results

OLS with Interaction:

$Redist_i = \alpha + \beta_1 Gender_i + \beta_2 Income_i$
$+ \beta_3 Gender_i * Income_i + \epsilon_i$

$$\frac{\partial Redist}{\partial Gender} = \beta_1 + \beta_3 * Income$$

$$\frac{\partial Redist}{\partial Income} = \beta_2 + \beta_3 * Gender$$



	Dependent variable:
	redist
gender1	−2.942614*** (0.700510)
income	0.079980*** (0.000110)
gender1:income	0.000986*** (0.000152)
Constant	0.112903 (0.454926)
Observations	1,000
Note:	*p<0.1; **p<0.05; ***p<0.01

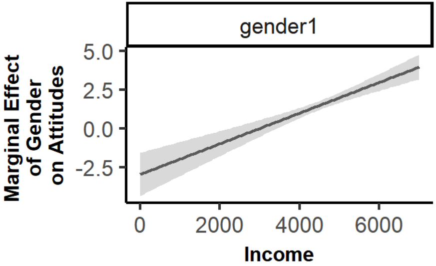
7. Interpreting Regression Results

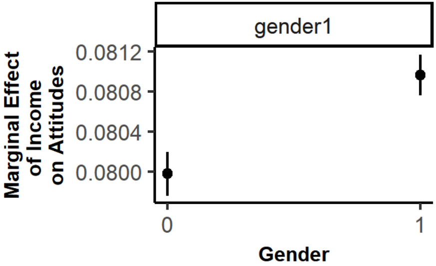
OLS with Interaction:

$Redist_i = \alpha + \beta_1 Gender_i + \beta_2 Income_i$
$+ \beta_3 Gender_i * Income_i + \epsilon_i$

$$\frac{\partial Redist}{\partial Gender} = \beta_1 + \beta_3 * Income$$

$$\frac{\partial Redist}{\partial Income} = \beta_2 + \beta_3 * Gender$$





	Dependent variable:
	redist
gender1	−2.942614*** (0.700510)
income	0.079980*** (0.000110)
gender1:income	0.000986*** (0.000152)
Constant	0.112903 (0.454926)
Observations	1,000
Note:	*p<0.1; **p<0.05; ***p<0.01

8. Predictions from Regressions

- The coefficient on the regression of income on attitude to redistribution is -0.000818

8. Predictions from Regressions

- The coefficient on the regression of income on attitude to redistribution is -0.000818
  - So??? What do we learn from this?

8. Predictions from Regressions

- The coefficient on the regression of income on attitude to redistribution is -0.000818
  - So??? What do we learn from this?
  - Coefficients are hard to interpret, and depend on how we measure each variable
  - And p-values are arbitrary (0.049 vs. 0.051)

8. Predictions from Regressions

- The coefficient on the regression of income on attitude to redistribution is -0.000818
  - So??? What do we learn from this?
  - Coefficients are hard to interpret, and depend on how we measure each variable
  - And p-values are arbitrary (0.049 vs. 0.051)
- Better to make specific *predictions* of how changes in D produce changes in Y

8. Predictions from Regressions

$$Attitude_i = \alpha + \beta_1 \text{ Income}_i + \epsilon_i$$

8. Predictions from Regressions

$$Attitude_i = \alpha + \beta_1 \text{ Income}_i + \epsilon_i$$

$$Attitude_i = 2.235 - 0.000818 \text{ Income}_i + N(0, 2.378)$$

8. Predictions from Regressions

$$Attitude_i = \alpha + \beta_1 \text{ Income}_i + \epsilon_i$$

$$Attitude_i = 2.235 - 0.000818 \text{ Income}_i + N(0, 2.378)$$

If Income is 3000:

$$Attitude_i = 2.235 - 0.000818 * 3000 + N(0, 2.378)$$

$$Attitude_i = -0.219 + N(0, 2.378)$$

8. Predictions from Regressions

$$Attitude_i = \alpha + \beta_1 \text{ Income}_i + \epsilon_i$$

$$Attitude_i = 2.235 - 0.000818 \text{ Income}_i + N(0, 2.378)$$

If Income is 6000:

$$Attitude_i = 2.235 - 0.000818 * 6000 + N(0, 2.378)$$

$$Attitude_i = -2.673 + N(0, 2.378)$$

8. Predictions from Regressions

$$Attitude_i = \alpha + \beta_1 \text{ Income}_i + \epsilon_i$$

$$Attitude_i = 2.235 - 0.000818 \text{ Income}_i + N(0, 2.378)$$

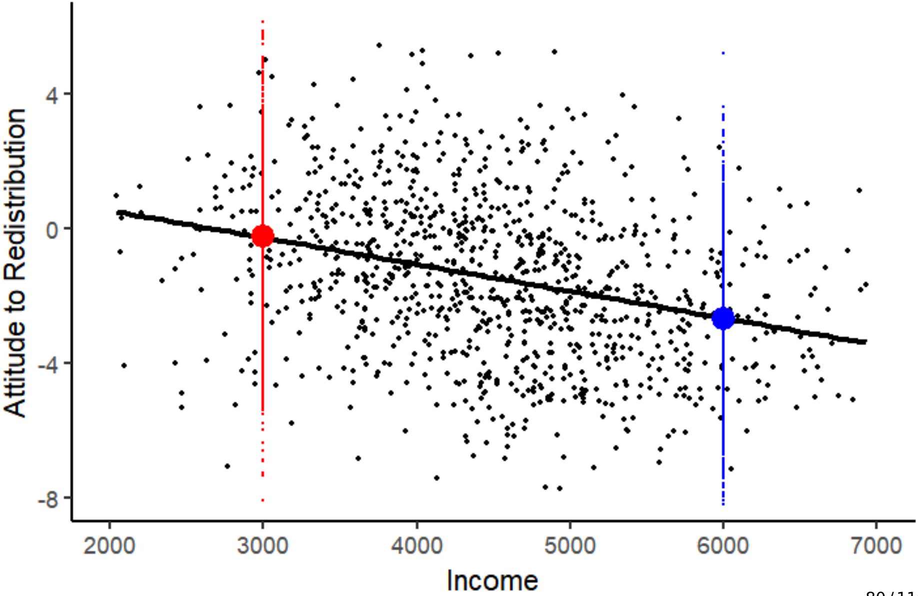
Increasing Income from 3000 to 6000:

$$\Delta Attitude_i = (2.235 - 0.000818 * 6000) - (2.235 - 0.000818 * 300$$

$$\Delta Attitude_i = -2.673 - (-0.219)$$

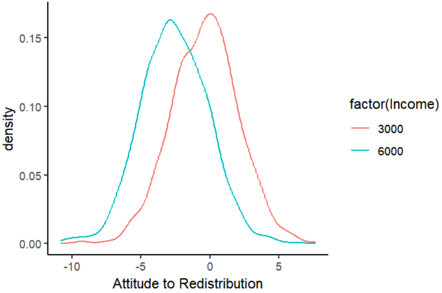
$$\Delta Attitude_i = -2.454$$

8. Predictions from Regressions



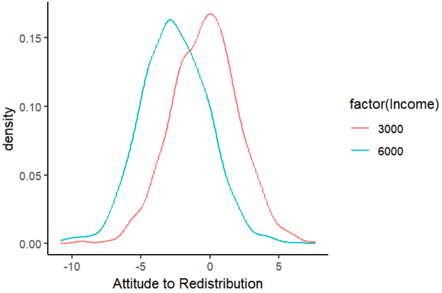
8. Predictions from Regressions

Predicted Values:

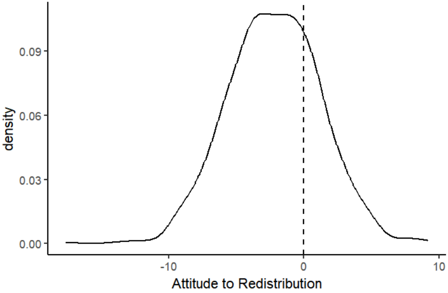


8. Predictions from Regressions

Predicted Values:



First Differences:



8. Predictions from Regressions

- The regression model matters because the wrong model makes non-sensical predictions

8. Predictions from Regressions

- The regression model matters because the wrong model makes non-sensical predictions
- Consider a binary outcome: $Gender_i = \alpha + \beta Income_i + \epsilon_i$

8. Predictions from Regressions

- The regression model matters because the wrong model makes non-sensical predictions
- Consider a binary outcome: $Gender_i = \alpha + \beta Income_i + \epsilon_i$
- Compare the OLS and Logit regression tables:

8. Predictions from Regressions

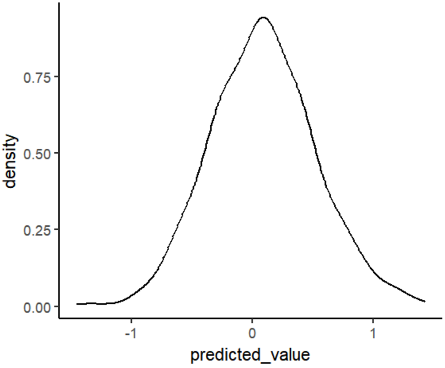
- The regression model matters because the wrong model makes non-sensical predictions
- Consider a binary outcome: $Gender_i = \alpha + \beta Income_i + \epsilon_i$
- Compare the OLS and Logit regression tables:

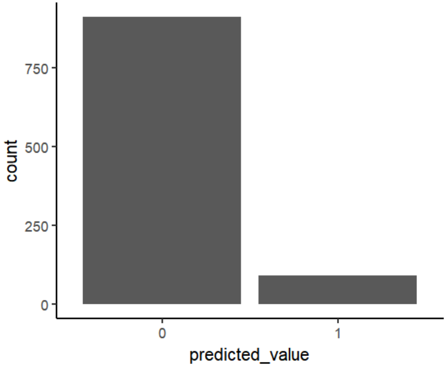
	Dependent variable:
	gender
income	0.0003*** (0.00001)
Constant	−0.696*** (0.066)
Observations	1,000
Note:	*p<0.1; **p<0.05; ***p<0.01

	Dependent variable:
	gender
income	0.001*** (0.0001)
Constant	−6.360*** (0.457)
Observations	1,000
Note:	*p<0.1; **p<0.05; ***p<0.01

8. Predictions from Regressions

- The regression model matters because the wrong model makes non-sensical predictions
- Consider a binary outcome: $Gender_i = \alpha + \beta Income_i + \epsilon_i$
- Compare the OLS and Logit **predictions** of gender for an income of R$3000:





Section 3

What Does Regression NOT Do?

What Does Regression NOT Do?

- Remember, regression is just fancy correlation

What Does Regression NOT Do?

- Remember, regression is just fancy correlation
- Even after following all this guidance, Regression does NOT:
  1. *Explain* anything
  2. Make bad data better
  3. Tell you which theory is 'correct'
  4. Make it clear what comparisons you are making

What Does Regression NOT Do?

- Remember, regression is just fancy correlation
- Even after following all this guidance, Regression does NOT:
  1. *Explain* anything
  2. Make bad data better
  3. Tell you which theory is 'correct'
  4. Make it clear what comparisons you are making
- These all require **research design**, **theory** and **assumptions**

What Does Regression NOT Do?

- **Correlation is not causation**

What Does Regression NOT Do?

- **Correlation is not causation**
  - If we look hard enough we can always find correlations

What Does Regression NOT Do?

- **Correlation is not causation**
  - If we look hard enough we can always find correlations
  - By chance...

What Does Regression NOT Do?

- **Correlation is not causation**
  - If we look hard enough we can always find correlations
  - By chance...
  - Due to complex social and historical patterns...

What Does Regression NOT Do?

- **Correlation is not causation**
  - If we look hard enough we can always find correlations
  - By chance...
  - Due to complex social and historical patterns...
  - But we cannot conclude that D causes or explains Y

What Does Regression NOT Do?

- **Correlation is not causation**
  - If we look hard enough we can always find correlations
  - By chance...
  - Due to complex social and historical patterns...
  - But we cannot conclude that D causes or explains Y
- *More* data will not help

What Does Regression NOT Do?

- **Correlation is not causation**
  - If we look hard enough we can always find correlations
  - By chance...
  - Due to complex social and historical patterns...
  - But we cannot conclude that D causes or explains Y
- *More* data will not help
- The problem is the *content* of data; it does not allow us to answer the causal question

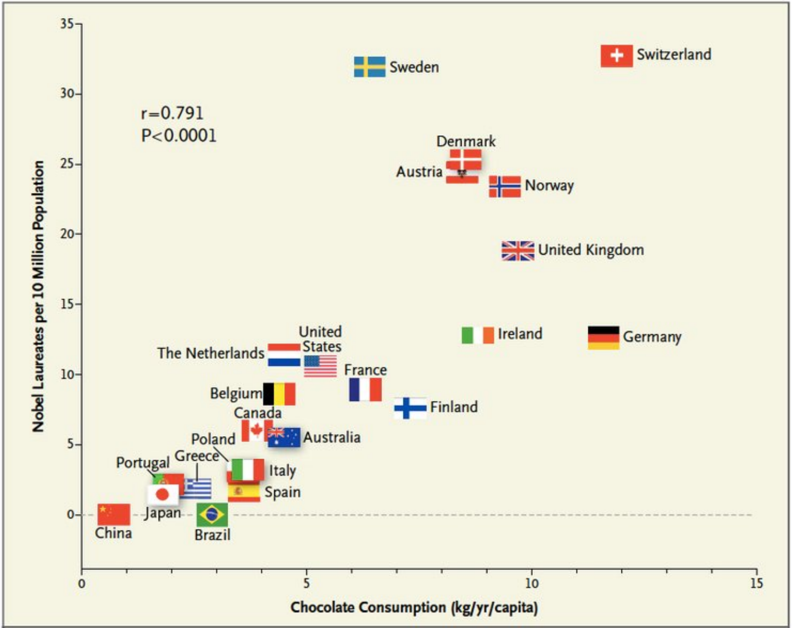
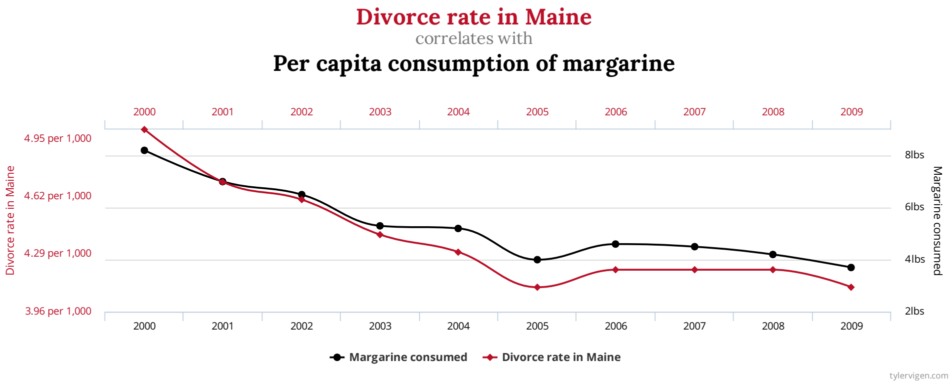
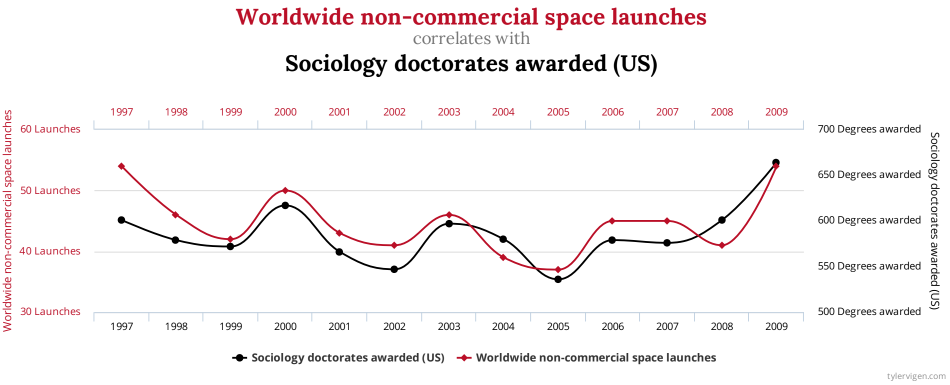
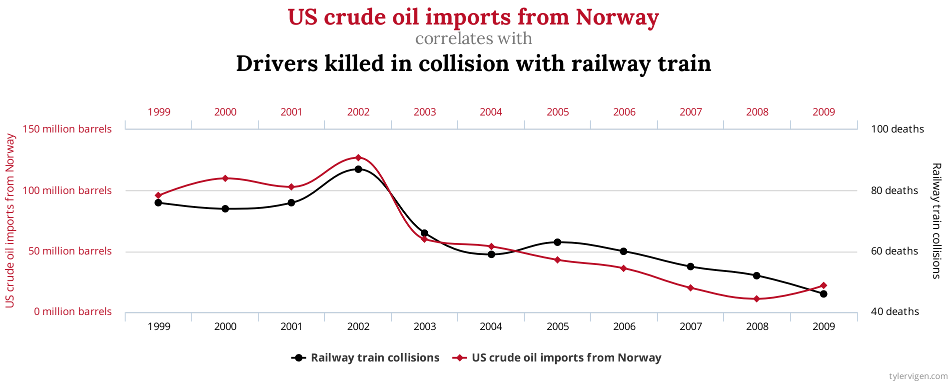
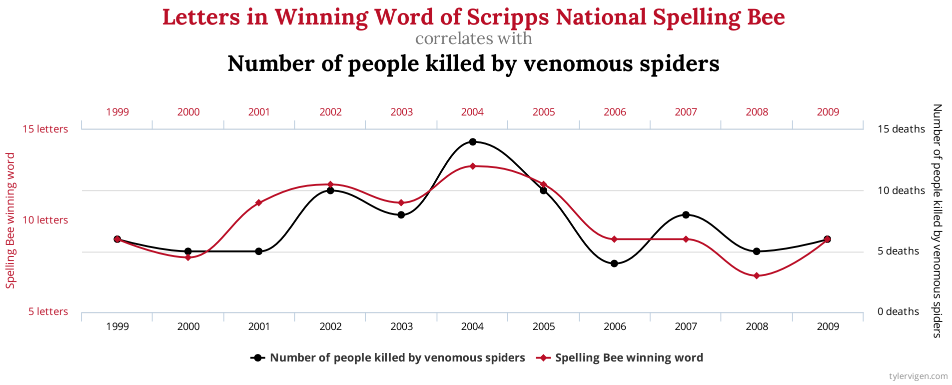


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.


Divorce rate in Maine
correlates with
Per capita consumption of margarine
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
4.95 per 1,000
4.62 per 1,000
4.29 per 1,000
3.96 per 1,000
Divorce rate in Maine
8lbs
6lbs
4lbs
2lbs
Margarine consumed
Margarine consumed
Divorce rate in Maine
tylervigen.com


Worldwide non-commercial space launches
correlates with
Sociology doctorates awarded (US)
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
60 Launches
50 Launches
40 Launches
30 Launches
Worldwide non-commercial space launches
700 Degrees awarded
650 Degrees awarded
600 Degrees awarded
550 Degrees awarded
500 Degrees awarded
Sociology doctorates awarded (US)
Sociology doctorates awarded (US)
Worldwide non-commercial space launches
tylervigen.com


US crude oil imports from Norway
correlates with
Drivers killed in collision with railway train
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
150 million barrels
100 million barrels
50 million barrels
0 million barrels
US crude oil imports from Norway
100 deaths
80 deaths
60 deaths
40 deaths
Railway train collisions
Railway train collisions
US crude oil imports from Norway
tylervigen.com


Letters in Winning Word of Scripps National Spelling Bee
correlates with
Number of people killed by venomous spiders
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
15 letters
10 letters
5 letters
Spelling Bee winning word
15 deaths
10 deaths
5 deaths
0 deaths
Number of people killed by venomous spiders
Number of people killed by venomous spiders
Spelling Bee winning word
tylervigen.com

What Does Regression NOT Do?

- Why is correlation (regression) not causation?

What Does Regression NOT Do?

- Why is correlation (regression) not causation?
  1. Omitted Variable Bias

What Does Regression NOT Do?

- Why is correlation (regression) not causation?
  1. Omitted Variable Bias
  2. Reverse Causation

What Does Regression NOT Do?

- Why is correlation (regression) not causation?
  1. Omitted Variable Bias
  2. Reverse Causation
  3. Selection Bias

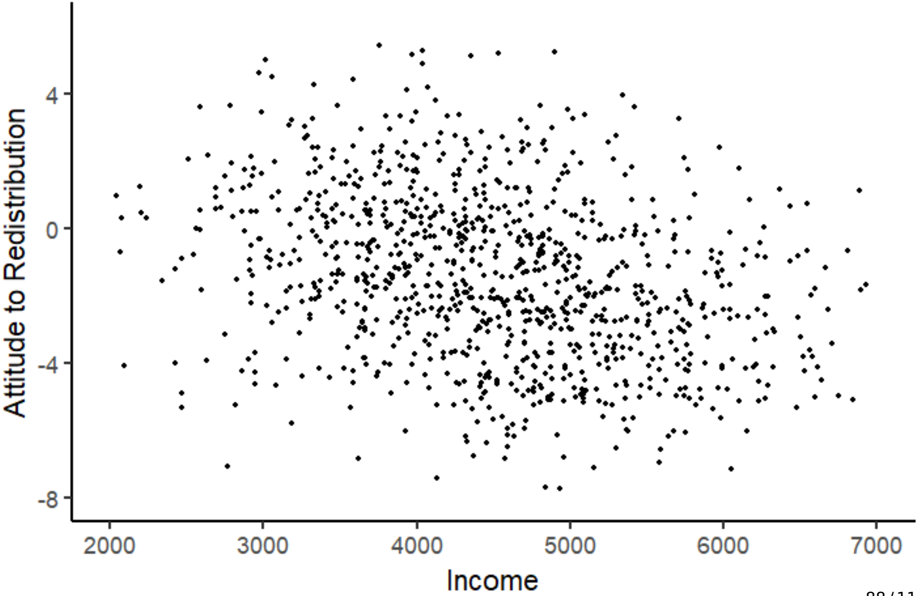
What Does Regression NOT Do?

- Why is correlation (regression) not causation?
  1. Omitted Variable Bias
  2. Reverse Causation
  3. Selection Bias
  4. Measurement Bias

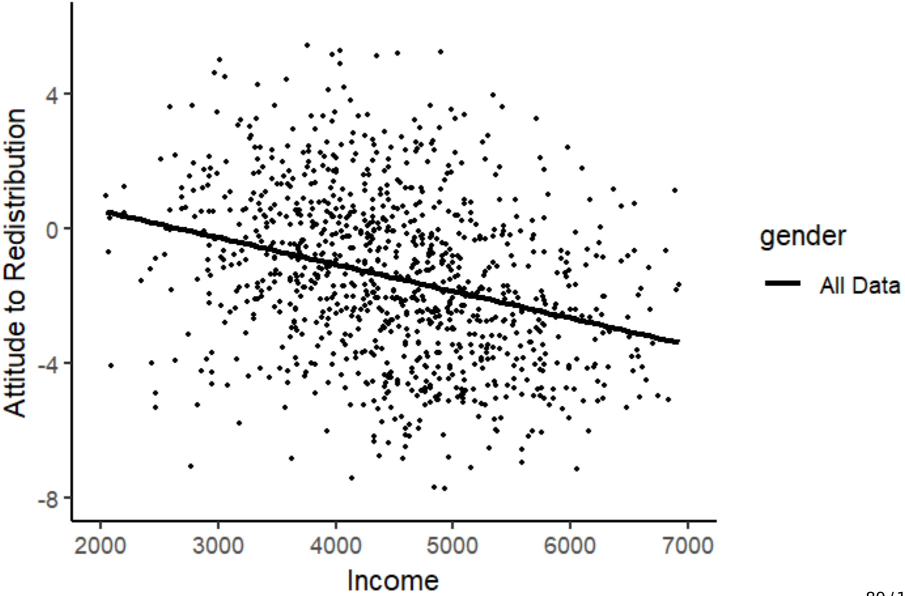
What Does Regression NOT Do?

- Why is correlation (regression) not causation?
  1. Omitted Variable Bias
  2. Reverse Causation
  3. Selection Bias
  4. Measurement Bias
  5. Lack of Overlap, Model Dependence

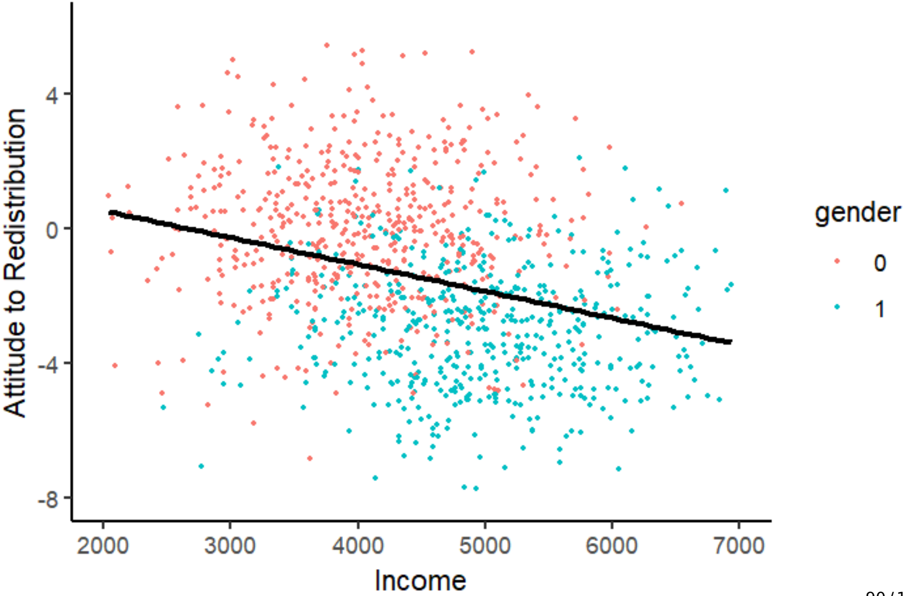
1. Omitted Variable Bias



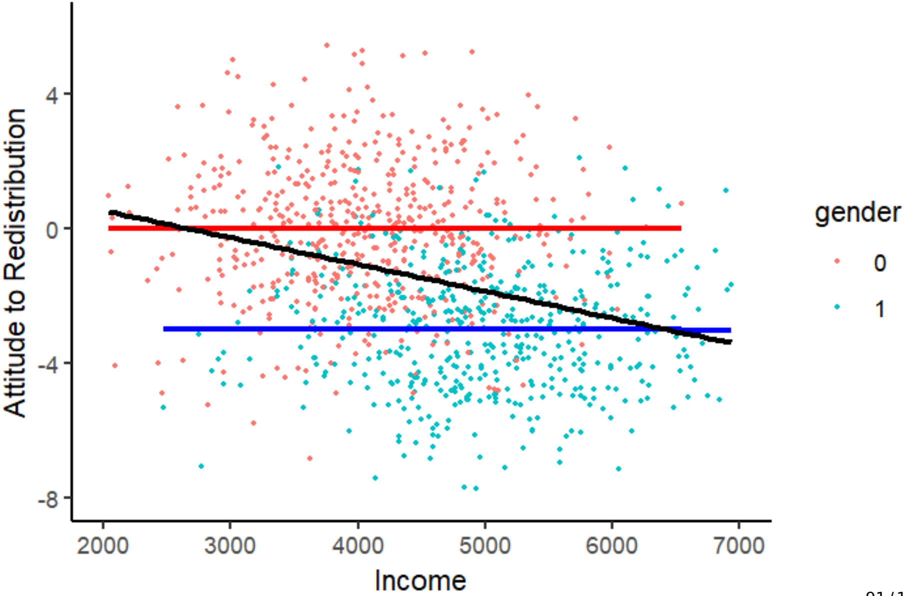
1. Omitted Variable Bias



1. Omitted Variable Bias



1. Omitted Variable Bias



2. Reverse Causation

- Significant regression coefficients just reflect the values in our dataset moving together

2. Reverse Causation

- Significant regression coefficients just reflect the values in our dataset moving together
- Does the 'direction' of regression matter? I.e. Does regression treat D and Y differently?

2. Reverse Causation

- Significant regression coefficients just reflect the values in our dataset moving together
- Does the 'direction' of regression matter? I.e. Does regression treat *D* and *Y* differently? Yes!

	Dependent variable:
	redist
income	−0.011 (0.029)
gender1	−1.201*** (0.058)
Constant	0.589*** (0.038)
Observations	1,000
Note:	*p<0.1; **p<0.05; ***p<0.01

	Dependent variable:
	income
redist	−0.013 (0.034)
gender1	0.993*** (0.069)
Constant	−0.487*** (0.043)
Observations	1,000
Note:	*p<0.1; **p<0.05; ***p<0.01

2. Reverse Causation

- Significant regression coefficients just reflect the values in our dataset moving together
- Does the 'direction' of regression matter? I.e. Does regression treat D and Y differently? Yes!

	Dependent variable:
	redist
income	−0.011 (0.029)
gender1	−1.201*** (0.058)
Constant	0.589*** (0.038)
Observations	1,000
Note:	*p<0.1; **p<0.05; ***p<0.01

	Dependent variable:
	income
redist	−0.013 (0.034)
gender1	0.993*** (0.069)
Constant	−0.487*** (0.043)
Observations	1,000
Note:	*p<0.1; **p<0.05; ***p<0.01

- Remember, regression measures the *vertical* (not diagonal) distances to the regression line
  - It minimizes the *prediction errors* for Y

2. Reverse Causation

- Significant regression coefficients just reflect the values in our dataset moving together
- Does the 'direction' of regression matter? I.e. Does regression treat D and Y differently? Yes!

	Dependent variable:
	redist
income	−0.011 (0.029)
gender1	−1.201*** (0.058)
Constant	0.589*** (0.038)
Observations	1,000
Note:	*p<0.1; **p<0.05; ***p<0.01

	Dependent variable:
	income
redist	−0.013 (0.034)
gender1	0.993*** (0.069)
Constant	−0.487*** (0.043)
Observations	1,000
Note:	*p<0.1; **p<0.05; ***p<0.01

- Remember, regression measures the *vertical* (not diagonal) distances to the regression line
  - It minimizes the *prediction errors* for Y
- But that doesn't mean it identifies the direction of causation!

2. Reverse Causation

- Higher income may lead to higher tax payments and therefore cause more negative attitudes to redistribution

2. Reverse Causation

- Higher income may lead to higher tax payments and therefore cause more negative attitudes to redistribution
- But negative attitudes to redistribution might also make you more likely to work in the private sector and cause you to receive a higher salary

2. Reverse Causation

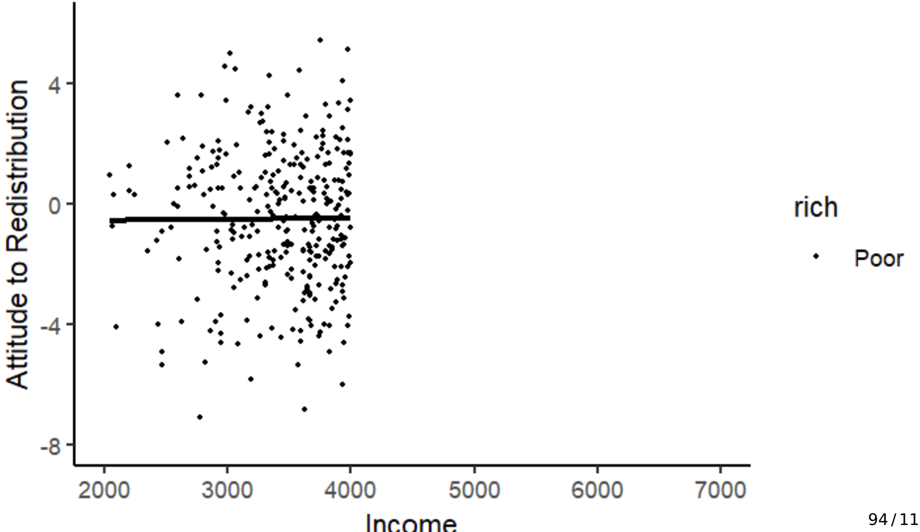
- Higher income may lead to higher tax payments and therefore cause more negative attitudes to redistribution
- But negative attitudes to redistribution might also make you more likely to work in the private sector and cause you to receive a higher salary
- Both would look the same in a regression

2. Reverse Causation

- Higher income may lead to higher tax payments and therefore cause more negative attitudes to redistribution
- But negative attitudes to redistribution might also make you more likely to work in the private sector and cause you to receive a higher salary
- Both would look the same in a regression
- We cannot *explain* the relationship with a regression

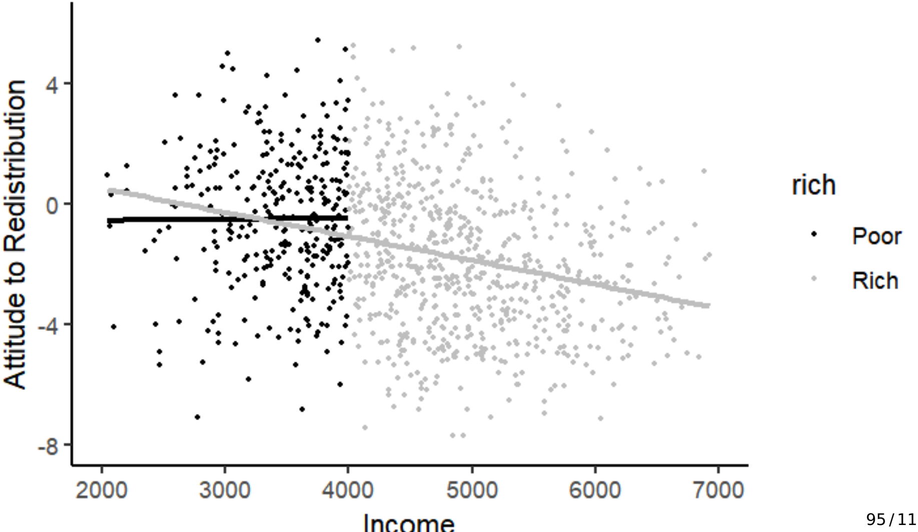
3. Selection Bias

- Imagine we do not see 'rich' units with high income (above R$4000)



3. Selection Bias

- Imagine we do not see 'rich' units with high income (above R$4000)



3. Selection Bias

- There are four selection risks:
  1. **Selection into existence**
  2. **Selection into survival**
  3. **Selection into the dataset**
  4. **Selection into treatment**

3. Selection Bias

- There are four selection risks:
  1. **Selection into existence**
  2. **Selection into survival**
  3. **Selection into the dataset**
  4. **Selection into treatment**
- In each case, we don't see the *full* relationship between D and Y

3. Selection Bias

- There are four selection risks:
    1. **Selection into existence**
    2. **Selection into survival**
    3. **Selection into the dataset**
    4. **Selection into treatment**
- In each case, we don't see the *full* relationship between D and Y
- So our regression estimates are biased

3. Selection Bias

- There are four selection risks:
  1. **Selection into existence:**

3. Selection Bias

- There are four selection risks:
  1. **Selection into existence:**
     - Where do units (eg. political parties) come from?

3. Selection Bias

- There are four selection risks:
  1. **Selection into existence:**
     - Where do units (eg. political parties) come from?
     - Probably only parties that have a chance of success are formed

3. Selection Bias

- There are four selection risks:
  1. **Selection into existence:**
     - Where do units (eg. political parties) come from?
     - Probably only parties that have a chance of success are formed
     - Does forming a party cause electoral success? Not for most people!

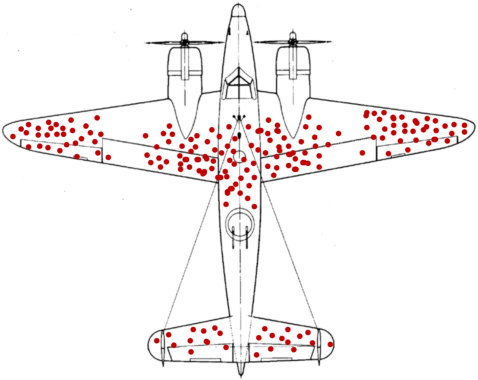
3. Selection Bias

- There are four selection risks:
    2. **Selection into survival:**

3. Selection Bias

- There are four selection risks:
  2. **Selection into survival:**
     - Certain types of units disappear, so the units we see don't tell the full story

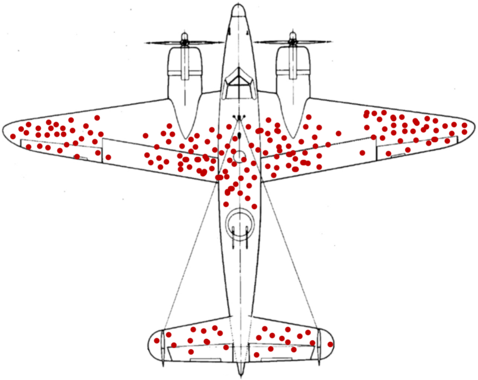
3. Selection Bias
   - There are four selection risks:
     2. **Selection into survival:**
        - Certain types of units disappear, so the units we see don't tell the full story



- Where would additional armour protect bombers?

3. Selection Bias

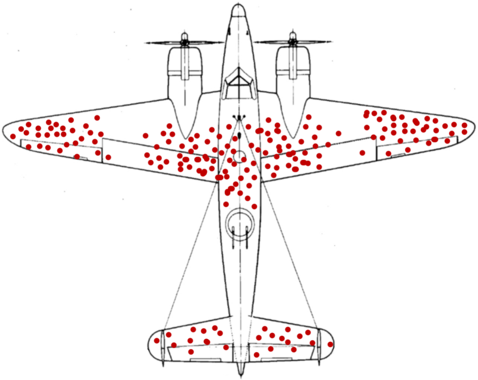
- There are four selection risks:
  2. **Selection into survival:**
     - Certain types of units disappear, so the units we see don't tell the full story



- Where would additional armour protect bombers?
- Returned bombers got hit

3. Selection Bias

- There are four selection risks:
  2. **Selection into survival:**
     - Certain types of units disappear, so the units we see don't tell the full story



- Where would additional armour protect bombers?
- Returned bombers got hit
- But we do not know where *bombers that did not return* got hit

3. Selection Bias

- There are four selection risks:
  3. **Selection into the dataset:**

3. Selection Bias

There are four selection risks:

3. **Selection into the dataset:**
    - Our dataset may not be representative

3. Selection Bias

There are four selection risks:

Selection into the dataset:

- Our dataset may not be representative
- Only units with particular values of D and Y enter the dataset

3. Selection Bias

There are four selection risks:

Selection into the dataset:

Our dataset may not be representative

Only units with particular values of D and Y enter the dataset

Eg. If survey respondents who refuse are different from those who respond

The anti-redistribution poor may dislike answering surveys

3. Selection Bias

There are four selection risks:

Selection into the dataset:

Our dataset may not be representative

Only units with particular values of D and Y enter the dataset

Eg. If survey respondents who refuse are different from those who respond

The anti-redistribution poor may dislike answering surveys

The rich refuse to answer surveys for fear of paying taxes

3. Selection Bias

There are four selection risks:

Selection into treatment:

3. Selection Bias

There are four selection risks:

Selection into treatment:

All units are in our dataset, but they *choose* their treatment value

3. Selection Bias

There are four selection risks:

Selection into treatment:

All units are in our dataset, but they *choose* their treatment value

Who chooses treatment? Those with the most to benefit, i.e. depending on Y!

3. Selection Bias

There are four selection risks:

Selection into treatment:

All units are in our dataset, but they *choose* their treatment value

Who chooses treatment? Those with the most to benefit, i.e. depending on Y!

Applying treatment to the others would probably have a very different effect

4. Measurement Bias

What happens if we measure our variables wrongly?

Very likely!

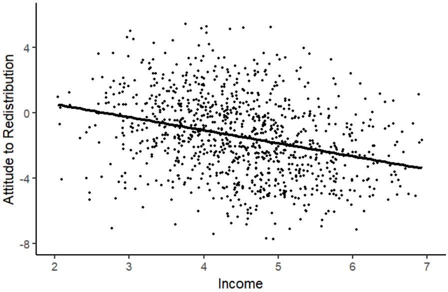
Effects of Measurement Error

	Measured with **Bias**	Measured with **Random Noise**
Outcome Variable	Coefficient biased	No bias but wider standard errors
Treatment Variable	Coefficient biased	Effect biased towards zero

4. Measurement Bias

What happens if we measure our variables wrongly?

Accurate Data:

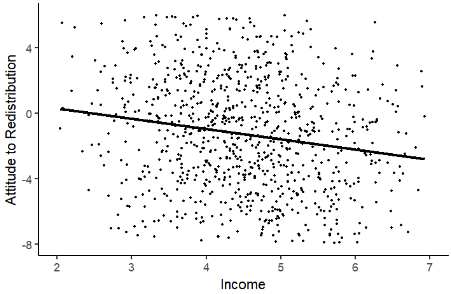


	Dependent variable:
	redist
income	−0.818*** (0.078)
Constant	2.235*** (0.361)
Observations	1,000
Note:	*p<0.1; **p<0.05; ***p<0.01

4. Measurement Bias

What happens if we measure our variables wrongly?

Noise in the **outcome variable**:

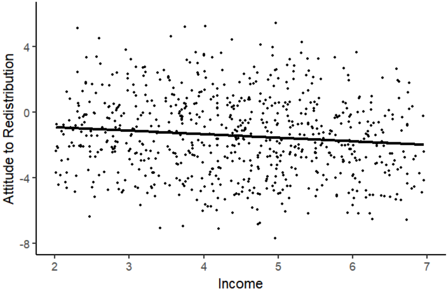


	Dependent variable:
	redist
income	-0.831^{***} (0.144)
Constant	2.272^{***} (0.665)
Observations	1,000
Note:	*p<0.1; **p<0.05; ***p<0.01

4. Measurement Bias

What happens if we measure our variables wrongly?

Noise in the **explanatory** variable:



	Dependent variable:
	redist
income	−0.187*** (0.037)
Constant	−0.620*** (0.183)
Observations	1,000
Note:	*p<0.1; **p<0.05; ***p<0.01

5. Lack of Overlap

Regression normally helps us pick appropriate comparisons

5. Lack of Overlap

Regression normally helps us pick appropriate comparisons

Eg. Controlling for gender, what is the effect of income on attitudes to redistribution?

5. Lack of Overlap

Regression normally helps us pick appropriate comparisons

Eg. Controlling for gender, what is the effect of income on attitudes to redistribution?

But what if there are no women with high income?

5. Lack of Overlap

Regression normally helps us pick appropriate comparisons

Eg. Controlling for gender, what is the effect of income on attitudes to redistribution?

But what if there are no women with high income?

Regression *creates* comparisons for us

How? Using the functional form of the regression

A linear regression interpolates/extrapolates *linearly* to 'create' comparison cases

5. Lack of Overlap

Regression normally helps us pick appropriate comparisons

Eg. Controlling for gender, what is the effect of income on attitudes to redistribution?

But what if there are no women with high income?

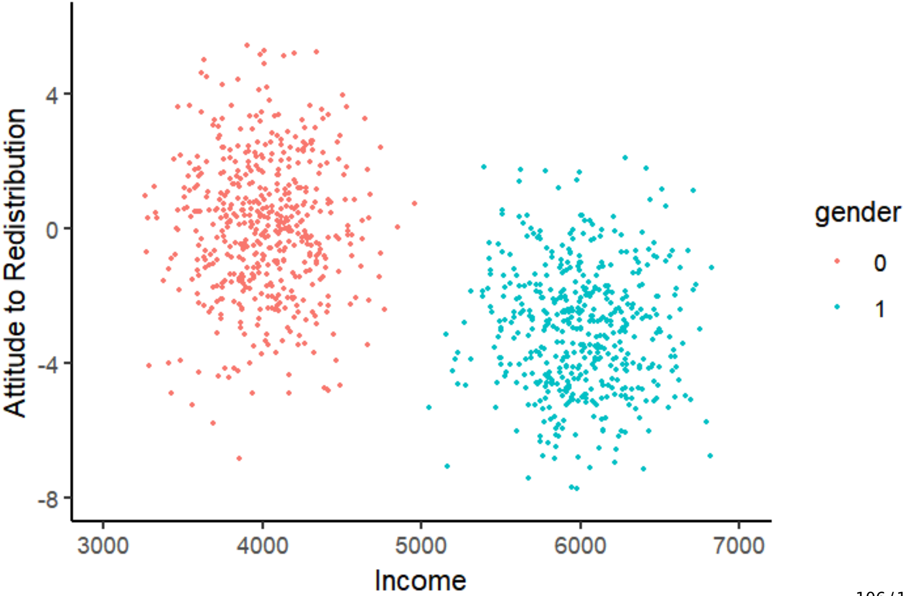
Regression *creates* comparisons for us

How? Using the functional form of the regression

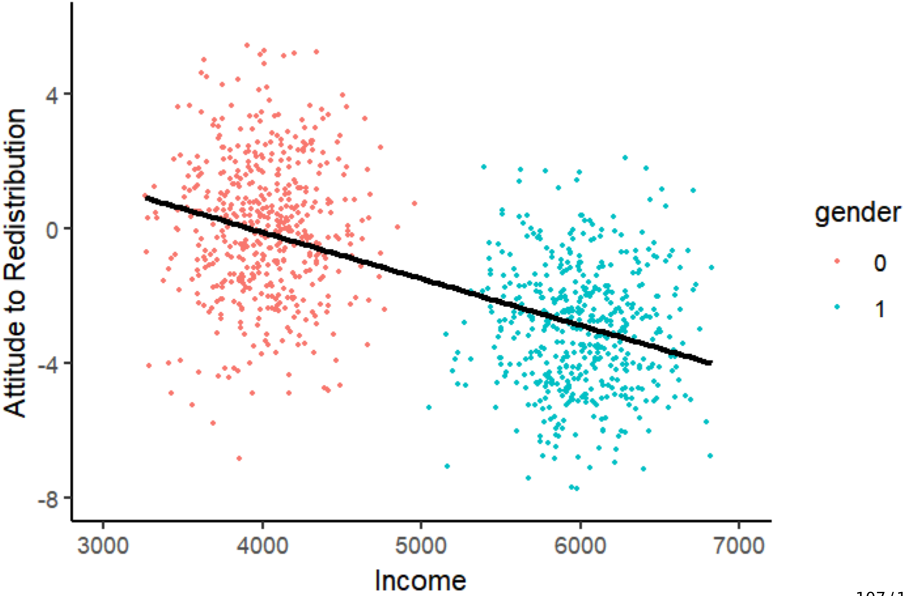
A linear regression interpolates/extrapolates *linearly* to 'create' comparison cases

Lack of overlap probably means we *cannot* explain outcomes with this data

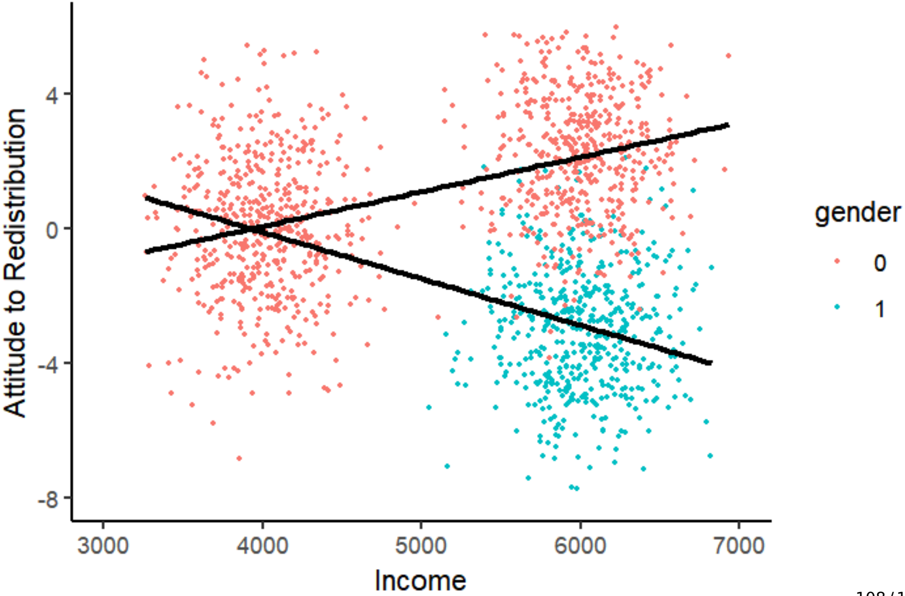
5. Lack of Overlap



5. Lack of Overlap



5. Lack of Overlap



5. Lack of Overlap

With more than a few variables, lack of overlap is *guaranteed*

5. Lack of Overlap

With more than a few variables, lack of overlap is *guaranteed*

6 variables with 10 categories each = $10^6 = 1,000,000$ possibilities

5. Lack of Overlap

With more than a few variables, lack of overlap is *guaranteed*

6 variables with 10 categories each = $10^6 = 1,000,000$ possibilities

Common datasets have 0% counterfactuals present in the data (King 2006)

5. Lack of Overlap

With more than a few variables, lack of overlap is *guaranteed*

6 variables with 10 categories each = $10^6 = 1,000,000$ possibilities

Common datasets have 0% counterfactuals present in the data (King 2006)

How many 45 year-old female accountants with a PhD and a cat who live in Centro are there?

5. Lack of Overlap

With more than a few variables, lack of overlap is *guaranteed*

6 variables with 10 categories each = $10^6 = 1,000,000$ possibilities

Common datasets have 0% counterfactuals present in the data (King 2006)

How many 45 year-old female accountants with a PhD and a cat who live in Centro are there?

And we need some that are low-income and some that are high-income

5. Lack of Overlap

With more than a few variables, lack of overlap is *guaranteed*

6 variables with 10 categories each = $10^6 = 1,000,000$ possibilities

Common datasets have 0% counterfactuals present in the data (King 2006)

How many 45 year-old female accountants with a PhD and a cat who live in Centro are there?

And we need some that are low-income and some that are high-income

A problem of **multi-dimensionality**

5. Lack of Overlap

With more than a few variables, lack of overlap is *guaranteed*

6 variables with 10 categories each = $10^6 = 1,000,000$ possibilities

Common datasets have 0% counterfactuals present in the data (King 2006)

How many 45 year-old female accountants with a PhD and a cat who live in Centro are there?

And we need some that are low-income and some that are high-income

A problem of **multi-dimensionality**

And of **model dependence** - our results depend on the functional form (linear, quadratic etc.) in our regression model

Summary

Regression is just fancy correlation

A **conditional expectation function**

Summary

Regression is just fancy correlation

A **conditional expectation function**

'Smart' regression pays more attention to what **comparisons** you want to make than to statistical tests

And to **interpretation/prediction** rather than p-values

Summary

Regression is just fancy correlation

A **conditional expectation function**

'Smart' regression pays more attention to what **comparisons** you want to make than to statistical tests

And to **interpretation/prediction** rather than p-values

Regression **cannot explain** relationships

Correlation is not causation

Summary

Regression is just fancy correlation

A **conditional expectation function**

'Smart' regression pays more attention to what **comparisons** you want to make than to statistical tests

And to **interpretation/prediction** rather than p-values

Regression **cannot explain** relationships

Correlation is not causation

We need to understand better **how the data were produced**

Summary

Regression is just fancy correlation

A **conditional expectation function**

'Smart' regression pays more attention to what **comparisons** you want to make than to statistical tests

And to **interpretation/prediction** rather than p-values

Regression **cannot explain** relationships

Correlation is not causation

We need to understand better **how the data were produced**

Explanation depends on research design, data selection, assumptions and qualitative evidence