FLS 6441 - Methods III: Explanation and Causation

Week 1 - Review of Regression

Jonathan Phillips

February 2019

Course Objectives

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

Course Website

Course Topics

1. Review of Regression
2. A Framework for Explanation
3. Field Experiments
4. Survey and Lab Experiments
5. Randomized Natural Experiments
6. Instrumental Variables
7. Discontinuities
8. Difference-in-Differences
9. Controlling for Confounding
10. Matching
11. Comparative Cases and Process Tracing
12. Generalizability, Reproducibility and Mechanisms

Course Schedule

- Wednesday 18h - Submit Replication Task
- Thursday 14h-16h - Class
- Thursday 16.15-17.30 - Lab
- Friday 10h-12h - Office Hours (DCP 2061)

Project

- Quality > Quantity
- Max 15 pages, English or Portuguese
- Submit paper and code by email to me by 30th June 2019
- Use at least one of the methods studied in class
- Pick a simple question and dataset

Today’s Objectives

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

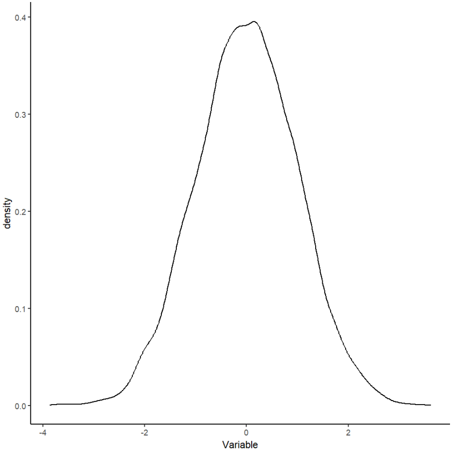
Section 1

What Does Regression Actually Do?

Data

1. We work with variables, which VARY!

	Variable
1	0.39
2	1.69
3	-1.05
4	-1.38
5	0.81
6	2.01
7	0.06
8	0.98
9	-0.98
10	-0.39



What Does Regression Actually Do?

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

Regression as Least Squares

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

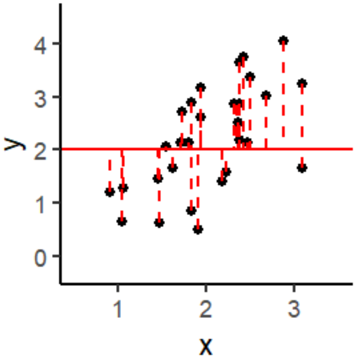
Regression as Least Squares

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

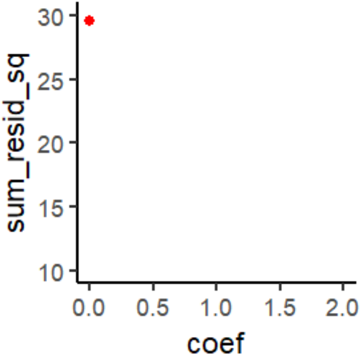
Regression as Least Squares

- 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



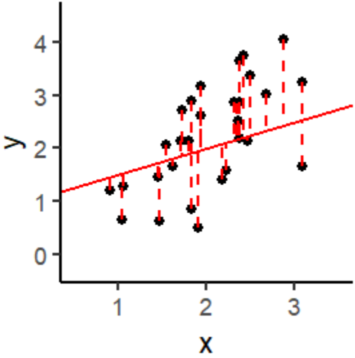
Sum of Squared Residuals = 29.6



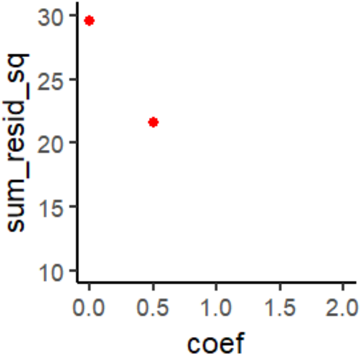
Regression as Least Squares

- 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



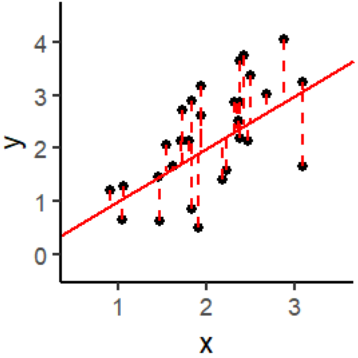
Sum of Squared Residuals = 21.6



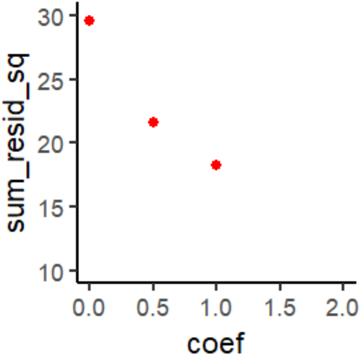
Regression as Least Squares

- 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



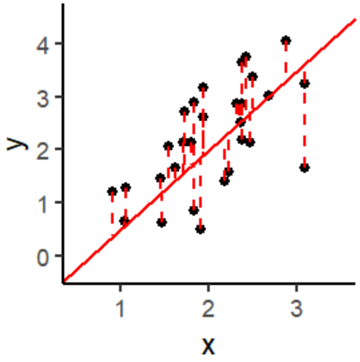
Sum of Squared Residuals = 18.3



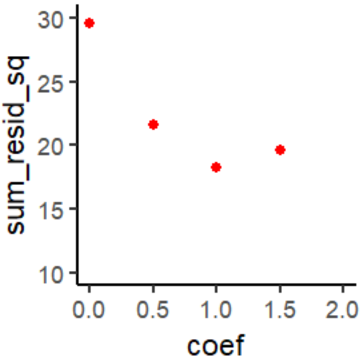
Regression as Least Squares

- 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



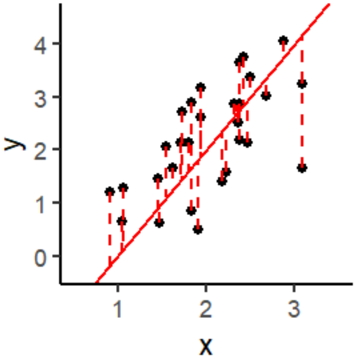
Sum of Squared Residuals = 19.6



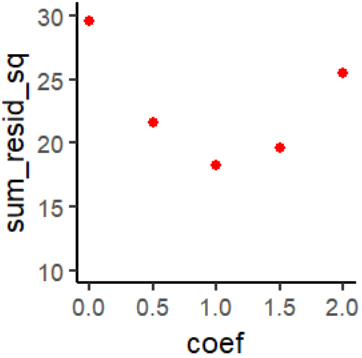
Regression as Least Squares

- 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



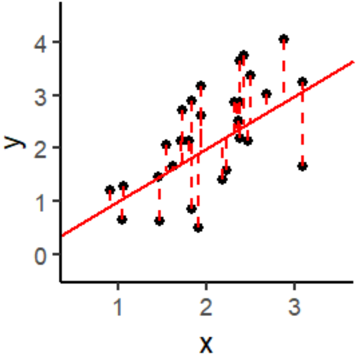
Sum of Squared Residuals = 25.5



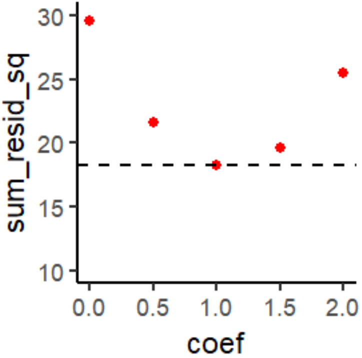
Regression as Least Squares

- 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



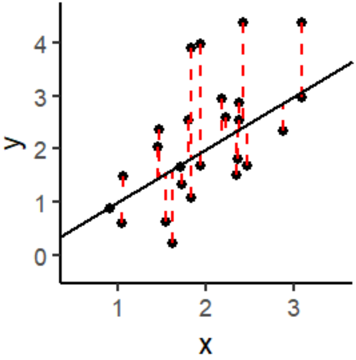
Sum of Squared Residuals = 18.3



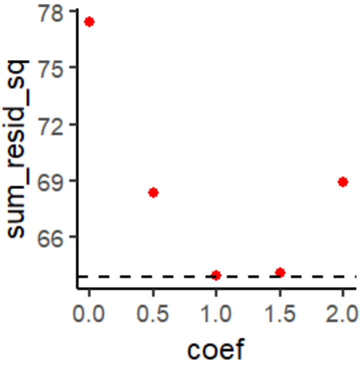
Regression as Least Squares

- 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

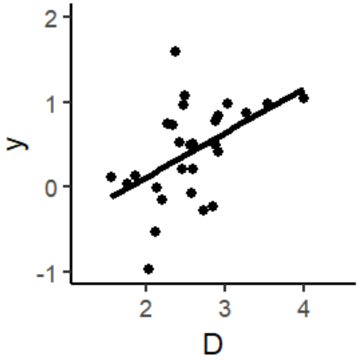


Sum of Squared Residuals = 63.9



Regression as Least Squares

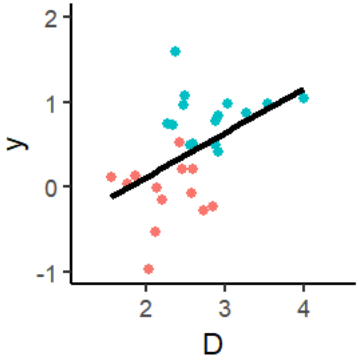
- Dummy control variables *remove variation* associated with specific levels or categories
  - The same for fixed effects
- $y_{ij} = \alpha + \beta_1 D_{ij} + \epsilon_i$



Ignoring the dummy control variable, the slope coefficient is 1

Regression as Least Squares

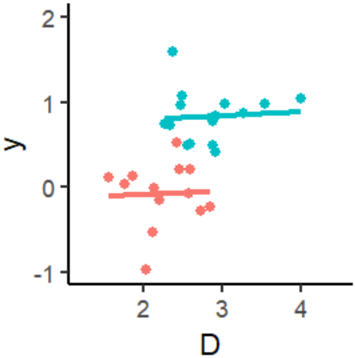
- Dummy control variables *remove variation* associated with specific levels or categories
  - The same for fixed effects
- $y_{ij} = \alpha + \beta_1 D_{ij} + \epsilon_i$



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

Regression as Least Squares

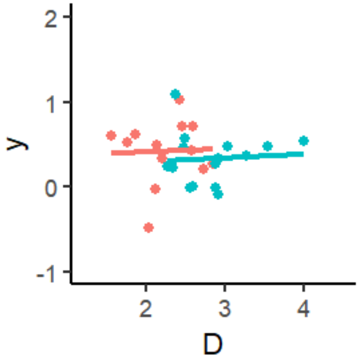
- Dummy control variables *remove variation* associated with specific levels or categories
  - The same for fixed effects
- $y_{ij} = \alpha + \beta_1 D_{ij} + \beta_2 X_j + \epsilon_i$



What if we treated each group *separately*?

Regression as Least Squares

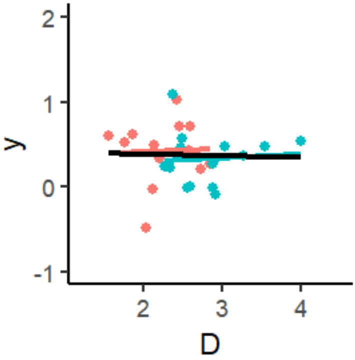
- Dummy control variables *remove variation* associated with specific levels or categories
  - The same for fixed effects
- $y_{ij} = \alpha + \beta_1 D_{ij} + \beta_2 X_j + \epsilon_i$



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

Regression as Least Squares

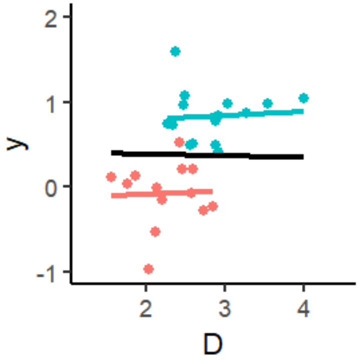
- Dummy control variables *remove variation* associated with specific levels or categories
  - The same for fixed effects
- $y_{ij} = \alpha + \beta_1 D_{ij} + \beta_2 X_j + \epsilon_i$



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

Regression as Least Squares

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

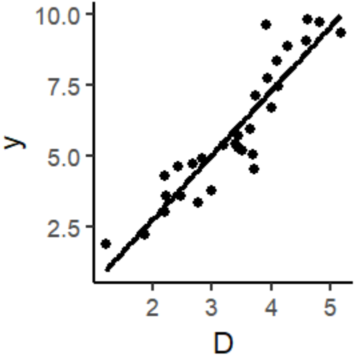


Equivalently, dummy control variables restrict comparisons to **within the same group**:

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

Regression as Least Squares

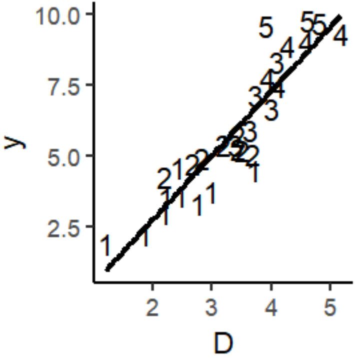
- Continuous control variables *remove variation* based on how much the control explains y
- $y_i = \alpha + \beta_1 D_i + \epsilon_i$



The coefficient β_1 is 2.267
Real effect = 1

Regression as Least Squares

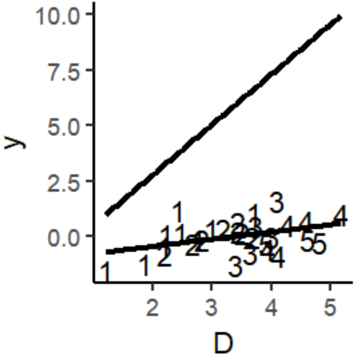
- Continuous control variables *remove variation* based on how much the control explains y
- $y_i = \alpha + \beta_1 D_i + \epsilon_i$



The coefficient β_1 is 2.267
Real effect = 1

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

Regression as Conditional Expectation

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

Regression as Conditional Expectation

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

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

Regression as Conditional Expectation

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

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

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

Regression as Conditional Expectation

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

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

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

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

Regression as Conditional Expectation

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

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

- $E(y|x)$,

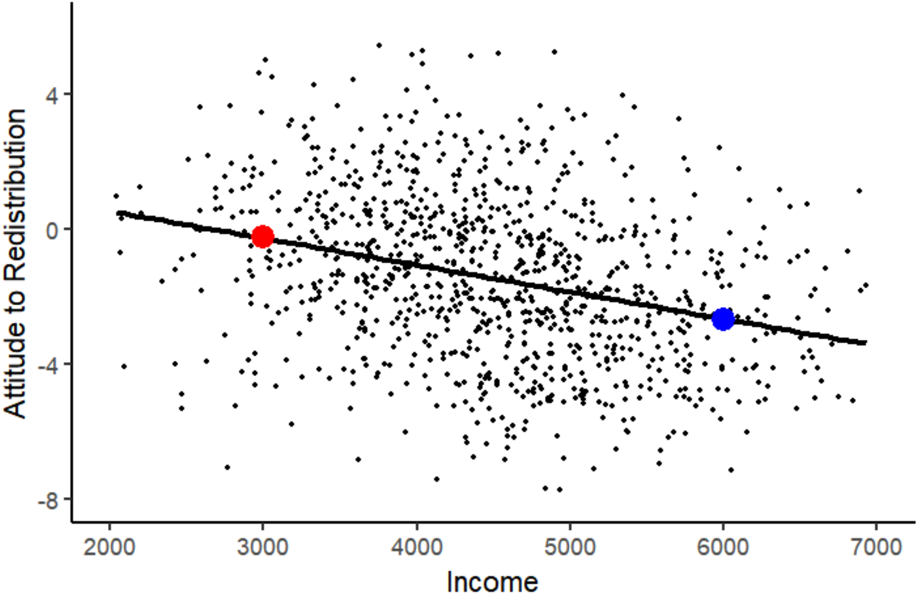
Regression as Conditional Expectation

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

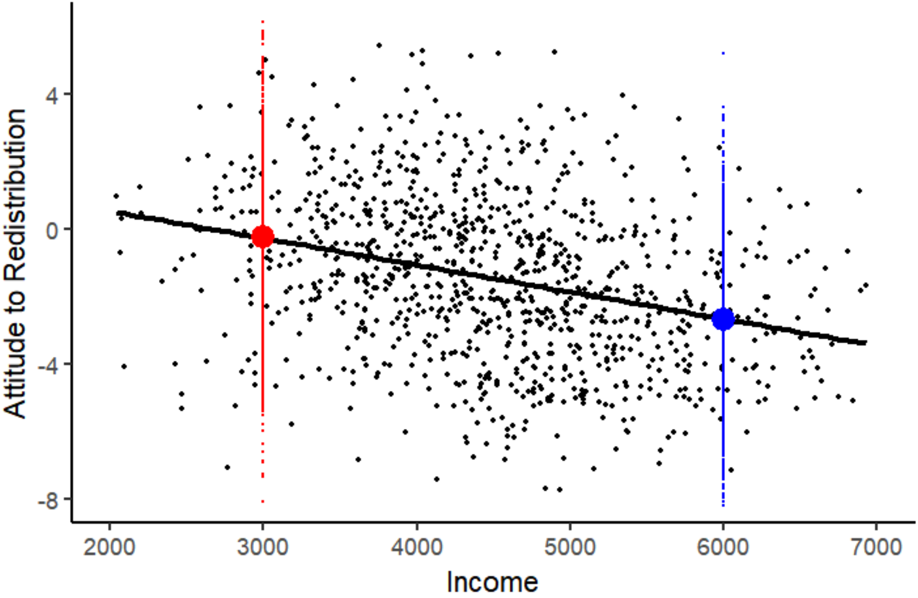
$$y_i = \alpha + \beta_1 D_i + \epsilon_i$$
$$Attitude_i = \alpha + \beta_1 Income_i + N(0, \sigma^2)$$
$$Attitude_i = 2.235 - 0.000818 * Income_i + N(0, 2.38)$$

- $E(y|x)$, $E(Attitude|Income)$
  - When income is 3000, the average attitude is -0.22
  - When income is 6000, the average attitude is -2.67
- $E(attitude|income, age, gender, municipality)$

Regression as Conditional Expectation



Regression as Conditional Expectation



Regression as Conditional Expectation

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

Regression as Conditional Expectation

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

Regression as Conditional Expectation

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

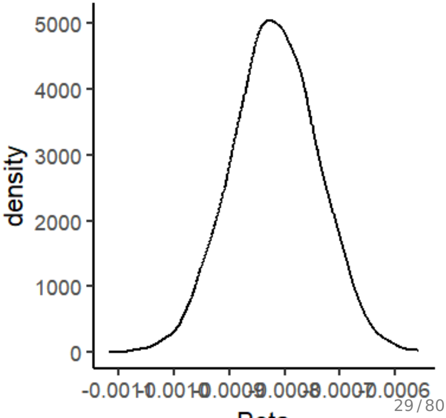
Regression as Conditional Expectation

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

Regression as Conditional Expectation

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

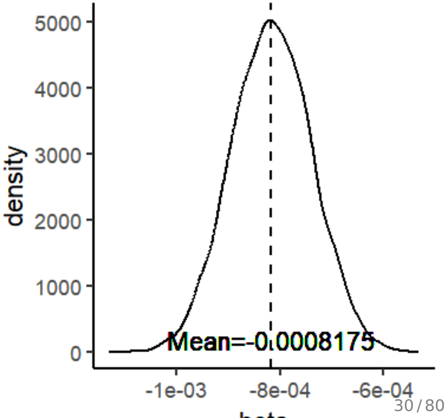
	Dependent variable:
	redist
income	-0.000818^{***} (0.000078)
Constant	2.234719^{***} (0.361477)
Observations	1,000
Note:	*p<0.1; **p<0.05; ***p<0.01



Regression as Conditional Expectation

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

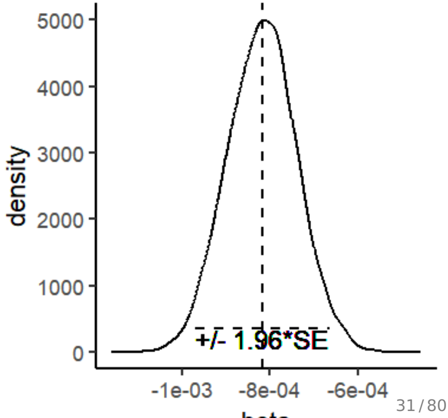
	Dependent variable:
	redist
income	-0.000818^{***} (0.000078)
Constant	2.234719^{***} (0.361477)
Observations	1,000
Note:	*p<0.1; **p<0.05; ***p<0.01



Regression as Conditional Expectation

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

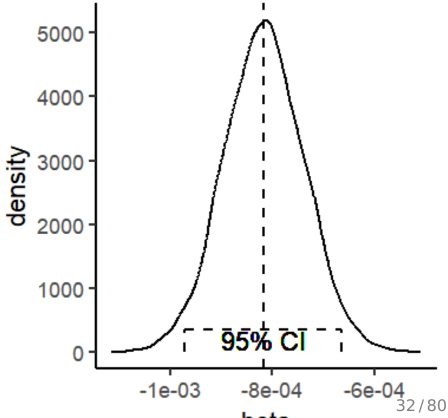
	Dependent variable:
	redist
income	-0.000818^{***} (0.000078)
Constant	2.234719^{***} (0.361477)
Observations	1,000
Note:	*p<0.1; **p<0.05; ***p<0.01



Regression as Conditional Expectation

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



Regression as (Partial) Correlation

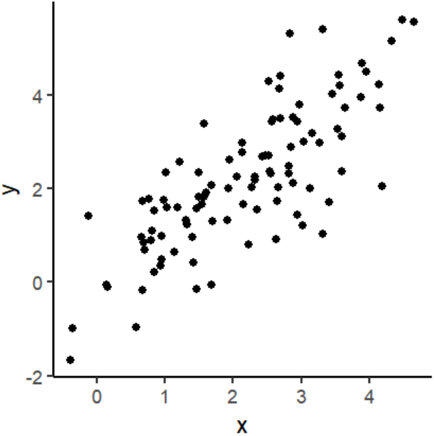
- Regression with two variables is very similar to calculating correlation

Regression as (Partial) Correlation

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

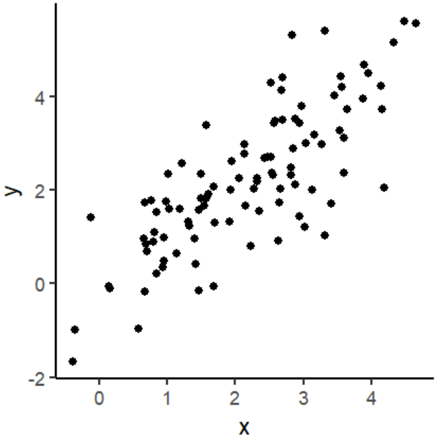
Regression as (Partial) Correlation

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



Regression as (Partial) Correlation

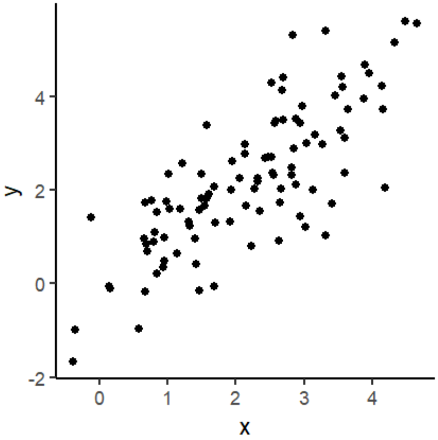
- Regression with two variables is very similar to calculating correlation:
- $\hat{\beta} = cor(x, y) * \frac{\sigma_Y}{\sigma_X}$



- Correlation is 0.781

Regression as (Partial) Correlation

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

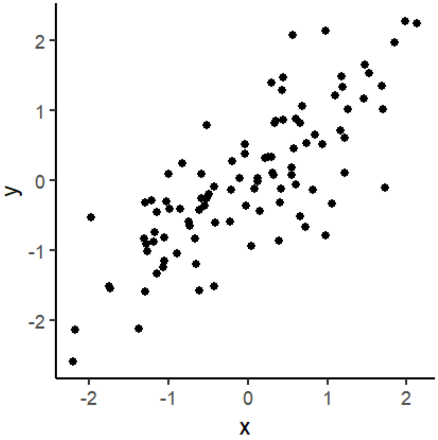


- Correlation is 0.781
- Regression Results:

	term	estimate
1	(Intercept)	0.006
2	x	1.008

Regression as (Partial) Correlation

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



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

	term	estimate
1	(Intercept)	0.000
2	x	0.781

Regression as (Partial) Correlation

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

Regression as (Partial) Correlation

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

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)

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)

$$\beta_{x_1} = \frac{r_{yx_1} - r_{yx_2} r_{x_1x_2}}{1 - r^2_{x_1x_2}}$$

$$r_{yx_1|x_2} = \frac{r_{yx_1} - r_{yx_2} r_{x_1x_2}}{\sqrt{(1 - r^2_{yx_2})(1 - r^2_{x_1x_2})}}$$

- **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 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 a Model/Link Function:** To match the data type of your outcome variable
3. **Choose Covariates:** To match your strategy of inference

Regression Guide

1. **Choose Variables and Measures:** To test a specific hypothesis
2. **Choose a Model/Link Function:** To match the data type of your outcome variable
3. **Choose Covariates:** To match your strategy of inference
4. **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 a Model/Link Function:** To match the data type of your outcome variable
3. **Choose Covariates:** To match your strategy of inference
4. **Choose Fixed Effects:** To focus on comparisons at a specific level
5. **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 a Model/Link Function:** To match the data type of your outcome variable
3. **Choose Covariates:** To match your strategy of inference
4. **Choose Fixed Effects:** To focus on comparisons at a specific level
5. **Choose Error Structure:** To match known dependencies/clustering in the data or sampling
6. **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 a Model/Link Function:** To match the data type of your outcome variable
3. **Choose Covariates:** To match your strategy of inference
4. **Choose Fixed Effects:** To focus on comparisons at a specific level
5. **Choose Error Structure:** To match known dependencies/clustering in the data or sampling
6. **Interpret the Coefficients:** To match the type/scale of the explanatory variable, outcome variable and model
7. **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. 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?

3. Covariates

- Which covariates should we include?

3. Covariates

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

3. 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

3. 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
- Most crucial where there is theory or evidence that this variable could be an **omitted variable**

4. Fixed Effects

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

4. 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

4. 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

4. 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
- Our question becomes: How do variations within income in the same country affect attitudes to redistribution?

5. Errors Structure

- After all of our controls and fixed effects, we need to think about the errors - the bit we cannot explain

5. Errors Structure

- After all of our controls and fixed effects, we need to think about the errors - the bit we cannot explain
- 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 your attitudes to redistribution are probably very similar to everyone you live with, even after controlling for income etc.

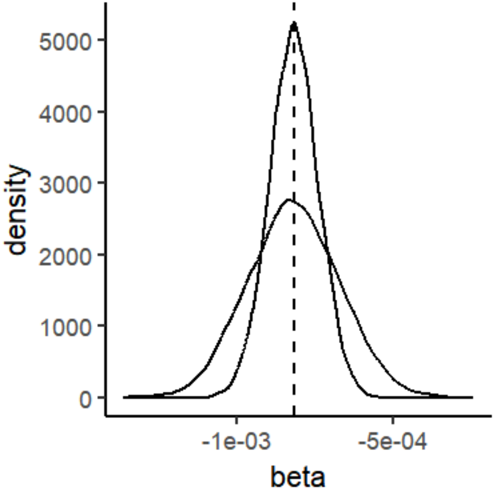
5. Errors Structure

- After all of our controls and fixed effects, we need to think about the errors - the bit we cannot explain
- 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 your attitudes to redistribution are probably very similar to everyone you live with, even after controlling for income etc.
- 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-optimistic* (too small)

5. Errors Structure

- After all of our controls and fixed effects, we need to think about the errors - the bit we cannot explain
- 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 your attitudes to redistribution are probably very similar to everyone you live with, even after controlling for income etc.
- 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-optimistic* (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

5. Errors Structure



6. Interpreting Regression Results

- Difficult! It depends on the scale of the explanatory variable, scale of the outcome, the regression model we used, and the presence of any interaction
- Basic OLS:
  - 1 [unit of explanatory variable] change in the explanatory variable is associated with a β [unit of outcome variable] change in the outcome

6. Interpreting Regression Results

- Difficult! It depends on the scale of the explanatory variable, scale of the outcome, the regression model we used, and the presence of any interaction
- Logit:
  - 1 [unit of explanatory variable] change in the explanatory variable is associated with a β [unit of outcome variable] change in the outcome

7. Predictions from Regressions

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

7. Predictions from Regressions

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

7. 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

7. 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
- Better to make specific *predictions* of how changes in X produce changes in Y

7. Predictions from Regressions

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

7. Predictions from Regressions

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

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

7. 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:

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

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

7. 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:

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

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

7. Predictions from Regressions

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

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

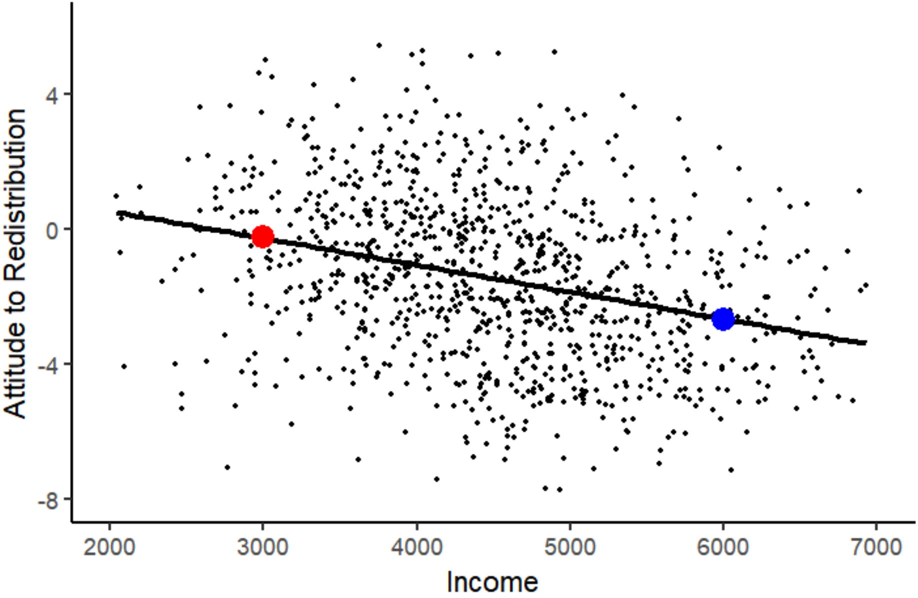
Increasing Income from 3000 to 6000:

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

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

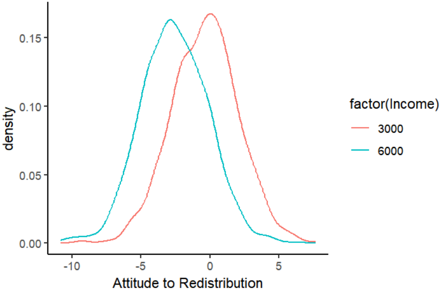
$$\Delta Attitude_i = -2.454$$

7. Predictions from Regressions

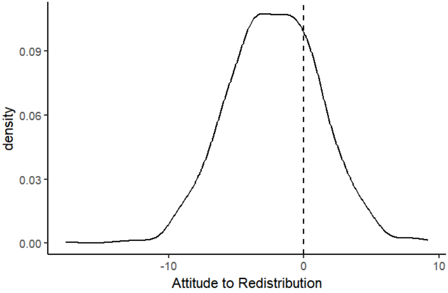


7. Predictions from Regressions

Predicted Values:



First Differences:



7. Predictions from Regressions

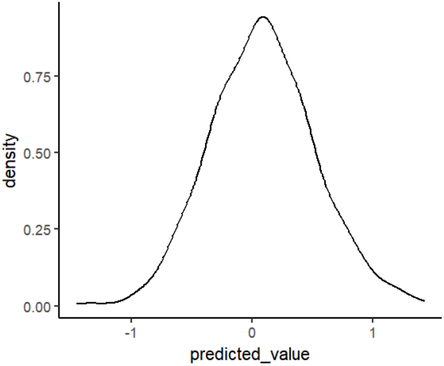
- The regression model matters because the wrong model makes non-sensical predictions
- Consider a binary outcome: $Gender_i = \alpha + \beta Income_i + \epsilon_i$
- Compare the OLS and Logit regression tables:

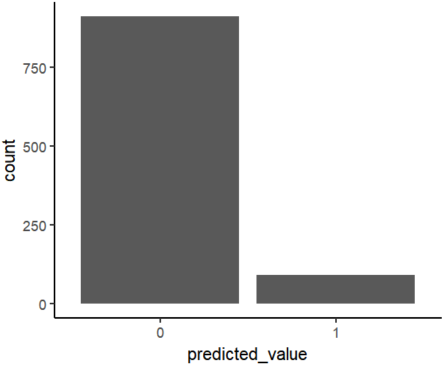
	Dependent variable:
	as.numeric(as.character(gender))
income	0.0003*** (0.00001)
Constant	−0.696*** (0.066)
Observations	1,000
Note:	*p<0.1; **p<0.05; ***p<0.01

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

7. 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
- Even after following all this guidance, Regression does NOT:
  1. *Explain* anything
  2. Make bad data better
  3. Tell you which model is ’best’
  4. Guarantee you are making sensible comparisons
- 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 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 patterns...
  - But we cannot conclude that D causes or explains Y

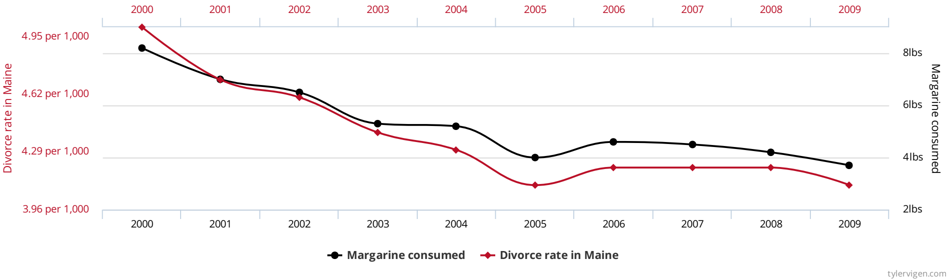
What Does Regression NOT Do?

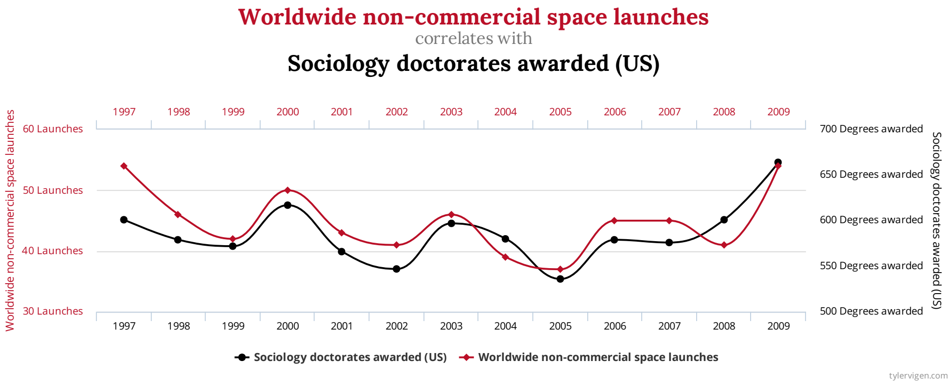
- **Correlation is not causation**
  - If we look hard enough we can always find correlations
  - By chance...
  - Due to complex social patterns...
  - But we cannot conclude that D causes or explains Y
- *More* data will not help

What Does Regression NOT Do?

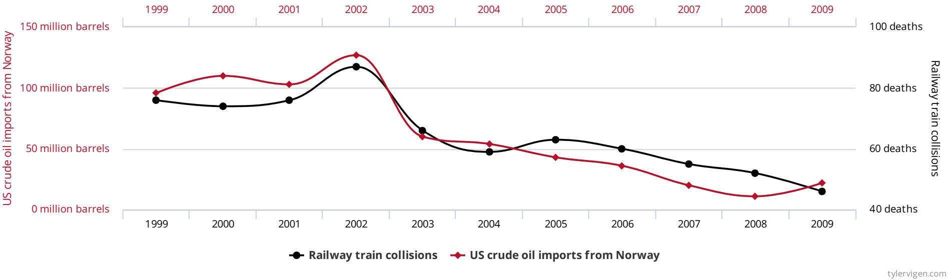
- **Correlation is not causation**
  - If we look hard enough we can always find correlations
  - By chance...
  - Due to complex social patterns...
  - But we cannot conclude that D causes or explains Y
- *More* data will not help
- The problem is the *type* of data; it does not allow us to answer the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

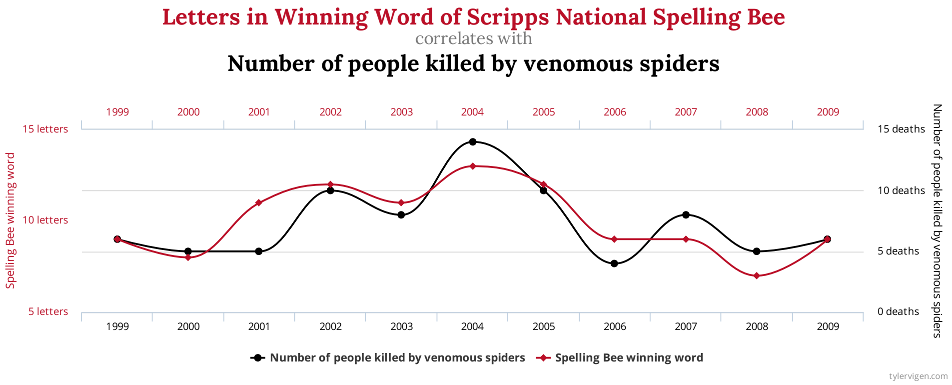
Divorce rate in Maine
correlates with
Per capita consumption of margarine




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





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

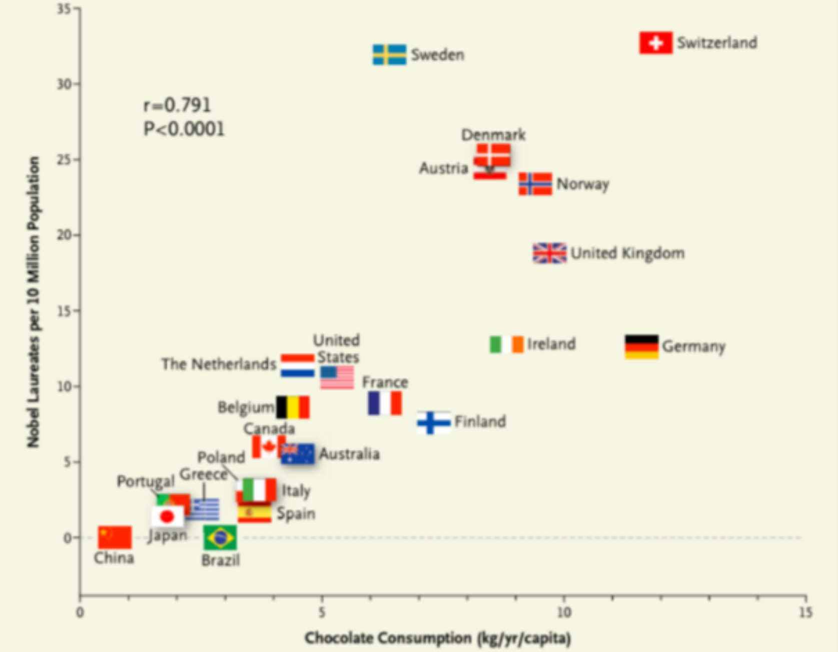
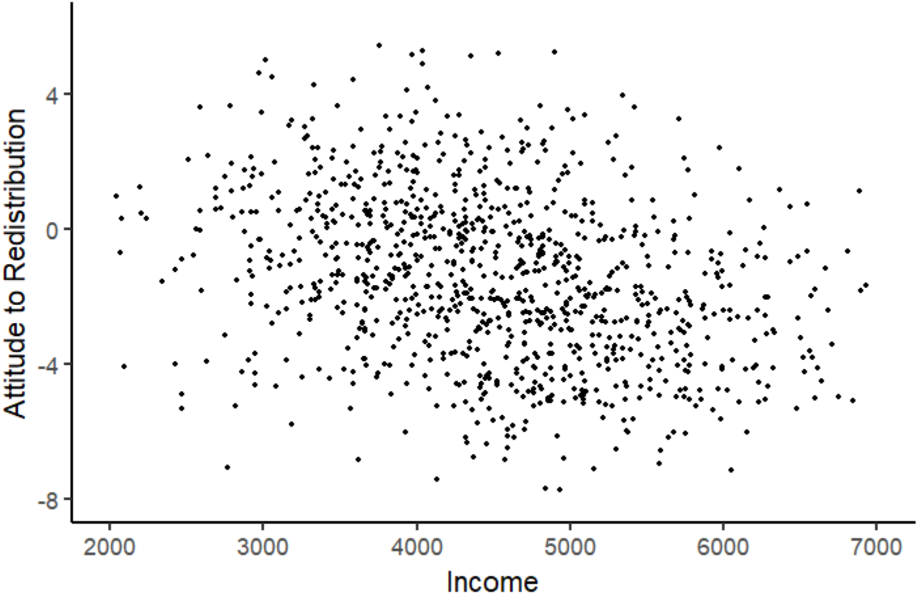


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

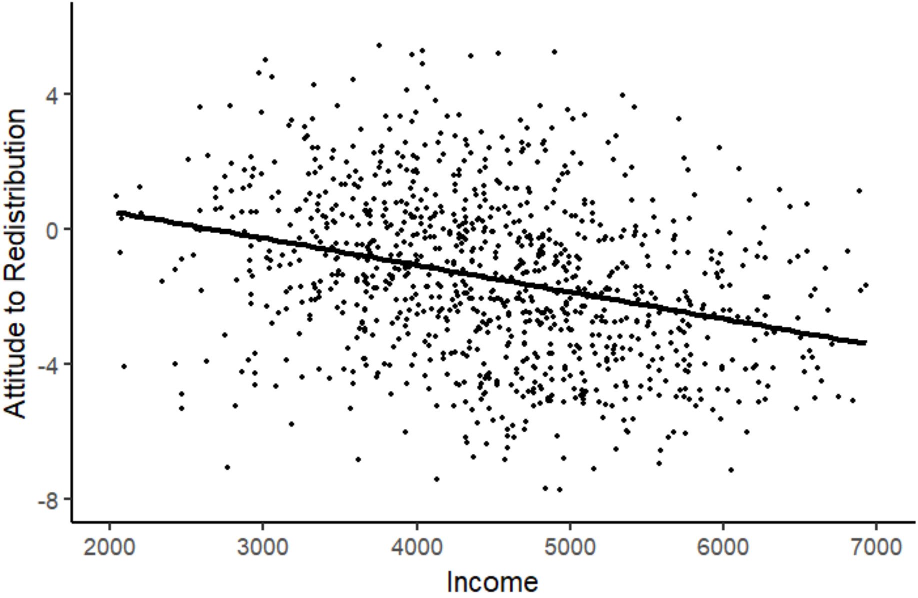
What Does Regression NOT Do?

- Lots of things can go ’wrong’ with regression:
    1. Omitted Variable Bias
    2. Reverse Causation
    3. Selection Bias
    4. Measurement Bias
    5. Lack of Overlap, Model Dependence

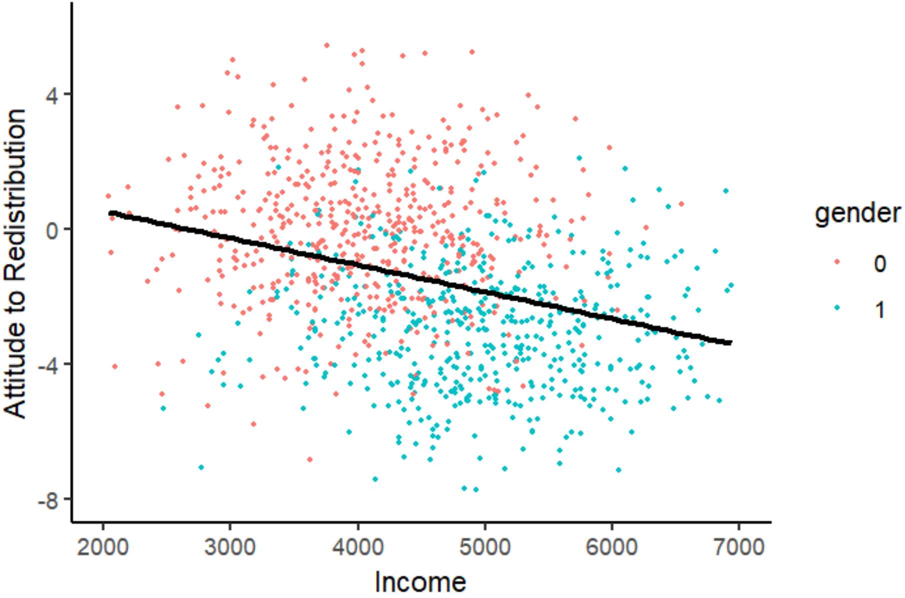
1. Omitted Variable Bias



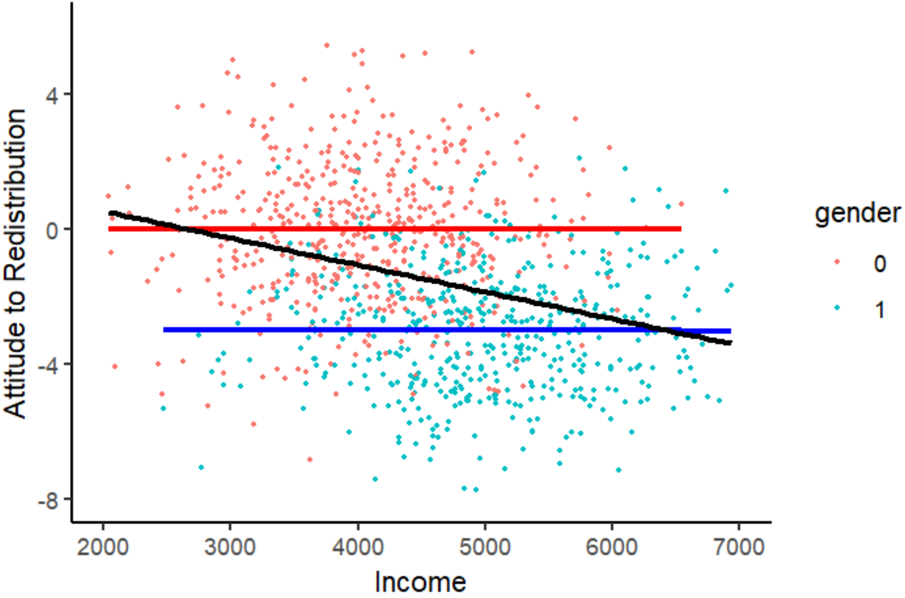
1. Omitted Variable Bias



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 X 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 X 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

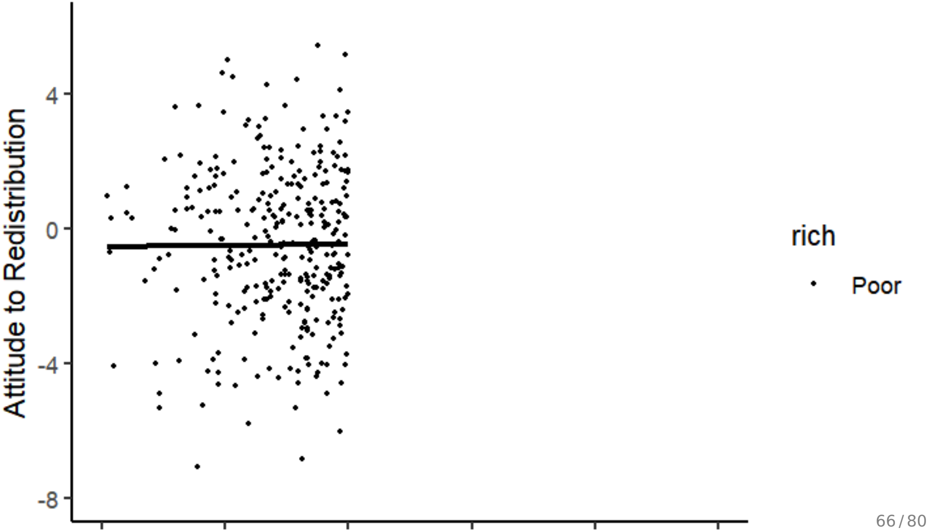
- Remember, regression measures the *vertical* (not diagonal) distances to the regression line
  - It minimizes the prediction errors for Y
- But that doesn't mean it identifies the direction of causation!

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 X and Y
- So our regression estimates are biased

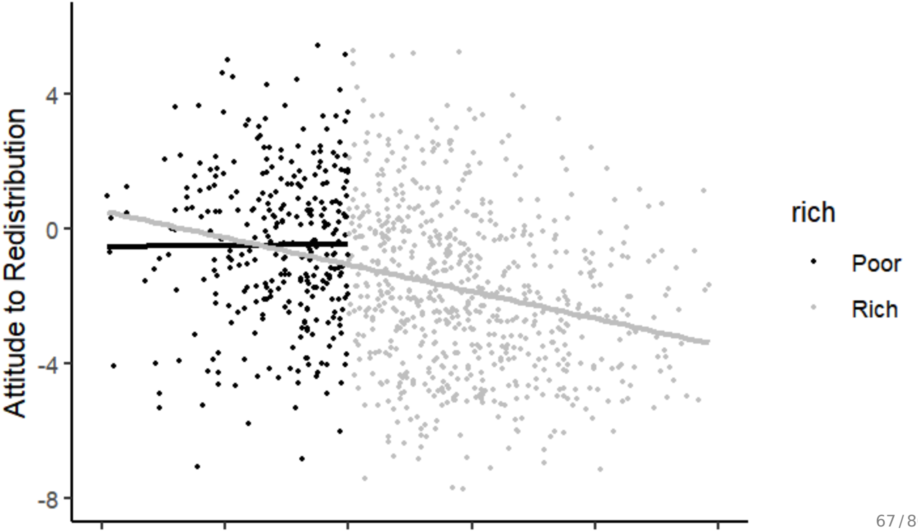
3. Selection Bias

- 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)

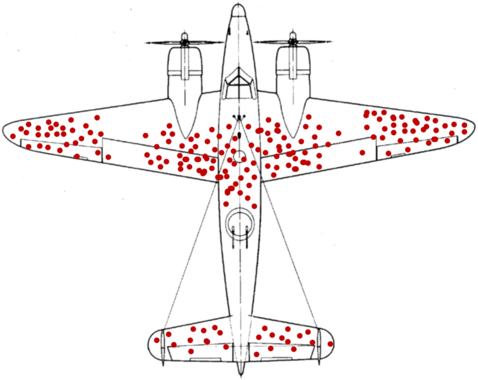


3. Selection Bias

- There are four selection risks:
  1. **Selection into existence:**
     - 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:**
     - 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:**
     - Our dataset may not be representative
     - Only units with particular values of X and Y enter the dataset
     - Eg. If survey respondents who refuse are different from those who respond

3. Selection Bias

- There are four selection risks:
  4. **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

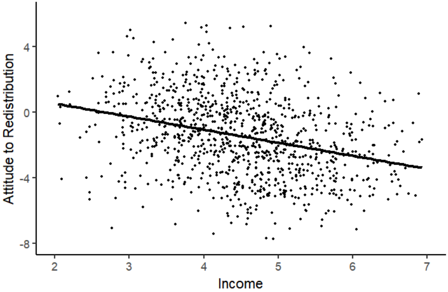
- What happens if we measure our variables wrongly?

Effects of Measurement Error

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

4. Measurement Bias

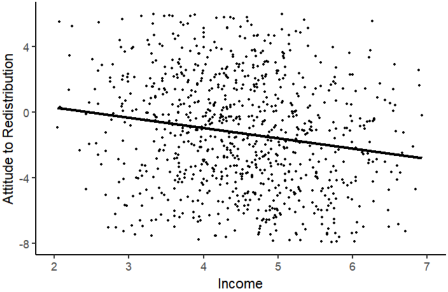
- What happens if we measure our variables wrongly?
- No extra noise:



	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

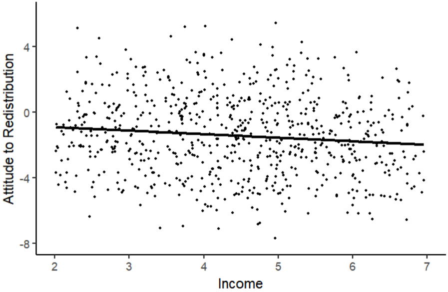
- What happens if we measure our variables wrongly?
- Noise in the **outcome variable**:



	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:

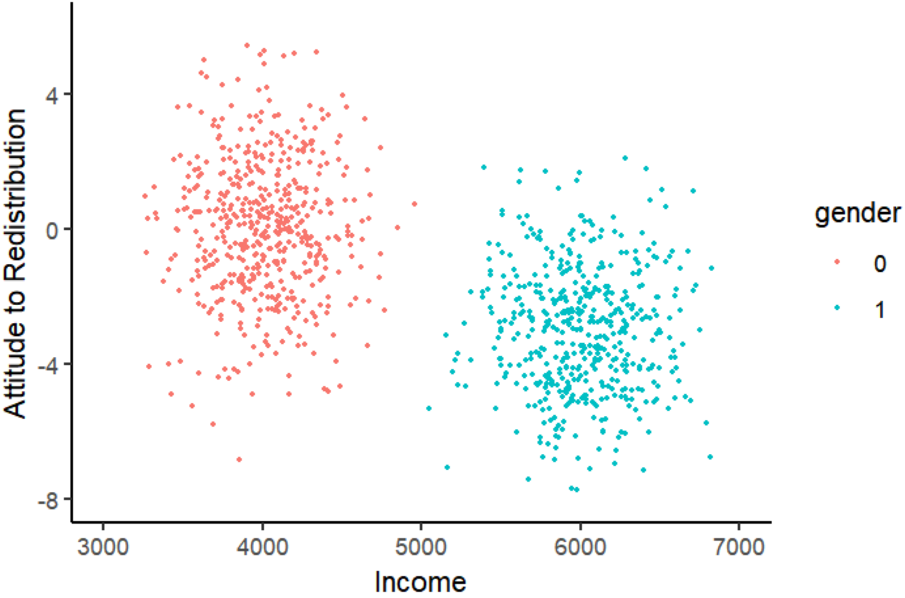


	Dependent variable:
	redist
income	-0.187^{***} (0.037)
Constant	-0.620^{***} (0.183)
Observations	1,000
Note:	*p<0.1; **p<0.05; ***p<0.01

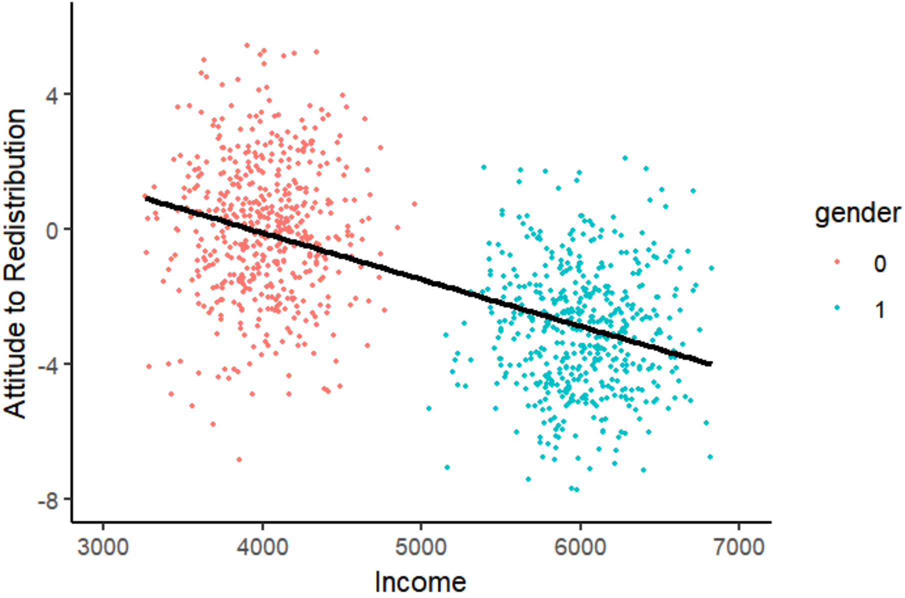
5. Lack of Overlap

- Regression normally helps us pick appropriate comparisons
  - Eg. Comparing only among men, 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? That's where the functional for of the regression comes in
  - A linear regression interpolates/extrapolates *linearly* to 'create' comparison cases
- Lack of overlap probably means we *cannot* explain outcomes with this data

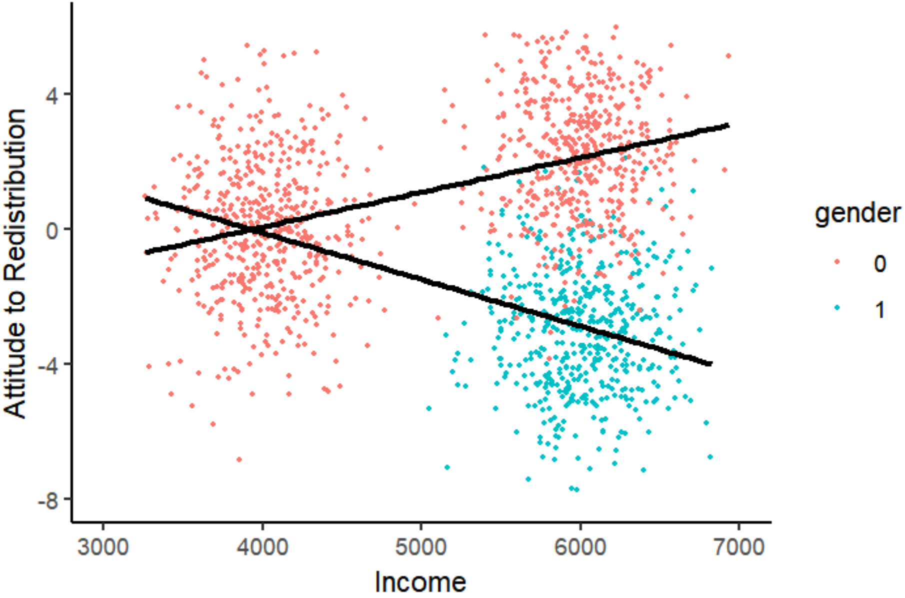
5. Lack of Overlap



5. Lack of Overlap



5. Lack of Overlap



5. Lack of Overlap

- With more than a few variables, lack of overlap is *guaranteed*
- 6 variables with 10 categories each = $10^6 = 1,000,000$ possibilities, and a sample of maybe 5,000?
- 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 in our regression model