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?

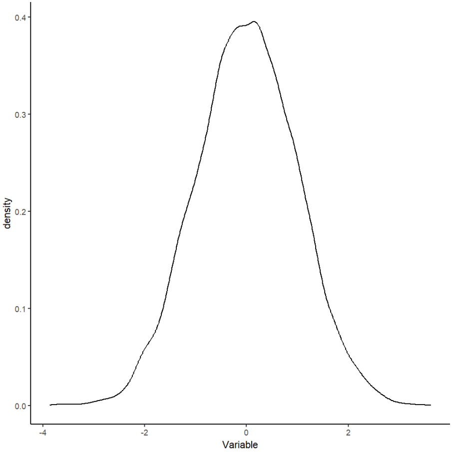
Section 1

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

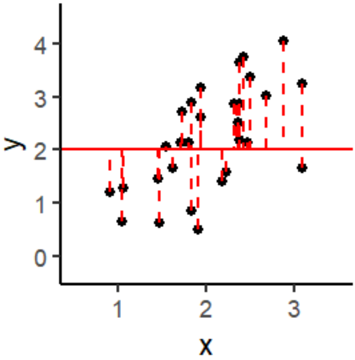
Regression as Least Squares

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

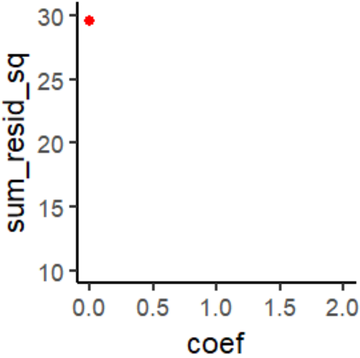
Regression as Least Squares

- 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



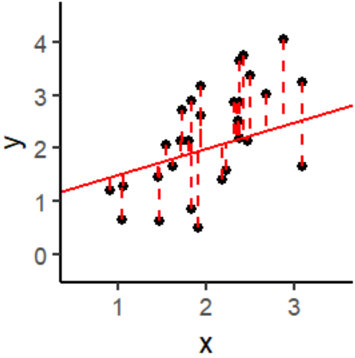
Sum of Squared Residuals = 29.6



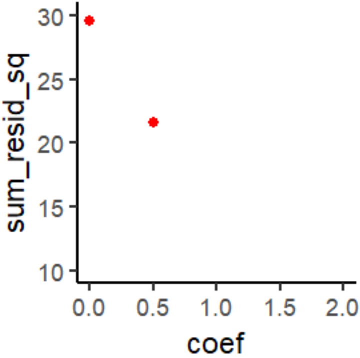
Regression as Least Squares

- 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



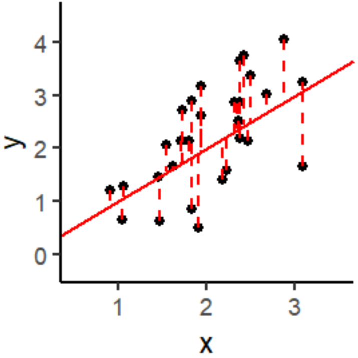
Sum of Squared Residuals = 21.6



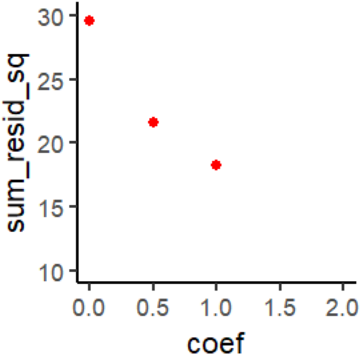
Regression as Least Squares

- 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



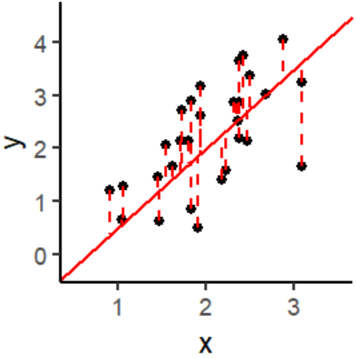
Sum of Squared Residuals = 18.3



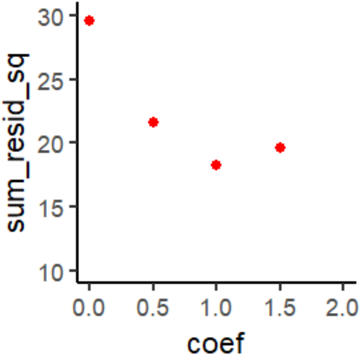
Regression as Least Squares

- 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



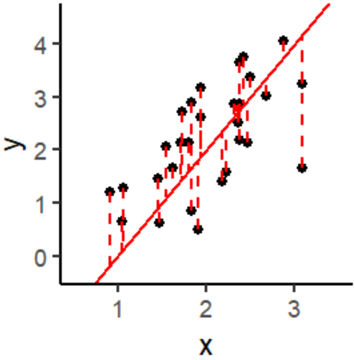
Sum of Squared Residuals = 19.6



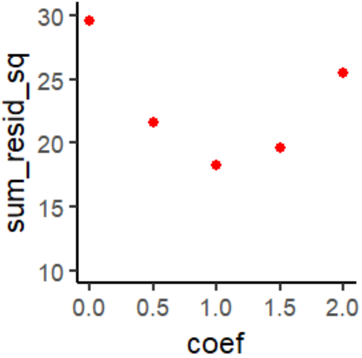
Regression as Least Squares

- 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



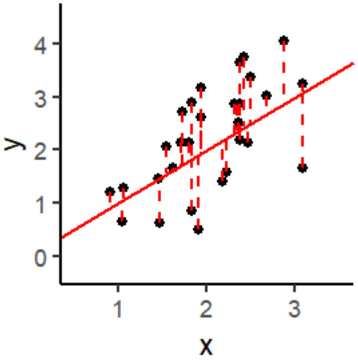
Sum of Squared Residuals = 25.5



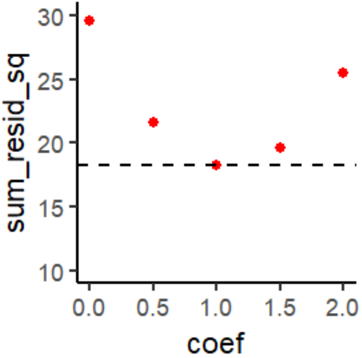
Regression as Least Squares

- 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



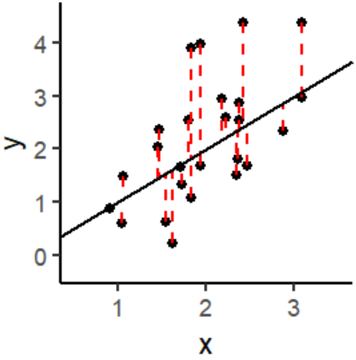
Sum of Squared Residuals = 18.3



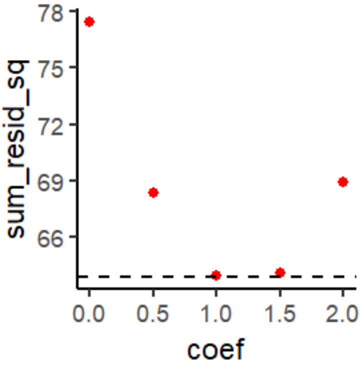
Regression as Least Squares

- 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

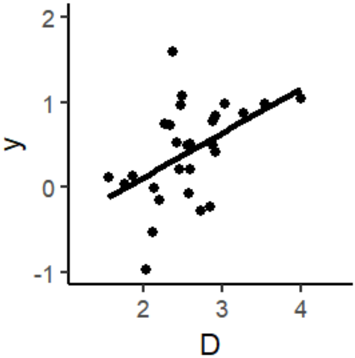


Sum of Squared Residuals = 63.9



Regression as Least Squares

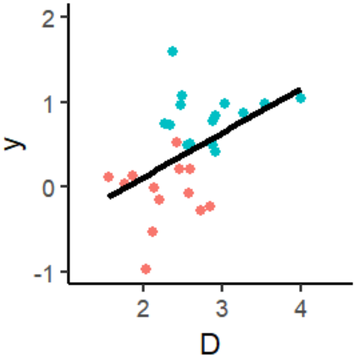
- Dummy control variables *remove variation* associated with specific levels or categories
  - The same for fixed effects
- $y_{ij} = \alpha + \beta_1 D_{ij} + \beta_2 X_j + \epsilon_i$



Ignoring the dummy control variable, the slope coefficient is 1

Regression as Least Squares

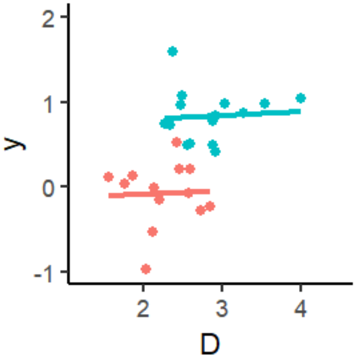
- Dummy control variables *remove variation* associated with specific levels or categories
  - The same for fixed effects
- $y_{ij} = \alpha + \beta_1 D_{ij} + \beta_2 X_j + \epsilon_i$



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

Regression as Least Squares

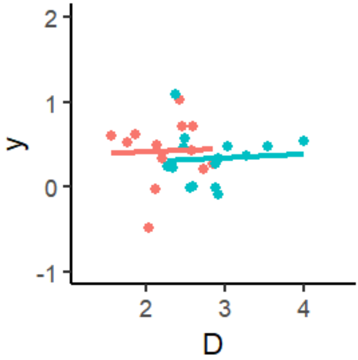
- Dummy control variables *remove variation* associated with specific levels or categories
  - The same for fixed effects
- $y_{ij} = \alpha + \beta_1 D_{ij} + \beta_2 X_j + \epsilon_i$



What if we treated each group *separately*?

Regression as Least Squares

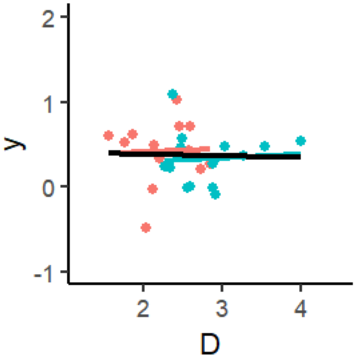
- Dummy control variables *remove variation* associated with specific levels or categories
  - The same for fixed effects
- $y_{ij} = \alpha + \beta_1 D_{ij} + \beta_2 X_j + \epsilon_i$



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

Regression as Least Squares

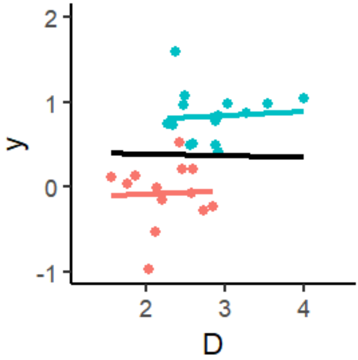
- Dummy control variables *remove variation* associated with specific levels or categories
  - The same for fixed effects
- $y_{ij} = \alpha + \beta_1 D_{ij} + \beta_2 X_j + \epsilon_i$



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
- $y_{ij} = \alpha + \beta_1 D_{ij} + \beta_2 X_j + \epsilon_i$

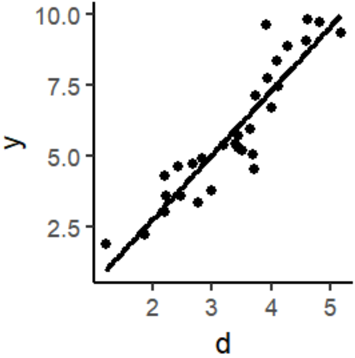


Equivalently, dummy control variables restrict comparisons to **within the same group**:

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

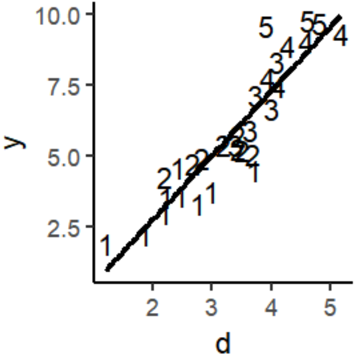
- Continuous control variables *remove variation* based on how much the control explains y
- $y_i = \alpha + \beta_1 D_i + \beta_2 X_i + \epsilon_i$



The coefficient β_1 is 2.267 Real effect = 1

Regression as Least Squares

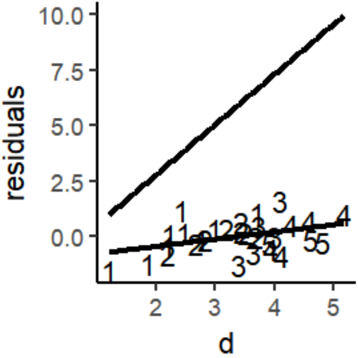
- Continuous control variables *remove variation* based on how much the control explains y
- $y_i = \alpha + \beta_1 D_i + \beta_2 X_i + \epsilon_i$



The coefficient β_1 is 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

- Regression is a **Conditional Expectation Function**

Regression

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

Regression

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

Regression

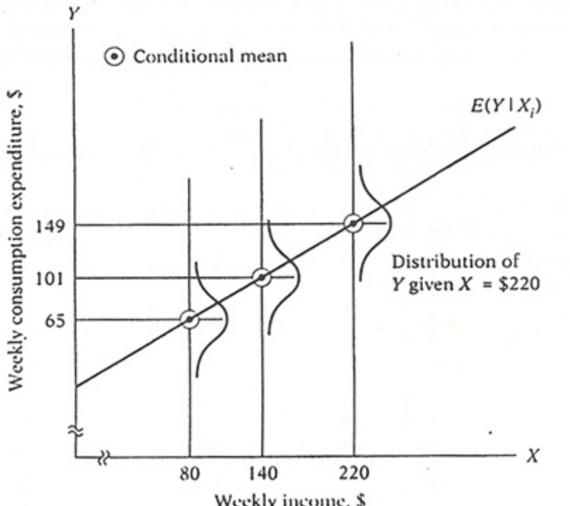
- Regression is a **Conditional Expectation Function**
- Conditional on x, what is our expectation (mean value) of y?
- $E(y|x)$
- When age is 20 ($x = 40$), the average salary is R1.000 ($y = 1.000$)
- When age is 40 ($x = 40$), the average salary is R2.000 ($y = 2.000$)

Regression

- Regression is a **Conditional Expectation Function**: $E(y|x)$

Regression

- Regression is a **Conditional Expectation Function**: $E(y|x)$
- It predicts the **mean**, not the median, not the minimum, not the maximum



Regression

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Regression

- Regression with two variables is very similar to calculating correlation

Regression

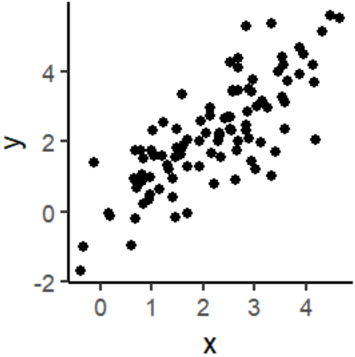
- Regression with two variables is very similar to calculating correlation
- $\hat{\beta} = cor(x, y) * \frac{\sigma_Y}{\sigma_X}$

Regression

- 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 ($\frac{(x-\bar{x})}{\sigma_x}$)

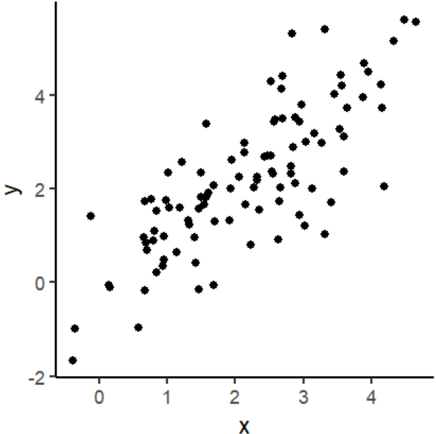
Regression

- 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 ($\frac{(x - \bar{x})}{\sigma_x}$)



Regression

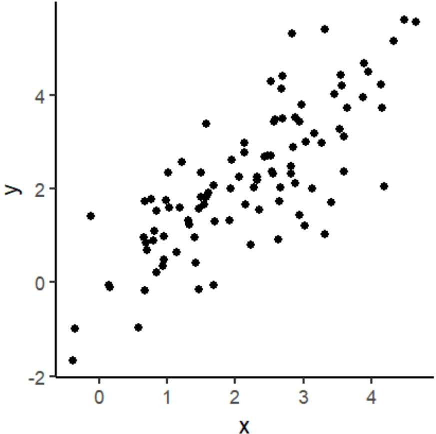
- Regression with two variables is very similar to calculating correlation:
- $\hat{\beta} = cor(x, y) * \frac{\sigma_Y}{\sigma_X}$
- It's *identical* if we standardize both variables first ($\frac{(x - \bar{x})}{\sigma_x}$)



- Correlation is 0.781

Regression

- 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 ($\frac{(x - \bar{x})}{\sigma_x}$)

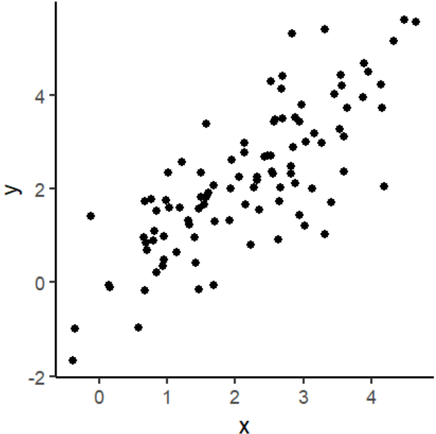


- Correlation is 0.781
- Regression Results:

	term	estimate
1	(Intercept)	0.006
2	x	1.008

Regression

- 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

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

Regression

- Regression with **multiple** variables is very similar to calculating **partial** correlation:
- Just a small difference in the denominator (how we standardize the measure)

Regression

- Regression with **multiple** variables is very similar to calculating **partial** correlation:
- Just a small difference in the denominator (how we standardize the measure)

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

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

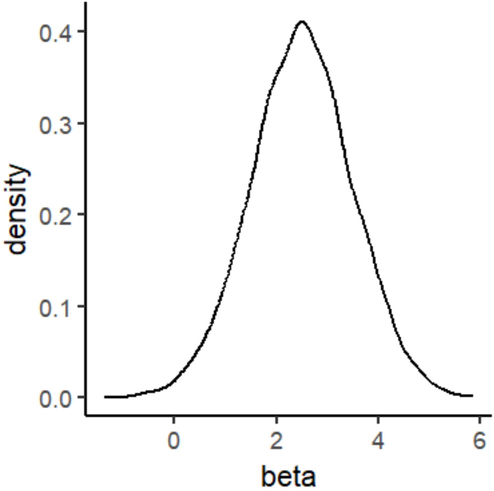
- **There is no magic in regression, it's just correlation 'extra'**

Regression

- We **NEVER** know the true value of β

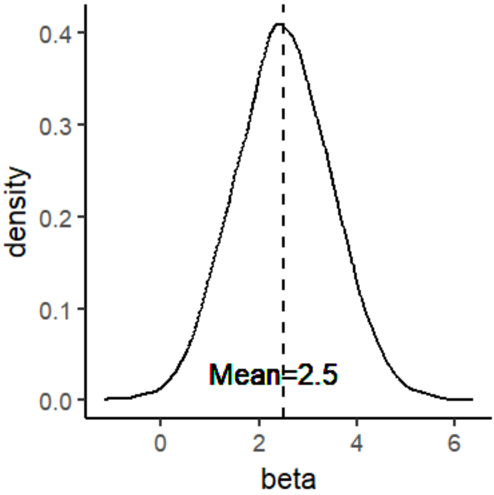
Regression

- We **NEVER** know the true value of β
- We **estimate a distribution** for β



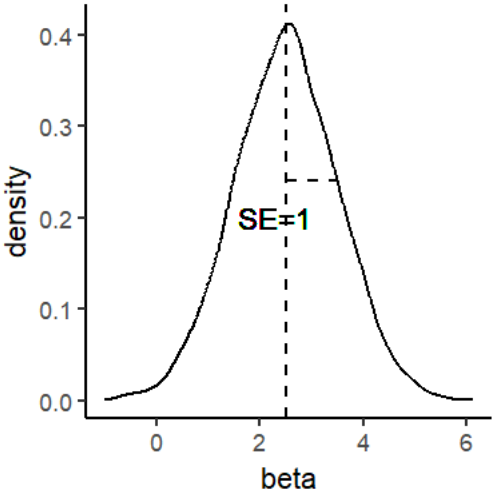
Regression

- We **NEVER** know the true value of β
- We **estimate a distribution** for β



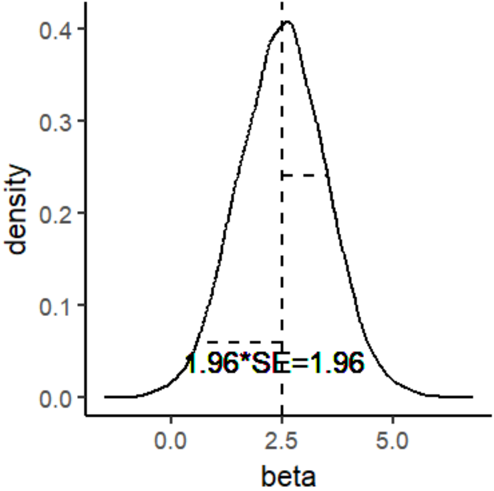
Regression

- We **NEVER** know the true value of β
- We **estimate a distribution** for β



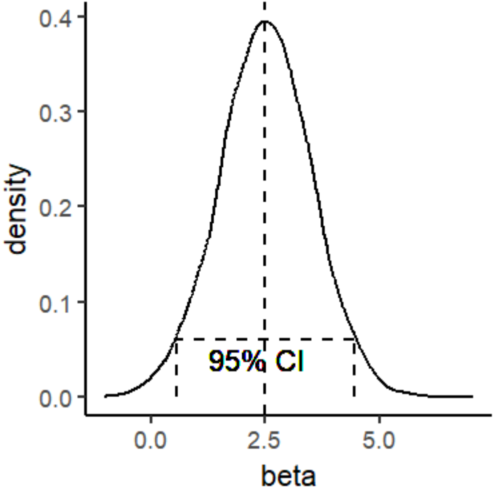
Regression

- We **NEVER** know the true value of β
- We **estimate a distribution** for β



Regression

- We **NEVER** know the true value of β
- We **estimate a distribution** for β



Section 2

Guide to 'Smart' Regression

Regression Guide

1. **Choose variables and measures:** To test a specific hypothesis
2. **Choose a Model/Link Function:** Should match the data type of your outcome variable
3. **Choose Covariates:** To match your strategy of inference
4. **Choose Fixed Effects:** To focus on a specific level of variation
5. **Choose Error Structure:** To match known dependencies/clustering in the data
6. **Interpret the coefficients:** Depending on the type/scale of the explanatory variable

2. Regression Models

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

- Continuous -> Ordinary Least Squares

```
zelig(Y   X,data=d,model="ls")
```

- Binary -> Logit

```
zelig(Y   X,data=d,model="logit")
```

- Unordered categories -> Multinomial logit

```
zelig(Y   X,data=d,model="mlogit")
```

- Ordered categories -> Ordered logit

```
zelig(Y   X,data=d,model="ologit")
```

- Count -> Poisson

```
zelig(Y   X,data=d,model="poisson")
```

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

Predictions from Regressions

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

Predictions from Regressions

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

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

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

Predictions from Regressions

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

Predictions from Regressions

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

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

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:

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

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

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:

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

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

Predictions from Regressions

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

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

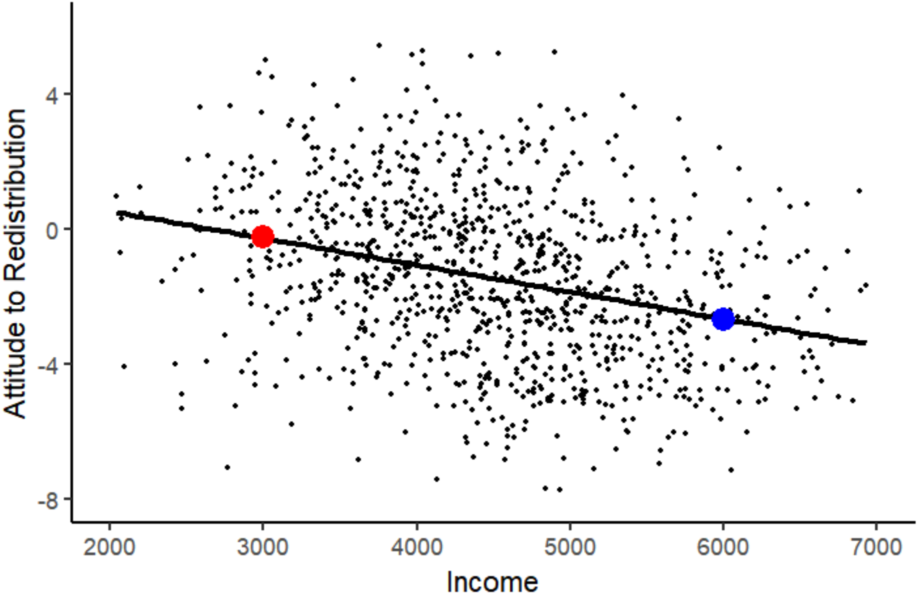
Increasing Income from 3000 to 6000:

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

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

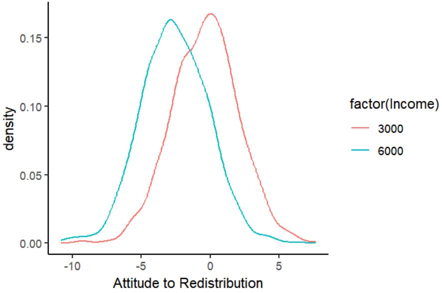
$$\Delta Attitude_i = -2.454$$

Predictions from Regressions

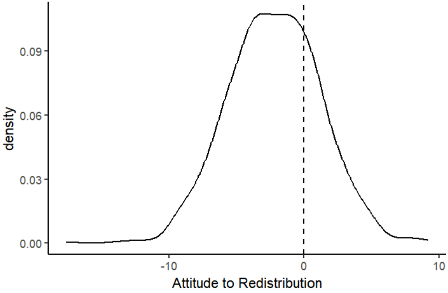


Predictions from Regressions

Predicted Values:



First Differences:



Predictions from Regressions

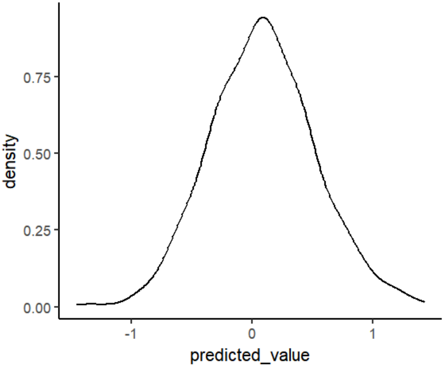
- The regression model matters because the wrong model makes non-sensical predictions
- Consider a binary outcome: $Gender_i = \alpha + \beta Income_i + \epsilon_i$
- Compare the OLS and Logit regression tables:

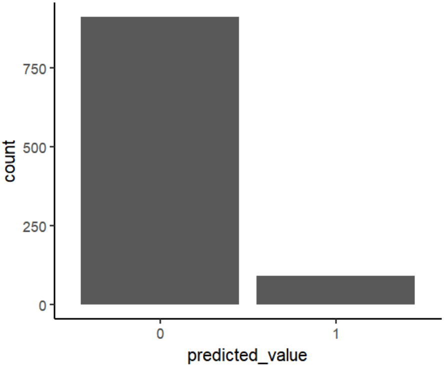
	Dependent variable:
	as.numeric(as.character(gender))
income	0.0003*** (0.00001)
Constant	−0.696*** (0.066)
Observations	1,000
Note:	*p<0.1; **p<0.05; ***p<0.01

	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

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:

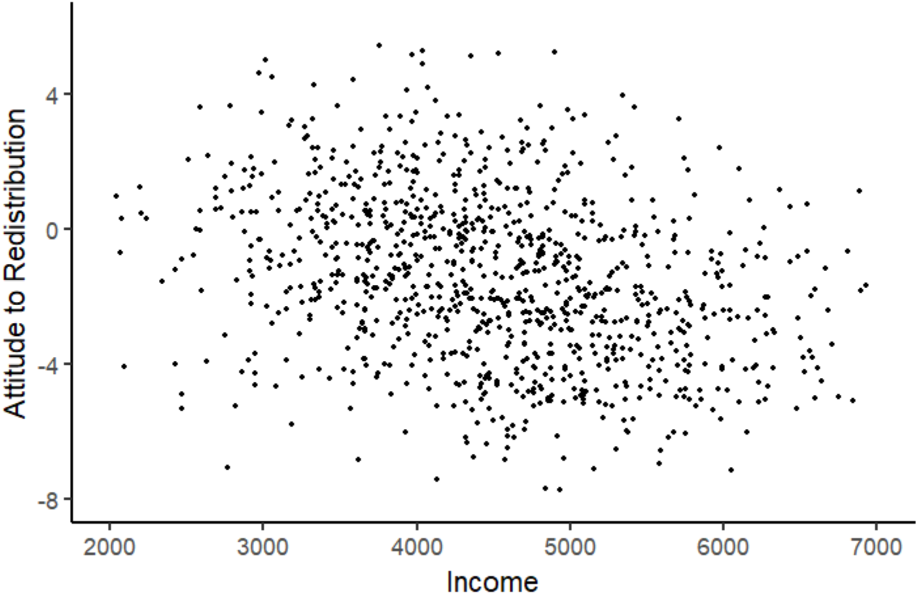




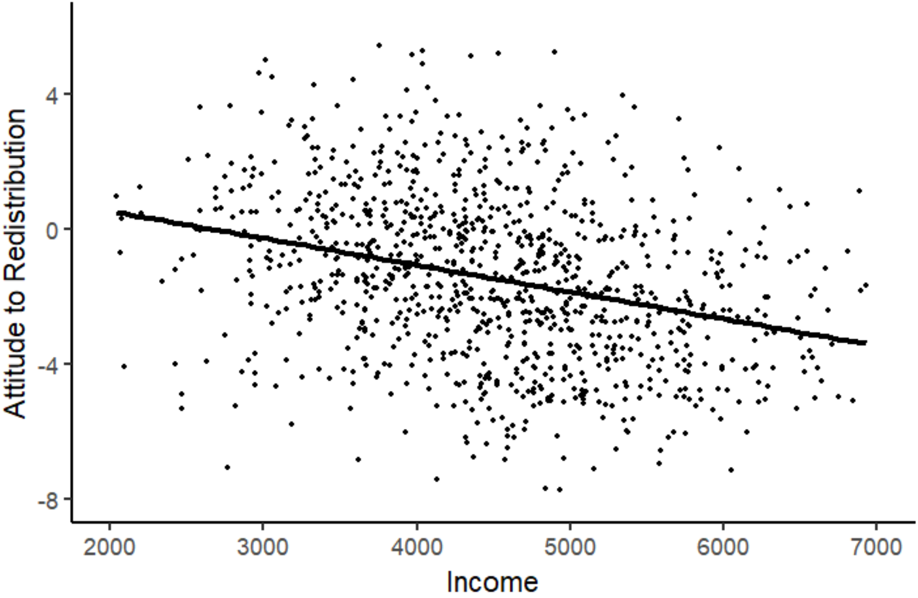
Section 3

What Does Regression NOT Do?

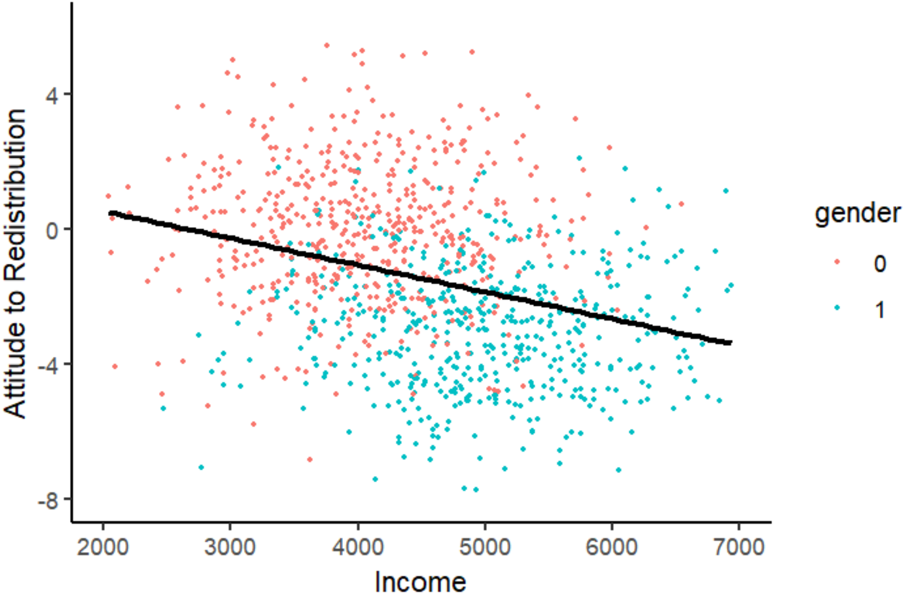
Omitted Variable Bias



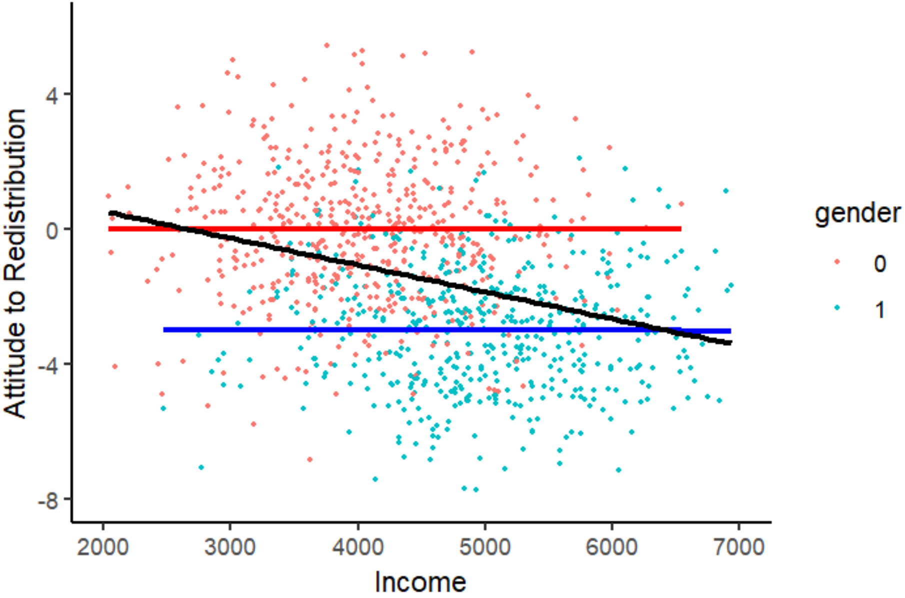
Omitted Variable Bias



Omitted Variable Bias



Omitted Variable Bias



Reverse Causation

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

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?

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

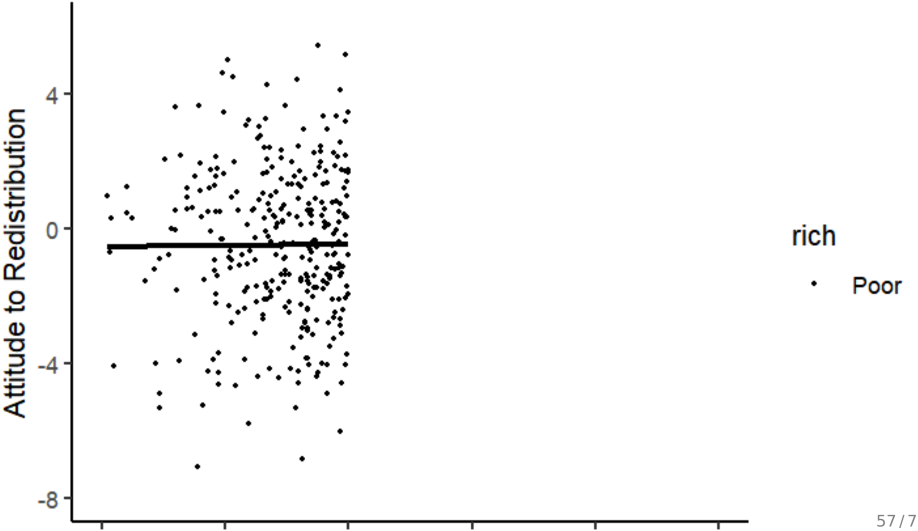
- Remember, regression measures the *vertical* (not diagonal) distances to the regression line
  - It minimizes the prediction errors for Y
- But that doesn't mean it identifies the direction of causation!

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

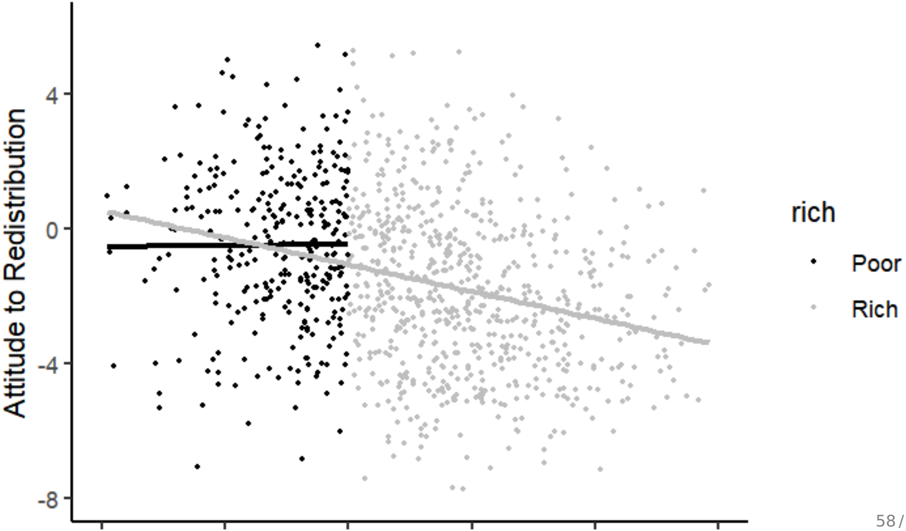
Selection Bias

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



Selection Bias

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

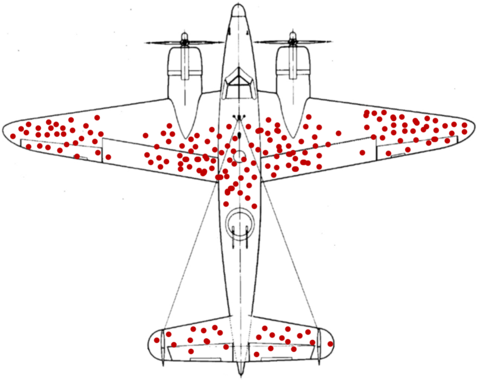


Selection Bias

- 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!

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

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

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

Measurement Bias

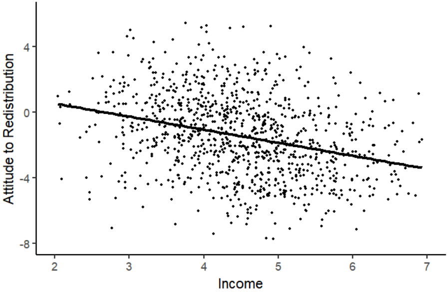
- What happens if we measure our variables wrongly?

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

Measurement Bias

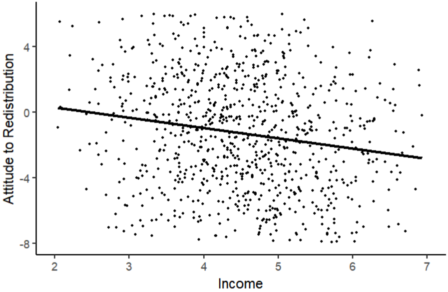
- What happens if we measure our variables wrongly?
- No extra noise:



	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

Measurement Bias

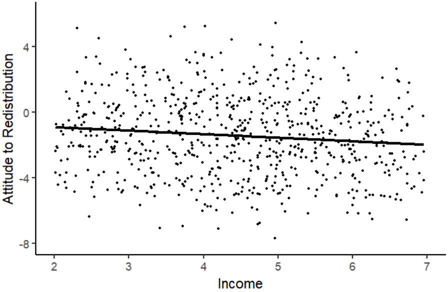
- What happens if we measure our variables wrongly?
- Noise in the **outcome variable**:



	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

Measurement Bias

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

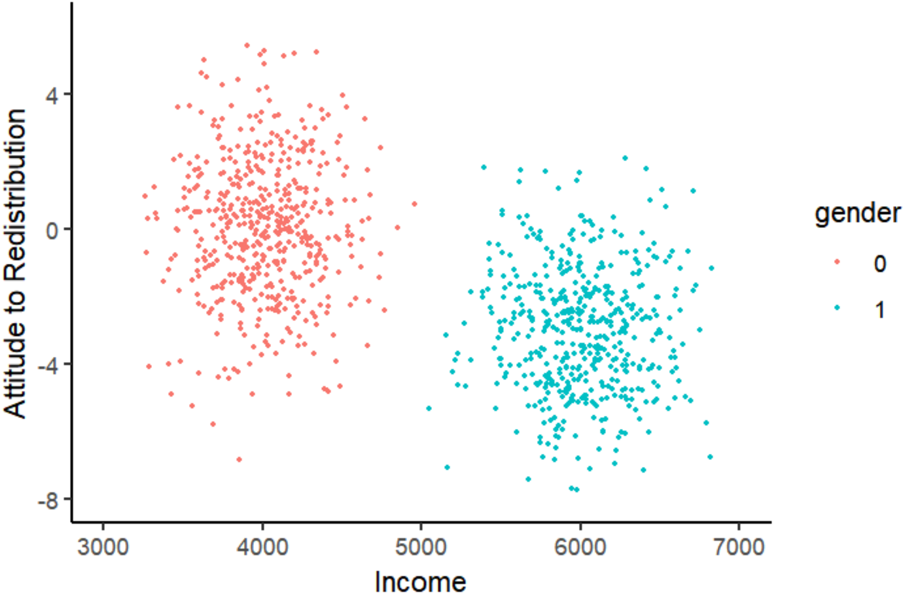


	Dependent variable:
	redist
income	-0.187^{***} (0.037)
Constant	-0.620^{***} (0.183)
Observations	1,000
Note:	*p<0.1; **p<0.05; ***p<0.01

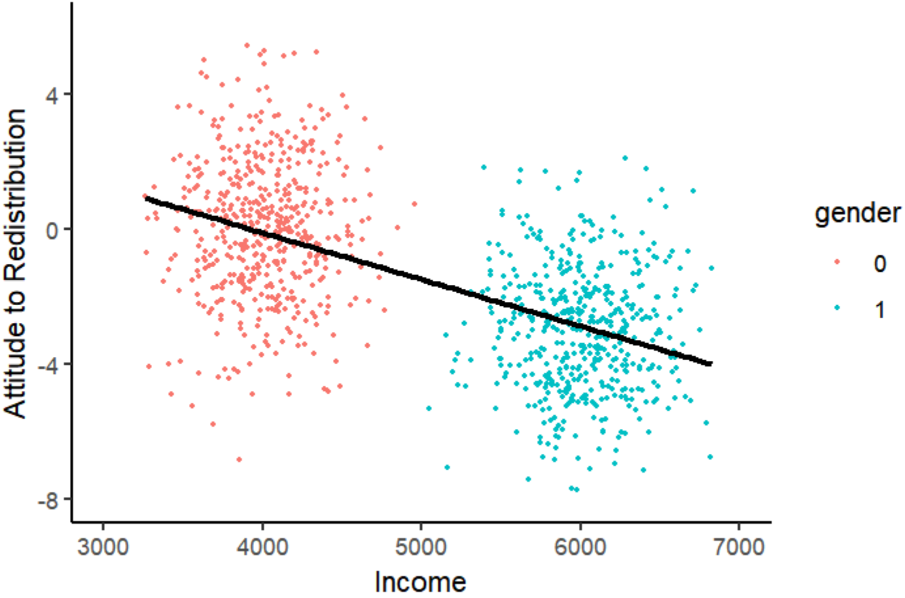
Lack of Overlap

- 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

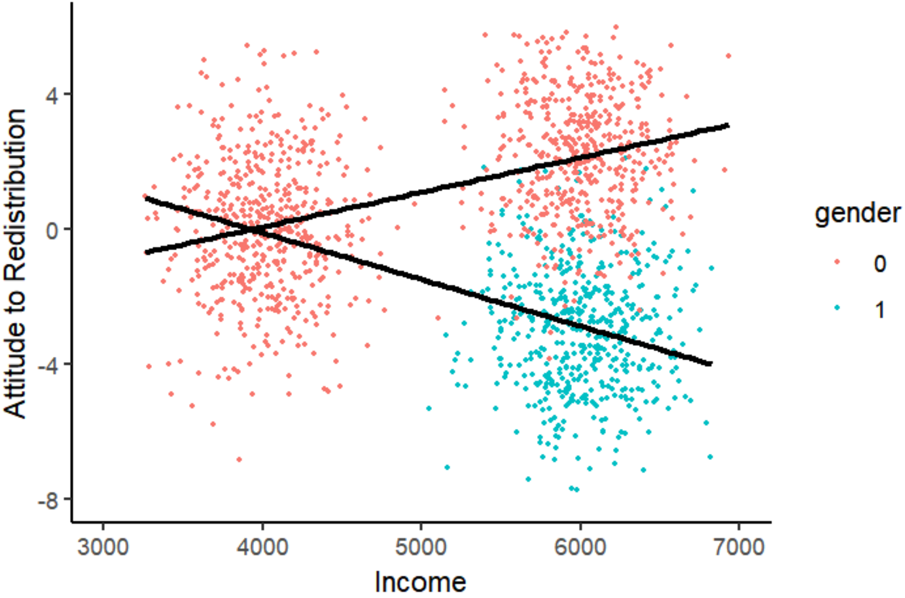
Lack of Overlap



Lack of Overlap



Lack of Overlap



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