FLS 6441 - Methods III: Explanation and Causation

Week 1 - Review

Jonathan Phillips

February 2019

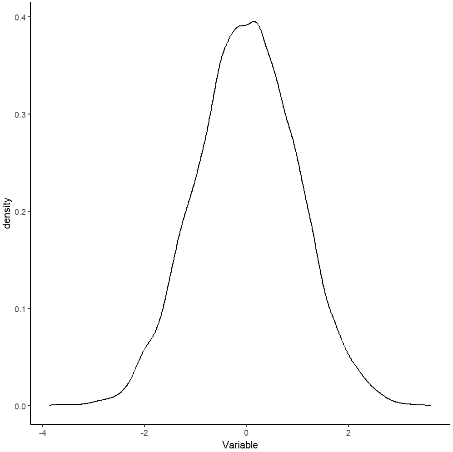
Course Objectives

1. temp

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



Regression

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

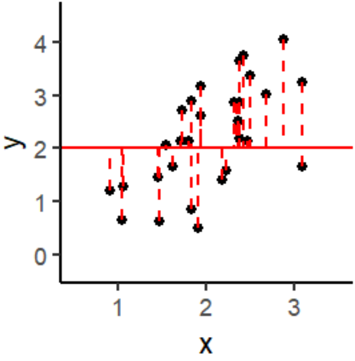
Regression

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

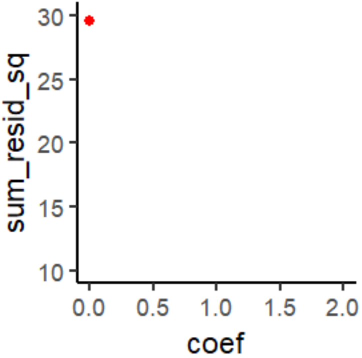
Regression

- 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



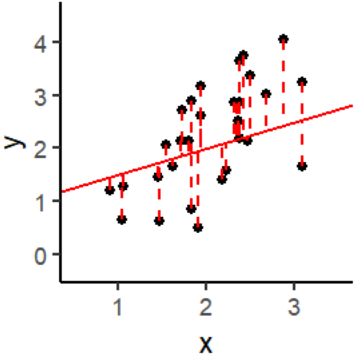
Sum of Squared Residuals = 29.6



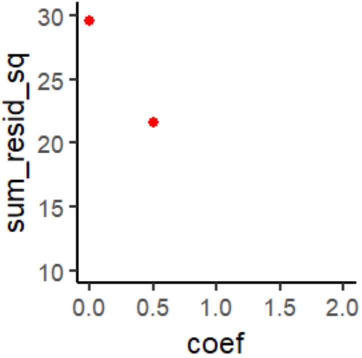
Regression

- 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



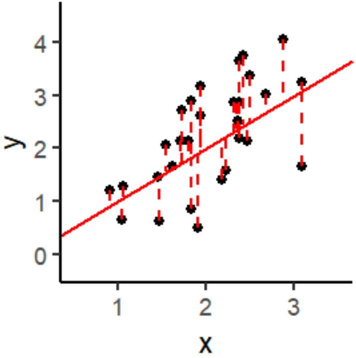
Sum of Squared Residuals = 21.6



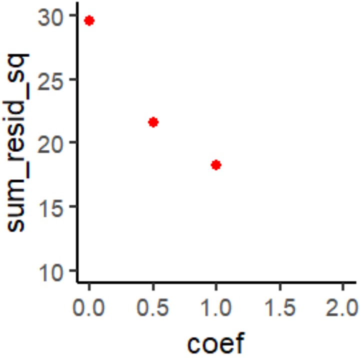
Regression

- 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



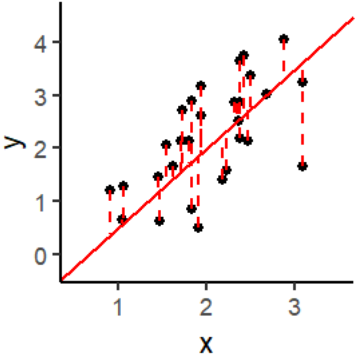
Sum of Squared Residuals = 18.3



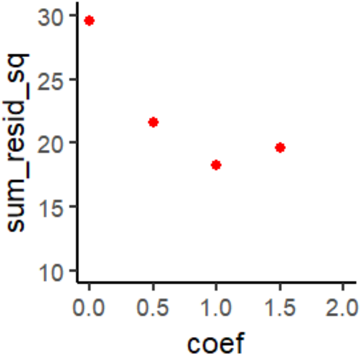
Regression

- 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



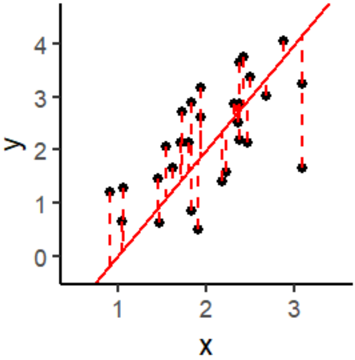
Sum of Squared Residuals = 19.6



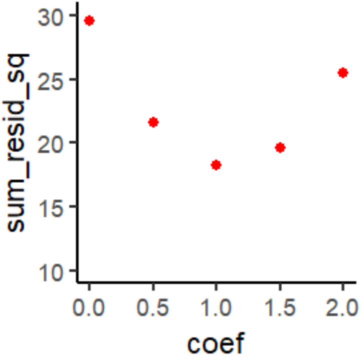
Regression

- 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



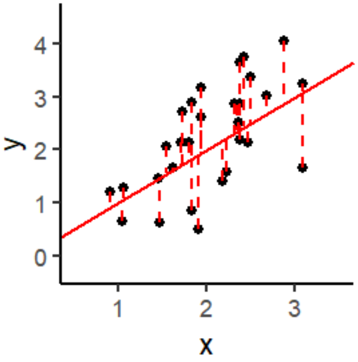
Sum of Squared Residuals = 25.5



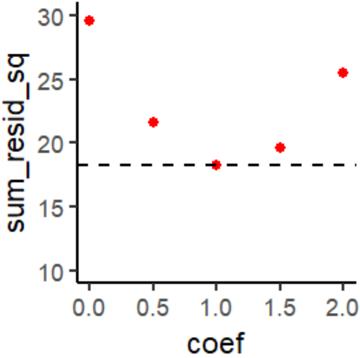
Regression

- 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



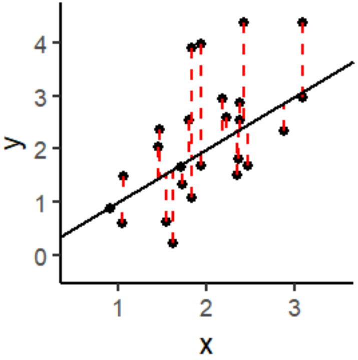
Sum of Squared Residuals = 18.3



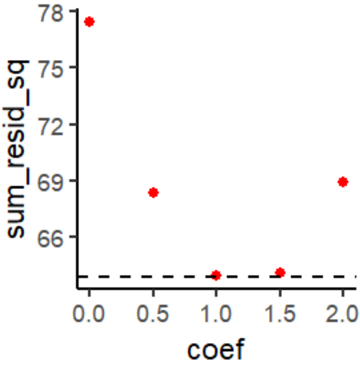
Regression

- 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

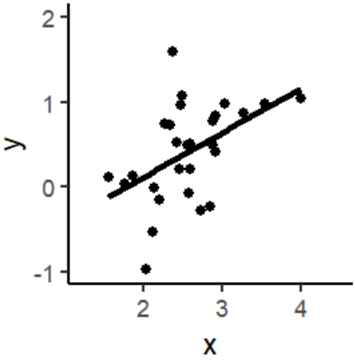


Sum of Squared Residuals = 63.9



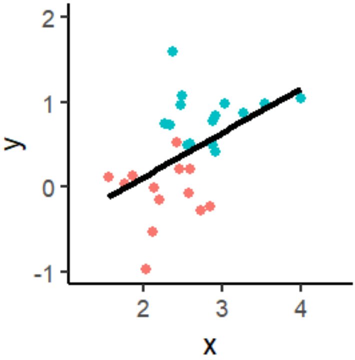
Regression

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



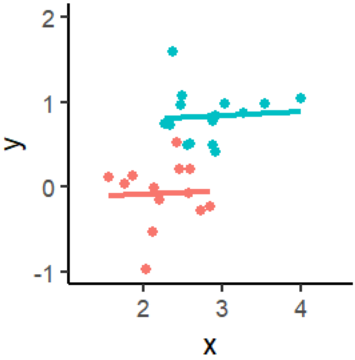
Regression

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



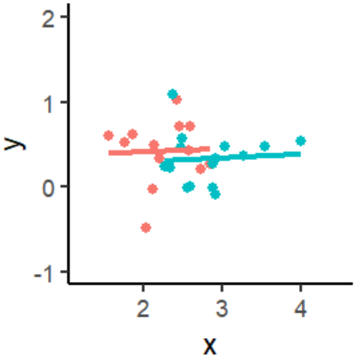
Regression

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



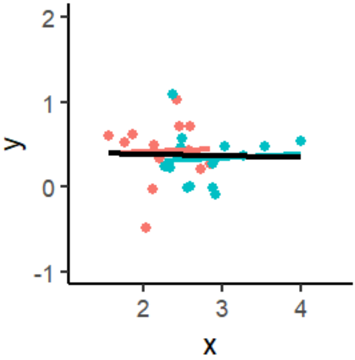
Regression

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



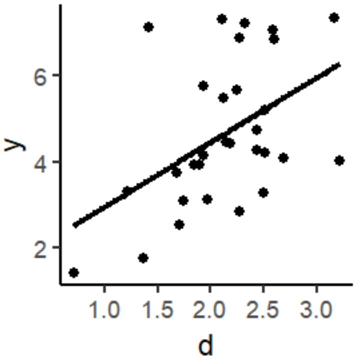
Regression

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



Regression

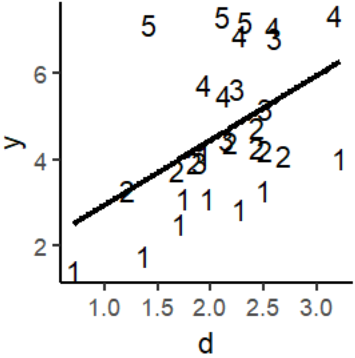
- Continuous control variables *remove variation* based on how much the control explains y
- $y_i = \alpha + \beta_1 D_i + \beta_2 X_i + \epsilon_i$



The coefficient on D is 1.503
Real coefficient = 1

Regression

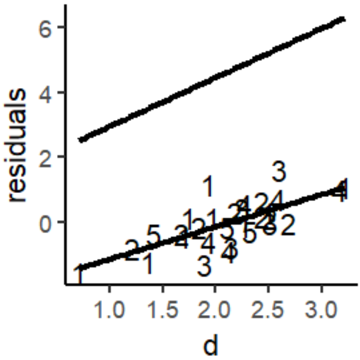
- Continuous control variables *remove variation* based on how much the control explains y
- $y_i = \alpha + \beta_1 D_i + \beta_2 X_i + \epsilon_i$



The coefficient on D is 1.503
Real coefficient = 1

Regression

- 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 on D is 0.991
Real coefficient = 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

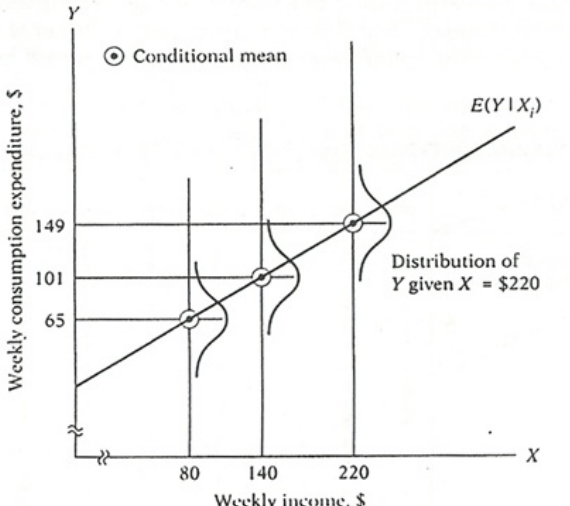
- Regression is a **Conditional Expectation Function**
- Conditional on x, what is our expectation (mean value) of y?
- $E(y|x)$
- When age is 20 ($x = 40$), the average salary is R1.000 ($y = 1.000$)
- When age is 40 ($x = 40$), the average salary is R2.000 ($y = 2.000$)

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

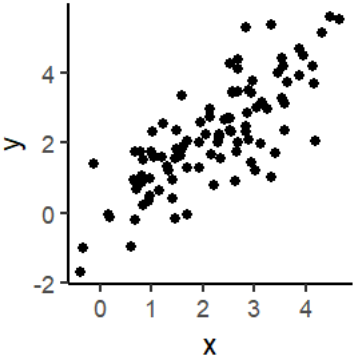
- Regression with two variables is very similar to calculating correlation
- $\hat{\beta} = cor(x, y) * \frac{\sigma_Y}{\sigma_X}$

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

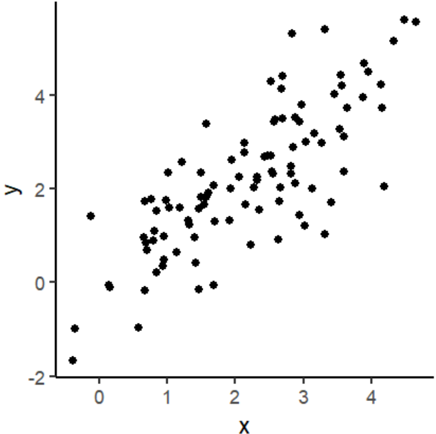
Regression

- 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

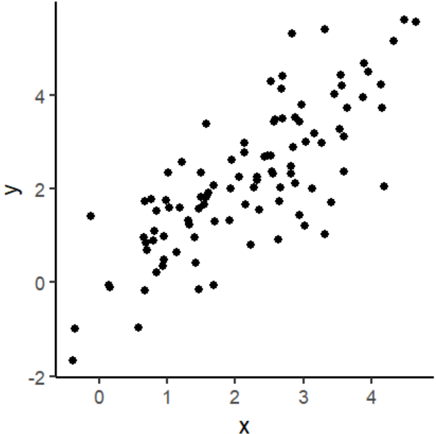
- Regression with two variables is very similar to calculating correlation:
- $\hat{\beta} = cor(x, y) * \frac{\sigma_Y}{\sigma_X}$
- It's *identical* if we standardize both variables first ($\frac{(x - \bar{x})}{\sigma_x}$)



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

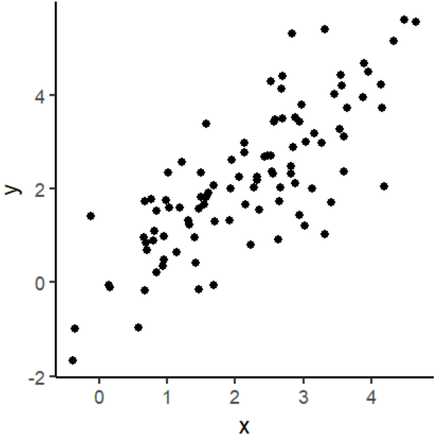


- Correlation is 0.781
- Regression Results:

	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)

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

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

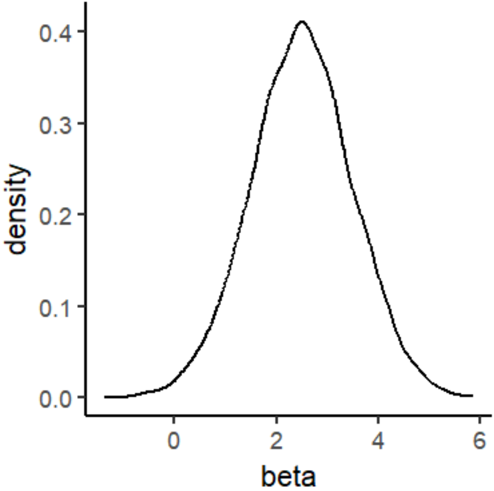
- **There is no magic in regression, it's just correlation 'extra'**

Regression

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

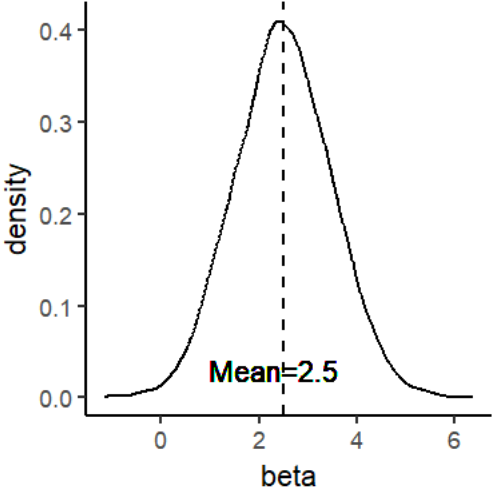
Regression

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



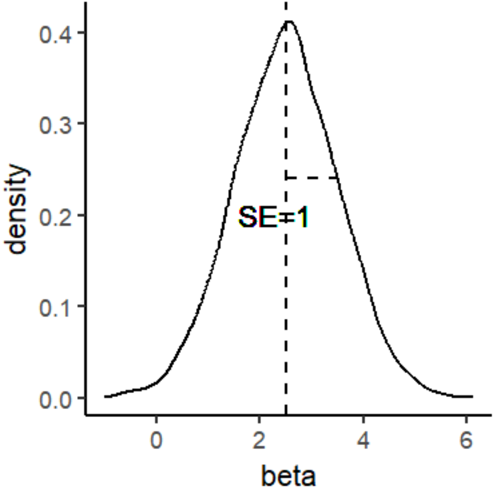
Regression

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



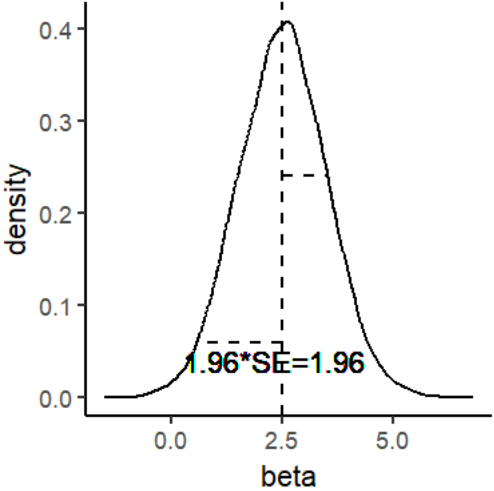
Regression

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



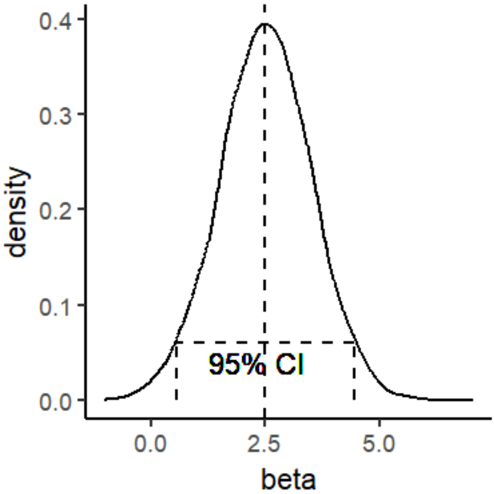
Regression

- 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



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")
```

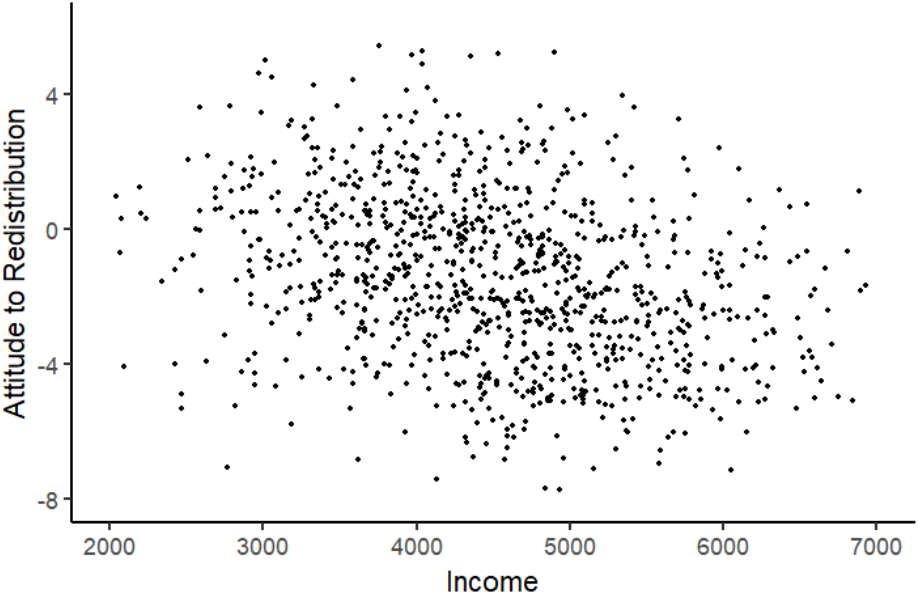
6. Interpreting Regression Results

- 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

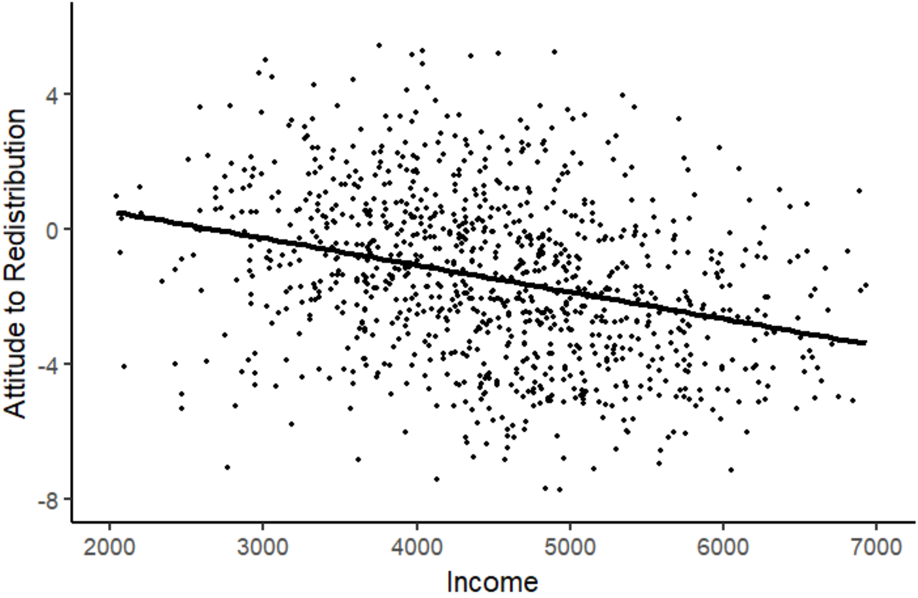
Predictions from Regressions

- temp

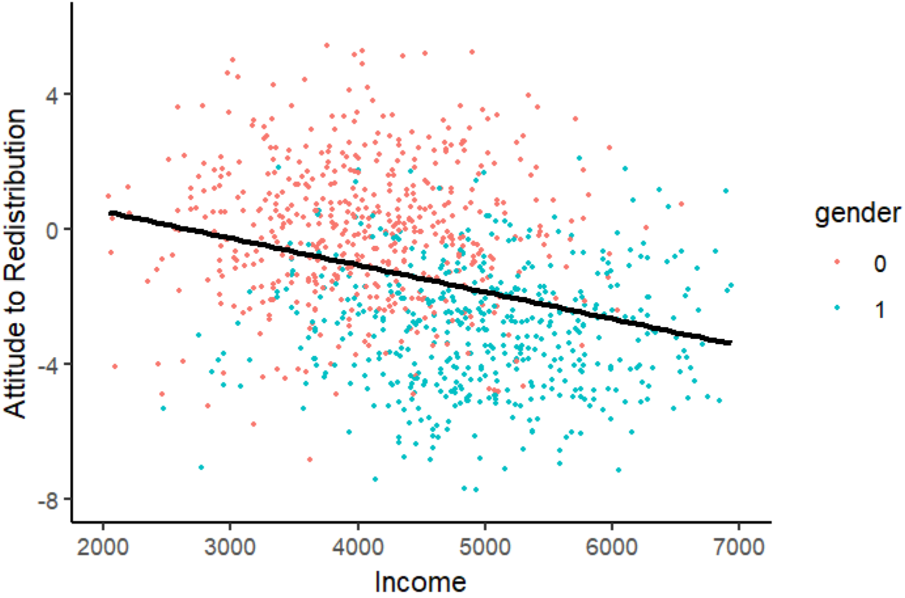
Omitted Variable Bias



Omitted Variable Bias



Omitted Variable Bias



Omitted Variable Bias