Regression and inference

Andy Eggers
Assoc. Professor
Department of Politics and International Relations

I want you to understand:

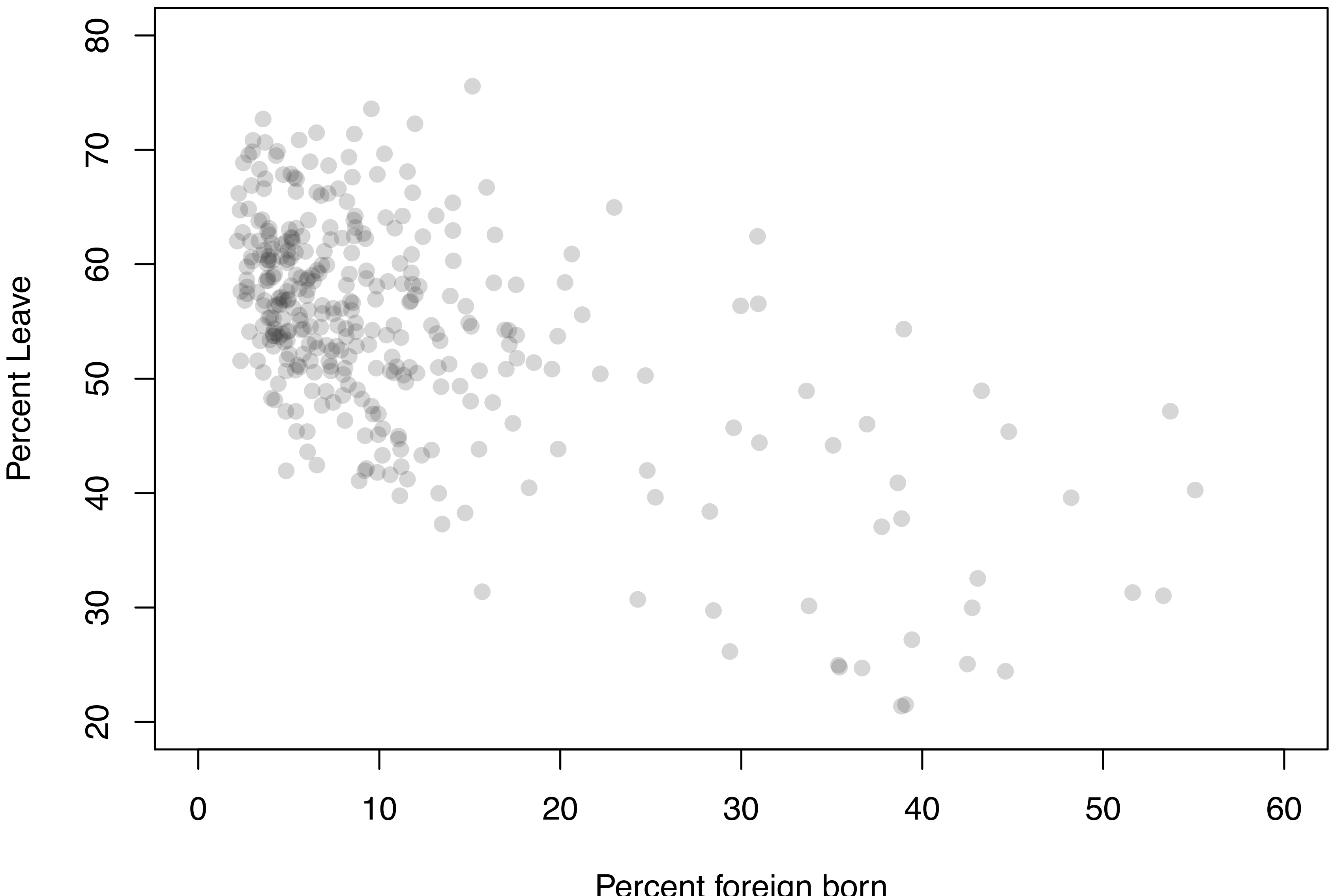
Dependent variable: Nobel Prizes awarded per capita (in log scale)

	(1)	(2)	(3)
Intercept	-1.629* (0.509)	-3.166* (0.511)	-2.982* (0.527)
Chocolate consumption per capita (log scale)	2.092* (0.298)	1.026* (0.326)	0.709 (0.415)
GDP/capita (thousands of USD)		0.105* (0.024)	0.106* (0.024)
NW Europe			0.549 (0.452)
R^2	0.70	0.85	0.86
N	34	34	34

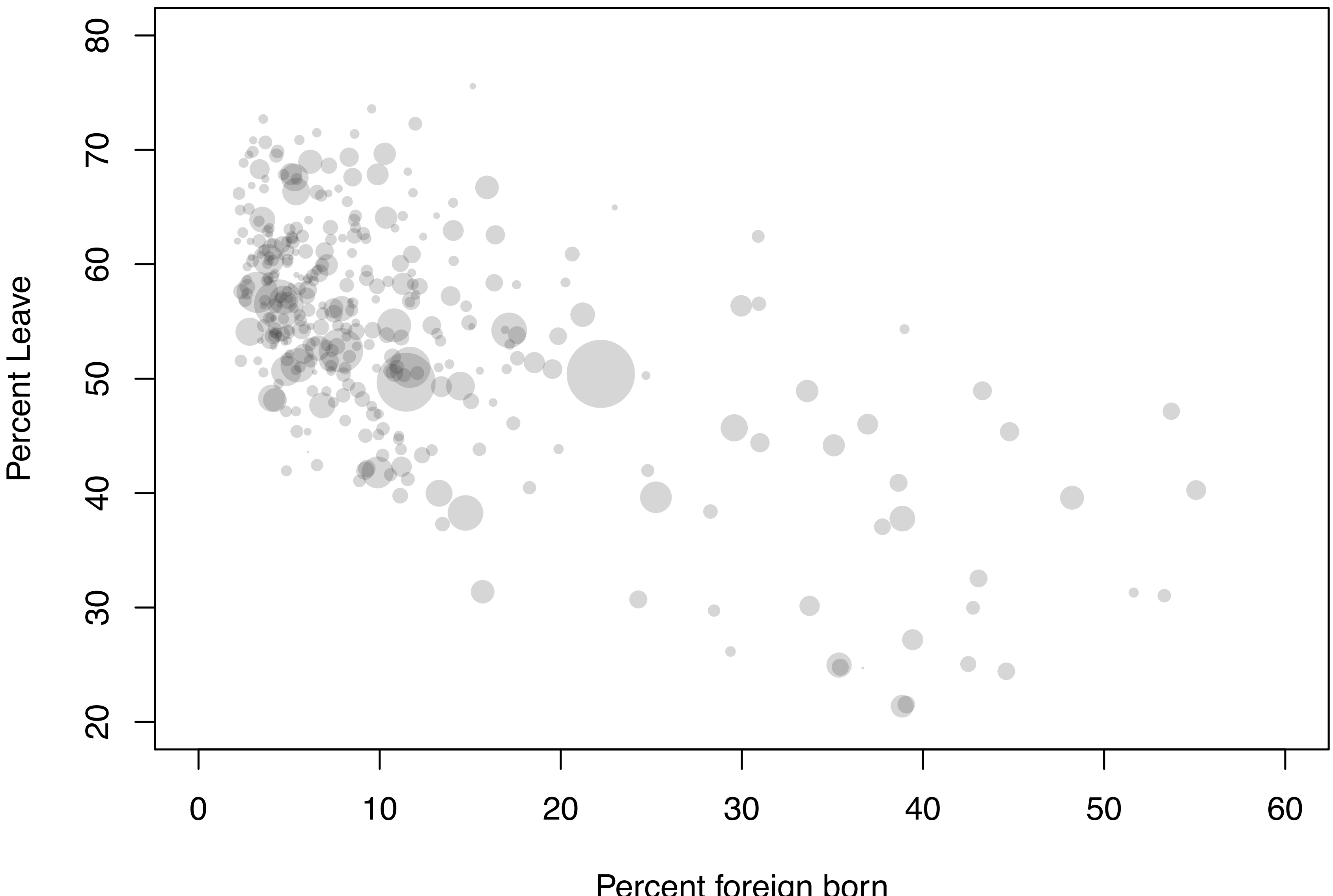
Standard errors in parentheses. * Indicates $p<0.05$

- what a dependent variable is
- what an independent variable is
- what the coefficients mean (intercept, slopes)
- what the stars mean (i.e. what $p<0.05$ means)
- what the standard errors mean

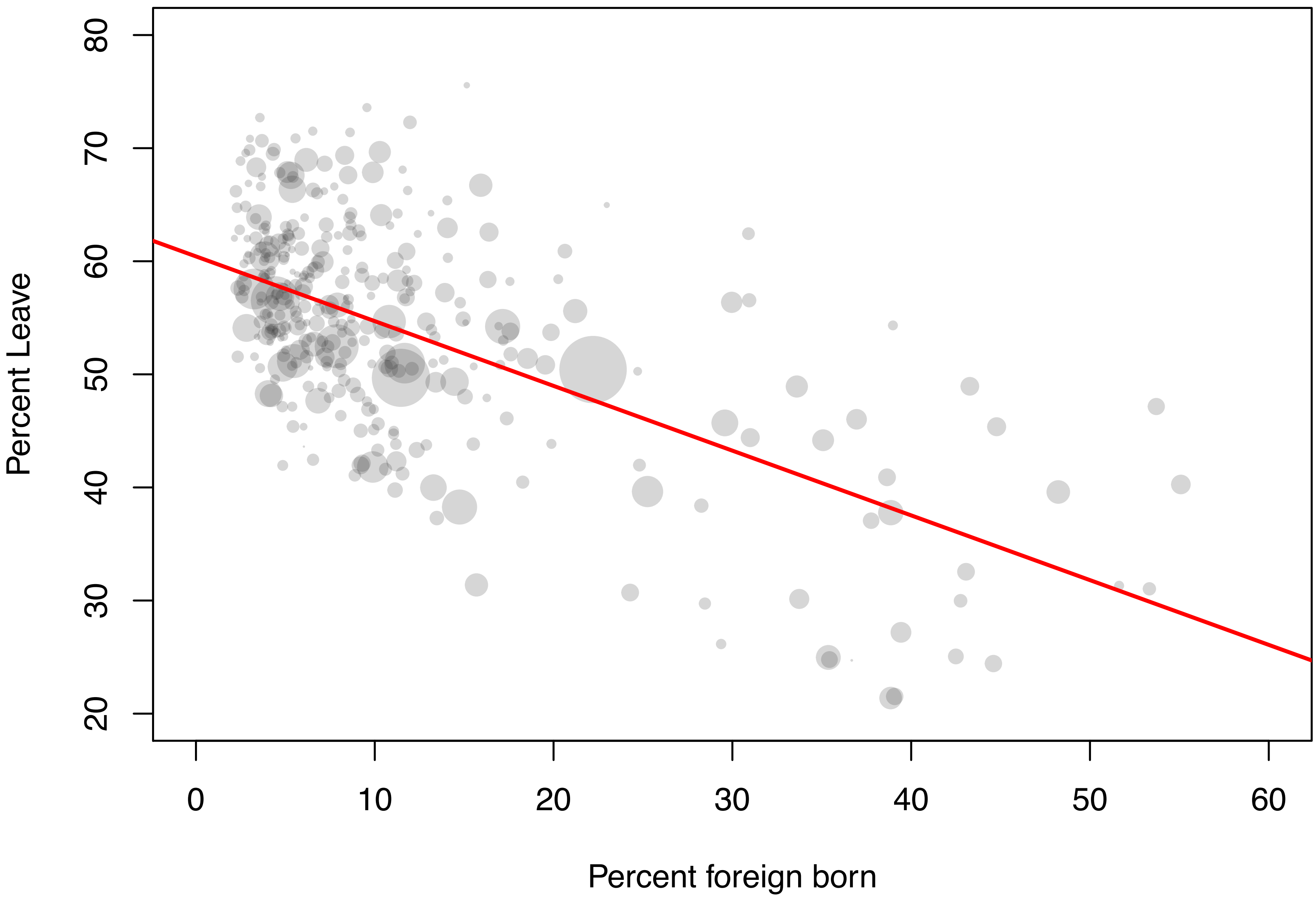
Local authorities with more foreign-born residents were less supportive of Brexit



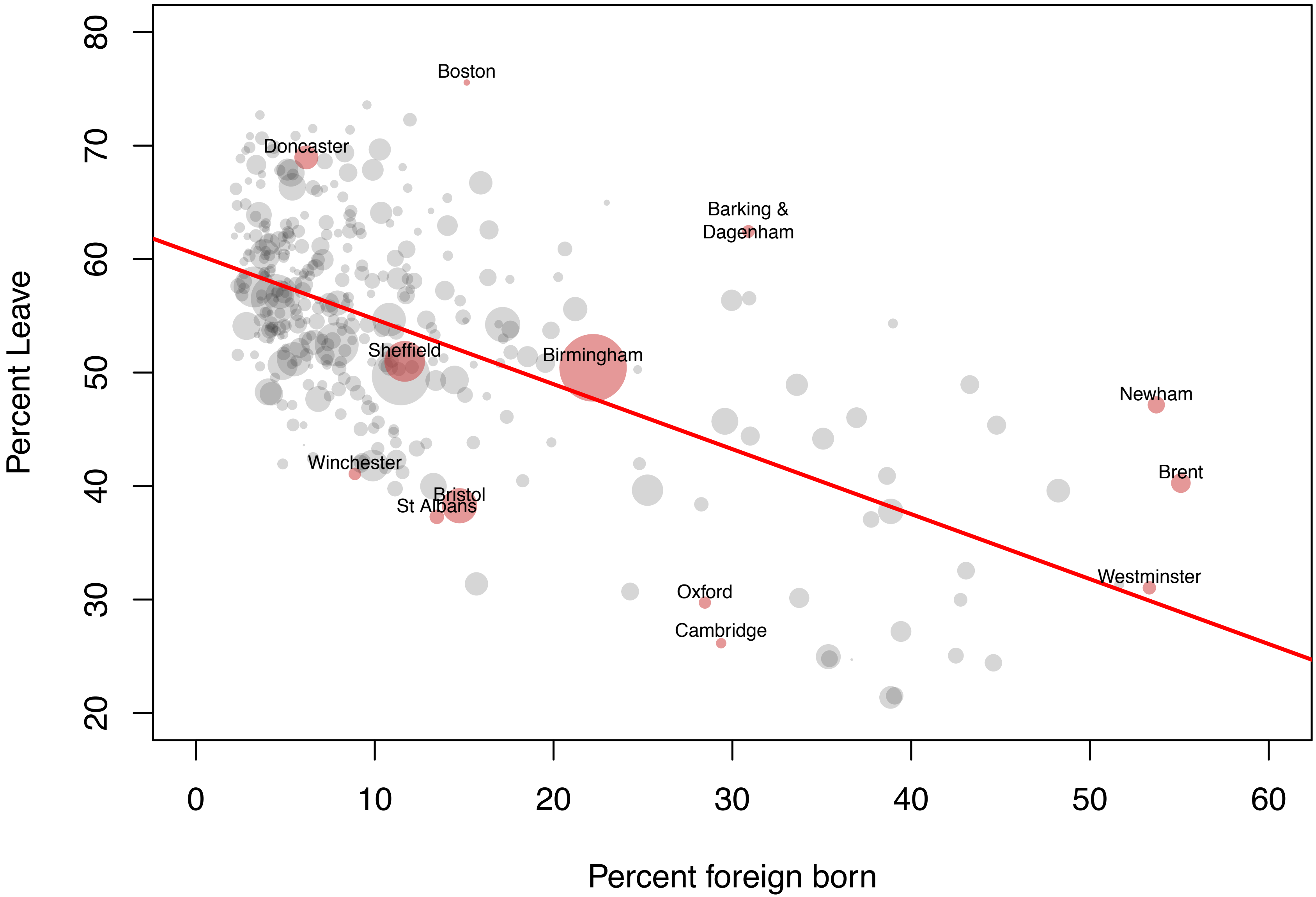
Local authorities with more foreign-born residents were less supportive of Brexit



Local authorities with more foreign-born residents were less supportive of Brexit



Local authorities with more foreign-born residents were less supportive of Brexit



Contact hypothesis

Contact hypothesis

"Prejudice (unless deeply rooted in the character structure of the individual) may be reduced by equal status contact between majority and minority groups in the pursuit of common goals. The effect is greatly enhanced if this contact is sanctioned by institutional supports (i.e., by law, custom or local atmosphere), and provided it is of a sort that leads to the perception of common interests and common humanity between members of the two groups."

— Gordon Allport (1954) *The Nature of Prejudice*

Question

Question

Brexit support is higher in places with fewer foreign-born residents. Does contact between immigrants and other local residents explain this pattern?

Question

Brexit support is higher in places with fewer foreign-born residents. Does contact between immigrants and other local residents explain this pattern?

- How could this pattern be explained by the contact hypothesis? (easy)

Question

Brexit support is higher in places with fewer foreign-born residents. Does contact between immigrants and other local residents explain this pattern?

- How could this pattern be explained by the contact hypothesis? (easy)
- How could this pattern be explained by other factors? (harder)

Today and tomorrow

Today and tomorrow

Running question: Why is there such a strong relationship between % foreign born and opposition to Brexit?

Today and tomorrow

Running question: Why is there such a strong relationship between % foreign born and opposition to Brexit?

Plan:

Today and tomorrow

Running question: Why is there such a strong relationship between % foreign born and opposition to Brexit?

Plan:

- How do we summarize the relationship between two variables?
  - bivariate OLS regression as main focus

Today and tomorrow

Running question: Why is there such a strong relationship between % foreign born and opposition to Brexit?

Plan:

- How do we summarize the relationship between two variables?
  - bivariate OLS regression as main focus
- How do we summarize the relationship between two variables controlling for a third variable?
  - multivariate OLS regression as main focus

Today and tomorrow

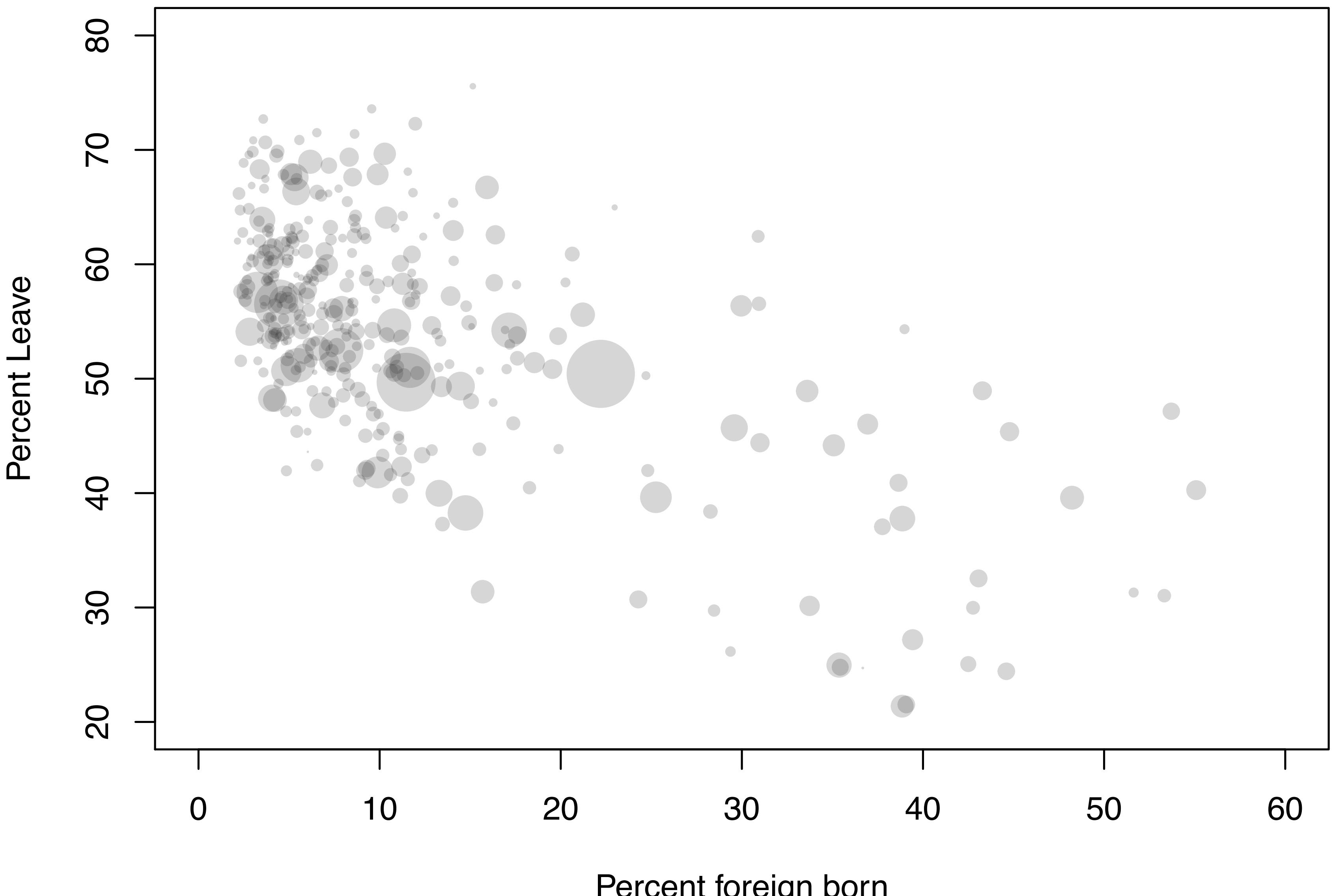
Running question: Why is there such a strong relationship between % foreign born and opposition to Brexit?

Plan:

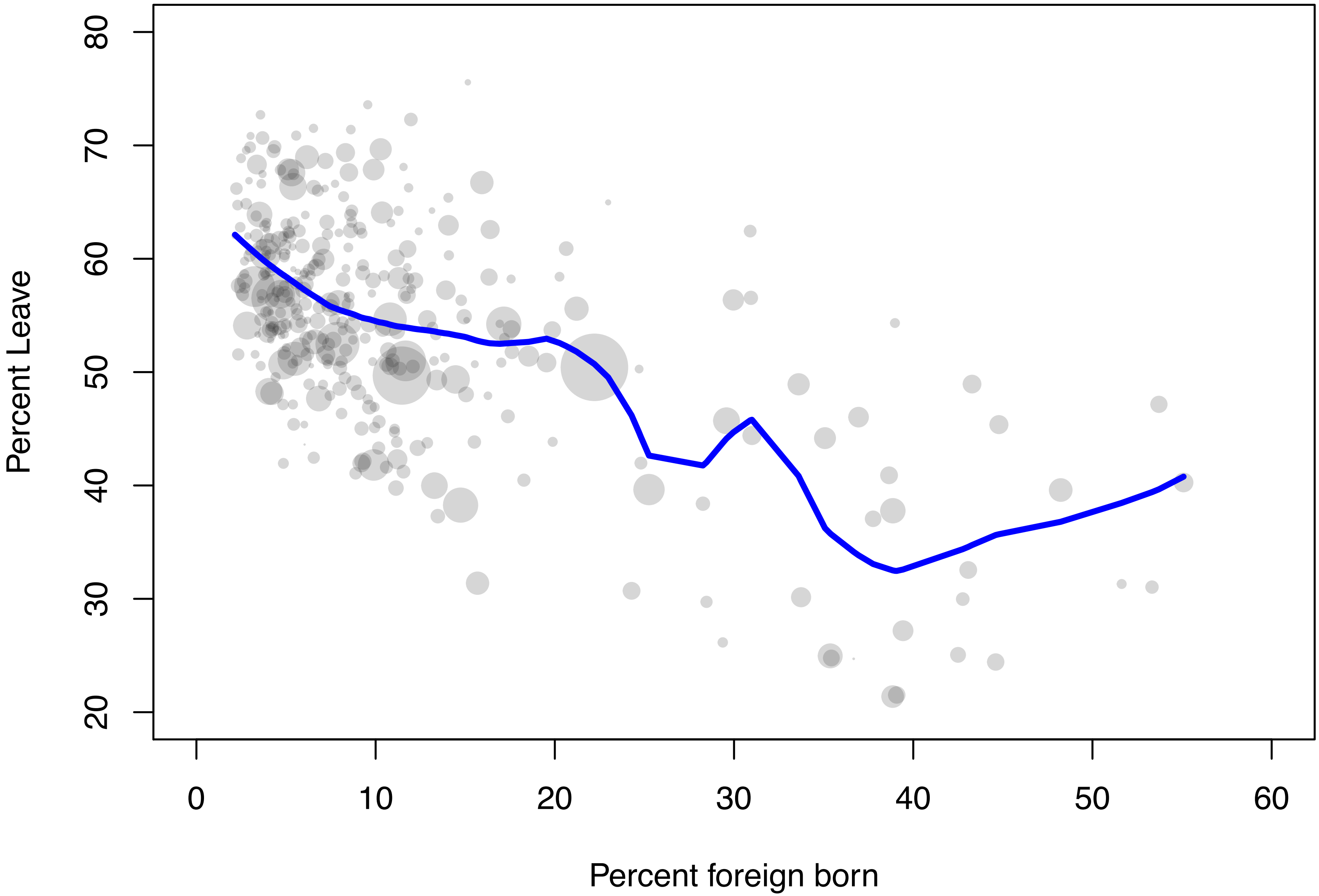
- How do we summarize the relationship between two variables?
  - bivariate OLS regression as main focus
- How do we summarize the relationship between two variables controlling for a third variable?
  - multivariate OLS regression as main focus
- How do we summarize our uncertainty about our conclusions?
  - standard errors, p-values, confidence intervals

Summarizing bivariate relationships: non-OLS options

Local authorities with more foreign-born residents were less supportive of Brexit



Kernel smoother (lokern function in R)



Single-number summaries: covariance

Single-number summaries: covariance

How do x and y tend to move together, i.e. how do they **covary?**

Single-number summaries: covariance

How do x and y tend to move together, i.e. how do they **covary**?

When x is above its mean, is y also above its mean? By how much?

Single-number summaries: covariance

How do x and y tend to move together, i.e. how do they **covary**?

When x is above its mean, is y also above its mean? By how much?

$$\mathrm{Cov}(x, y) = \frac{\Sigma_i (x_i - \overline{x})(y_i - \overline{y})}{n - 1}$$

Single-number summaries: covariance

How do x and y tend to move together, i.e. how do they **covary**?

When x is above its mean, is y also above its mean? By how much?

$$\mathrm{Cov}(x,y) = \frac{\Sigma_i (x_i - \overline{x})(y_i - \overline{y})}{n-1}$$

```
> cov(d$Percent_foreign_born, d$Percent_Leave, use = "complete")
[1] -62.17755
```

Single-number summaries: correlation

If you plot x and y, how closely are the points arranged on a line (and is the slope of that line positive or negative)?

Single-number summaries: correlation

If you plot x and y, how closely are the points arranged on a line (and is the slope of that line positive or negative)?

$$\mathrm{Cor}(x, y) = \frac{\mathrm{Cov}(x, y)}{\mathrm{sd}(x)\mathrm{sd}(y)}$$

Single-number summaries: correlation

If you plot x and y, how closely are the points arranged on a line (and is the slope of that line positive or negative)?

$$\mathrm{Cor}(x,y) = \frac{\mathrm{Cov}(x,y)}{\mathrm{sd}(x)\mathrm{sd}(y)}$$

"standard deviation of y"

Single-number summaries: correlation

If you plot x and y, how closely are the points arranged on a line (and is the slope of that line positive or negative)?

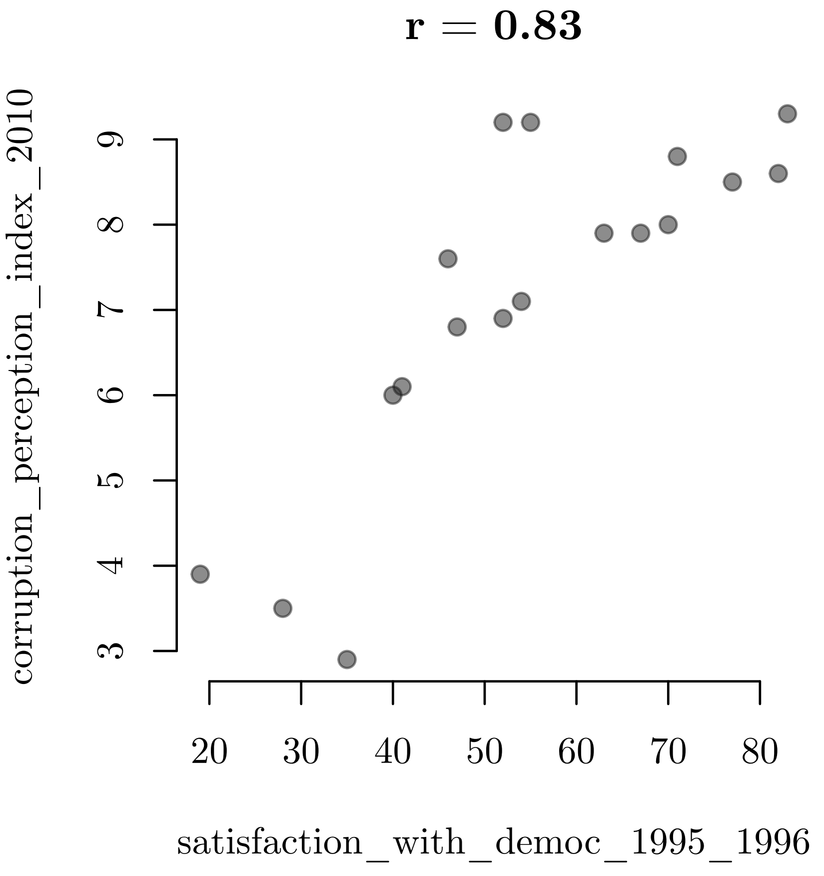
$$\mathrm{Cor}(x, y) = \frac{\mathrm{Cov}(x, y)}{\mathrm{sd}(x)\mathrm{sd}(y)}$$

"standard deviation of y"

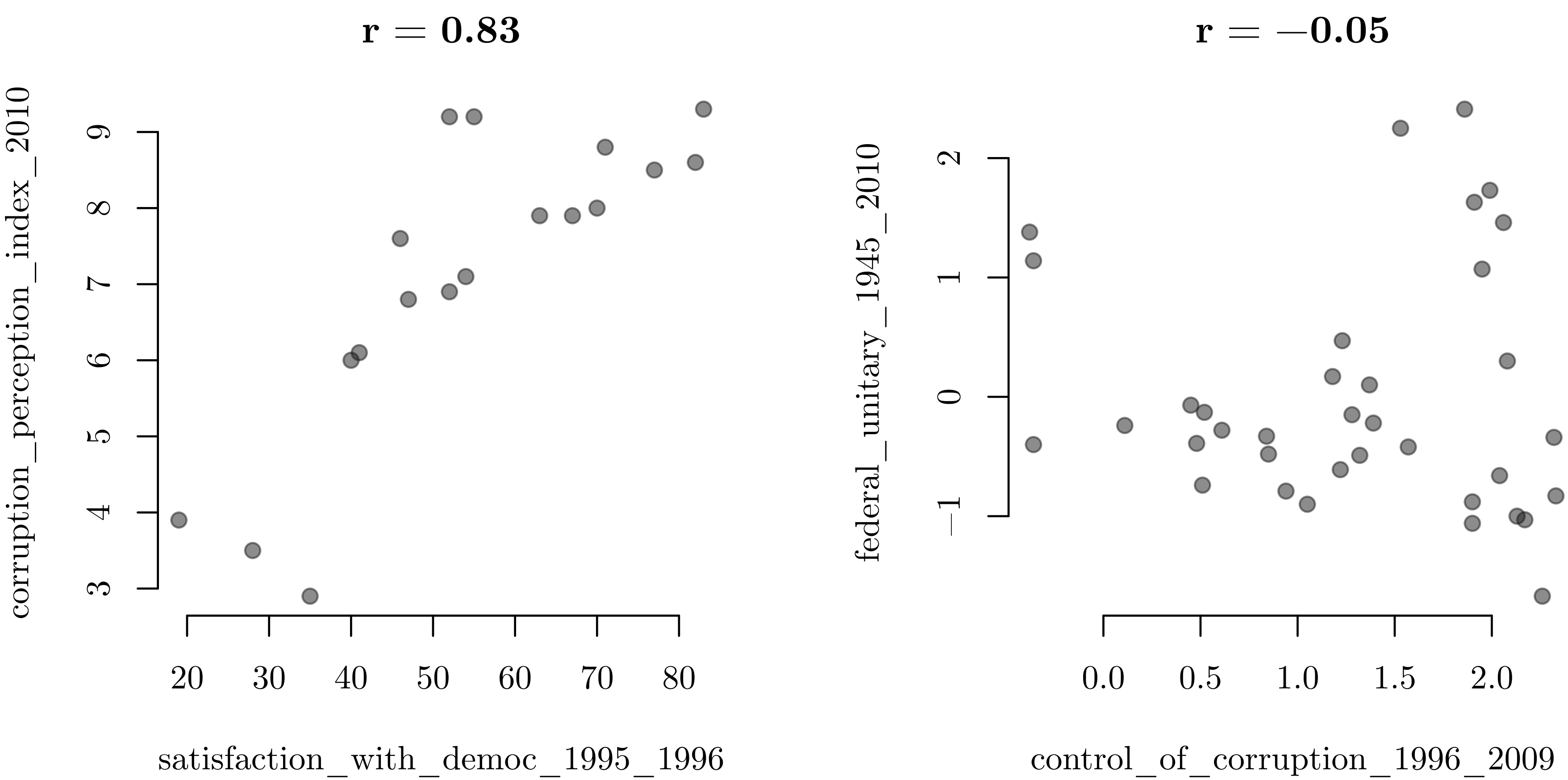
```
> cor(d$Percent_foreign_born, d$Percent_Leave, use = "complete")
[1] -0.6125353
```

Correlation examples

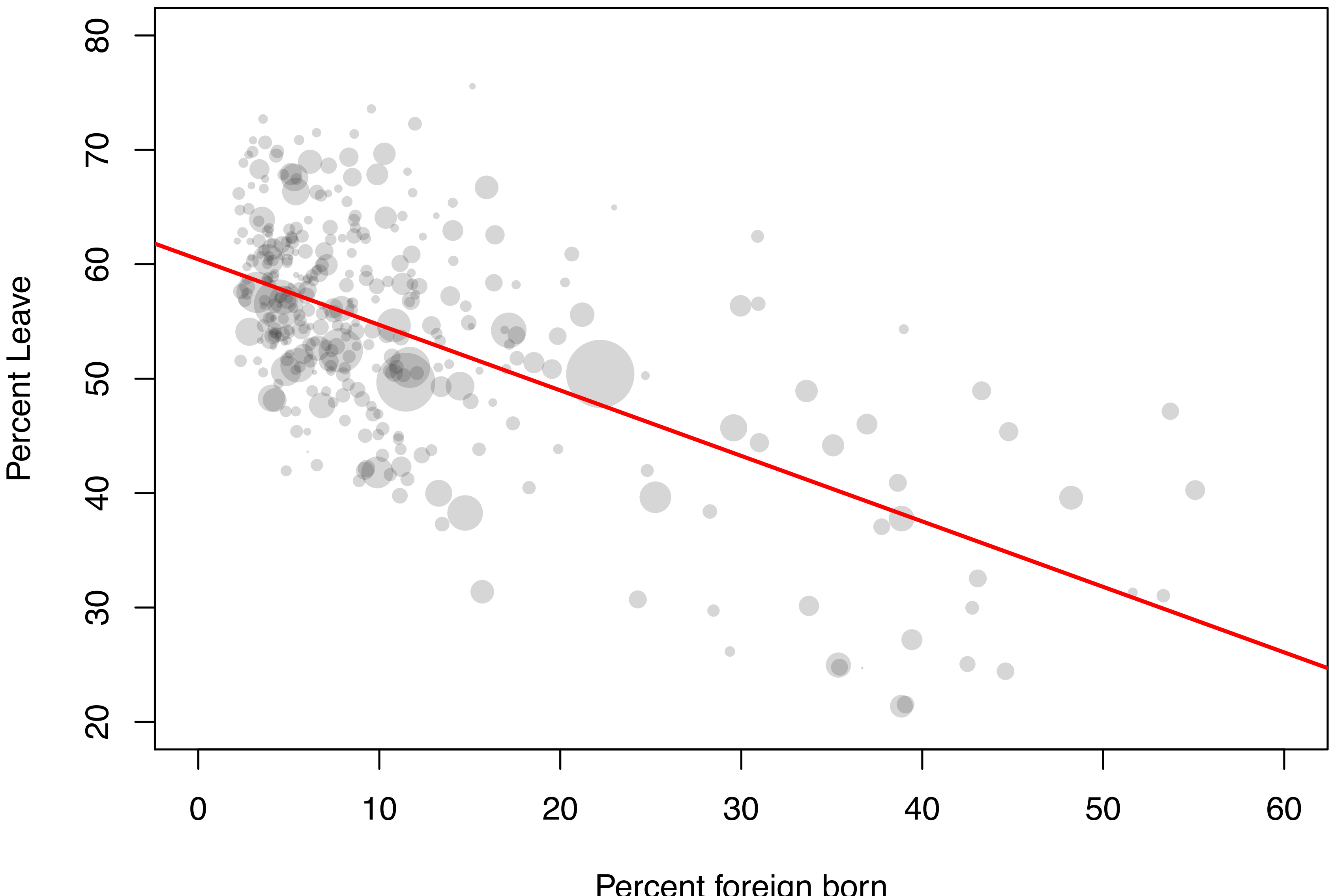
Correlation examples



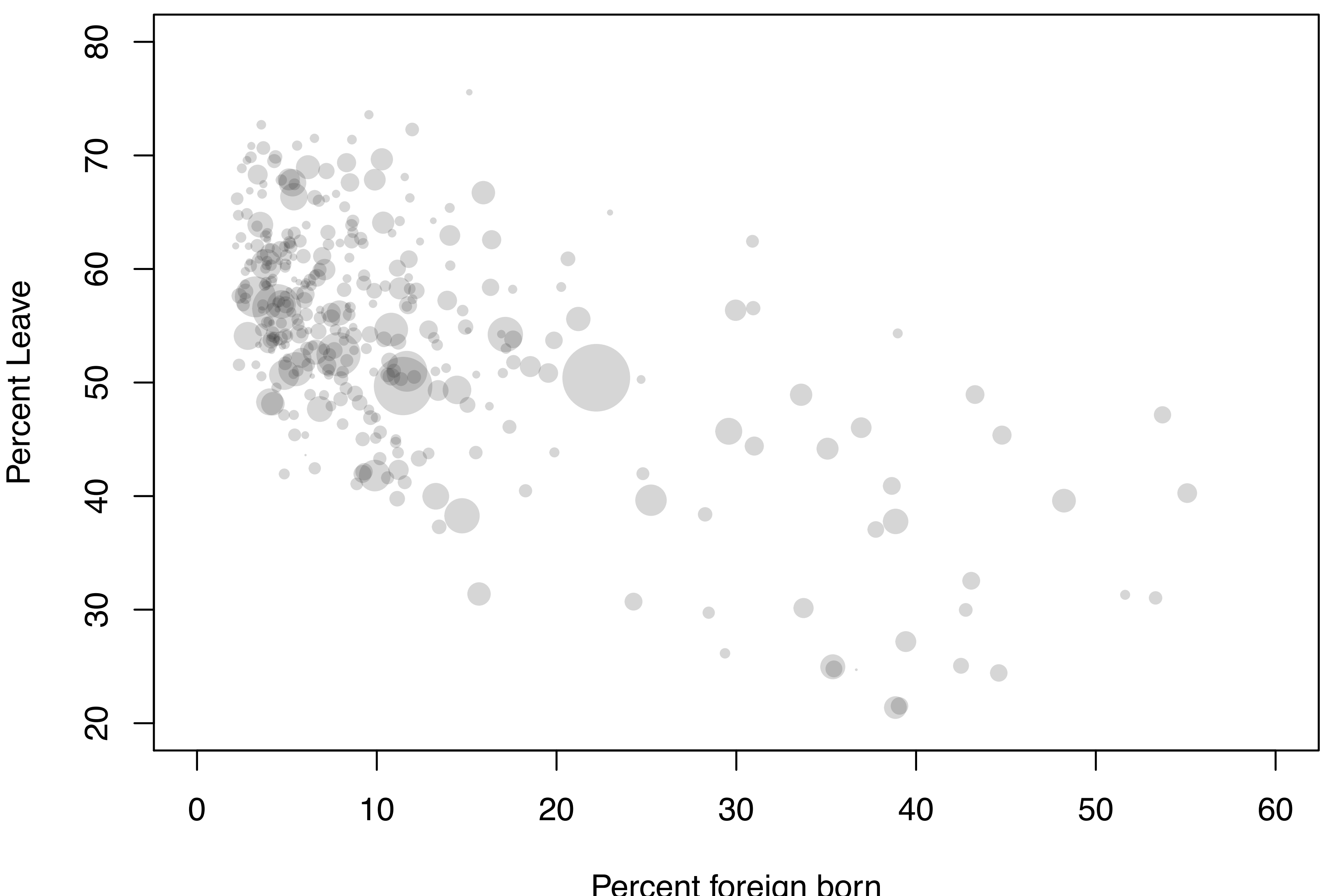
Correlation examples



The most important summary: OLS regression

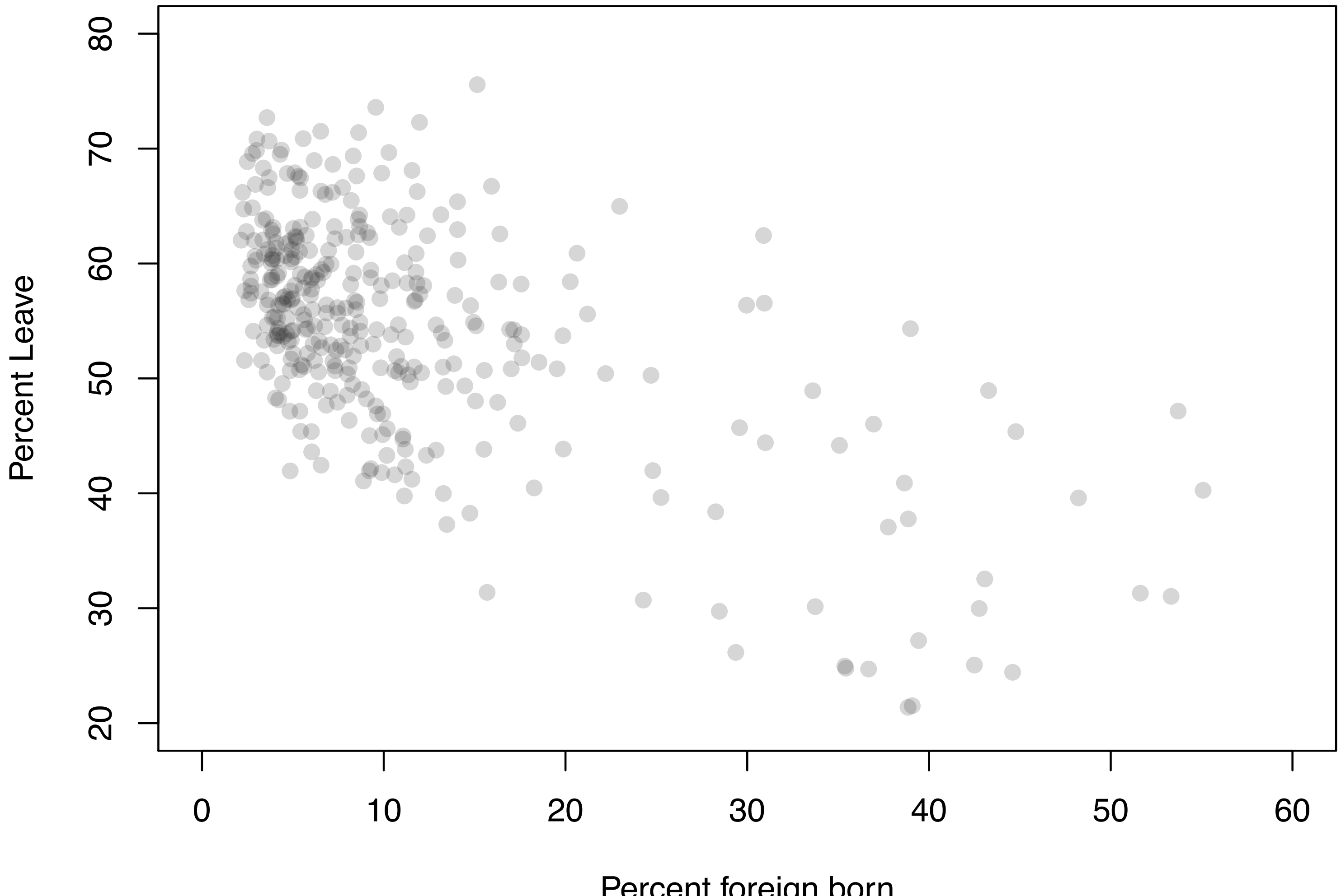


The most important summary: OLS regression



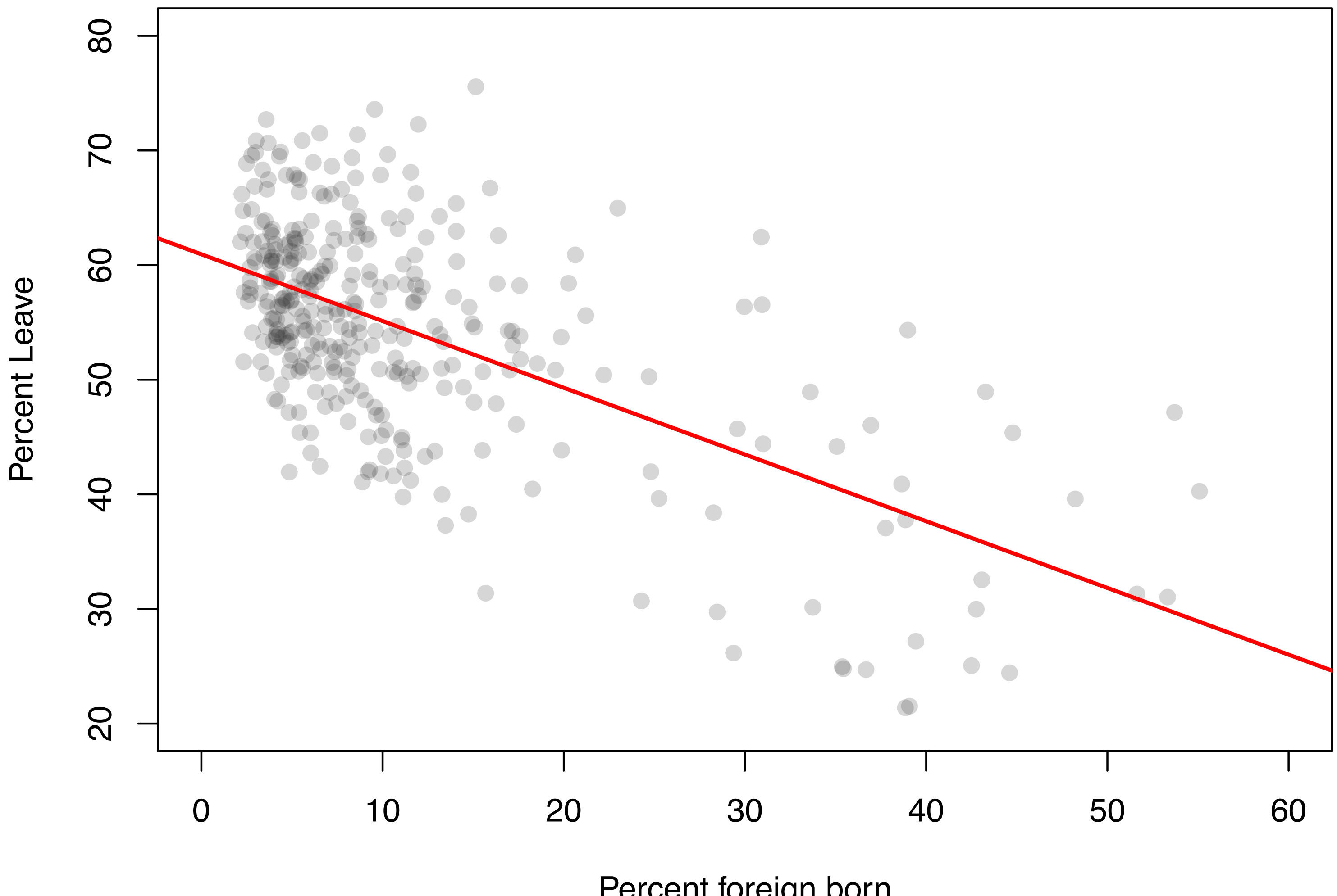
Step 1 for understanding OLS: residuals

A **residual** is the difference between the *actual* y-value and the *predicted* y-value (i.e. vertical diff. btw point and line).



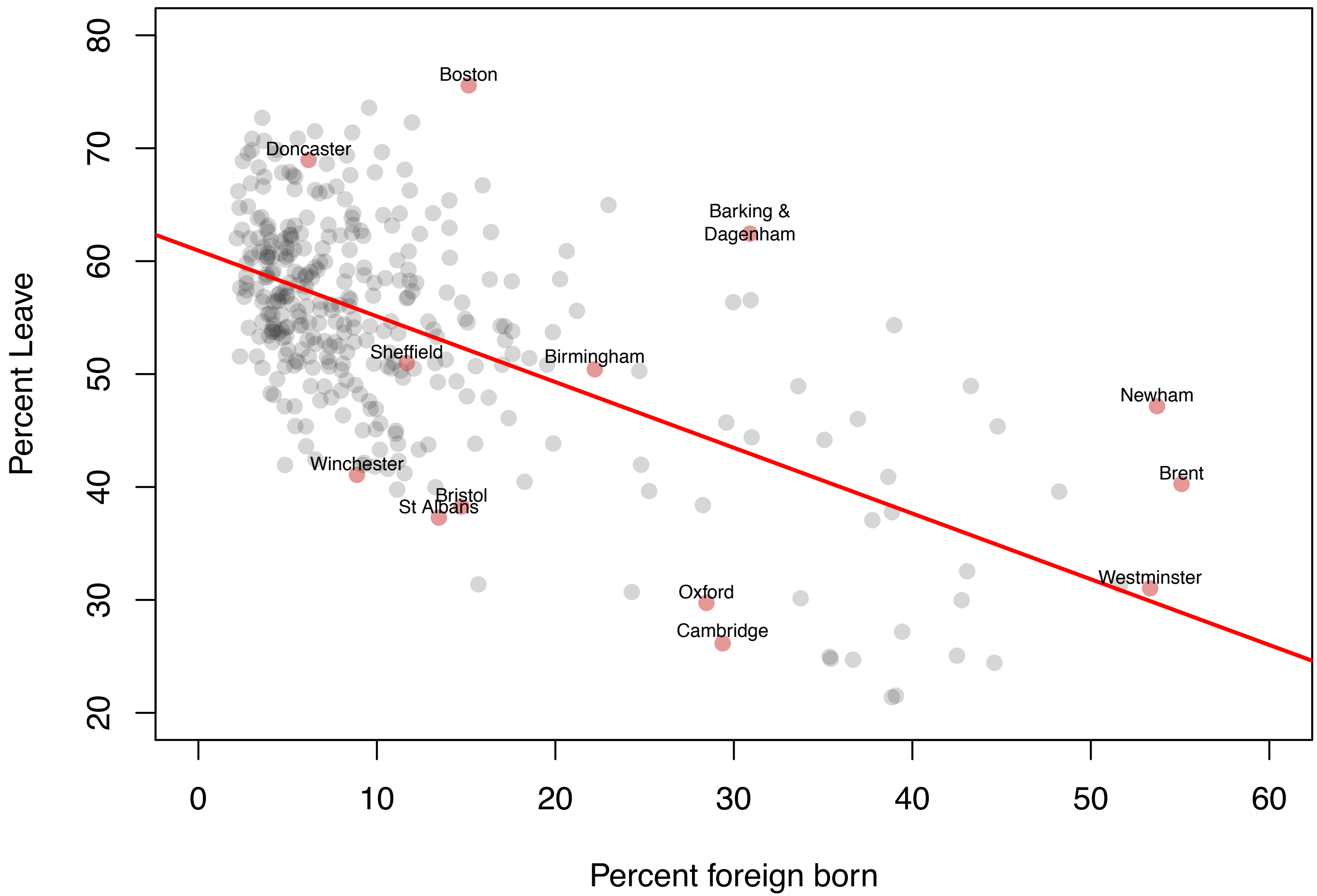
Step 1 for understanding OLS: residuals

A **residual** is the difference between the *actual* y-value and the *predicted* y-value (i.e. vertical diff. btw point and line).



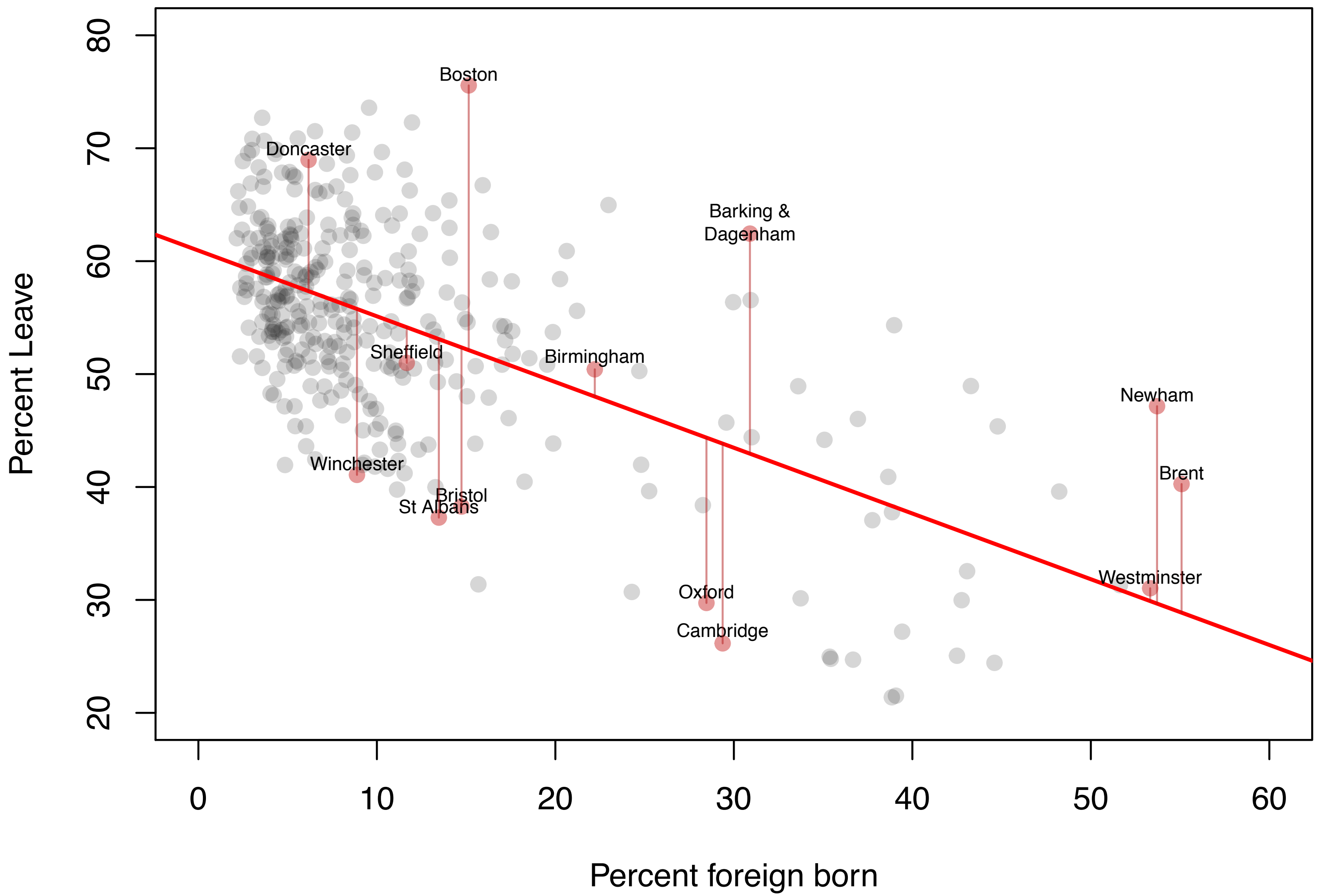
Step 1 for understanding OLS: residuals

A **residual** is the difference between the *actual* y-value and the *predicted* y-value (i.e. vertical diff. btw point and line).



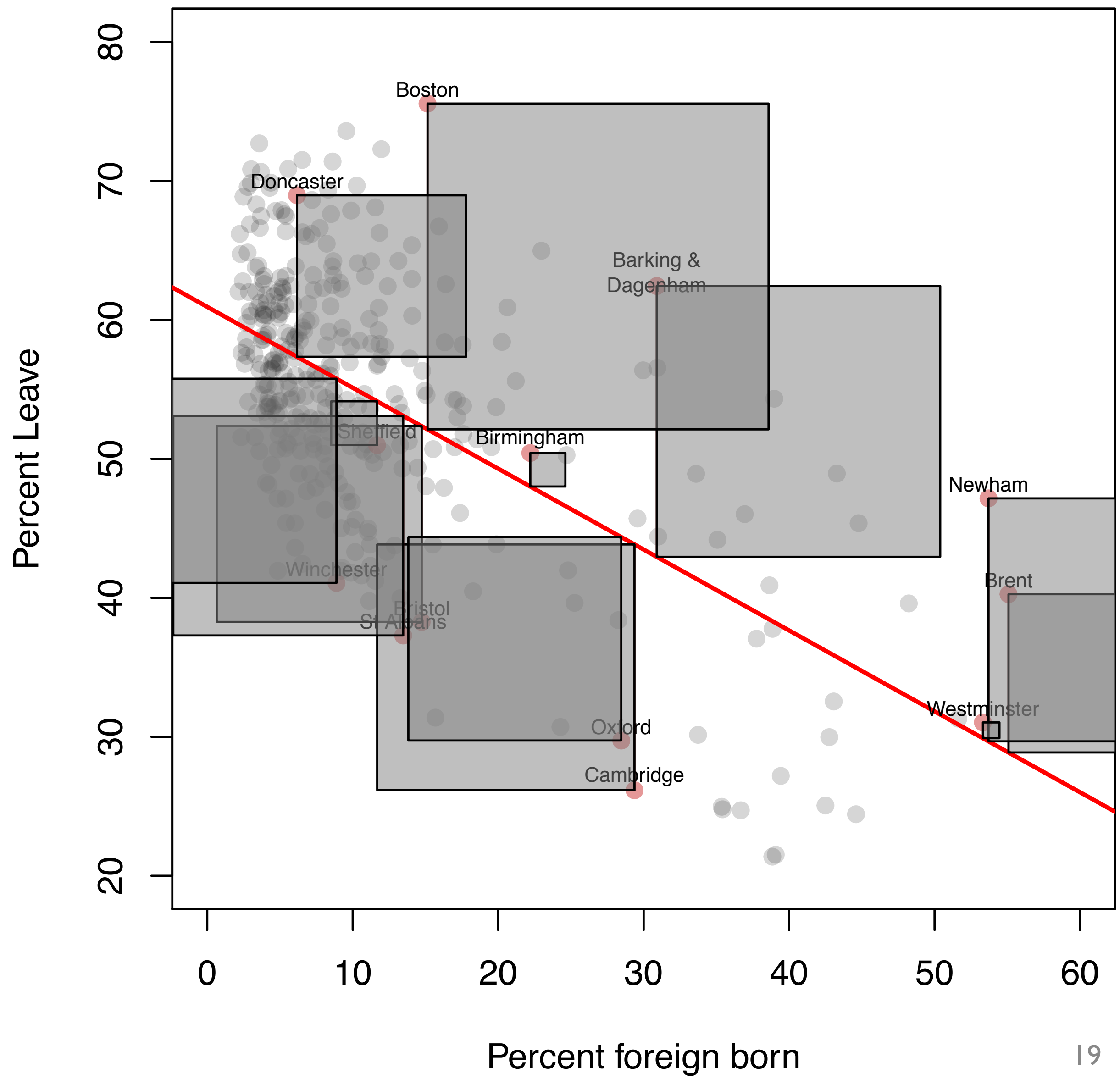
Step 1 for understanding OLS: residuals

A **residual** is the difference between the *actual* y-value and the *predicted* y-value (i.e. vertical diff. btw point and line).



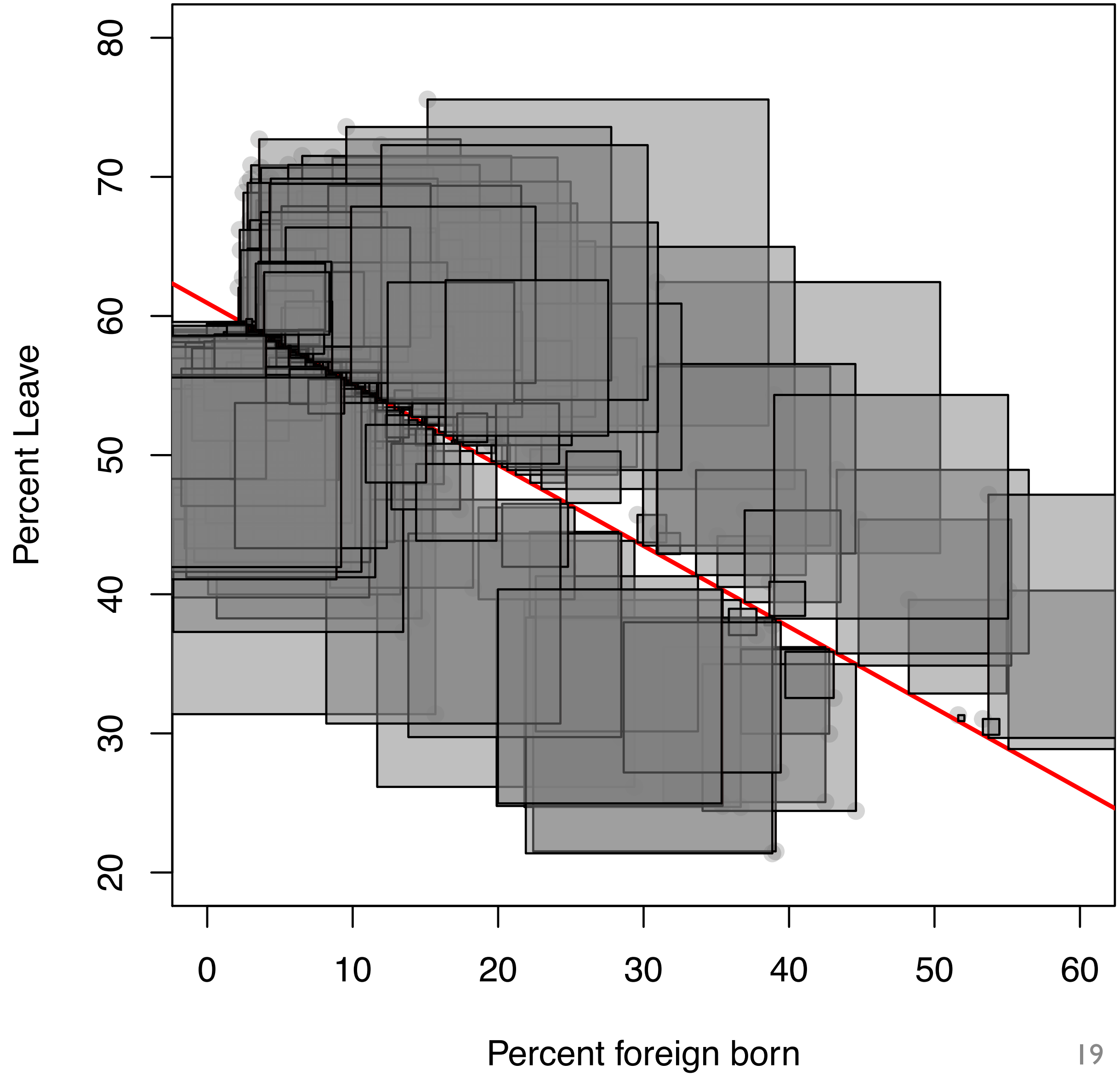
Step 2 for understanding OLS: sum of squared residuals

For any prediction line you draw, you can calculate residuals, square them, and sum them.



Step 2 for understanding OLS: sum of squared residuals

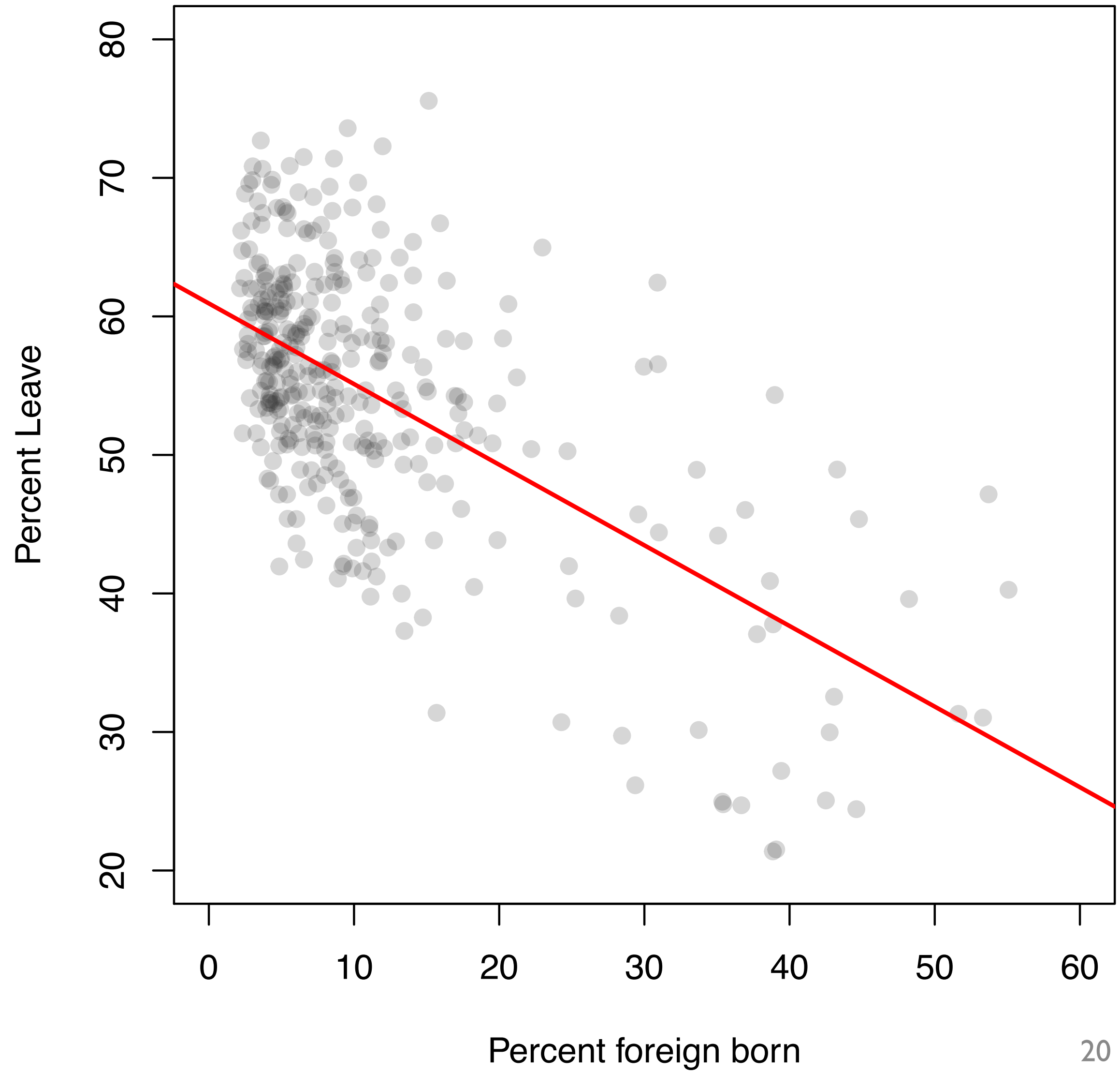
For any prediction line you draw, you can calculate residuals, square them, and sum them.



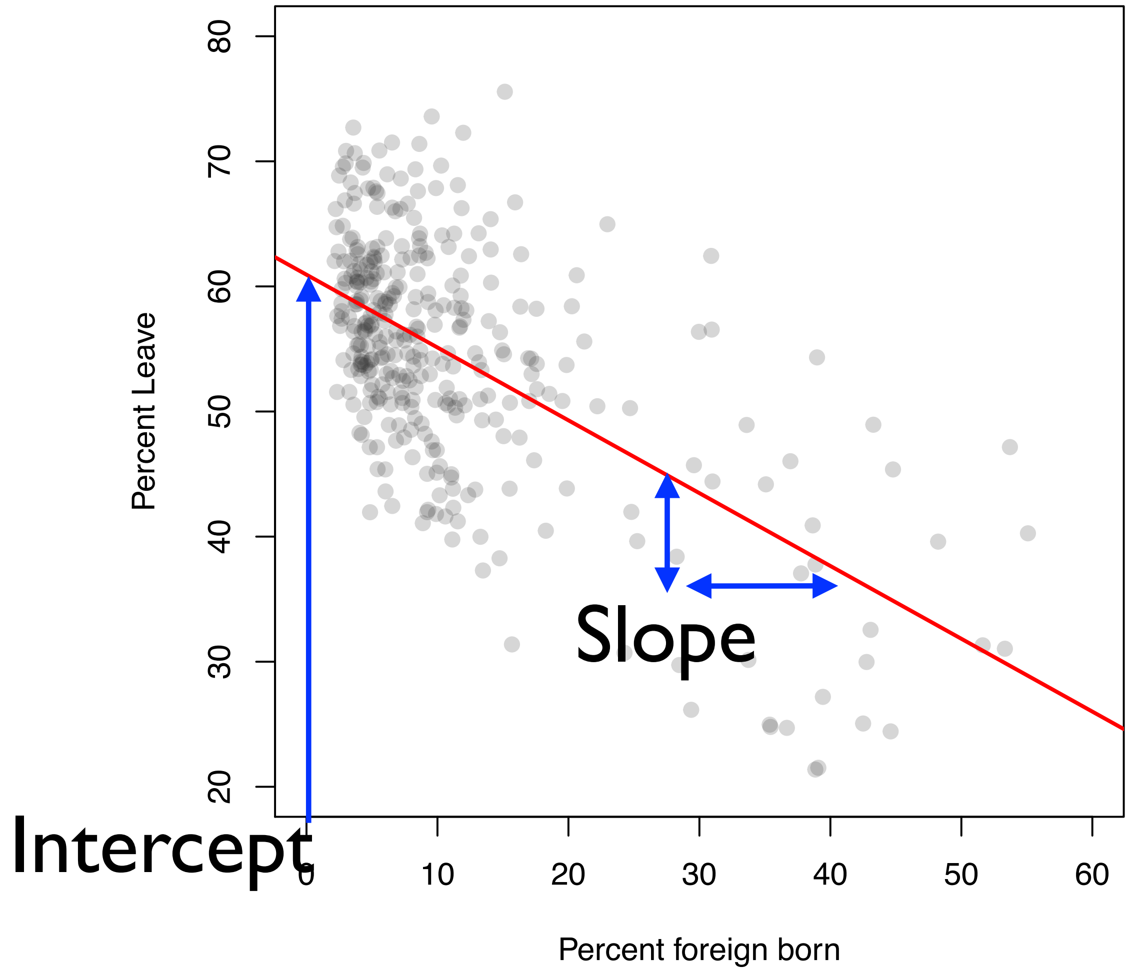
Step 3 for understanding OLS: minimizing the sum of squared residuals

The OLS regression line minimizes the sum of squared residuals (SSR).

Hence ordinary **least squares.**



Any line can be summarized by an **intercept** and a **slope**.



The intercept and slope of a bivariate regression are called the regression **coefficients.**

Implementing OLS

Implementing OLS

Some options:

Implementing OLS

Some options:

1. Use R to try every combination of slope and intercept; choose the combination that has the lowest sum of squared residuals.

Implementing OLS

Some options:

1. Use R to try every combination of slope and intercept; choose the combination that has the lowest sum of squared residuals.
2. Use calculus to find the slope and intercept that minimize the sum of squared residuals.

Implementing OLS

Some options:

1. Use R to try every combination of slope and intercept; choose the combination that has the lowest sum of squared residuals.
2. Use calculus to find the slope and intercept that minimize the sum of squared residuals.
3. Use lm() function in R:

Implementing OLS

Some options:

1. Use R to try every combination of slope and intercept; choose the combination that has the lowest sum of squared residuals.
2. Use calculus to find the slope and intercept that minimize the sum of squared residuals.
3. Use lm() function in R:

```
> lm(d$Percent_Leave ~ d$Percent_foreign_born)

Call:
lm(formula = d$Percent_Leave ~ d$Percent_foreign_born)

Coefficients:
           (Intercept)  d$Percent_foreign_born
               60.9373                 -0.5821
```

A (surprising?) fact about the slope coefficient

A (surprising?) fact about the slope coefficient

Covariance of x and y:

$$\mathrm{Cov}(x, y) = \frac{\Sigma_i (x_i - \overline{x})(y_i - \overline{y})}{n - 1}$$

A (surprising?) fact about the slope coefficient

Covariance of x and y:

$$\mathrm{Cov}(x, y) = \frac{\Sigma_i (x_i - \overline{x})(y_i - \overline{y})}{n-1}$$

Variance of x:

$$\mathrm{Var}(x) = \frac{\Sigma_i (x_i - \overline{x})^2}{n-1}$$

A (surprising?) fact about the slope coefficient

Covariance of x and y:

$$\text{Cov}(x,y) = \frac{\Sigma_i (x_i - \overline{x})(y_i - \overline{y})}{n-1}$$

Variance of x:

$$\text{Var}(x) = \frac{\Sigma_i (x_i - \overline{x})^2}{n-1}$$

Slope from OLS regression of y on x:

$$\hat{\beta} = \frac{\text{Cov}(x,y)}{\text{Var}(x)}$$

A (surprising?) fact about the slope coefficient

Covariance of x and y:

$$\mathrm{Cov}(x,y) = \frac{\Sigma_i (x_i - \overline{x})(y_i - \overline{y})}{n-1}$$

Variance of x:

$$\mathrm{Var}(x) = \frac{\Sigma_i (x_i - \overline{x})^2}{n-1}$$

Slope from OLS regression of y on x:

$$\hat{\beta} = \frac{\mathrm{Cov}(x,y)}{\mathrm{Var}(x)}$$

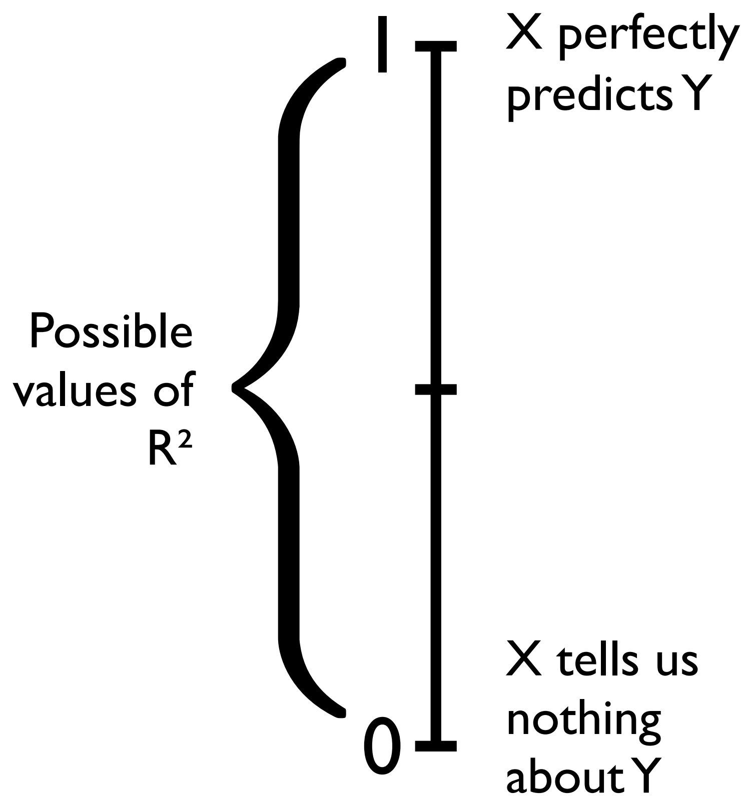
```
> cov(d$Percent_Leave, d$Percent_foreign_born, use = "complete")/var(d
$Percent_foreign_born, na.rm = T)
[1] -0.582101
> lm(d$Percent_Leave ~ d$Percent_foreign_born)

Call:
lm(formula = d$Percent_Leave ~ d$Percent_foreign_born)

Coefficients:
           (Intercept)  d$Percent_foreign_born
               60.9373                 -0.5821
```

How well does our regression line predict the outcome? R^2

```
> summary(lm(d$Percent_Leave ~ d$Percent_foreign_born))

Call:
lm(formula = d$Percent_Leave ~ d$Percent_foreign_born)

Residuals:
     Min       1Q   Median       3Q      Max 
-20.4253  -4.7247  -0.0025   4.4336  23.4417 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)            60.93732    0.61845   98.53   <2e-16 ***
d$Percent_foreign_born -0.58210    0.04062  -14.33   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.775 on 342 degrees of freedom
  (38 observations deleted due to missingness)
Multiple R-squared:  0.3752,Adjusted R-squared:  0.3734 
F-statistic: 205.4 on 1 and 342 DF,  p-value: < 2.2e-16
```

R^2: intuition

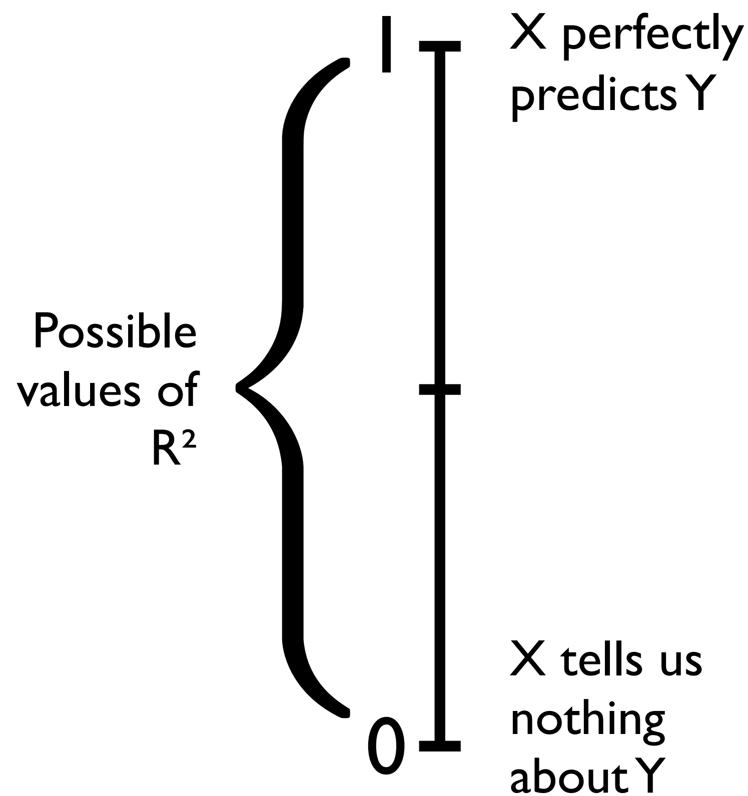
How much better are the predictions from our OLS regression line than the predictions from a flat line (i.e. not using X at all)?



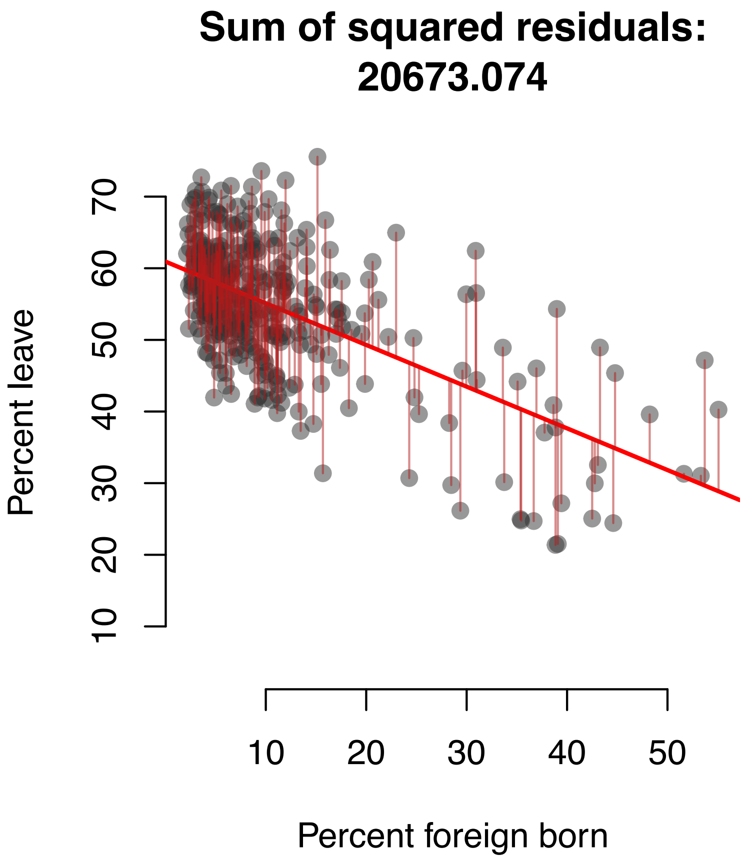
R^2: intuition

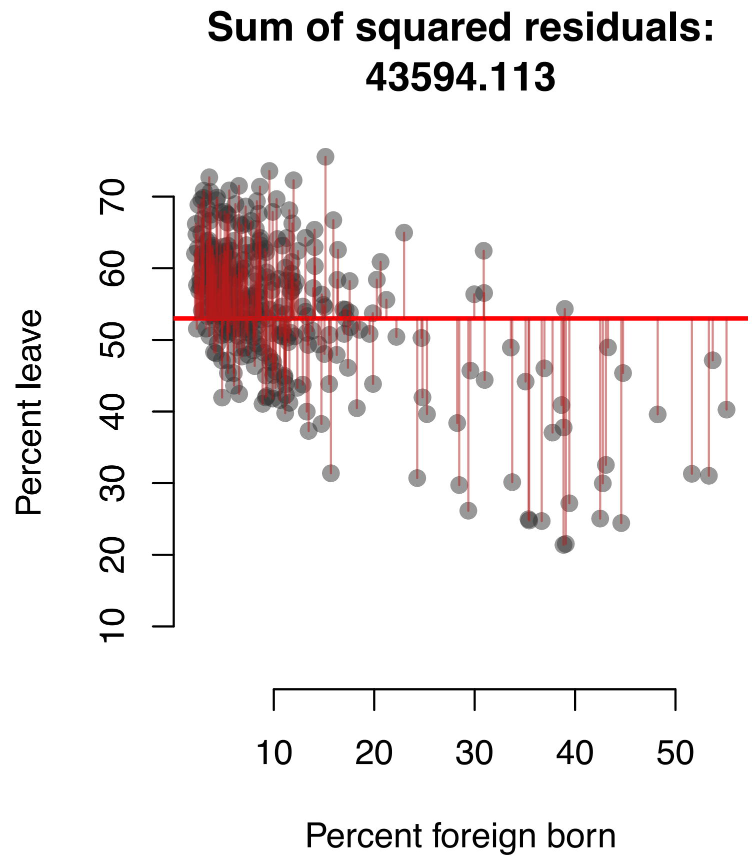
How much better are the predictions from our OLS regression line than the predictions from a flat line (i.e. not using X at all)?

How much of the variation in Y is "explained" by the variation in X?



R^2: calculation






```
> 1 - (20673.074/33087.482)
[1] 0.3751995
```

Connections between measures of bivariate relationships

Key measures:

- covariance
- correlation
- OLS regression output:
  - intercept
  - slope
  - R^2

Connections between measures of bivariate relationships

Key measures:

- covariance
- correlation
- OLS regression output:
  - intercept
  - slope
  - R^2

For any two variables, covariance, correlation, and regression slope will all have the same sign.

Connections between measures of bivariate relationships

Key measures:

- covariance
- correlation
- OLS regression output:
  - intercept
  - slope
  - R^2

For any two variables, covariance, correlation, and regression slope will all have the same sign.

For bivariate relationships, R^2 = correlation^2

Connections between measures of bivariate relationships

Key measures:

- covariance
- correlation
- OLS regression output:
  - intercept
  - slope
  - R^2

For any two variables, covariance, correlation, and regression slope will all have the same sign.

For bivariate relationships, R^2 = correlation^2

Covariance and regression slope (but not correlation) depend on the units

Connections between measures of bivariate relationships

Key measures:

- covariance
- correlation
- OLS regression output:
  - intercept
  - slope
  - R^2

For any two variables, covariance, correlation, and regression slope will all have the same sign.

For bivariate relationships, R^2 = correlation^2

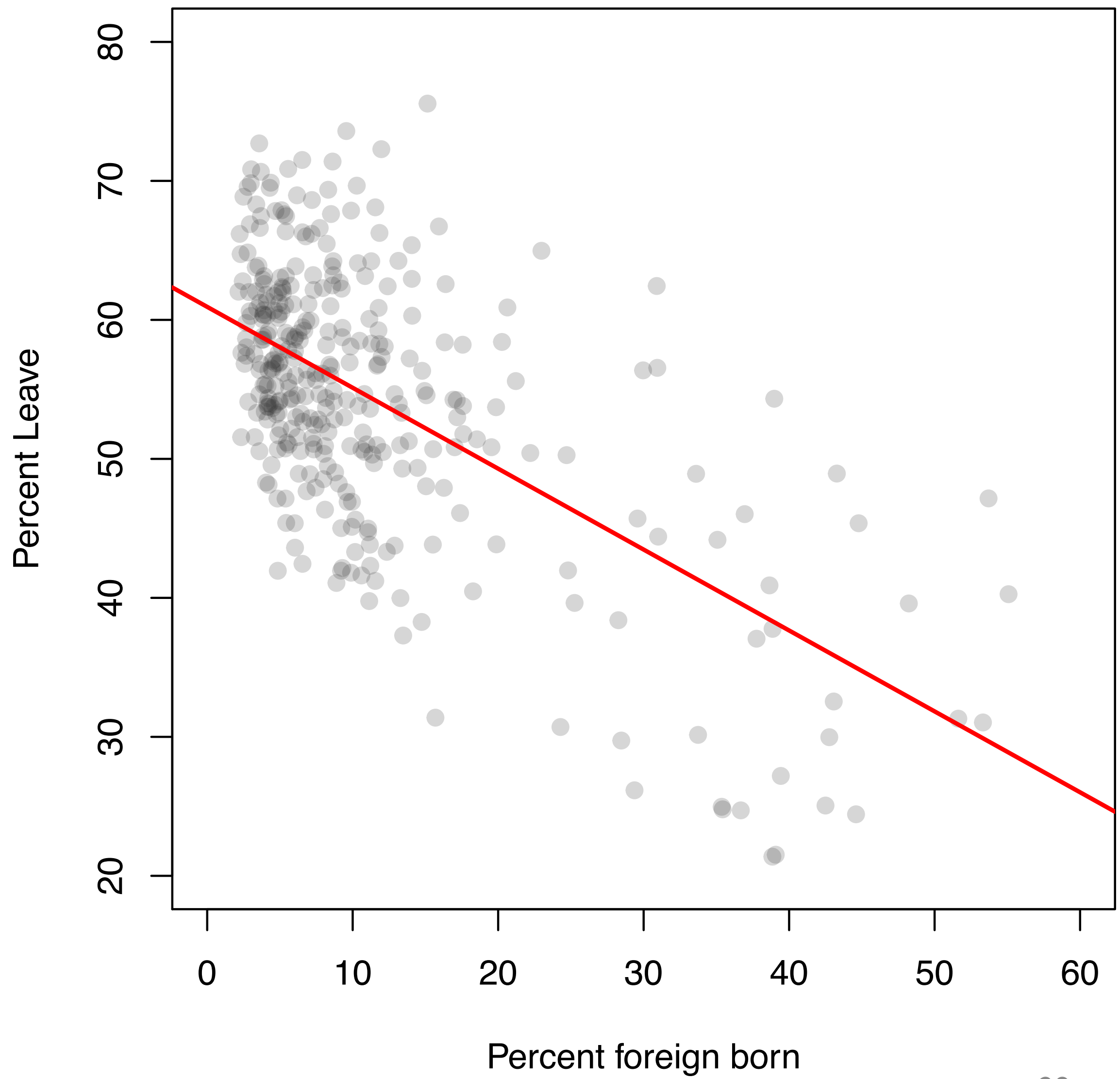
Regression slope (but not covariance or correlation) depends on which is Y and which is X

Covariance and regression slope (but not correlation) depend on the units

To discuss

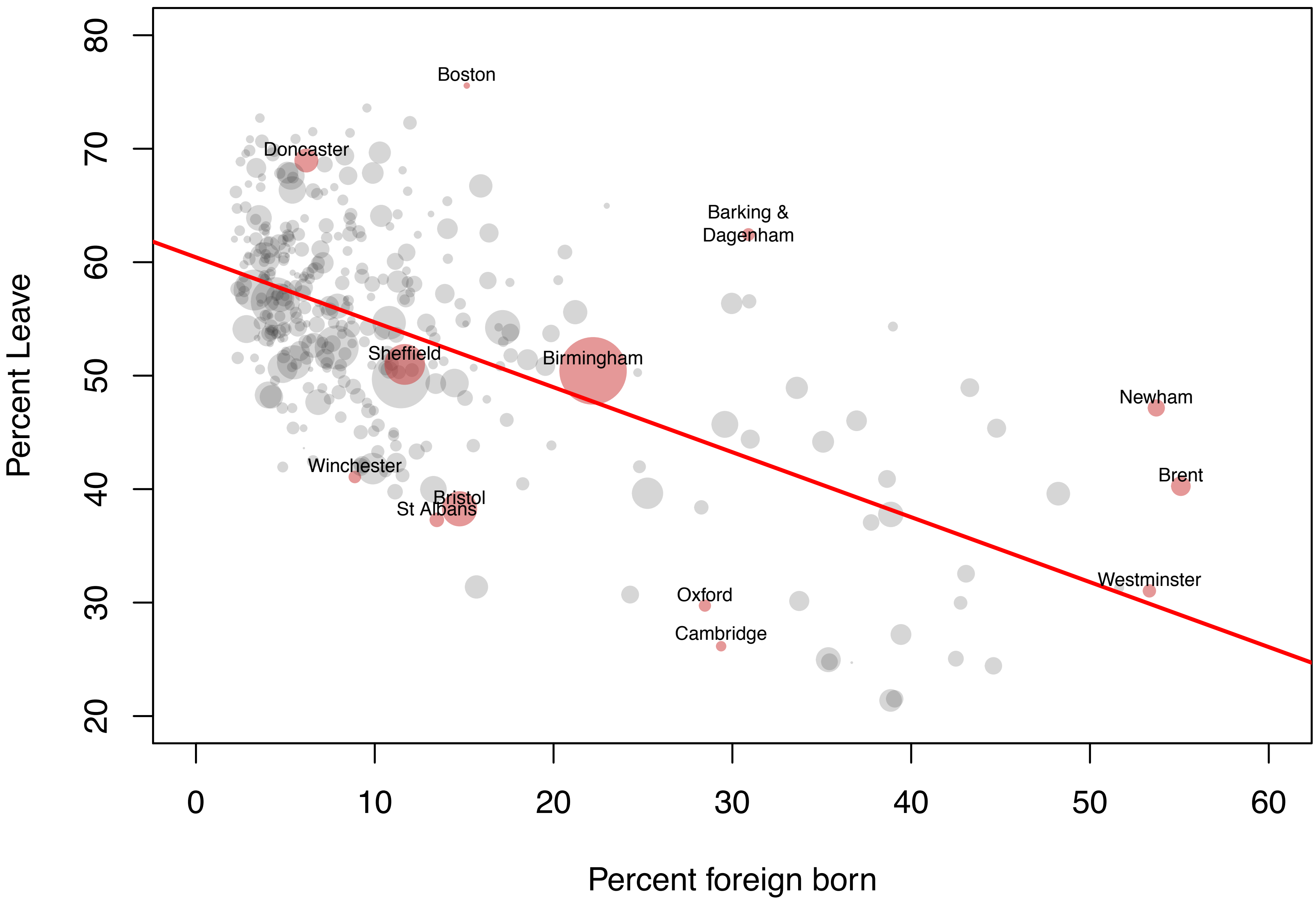
What other options could you imagine for deciding on a predictive line?

What are the advantages of OLS?



Summarizing multivariate relationships: motivation and one non-OLS solution

Did this pattern arise because contact with immigrants makes people less opposed to immigration?



Confounders

Confounders

One reason why two phenomena can be correlated is the presence of a **confounder**.

Confounders

One reason why two phenomena can be correlated is the presence of a **confounder**.

Ice cream
consumption

Confounders

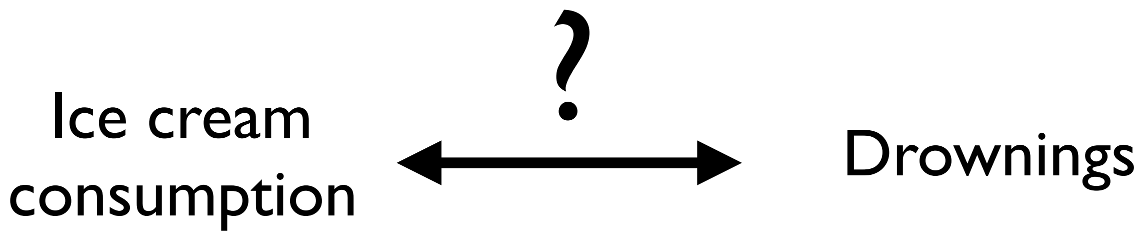
One reason why two phenomena can be correlated is the presence of a **confounder**.

Ice cream consumption

Drownings

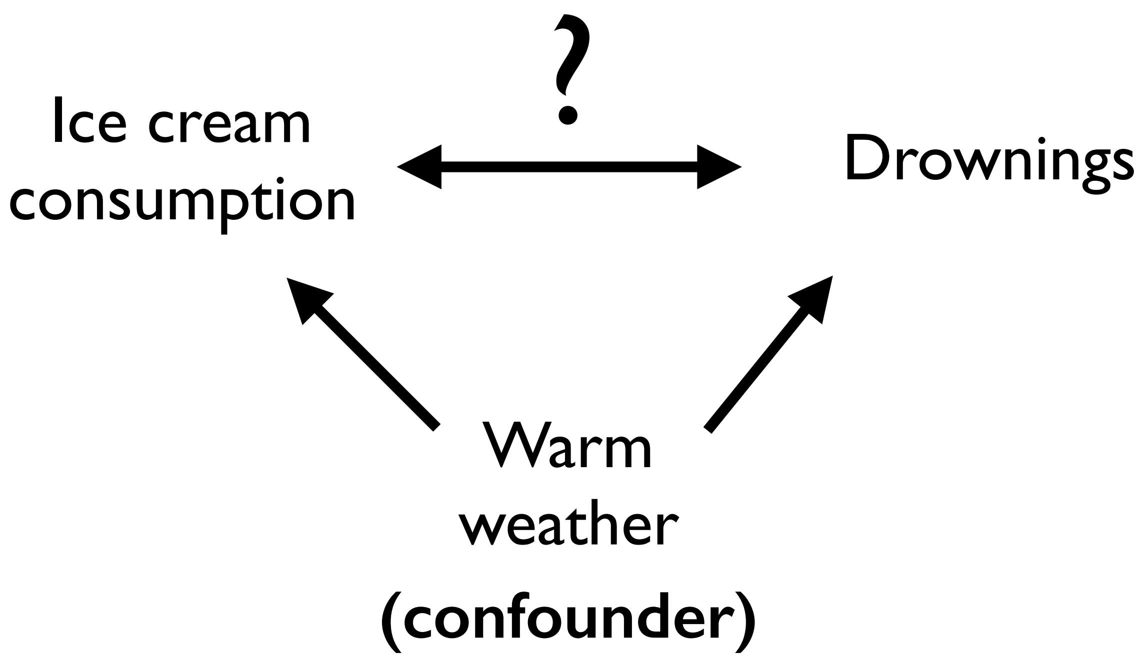
Confounders

One reason why two phenomena can be correlated is the presence of a **confounder**.



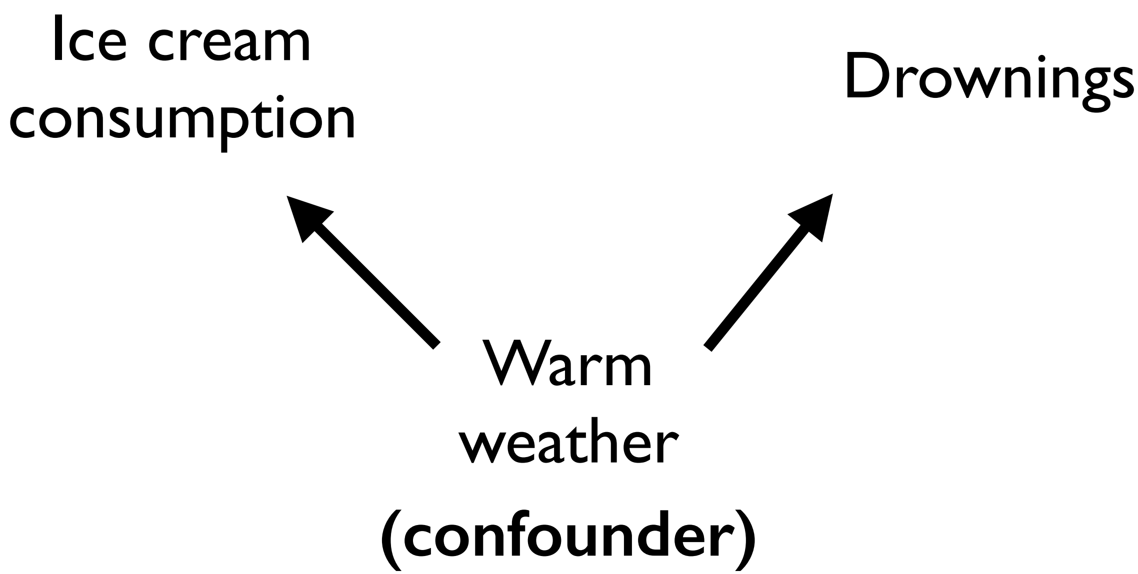
Confounders

One reason why two phenomena can be correlated is the presence of a **confounder**.



Confounders

One reason why two phenomena can be correlated is the presence of a **confounder**.



Confounders (2)

Confounders (2)

What are possible confounders in the relationship between percent of foreign-born residents and support for Brexit?

Confounders (2)

What are possible confounders in the relationship between percent of foreign-born residents and support for Brexit?

More foreign-born residents

Confounders (2)

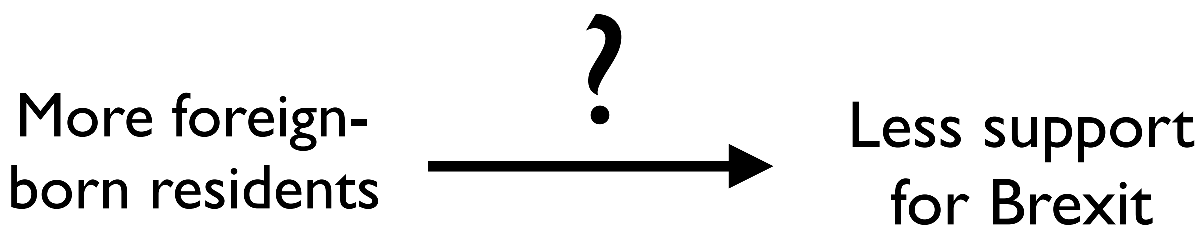
What are possible confounders in the relationship between percent of foreign-born residents and support for Brexit?

More foreign-born residents

Less support for Brexit

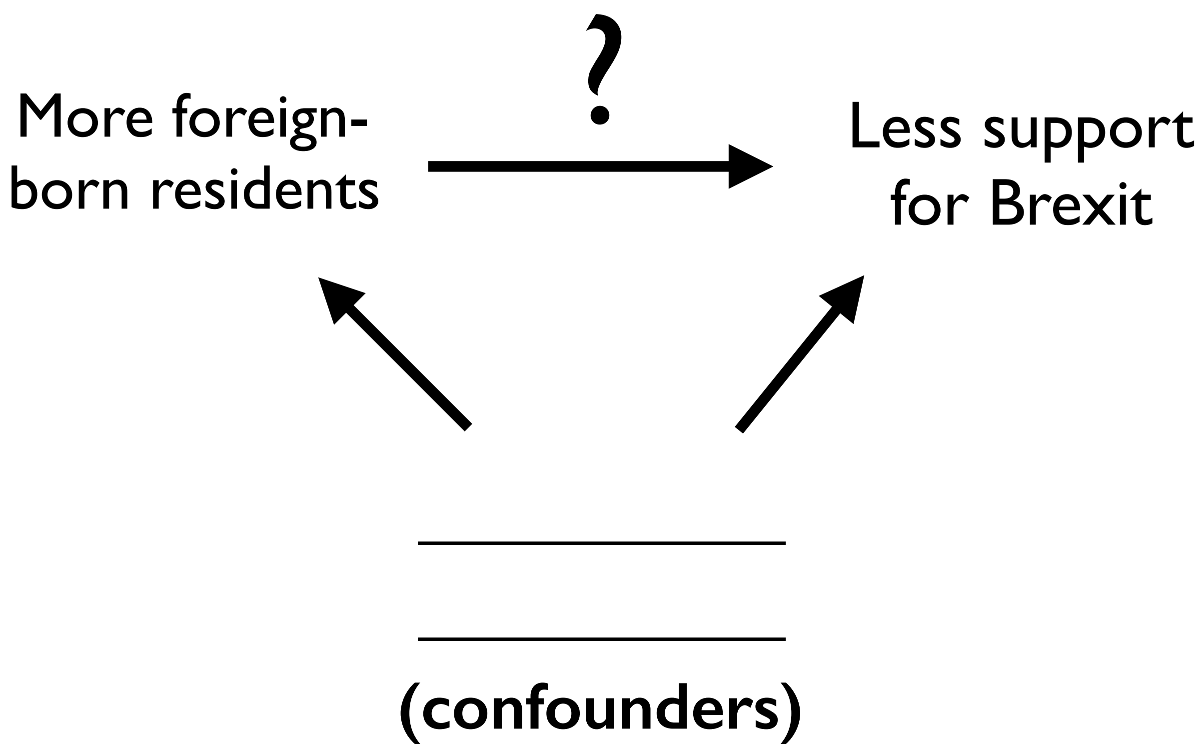
Confounders (2)

What are possible confounders in the relationship between percent of foreign-born residents and support for Brexit?



Confounders (2)

What are possible confounders in the relationship between percent of foreign-born residents and support for Brexit?



Confounders (3)

Confounders (3)

What are possible confounders in the relationship between exercise in your 40s and health in your 60s?

Confounders (3)

What are possible confounders in the relationship between exercise in your 40s and health in your 60s?

Exercising in
your 40s

Confounders (3)

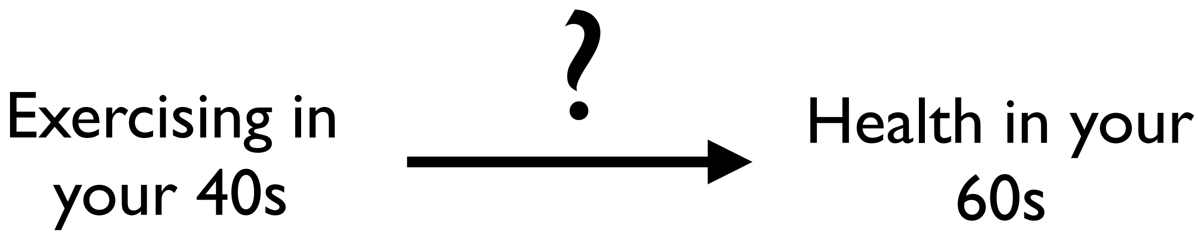
What are possible confounders in the relationship between exercise in your 40s and health in your 60s?

Exercising in your 40s

Health in your 60s

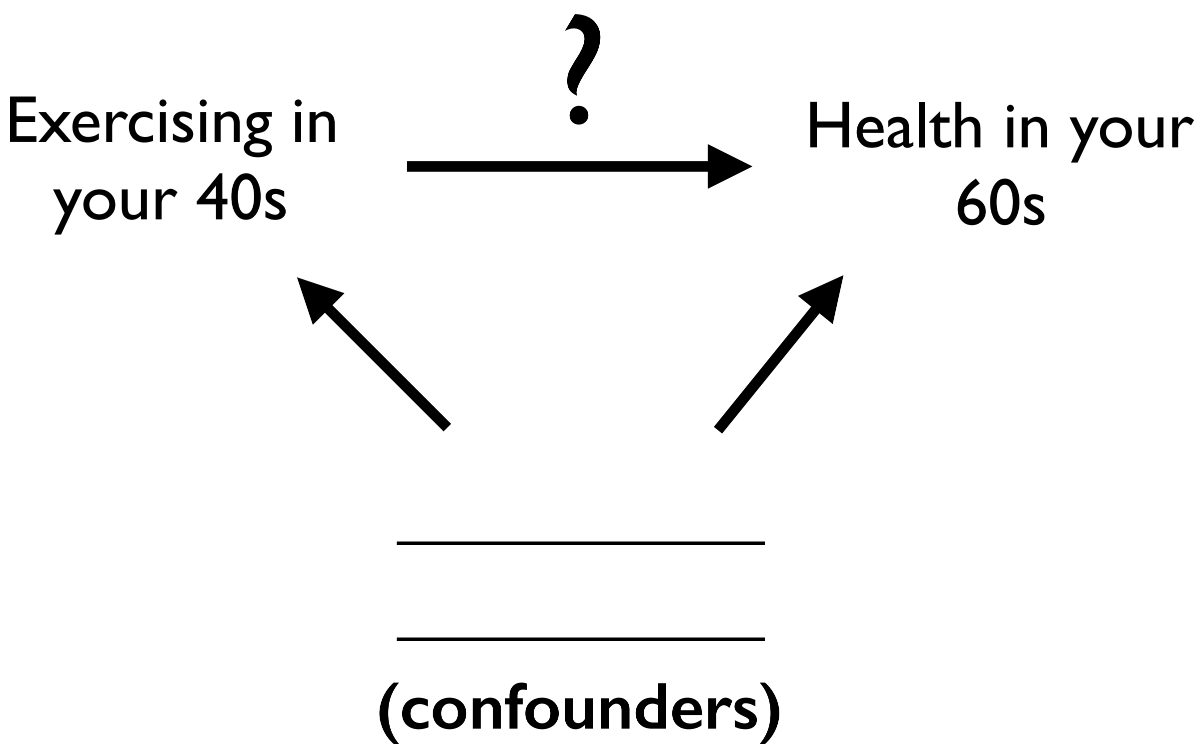
Confounders (3)

What are possible confounders in the relationship between exercise in your 40s and health in your 60s?



Confounders (3)

What are possible confounders in the relationship between exercise in your 40s and health in your 60s?



Controlling for confounders

Controlling for confounders

In many cases we want to measure the relationship between two phenomena **controlling for** (i.e. *holding constant*) one or more confounders.

Controlling for confounders

In many cases we want to measure the relationship between two phenomena **controlling for** (i.e. *holding constant*) one or more confounders.

- Are people who exercise less likely to develop dementia, controlling for diet and age?

Controlling for confounders

In many cases we want to measure the relationship between two phenomena **controlling for** (i.e. *holding constant*) one or more confounders.

- Are people who exercise less likely to develop dementia, controlling for diet and age?
- Are countries with more inclusive political systems less likely to experience violence, controlling for economic development and the number of ethnic groups?

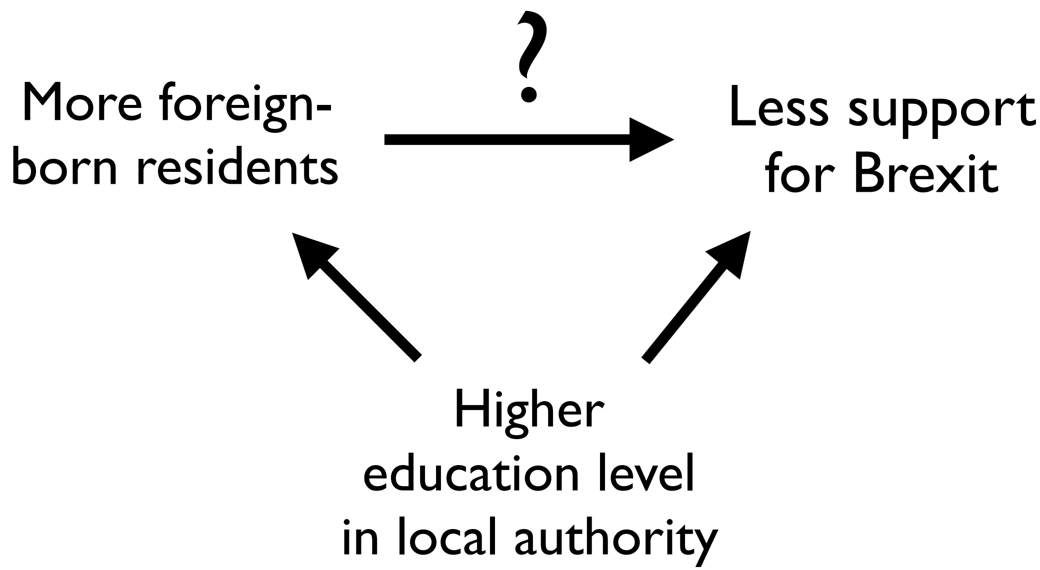
Controlling for confounders

In many cases we want to measure the relationship between two phenomena **controlling for** (i.e. *holding constant*) one or more confounders.

- Are people who exercise less likely to develop dementia, controlling for diet and age?
- Are countries with more inclusive political systems less likely to experience violence, controlling for economic development and the number of ethnic groups?
- Are local authorities with more foreign-born residents less likely to support Brexit, controlling for __________?

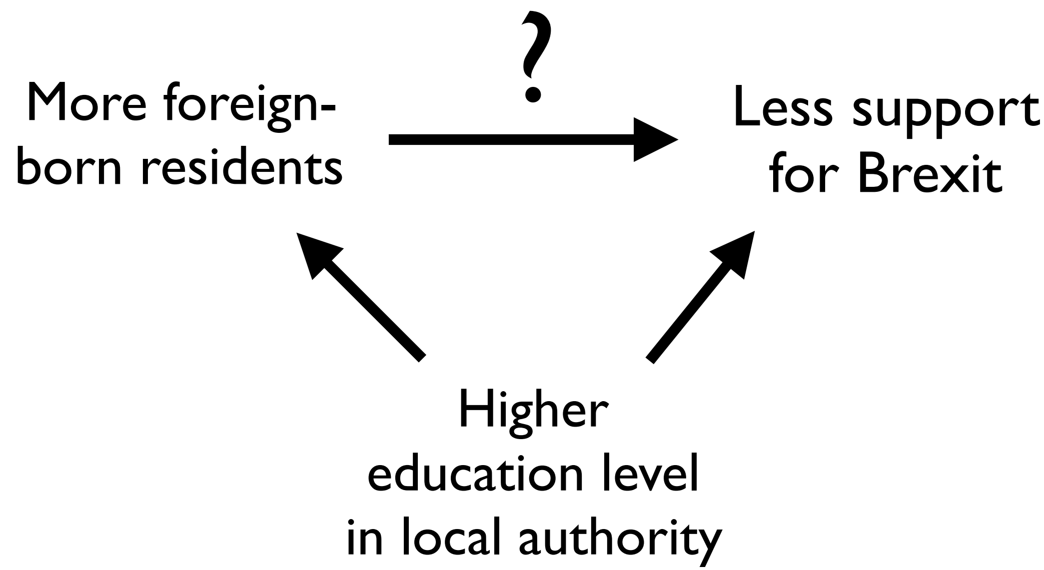
How do we control for confounders?

Let’s focus on education as a confounder in our Brexit example:



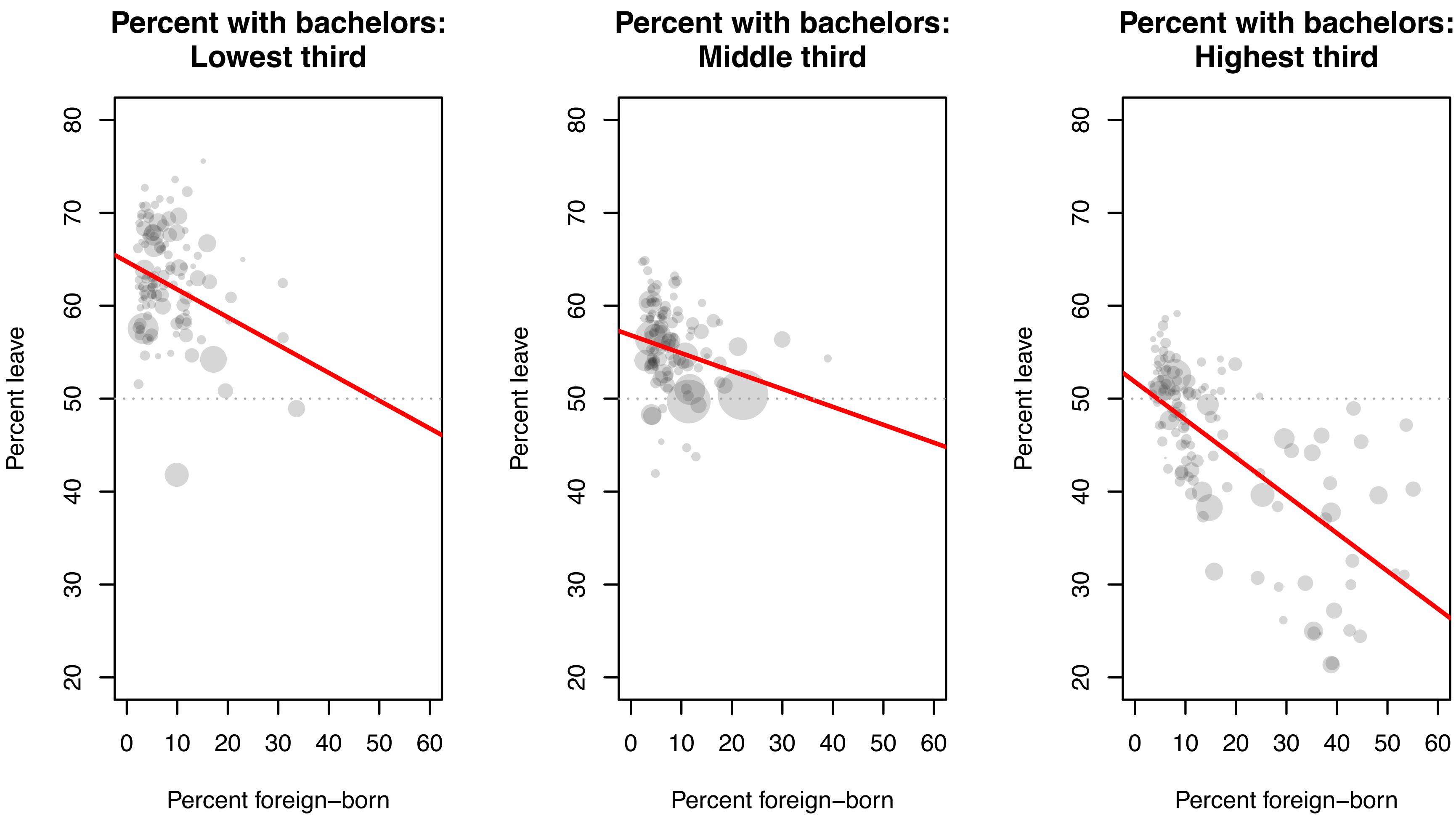
How do we control for confounders?

Let's focus on education as a confounder in our Brexit example:

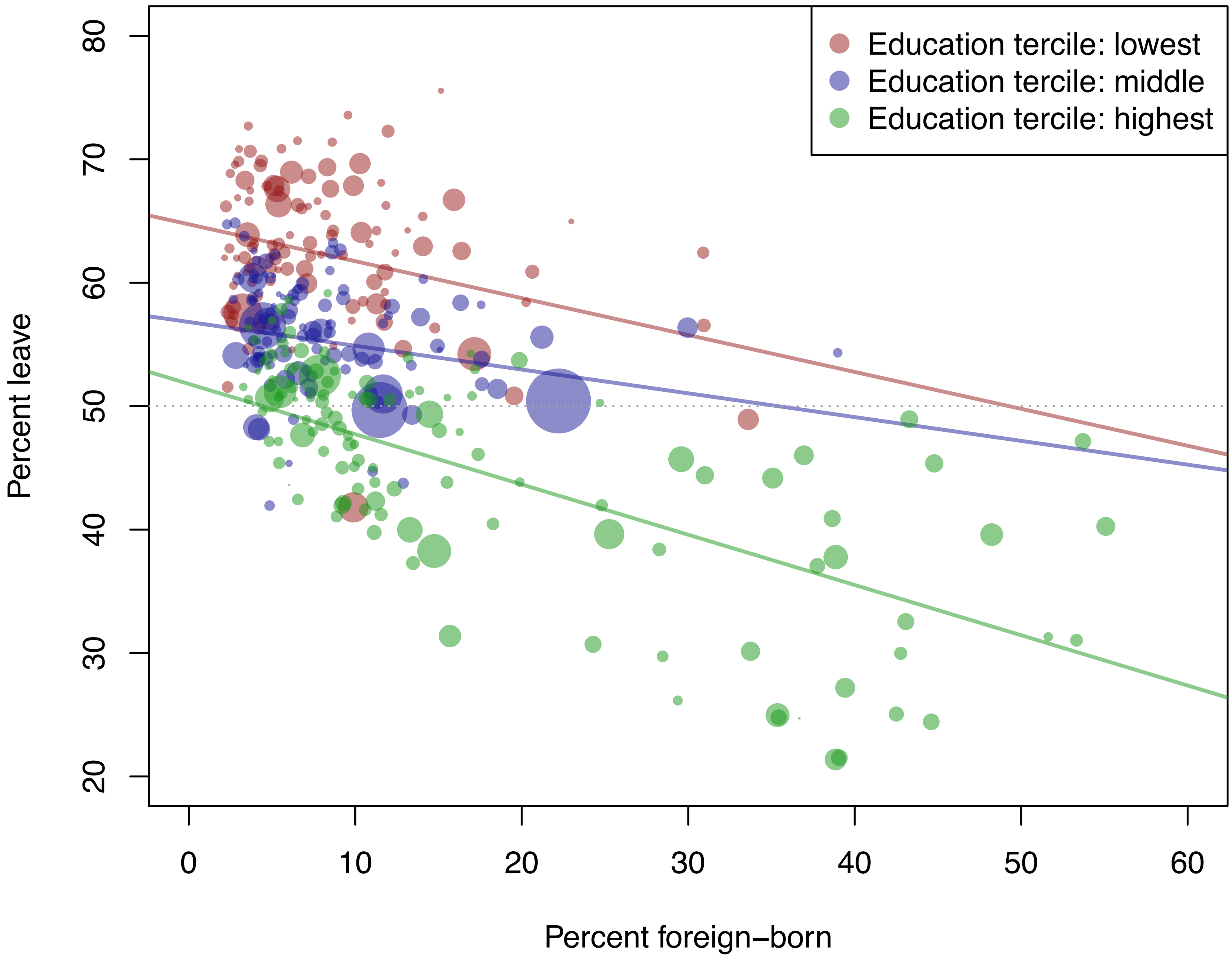


How can we measure the relationship between a local authority's proportion of foreign-born residents and its support for Brexit, controlling for its education level?

One idea: stratify by education level



Same thing in one plot



Summarizing multivariate relationships:
multivariate regression

A more general approach: multivariate regression

A more general approach: multivariate regression

Goal: measure relationship between
- “support for Leave” and
- “% foreign-born”

controlling for “% bachelors degree”.

A more general approach: multivariate regression

Goal: measure relationship between

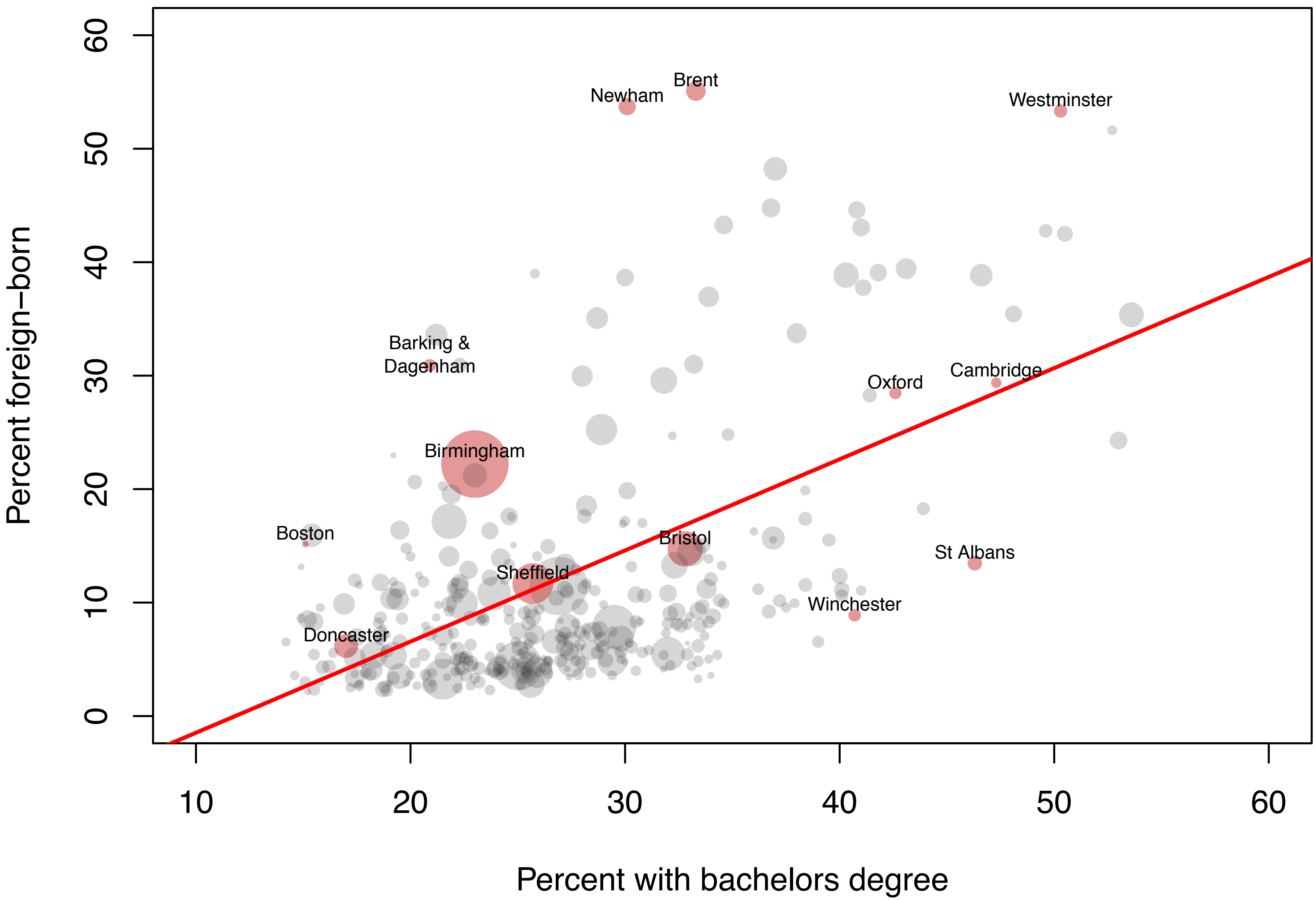
- "support for Leave" and
- "% foreign-born"

controlling for "% bachelors degree".

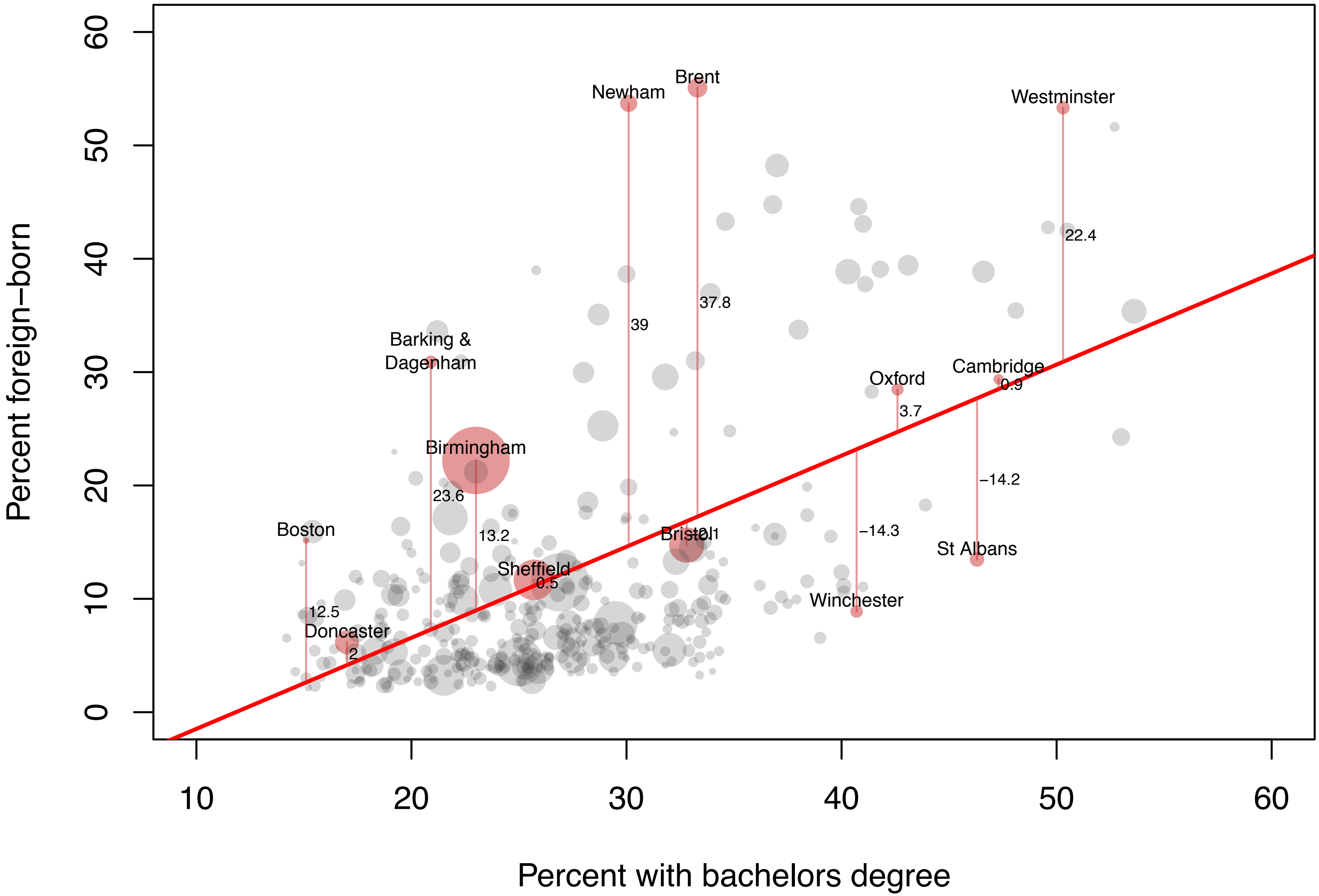
Basic idea: measure relationship between

- "support for Leave" and
- the part of "% foreign-born" that is not explained by "% bachelors degree"

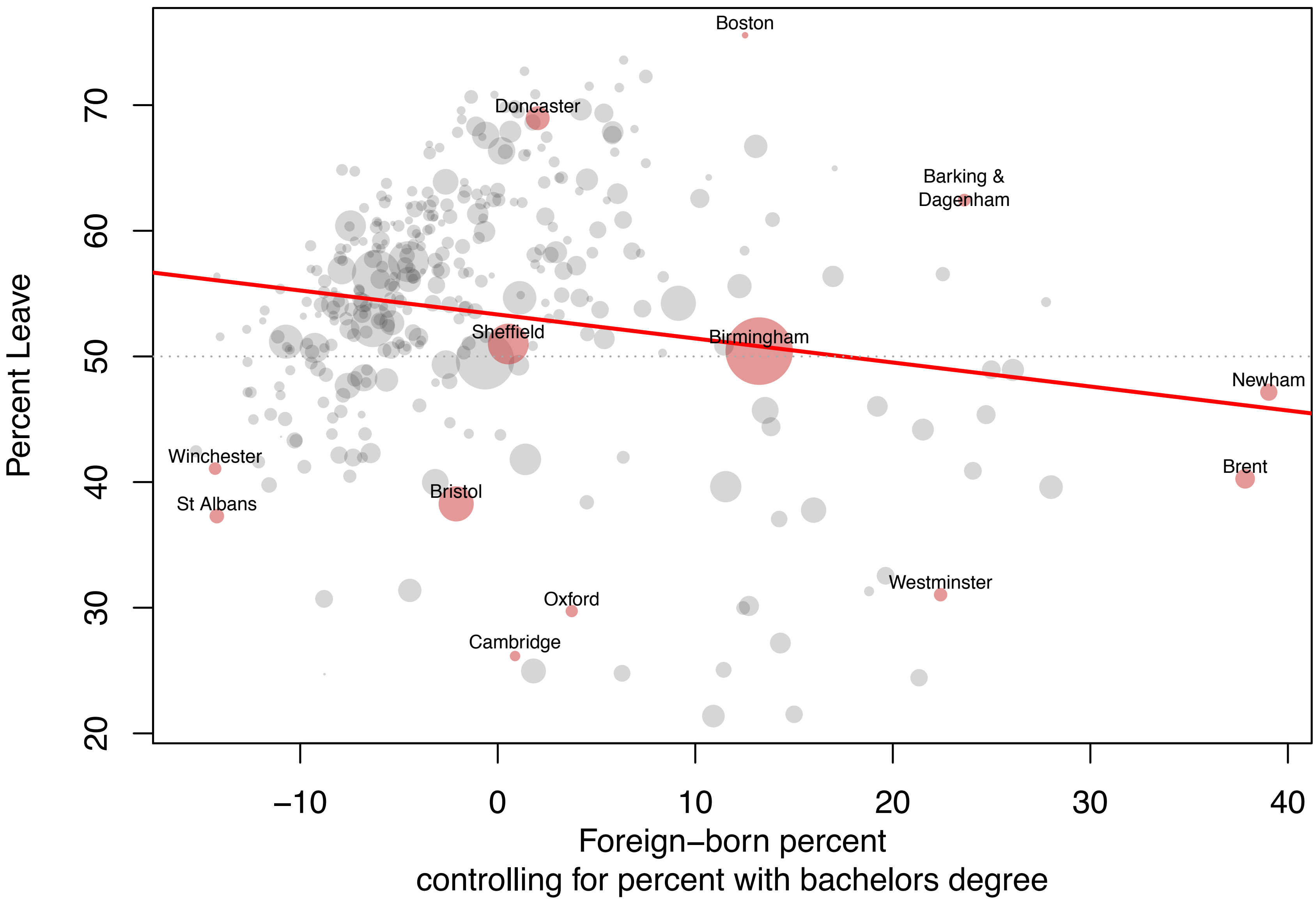
Step 1: regress explanatory variable (%foreign-born) on confounder (education)



Step 2: calculate residuals, i.e. the part of %foreign-born not "explained" by education



Step 3: regress outcome (%leave) on those residuals



Group activity

Group activity

Using the data sheet on the handout, find each group member's local authority (or another if we've already highlighted yours!) on Figures 1, 2, and 3.

1. Was your local authority more supportive of Brexit than would be expected given its % foreign born? or less?
2. Does your local authority have a higher % foreign born than would be expected given its % with bachelors? or lower?
3. Was your local authority more supportive of Brexit than would be expected given its % foreign born, controlling for its % with bachelors? or less?

Two ways to get the same answer

```
> lm(Percent_Leave ~ Percent_foreign_born + Bachelors_deg_percent, data = d)

Call:
lm(formula = Percent_Leave ~ Percent_foreign_born + Bachelors_deg_percent,
    data = d)

Coefficients:
          (Intercept)   Percent_foreign_born  Bachelors_deg_percent
              83.1386                -0.1742                -0.9875
```

```
> d$resids = resid(lm(Percent_foreign_born ~ Bachelors_deg_percent, data = d,
na.action = "na.exclude"))
> lm(Percent_Leave ~ resids, data = d)

Call:
lm(formula = Percent_Leave ~ resids, data = d)

Coefficients:
(Intercept)       resids
    54.4209      -0.1742
```

The usual way to think about multivariate regression

The usual way to think about multivariate regression

Above we showed how to interpret and estimate the regression coefficient on one variable controlling for the other.

The usual way to think about multivariate regression

Above we showed how to interpret and estimate the regression coefficient on one variable controlling for the other.

Can also just think about minimizing sum of squared residuals for a different prediction equation.

The usual way to think about multivariate regression

Above we showed how to interpret and estimate the regression coefficient on one variable controlling for the other.

Can also just think about minimizing sum of squared residuals for a different prediction equation.

Bivariate:

$\text{PercentLeave} = \beta_0 + \beta_1 \text{PercentForeignBorn}$

The usual way to think about multivariate regression

Above we showed how to interpret and estimate the regression coefficient on one variable controlling for the other.

Can also just think about minimizing sum of squared residuals for a different prediction equation.

Bivariate:

PercentLeave = β_0 + β_1PercentForeignBorn

Multivariate:

PercentLeave = β_0 + β_1PercentForeignBorn + β_2Education

Implementing multivariate regression

Implementing multivariate regression

Some options:

Implementing multivariate regression

Some options:

1. Use R to try every combination of 2 slopes and 1 intercept; choose the combination that has the lowest sum of squared residuals.

Implementing multivariate regression

Some options:

1. Use R to try every combination of 2 slopes and 1 intercept; choose the combination that has the lowest sum of squared residuals.
2. Use calculus to find the slope and intercept that minimize the sum of squared residuals.

Implementing multivariate regression

Some options:

1. Use R to try every combination of 2 slopes and 1 intercept; choose the combination that has the lowest sum of squared residuals.
2. Use calculus to find the slope and intercept that minimize the sum of squared residuals.
3. Use lm() function in R:

Implementing multivariate regression

Some options:

1. Use R to try every combination of 2 slopes and 1 intercept; choose the combination that has the lowest sum of squared residuals.
2. Use calculus to find the slope and intercept that minimize the sum of squared residuals.
3. Use lm() function in R:

```
> lm(d$Percent_Leave ~ d$Percent_foreign_born + d$Bachelors_deg_percent)

Call:
lm(formula = d$Percent_Leave ~ d$Percent_foreign_born + d$Bachelors_deg_percent)

Coefficients:
            (Intercept)   d$Percent_foreign_born  d$Bachelors_deg_percent
                83.1386                  -0.1742                  -0.9875
```

Presenting and interpreting results

	Dependent variable:		
	Percent supporting Leave		
	(1)	(2)	(3)
Percent foreign born	−0.582***	−0.174***	0.013
	(0.041)	(0.027)	(0.038)
Percent w. bachelors degree		−0.988***	−1.066***
		(0.035)	(0.035)
Mean age			0.755***
			(0.115)
Constant	60.937***	83.139***	52.808***
	(0.618)	(0.857)	(4.701)
Observations	344	344	344
R^2	0.375	0.813	0.834
Note:			*p<0.1; **p<0.05; ***p<0.01

Using “stargazer” package to present results

```
reg.1 = lm(Percent_Leave ~ Percent_foreign_born, data = d)
reg.2 = lm(Percent_Leave ~ Percent_foreign_born + Bachelors_deg_percent, data = d)
reg.3 = lm(Percent_Leave ~ Percent_foreign_born + Bachelors_deg_percent + age_mean, data = d)
```

```
stargazer(reg.1, reg.2, reg.3, dep.var.labels = "Percent supporting Leave", covariate.labels =
c("Percent foreign born", "Percent w. bachelors degree", "Mean age"), out = "brexit.html")
```

Using "stargazer" package to present results

To include table in Word document:

```
reg.1 = lm(Percent_Leave ~ Percent_foreign_born, data = d)
reg.2 = lm(Percent_Leave ~ Percent_foreign_born + Bachelors_deg_percent, data = d)
reg.3 = lm(Percent_Leave ~ Percent_foreign_born + Bachelors_deg_percent + age_mean, data = d)
```

```
stargazer(reg.1, reg.2, reg.3, dep.var.labels = "Percent supporting Leave", covariate.labels =
c("Percent foreign born", "Percent w. bachelors degree", "Mean age"), out = "brexit.html")
```

Using "stargazer" package to present results

To include table in Word document:

1. Install `stargazer` package and load it with `library(stargazer)`

```
reg.1 = lm(Percent_Leave ~ Percent_foreign_born, data = d)
reg.2 = lm(Percent_Leave ~ Percent_foreign_born + Bachelors_deg_percent, data = d)
reg.3 = lm(Percent_Leave ~ Percent_foreign_born + Bachelors_deg_percent + age_mean, data = d)
```

```
stargazer(reg.1, reg.2, reg.3, dep.var.labels = "Percent supporting Leave", covariate.labels =
c("Percent foreign born", "Percent w. bachelors degree", "Mean age"), out = "brexit.html")
```

Using "stargazer" package to present results

To include table in Word document:

1. Install `stargazer` package and load it with `library(stargazer)`

2. Run the regressions and store results

```
reg.1 = lm(Percent_Leave ~ Percent_foreign_born, data = d)
reg.2 = lm(Percent_Leave ~ Percent_foreign_born + Bachelors_deg_percent, data = d)
reg.3 = lm(Percent_Leave ~ Percent_foreign_born + Bachelors_deg_percent + age_mean, data = d)
```

```
stargazer(reg.1, reg.2, reg.3, dep.var.labels = "Percent supporting Leave", covariate.labels =
c("Percent foreign born", "Percent w. bachelors degree", "Mean age"), out = "brexit.html")
```

Using "stargazer" package to present results

To include table in Word document:

1. Install `stargazer` package and load it with `library(stargazer)`

2. Run the regressions and store results

```
reg.1 = lm(Percent_Leave ~ Percent_foreign_born, data = d)
reg.2 = lm(Percent_Leave ~ Percent_foreign_born + Bachelors_deg_percent, data = d)
reg.3 = lm(Percent_Leave ~ Percent_foreign_born + Bachelors_deg_percent + age_mean, data = d)
```

3. Use stargazer() command to output table to an html file

```
stargazer(reg.1, reg.2, reg.3, dep.var.labels = "Percent supporting Leave", covariate.labels =
c("Percent foreign born", "Percent w. bachelors degree", "Mean age"), out = "brexit.html")
```

Using “stargazer” package to present results

To include table in Word document:

1. Install `stargazer` package and load it with `library(stargazer)`

2. Run the regressions and store results

```
reg.1 = lm(Percent_Leave ~ Percent_foreign_born, data = d)
reg.2 = lm(Percent_Leave ~ Percent_foreign_born + Bachelors_deg_percent, data = d)
reg.3 = lm(Percent_Leave ~ Percent_foreign_born + Bachelors_deg_percent + age_mean, data = d)
```

3. Use stargazer() command to output table to an html file

```
stargazer(reg.1, reg.2, reg.3, dep.var.labels = "Percent supporting Leave", covariate.labels =
c("Percent foreign born", "Percent w. bachelors degree", "Mean age"), out = "brexit.html")
```

4. Open html file in browser (explorer/safari, not chrome), copy and paste table into word document

Inference

i.e. making claims beyond your sample

Sample vs population (1)

Sample vs population (1)

Say you have the data from a May 2016 survey asking how respondents plan to vote in the referendum and whether they went to university or not.

Sample vs population (1)

Say you have the data from a May 2016 survey asking how respondents plan to vote in the referendum and whether they went to university or not.

Your research questions are:

Sample vs population (1)

Say you have the data from a May 2016 survey asking how respondents plan to vote in the referendum and whether they went to university or not.

Your research questions are:

1. How much support is there for Brexit?

Sample vs population (1)

Say you have the data from a May 2016 survey asking how respondents plan to vote in the referendum and whether they went to university or not.

Your research questions are:

1. How much support is there for Brexit?
2. How is support for Brexit related to education?

Sample vs population (1)

Say you have the data from a May 2016 survey asking how respondents plan to vote in the referendum and whether they went to university or not.

Your research questions are:

1. How much support is there for Brexit?
2. How is support for Brexit related to education?

How would you answer these questions?

Sample vs population (1)

Say you have the data from a May 2016 survey asking how respondents plan to vote in the referendum and whether they went to university or not.

Your research questions are:

1. How much support is there for Brexit?
2. How is support for Brexit related to education?

How would you answer these questions?

Is there any **uncertainty** in your answers?

Sample vs population (2)

Sample vs population (2)

Is there uncertainty? Depends on who you are asking about.

Sample vs population (2)

Is there uncertainty? Depends on who you are asking about.

1. How much support is there for Brexit *among respondents to this survey*?
2. How is support for Brexit related to education *among respondents to this survey*?

Sample vs population (2)

Is there uncertainty? Depends on who you are asking about.

1. How much support is there for Brexit *among respondents to this survey*?
2. How is support for Brexit related to education *among respondents to this survey*?

(About the sample)

Sample vs population (2)

Is there uncertainty? Depends on who you are asking about.

1. How much support is there for Brexit *among respondents to this survey*?
2. How is support for Brexit related to education *among respondents to this survey*?

(About the sample)

1. How much support is there for Brexit *among all voters*?
2. How is support for Brexit related to education *among all voters*?

Sample vs population (2)

Is there uncertainty? Depends on who you are asking about.

1. How much support is there for Brexit *among respondents to this survey*?
2. How is support for Brexit related to education *among respondents to this survey*?

(About the sample)

1. How much support is there for Brexit *among all voters*?
2. How is support for Brexit related to education *among all voters*?

(About the population)

Sample vs population (2)

Is there uncertainty? Depends on who you are asking about.

(About the sample)	(About the population)
1. How much support is there for Brexit *among respondents to this survey*? 2. How is support for Brexit related to education *among respondents to this survey*?	1. How much support is there for Brexit *among all voters*? 2. How is support for Brexit related to education *among all voters*?

(About the sample)

(About the population)

No real uncertainty.

(Maybe about measurement.)

Sample vs population (2)

Is there uncertainty? Depends on who you are asking about.

1. How much support is there for Brexit *among respondents to this survey*?
2. How is support for Brexit related to education *among respondents to this survey*?

(About the sample)

No real uncertainty.
(Maybe about measurement.)

1. How much support is there for Brexit *among all voters*?
2. How is support for Brexit related to education *among all voters*?

(About the population)

Uncertainty due to **sampling variation**.

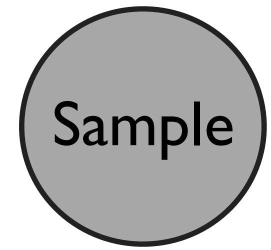
Sample vs population (3)

Sample vs population (3)

Generally, we have data from a **sample**

Sample vs population (3)

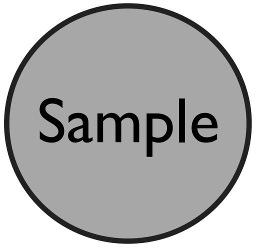
Generally, we have data from a **sample**



Sample vs population (3)

Generally, we have data from a **sample**

but we want to say something about a (larger) **population**.

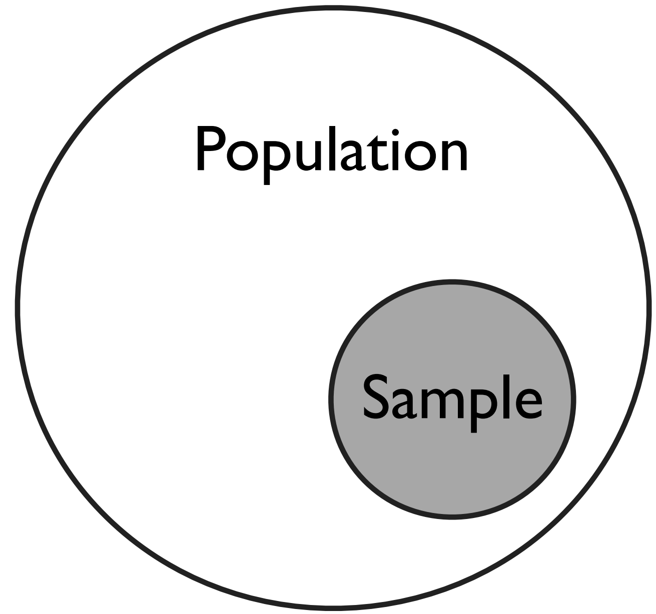


Sample vs population (3)

Generally, we have data from a **sample**

but we want to say something about a (larger) **population**.

This is **statistical inference**.

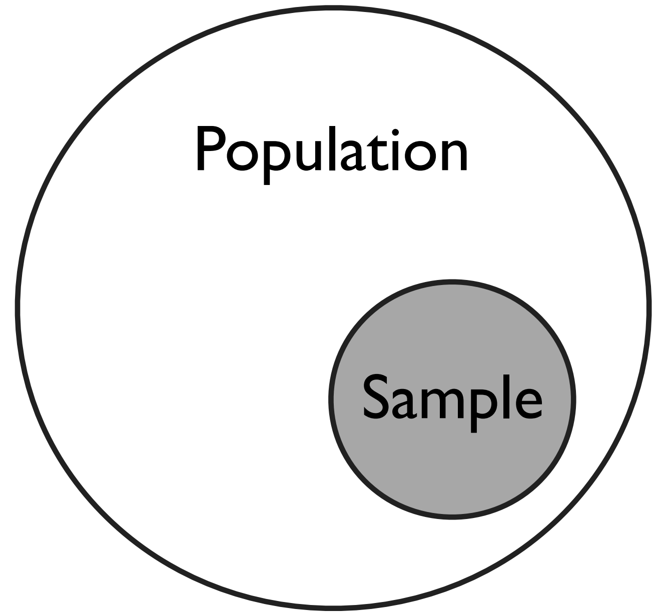


Sample vs population (3)

Generally, we have data from a **sample**

but we want to say something about a (larger) **population**.

This is **statistical inference**.



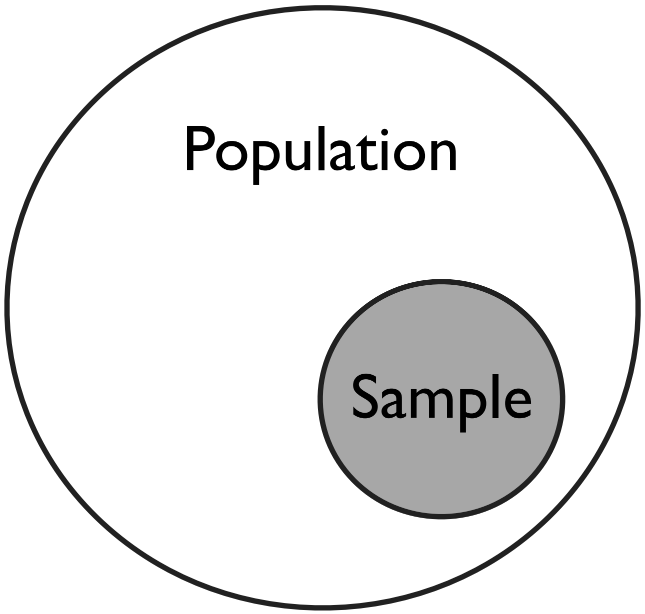
In hypothesis testing, we use data from a **sample** to assess conjectures about the **population**.

Sample vs population (3)

Generally, we have data from a **sample**

but we want to say something about a (larger) **population**.

This is **statistical inference**.



In hypothesis testing, we use data from a **sample** to assess conjectures about the **population**.

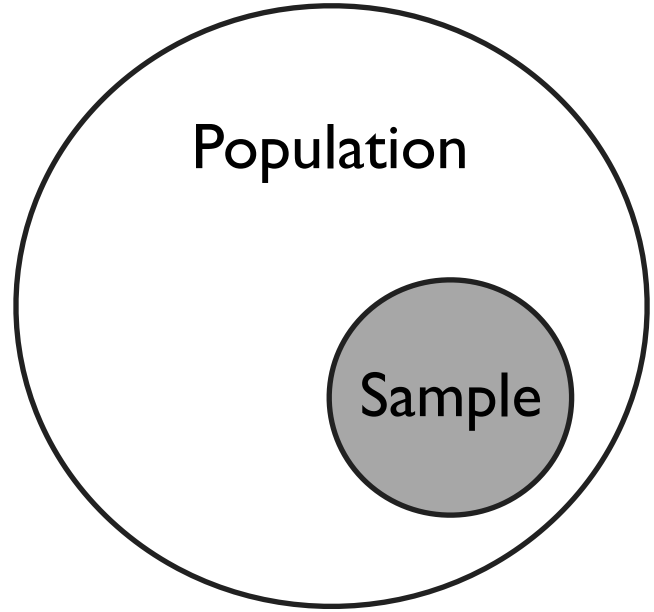
Because the sample is not the population:

Sample vs population (3)

Generally, we have data from a **sample**

but we want to say something about a (larger) **population**.

This is **statistical inference**.



In hypothesis testing, we use data from a **sample** to assess conjectures about the **population**.

Because the sample is not the population:

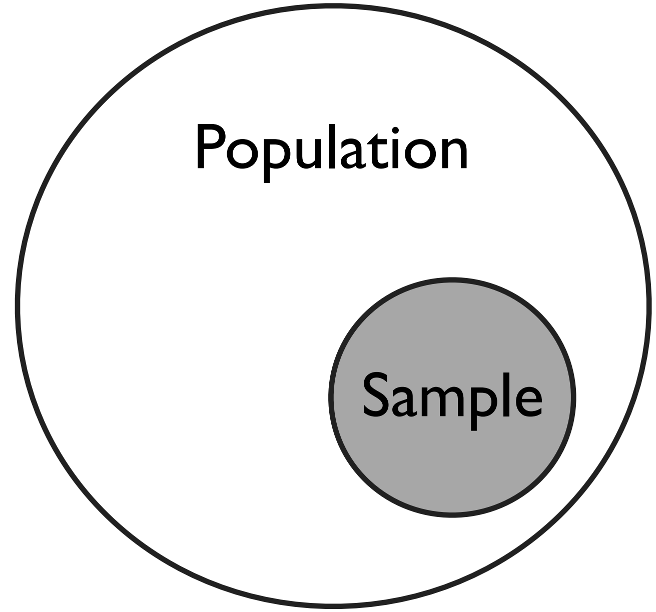
- polls have a **margin of error**

Sample vs population (3)

Generally, we have data from a **sample**

but we want to say something about a (larger) **population**.

This is **statistical inference**.



In hypothesis testing, we use data from a **sample** to assess conjectures about the **population**.

Because the sample is not the population:

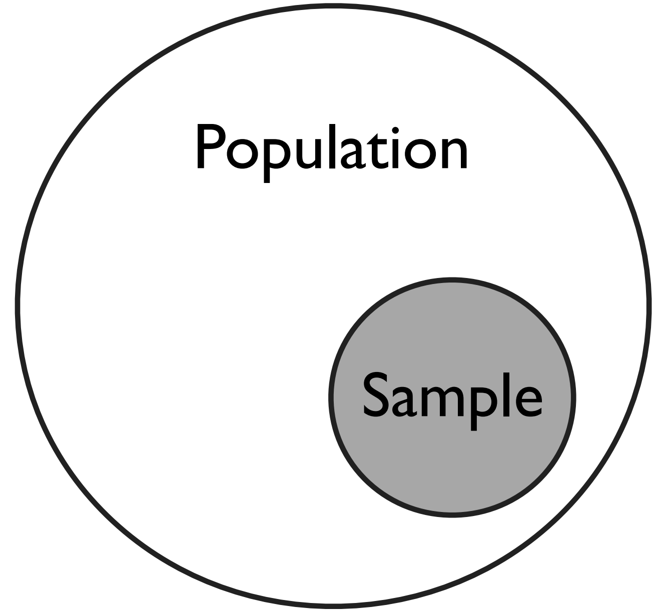
- polls have a **margin of error**
- regression coefficients have **standard errors**

Sample vs population (3)

Generally, we have data from a **sample**

but we want to say something about a (larger) **population**.

This is **statistical inference**.



In hypothesis testing, we use data from a **sample** to assess conjectures about the **population**.

Because the sample is not the population:

- polls have a **margin of error**
- regression coefficients have **standard errors**
- our conclusions in hypothesis testing are guesses, with confidence summarized by **p-values**

Thought experiment

Thought experiment

Imagine that on June 22, 2016, you take a random sample of people who will vote in the EU referendum and ask whether they support "Leave".

Will the level of support in your sample be close to the true average support?

Thought experiment

Imagine that on June 22, 2016, you take a random sample of people who will vote in the EU referendum and ask whether they support "Leave".

Will the level of support in your sample be close to the true average support?

If truly a random sample, there is no **bias**: you should expect to get the true value on average.

Why would the level of support in your sample differ from the true value? By how much?

Thought experiment

Imagine that on June 22, 2016, you take a random sample of people who will vote in the EU referendum and ask whether they support "Leave".

Will the level of support in your sample be close to the true average support?

If truly a random sample, there is no **bias**: you should expect to get the true value on average.

Why would the level of support in your sample differ from the true value? By how much?

What would the magnitude of this **random error** depend on?

- size of sample (1,006 GB adults vs. 10,000,000)
- true level of support (what if 100% supported remaining in EU?)

Simulating the thought experiment in R

Simulating the thought experiment in R

We know that 52% of all voters supported **Leave. We want to know** how much the result of a poll might deviate from the true level of support.

Let's find out using R!

Simulating the thought experiment in R

We know that 52% of all voters supported **Leave. We want to know** how much the result of a poll might deviate from the true level of support.

Let's find out using R!

Using R, I can randomly draw 10 ones and zeros, where the probability of drawing a one is 0.52:

Simulating the thought experiment in R

We know that 52% of all voters supported **Leave. We want to know** how much the result of a poll might deviate from the true level of support.

Let's find out using R!

Using R, I can randomly draw 10 ones and zeros, where the probability of drawing a one is 0.52:

```
> sample(x = c(0,1), size = 10, replace = T, prob = c(.48, .52))
 [1] 1 1 0 0 1 1 0 0 0 0
```

Simulating the thought experiment in R

We know that 52% of all voters supported **Leave. We want to know** how much the result of a poll might deviate from the true level of support.

Let's find out using R!

Using R, I can randomly draw 10 ones and zeros, where the probability of drawing a one is 0.52:

```
> sample(x = c(0,1), size = 10, replace = T, prob = c(.48, .52))
 [1] 1 1 0 0 1 1 0 0 0 0
```

I can do it again:

```
> sample(x = c(0,1), size = 10, replace = T, prob = c(.48, .52))
 [1] 1 1 0 0 1 1 0 1 1 0
```

Simulating the thought experiment in R

We know that 52% of all voters supported **Leave. We want to know** how much the result of a poll might deviate from the true level of support.

Let's find out using R!

Using R, I can randomly draw 10 ones and zeros, where the probability of drawing a one is 0.52:

```
> sample(x = c(0,1), size = 10, replace = T, prob = c(.48, .52))
 [1] 1 1 0 0 1 1 0 0 0 0
```

I can do it again:

```
> sample(x = c(0,1), size = 10, replace = T, prob = c(.48, .52))
 [1] 1 1 0 0 1 1 0 1 1 0
```

I can increase the number of "respondents" to 1,006:

```
> sample(x = c(0,1), size = 1006, replace = T, prob = c(.48, .52))
   [1] 0 0 0 0 0 0 0 1 1 1 1 1 0 1 1 1 1 1 1 1 1 0 0 1 1 0 0 0 0 0 0 0 1
  [39] 0 0 0 1 1 1 1 0 0 0 1 0 0 1 0 1 1 1 0 0 0 1 1 0 0 0 0 0 0 0 0 0 1
  [77] 1 0 0 1 1 1 0 0 0 0 0 1 1 1 1 1 1 1 0 0 0 0 1 0 0 0 0 0 1 1 0 0 1
 [115] 1 0 1 1 0 1 0 0 1 0 0 0 0 1 1 0 0 0 0 1 1 1 1 0 1 0 1 0 1 0 0 1 0
 [153] 0 1 1 0 0 1 1 1 0 0 1 0 1 0 1 1 1 1 0 0 0 1 0 1 0 1 1 1 1 1 1 0 1
 [191] 1 0 0 1 1 1 1 0 1 0 1 0 0 0 1 0 0 1 0 0 0 1 1 1 0 0 0 0 0 1 0 1 1
```

Simulating the thought experiment (2)

Simulating the thought experiment (2)

I can store the sample and take the mean:

```
> samp = sample(x = c(0,1), size = 1006, replace = T, prob = c(.48, .52))
> mean(samp)
[1] 0.5318091
```

Simulating the thought experiment (2)

I can store the sample and take the mean:

```
> samp = sample(x = c(0,1), size = 1006, replace = T, prob = c(.48, .52))
> mean(samp)
[1] 0.5318091
```

I can do it again:

```
> samp = sample(x = c(0,1), size = 1006, replace = T, prob = c(.48, .52))
> mean(samp)
[1] 0.5119284
```

Simulating the thought experiment (2)

I can store the sample and take the mean:

```
> samp = sample(x = c(0,1), size = 1006, replace = T, prob = c(.48, .52))
> mean(samp)
[1] 0.5318091
```

I can do it again:

```
> samp = sample(x = c(0,1), size = 1006, replace = T, prob = c(.48, .52))
> mean(samp)
[1] 0.5119284
```

I can do it 10,000 times and look at the histogram of support:

Simulating the thought experiment (2)

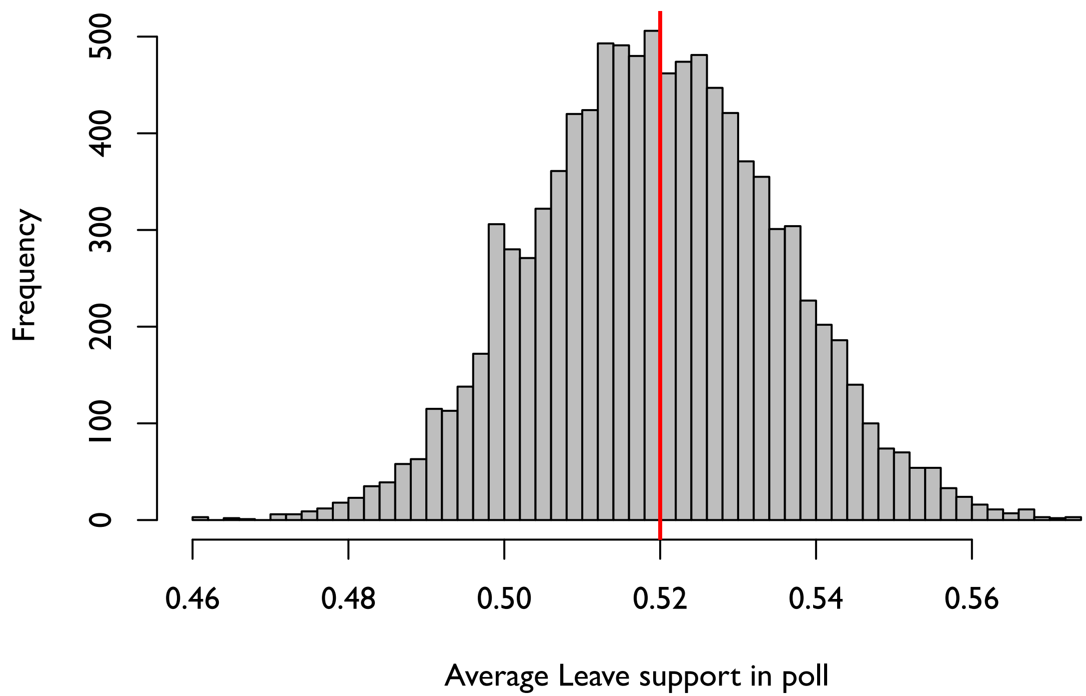
I can store the sample and take the mean:

```
> samp = sample(x = c(0,1), size = 1006, replace = T, prob = c(.48, .52))
> mean(samp)
[1] 0.5318091
```

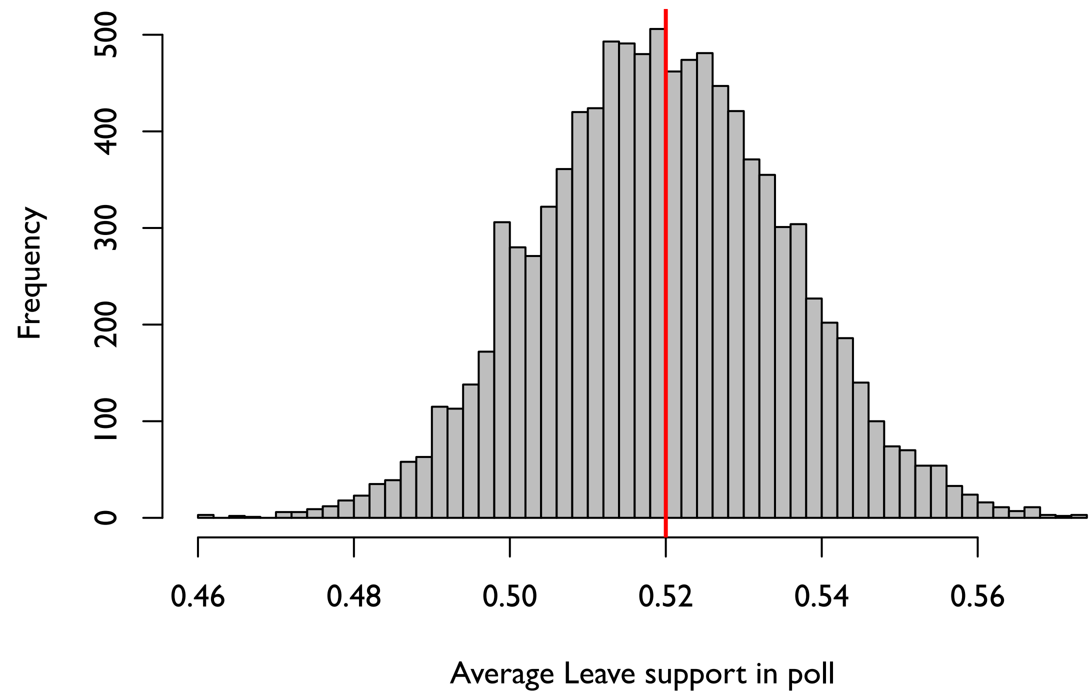
I can do it again:

```
> samp = sample(x = c(0,1), size = 1006, replace = T, prob = c(.48, .52))
> mean(samp)
[1] 0.5119284
```

I can do it 10,000 times and look at the histogram of support:

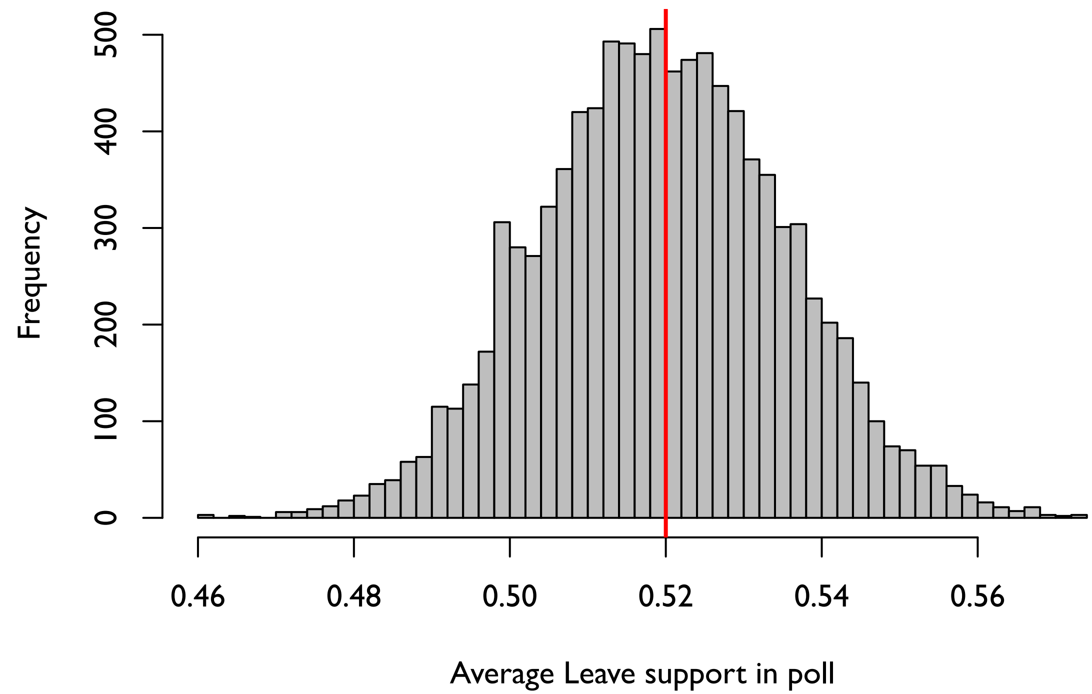


Simulating the thought experiment (3)



Simulating the thought experiment (3)

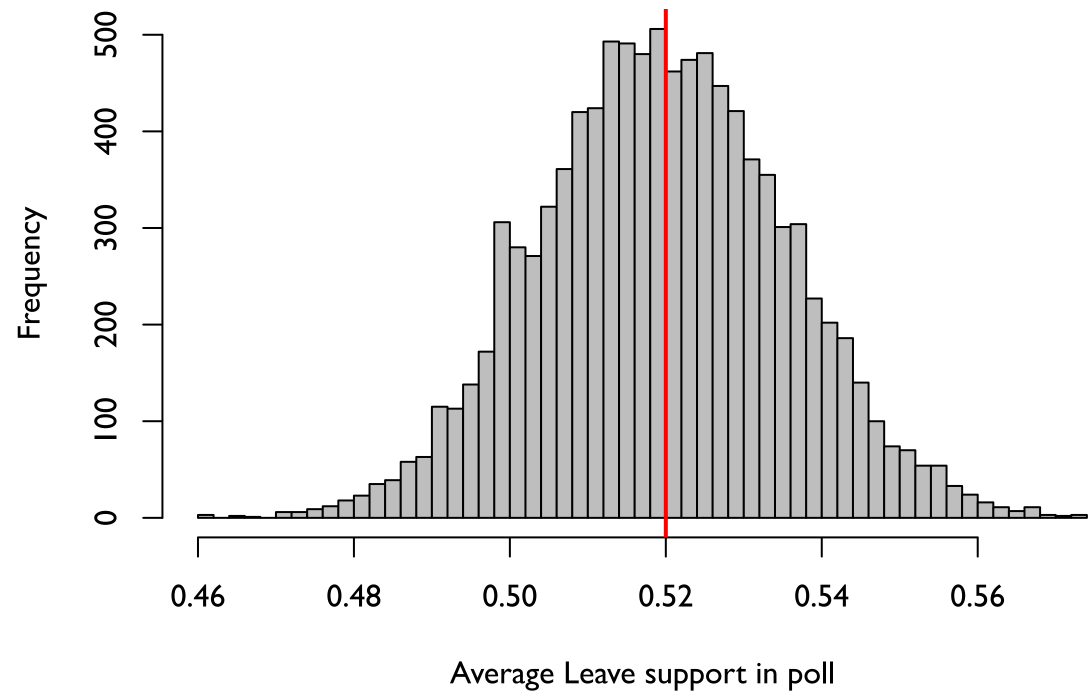
The results vary across our 10,000 "surveys" because of **sampling error**.



Simulating the thought experiment (3)

The results vary across our 10,000 "surveys" because of **sampling error**.

How much sampling error is there in our simulation?



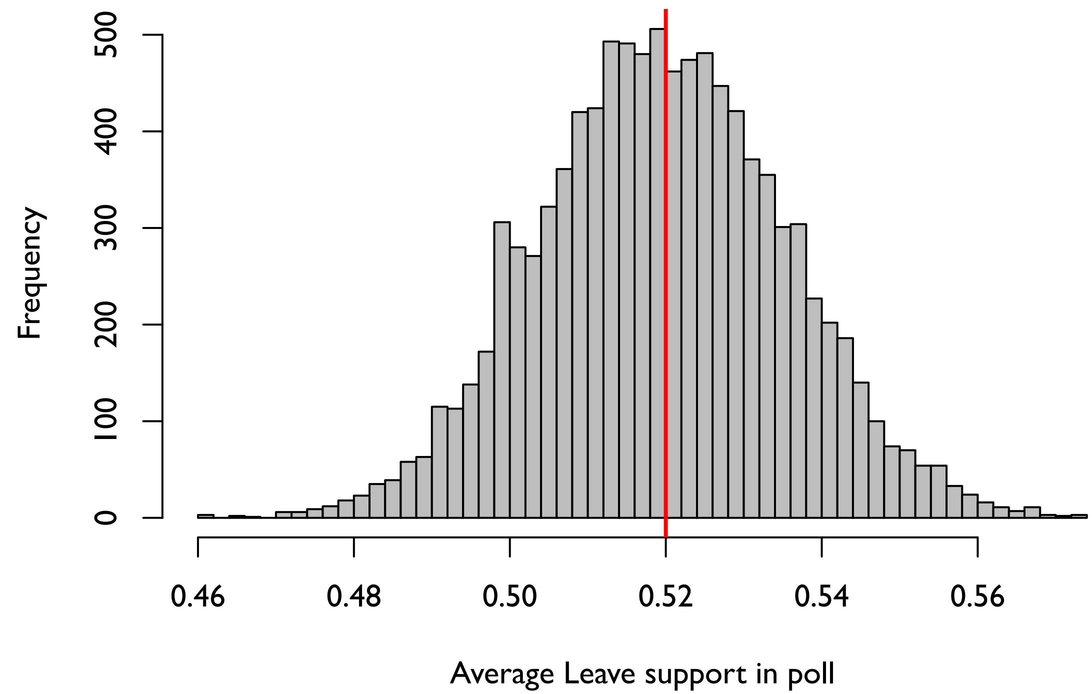
Simulating the thought experiment (3)

The results vary across our 10,000 "surveys" because of **sampling error**.

How much sampling error is there in our simulation?

The standard deviation:

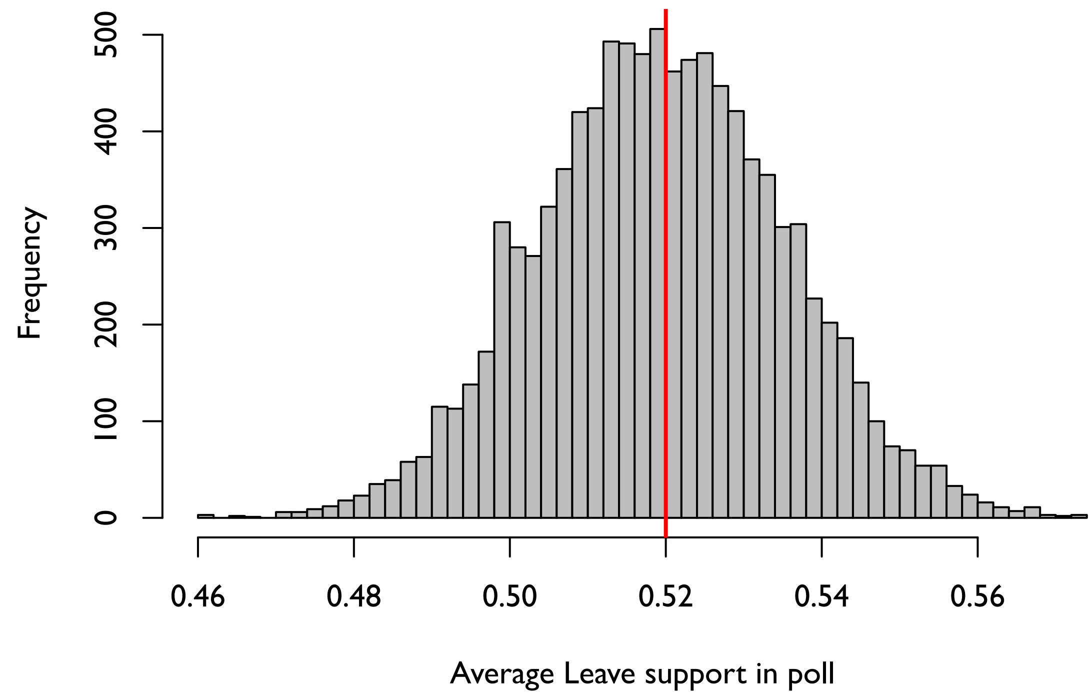
```
> sd(poll.results)
[1] 0.01572411
```



Simulating the thought experiment (3)

The results vary across our 10,000 "surveys" because of **sampling error.**

How much sampling error is there in our simulation?



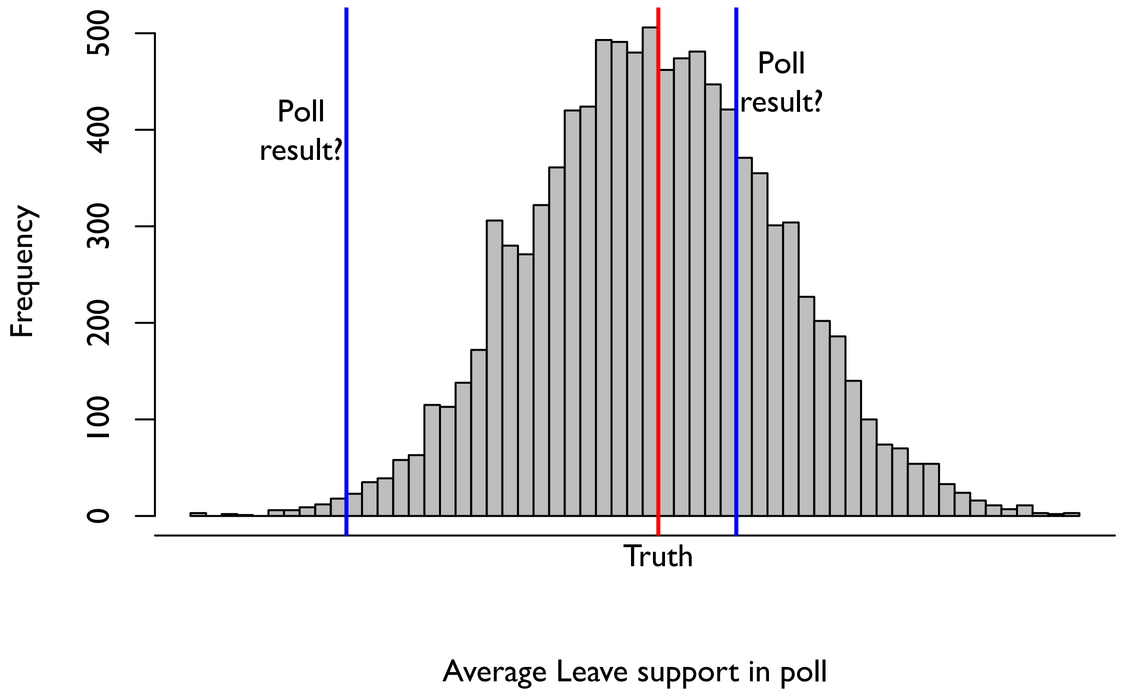
The standard deviation:

```
> sd(poll.results)
[1] 0.01572411
```

95% of the samples had a mean between 0.49 and 0.55:

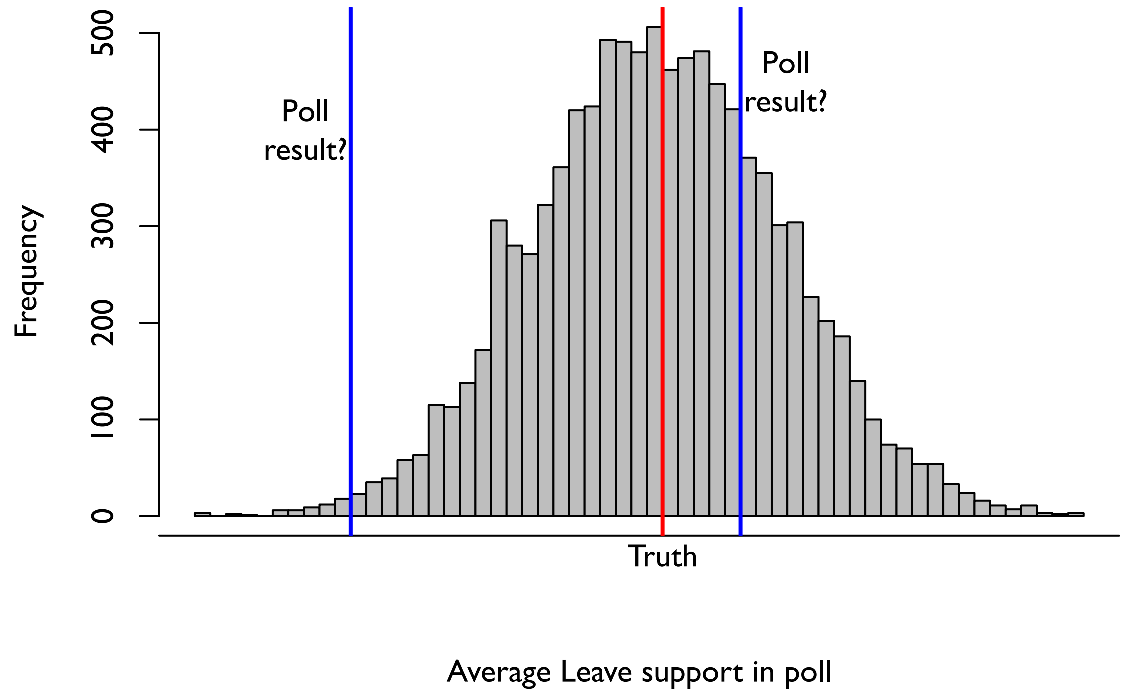
```
> quantile(poll.results, c(.025, .975))
     2.5%     97.5%
0.4890656 0.5497018
```

From thought experiment to margin of error



From thought experiment to margin of error

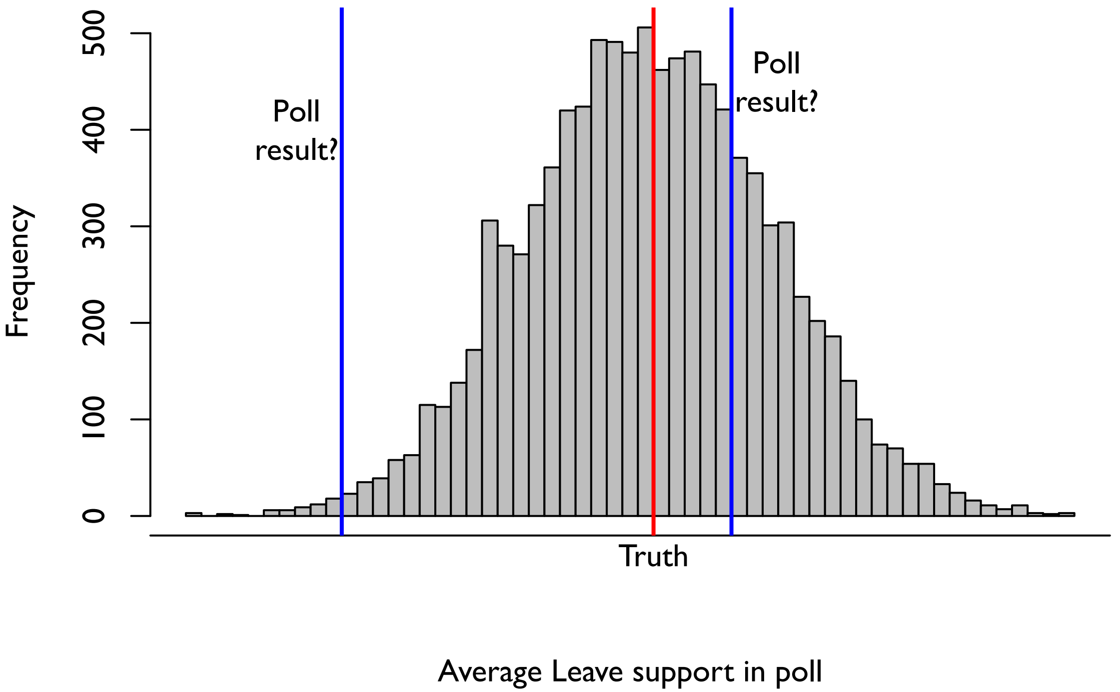
In a real survey, you don't know the answer; all you get is a **single number,** i.e. your poll result.



From thought experiment to margin of error

In a real survey, you don't know the answer; all you get is a **single number,** i.e. your poll result.

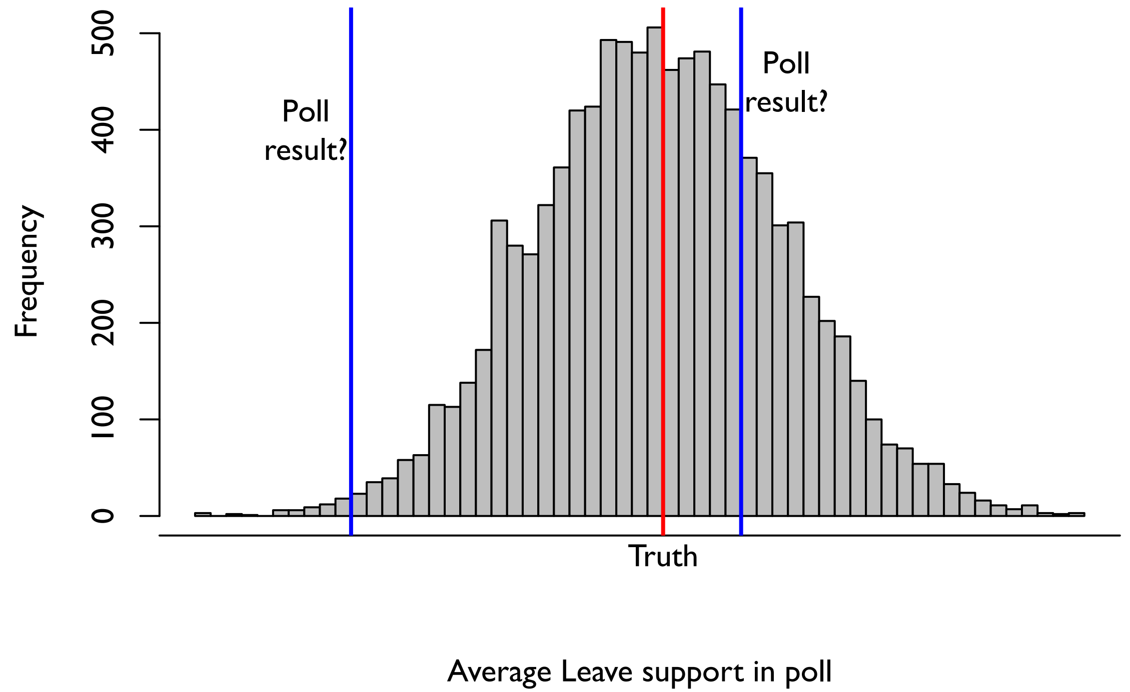
But the histogram from the thought experiment gives you a clue how close your number is to the "Truth".



From thought experiment to margin of error

In a real survey, you don't know the answer; all you get is a **single number,** i.e. your poll result.

But the histogram from the thought experiment gives you a clue how close your number is to the "Truth".

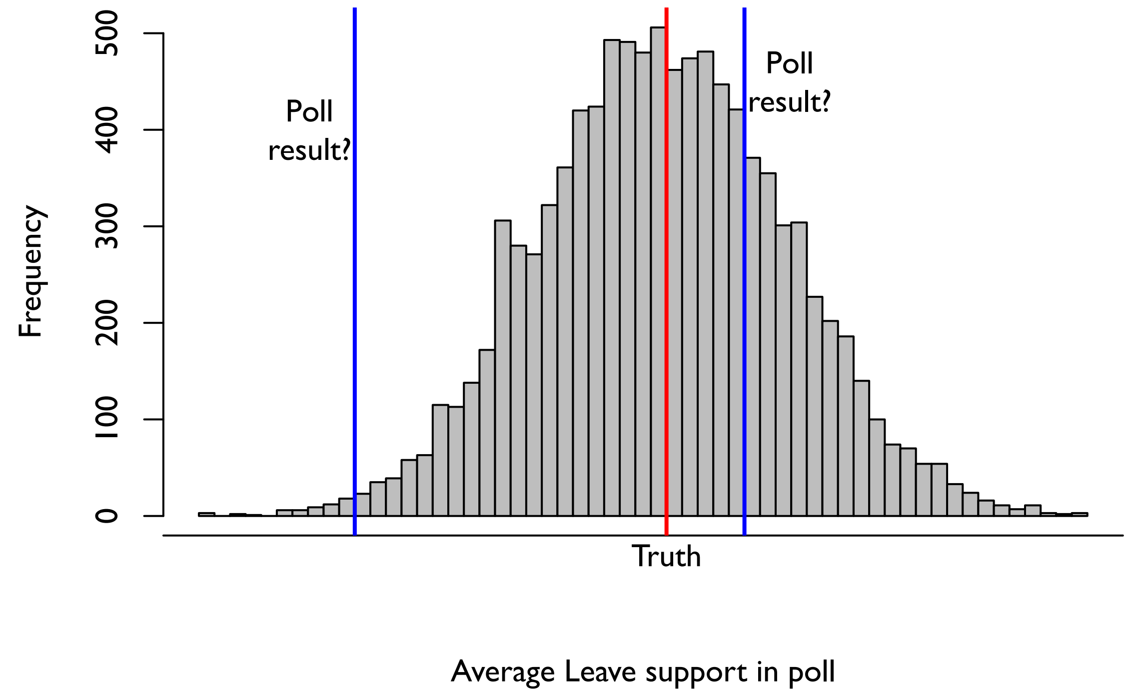


In our thought experiment (where we **know** the truth), 95% of the samples were within 0.031 of the truth.

From thought experiment to margin of error

In a real survey, you don't know the answer; all you get is a **single number,** i.e. your poll result.

But the histogram from the thought experiment gives you a clue how close your number is to the "Truth".



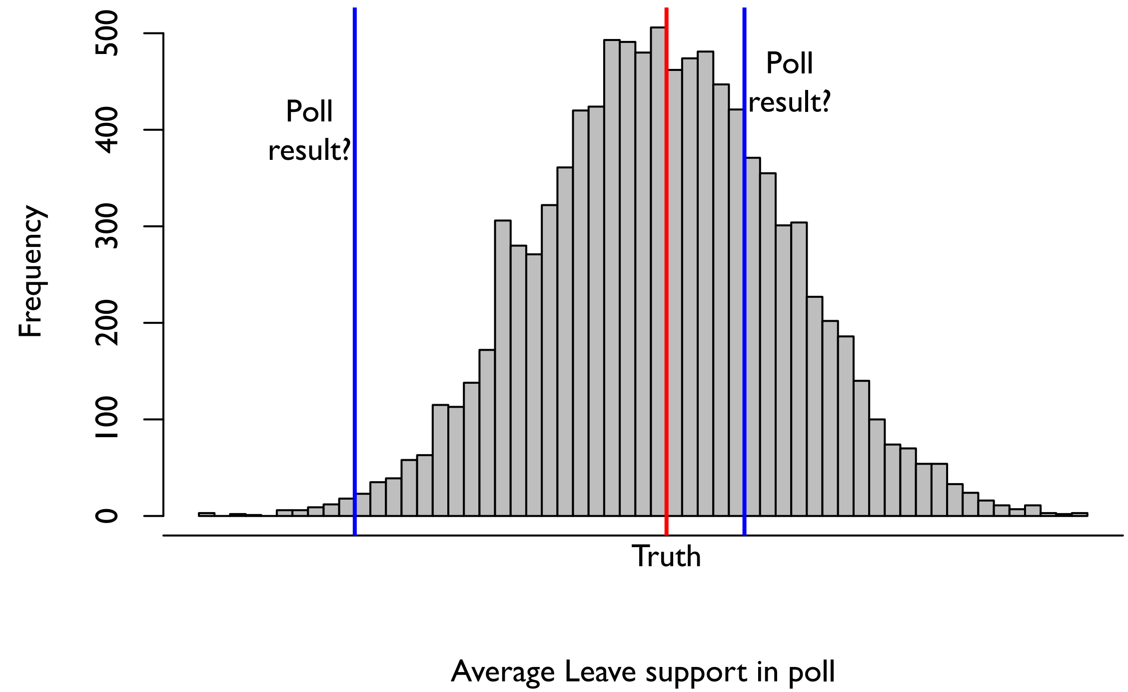
In our thought experiment (where we **know** the truth), 95% of the samples were within 0.031 of the truth.

In an actual survey (where **don't know** the truth), we have 95% confidence that our estimate is within 0.031 of the truth.

From thought experiment to margin of error

In a real survey, you don't know the answer; all you get is a **single number,** i.e. your poll result.

But the histogram from the thought experiment gives you a clue how close your number is to the "Truth".



In our thought experiment (where we **know** the truth), 95% of the samples were within 0.031 of the truth.

In an actual survey (where **don't know** the truth), we have 95% confidence that our estimate is within 0.031 of the truth.

Margin of error

From thought experiment to margin of error (2)

From thought experiment to margin of error (2)

So when we do a survey, we get:

From thought experiment to margin of error (2)

So when we do a survey, we get:

An estimate for Leave support (e.g. 49%)

From thought experiment to margin of error (2)

So when we do a survey, we get:

An estimate for Leave support (e.g. 49%)

(From the thought experiment:) An estimate of the standard deviation of poll results across samples: 0.0157 (called the **standard error** of the poll)

From thought experiment to margin of error (2)

So when we do a survey, we get:

An estimate for Leave support (e.g. 49%)

(From the thought experiment:) An estimate of the standard deviation of poll results across samples: 0.0157 (called the **standard error** of the poll)

(Combining the two:) A 95% confidence interval, which we expect to include the truth in 95% of samples: e.g. 49% ± 3.1% (3.1% is the **margin of error** of the poll)

Another way to get the margin of error (1)

Another way to get the margin of error (1)

Another way to get the margin
of error from a single sample:

Another way to get the margin of error (1)

Another way to get the margin of error from a single sample:

The **central limit theorem** says that the proportion of support in samples of size n will follow a Normal distribution centered on the truth with approximate standard deviation:

Another way to get the margin of error (1)

Another way to get the margin of error from a single sample:

The **central limit theorem** says that the proportion of support in samples of size n will follow a Normal distribution centered on the truth with approximate standard deviation:

$$\sqrt{\frac{\text{Variance of sample}}{n}}$$

Another way to get the margin of error (1)

Another way to get the margin of error from a single sample:

The **central limit theorem** says that the proportion of support in samples of size n will follow a Normal distribution centered on the truth with approximate standard deviation:

$$\sqrt{\frac{\text{Variance of sample}}{n}}$$

Our sample of e.g. 546 "Leaves" and 460 "Remains" has a variance of .248.

Another way to get the margin of error (1)

Another way to get the margin of error from a single sample:

The **central limit theorem** says that the proportion of support in samples of size n will follow a Normal distribution centered on the truth with approximate standard deviation:

$$\sqrt{\frac{\text{Variance of sample}}{n}}$$

Our sample of e.g. 546 "Leaves" and 460 "Remains" has a variance of .248.

So the estimated standard deviation (**standard error**) of our estimate is:

$$\sqrt{\frac{.248}{1006}} = 0.0157$$

Another way to get the margin of error (1)

Another way to get the margin of error from a single sample:

The **central limit theorem** says that the proportion of support in samples of size n will follow a Normal distribution centered on the truth with approximate standard deviation:

$$\sqrt{\frac{\text{Variance of sample}}{n}}$$

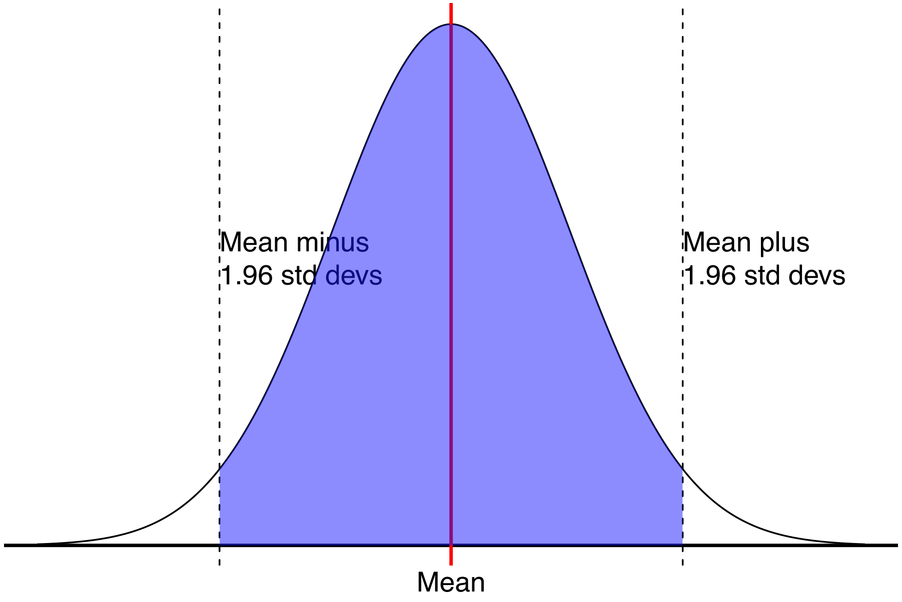
Our sample of e.g. 546 "Leaves" and 460 "Remains" has a variance of .248.

So the estimated standard deviation (**standard error**) of our estimate is:

$$\sqrt{\frac{.248}{1006}} = 0.0157$$

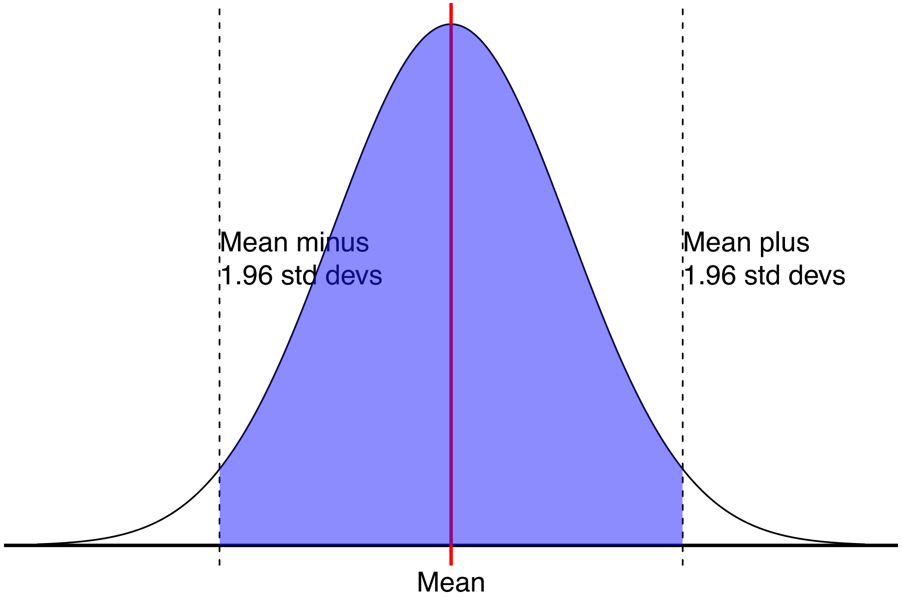
Compare: the standard deviation of our simulations was 0.0157

Another way to get the margin of error (2)



Another way to get the margin of error (2)

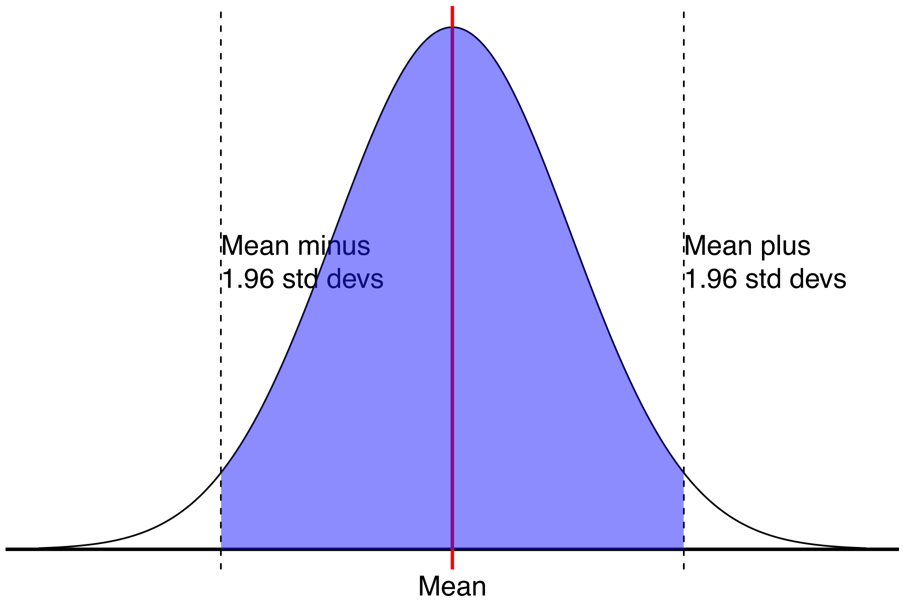
In a Normal distribution, about 95% of the draws are within 1.96 standard deviations of the mean.



Another way to get the margin of error (2)

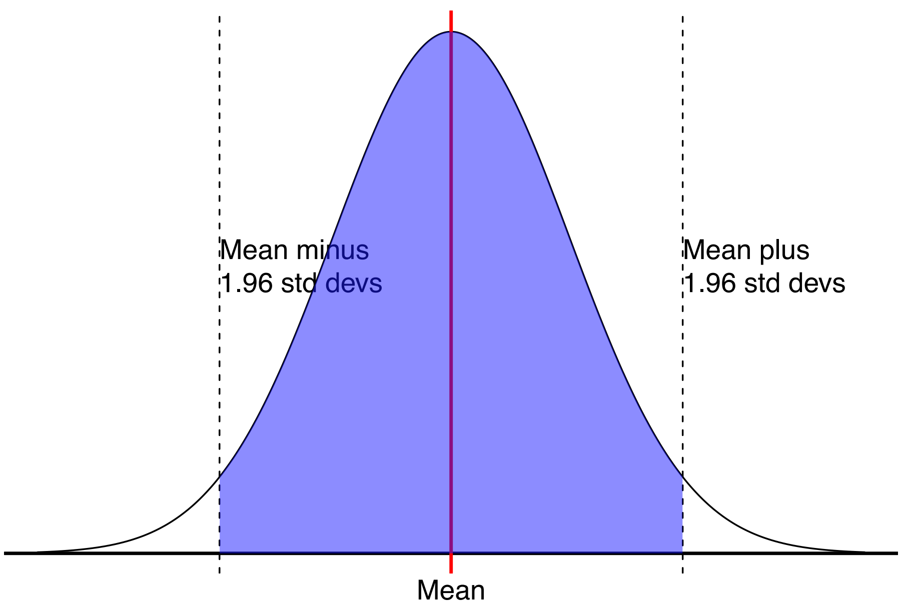
In a Normal distribution, about 95% of the draws are within 1.96 standard deviations of the mean.

This indicates that in 95% of surveys we run, our answer should be within 1.96 standard deviations of the truth.



Another way to get the margin of error (2)

In a Normal distribution, about 95% of the draws are within 1.96 standard deviations of the mean.

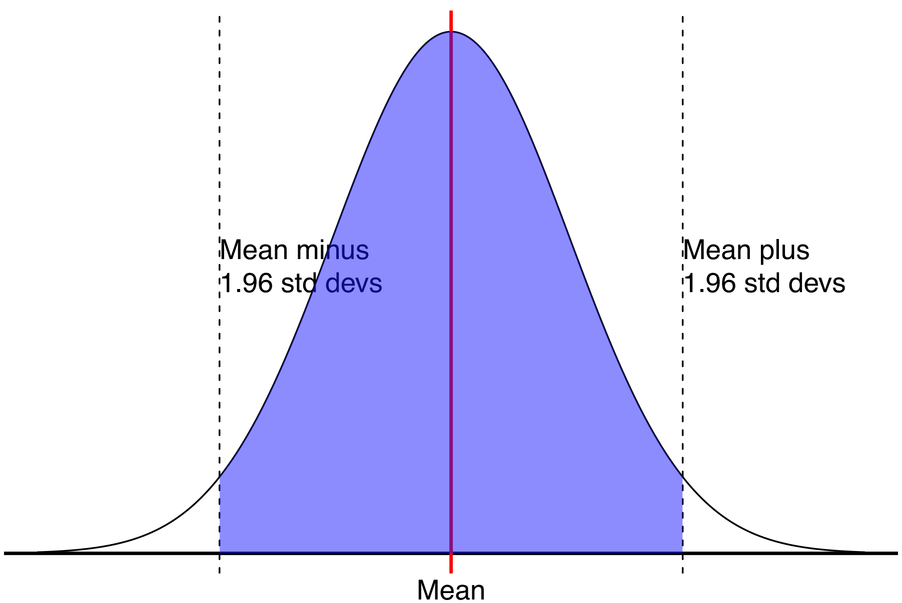


This indicates that in 95% of surveys we run, our answer should be within 1.96 standard deviations of the truth.

Given estimated standard deviation (standard error) of 0.0157, we have a **margin of error** (1.96 times standard error) of .031.

Another way to get the margin of error (2)

In a Normal distribution, about 95% of the draws are within 1.96 standard deviations of the mean.



This indicates that in 95% of surveys we run, our answer should be within 1.96 standard deviations of the truth.

Given estimated standard deviation (standard error) of 0.0157, we have a **margin of error** (1.96 times standard error) of .031.

Compare: our simulations implied a margin of error of 0.031.

Quick recap: survey part

Quick recap: survey part

- **Standard error of poll**: an estimate of how much the estimate might vary due to random error (**sampling error**)

Quick recap: survey part

- **Standard error of poll**: an estimate of how much the estimate might vary due to random error (**sampling error**)
- In 95% of polls, the true value should be within **margin of error** (≈ 2 × standard error) of the estimate (assuming no bias)

Quick recap: survey part

- **Standard error of poll**: an estimate of how much the estimate might vary due to random error (**sampling error**)
- In 95% of polls, the true value should be within **margin of error** ($\approx 2 \times$ standard error) of the estimate (assuming no bias)
- Two ways we got the margin of error:

Quick recap: survey part

- **Standard error of poll**: an estimate of how much the estimate might vary due to random error (**sampling error**)
- In 95% of polls, the true value should be within **margin of error** (≈ 2 × standard error) of the estimate (assuming no bias)
- Two ways we got the margin of error:
  - **Simulation** in R of 10,000 random samples of size 1,006 given a known level of support for "Leave"

Quick recap: survey part

- **Standard error of poll**: an estimate of how much the estimate might vary due to random error (**sampling error**)
- In 95% of polls, the true value should be within **margin of error** (≈ 2 × standard error) of the estimate (assuming no bias)
- Two ways we got the margin of error:
  - **Simulation** in R of 10,000 random samples of size 1,006 given a known level of support for "Leave"
  - **Central limit theorem:** approximation to a normal distribution

Thought experiment (2)

Thought experiment (2)

Suppose we know that

Thought experiment (2)

Suppose we know that

Support for Leave

57.9% of those who did not attend university

41.5% of those who did attend university

Thought experiment (2)

Suppose we know that

Support for Leave

57.9% of those who did not attend university

41.5% of those who did attend university

Thus if we ran the regression

$$\text{LeaveSupport} = \beta_0 + \beta_1 \text{ AttendedUniversity}$$

in the full population data, the coefficients will be:

Thought experiment (2)

Suppose we know that

Support for Leave

57.9% of those who did not attend university

41.5% of those who did attend university

Thus if we ran the regression

$$\text{LeaveSupport} = \beta_0 + \beta_1 \text{ AttendedUniversity}$$

in the full population data, the coefficients will be:

$$\beta_0 = .579, \beta_1 = -.164$$

Thought experiment (2)

Suppose we know that

Support for Leave

57.9% of those who did not attend university

41.5% of those who did attend university

Thus if we ran the regression

$$\text{LeaveSupport} = \beta_0 + \beta_1 \text{ AttendedUniversity}$$

in the full population data, the coefficients will be:

$$\beta_0 = .579, \beta_1 = -.164$$

But what if we draw a random sample and run this regression in our sample? How far off might the coefficients be?

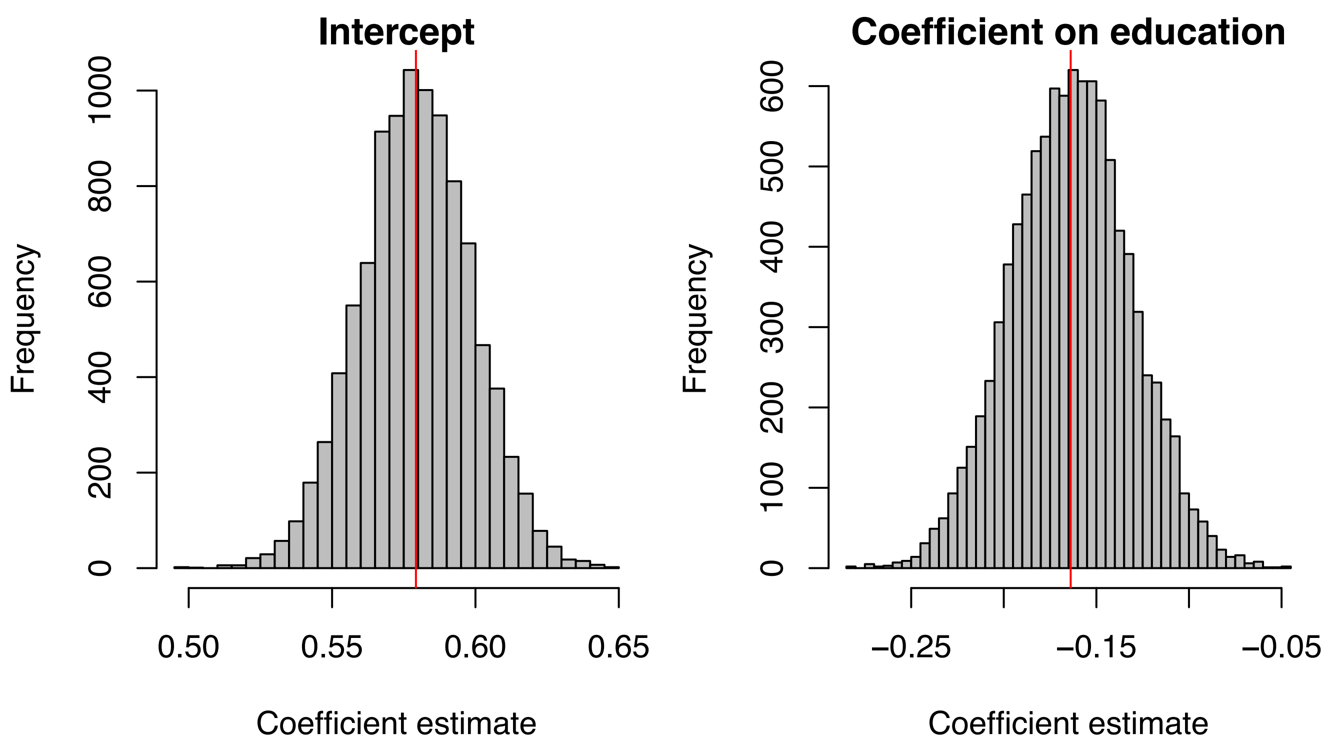
Thought experiment (2.1)

Thought experiment (2.1)

Across 10,000 simulated samples of size 1,006, the histograms for the two coefficients look like:

Thought experiment (2.1)

Across 10,000 simulated samples of size 1,006, the histograms for the two coefficients look like:



Thought experiment (2.2)

Thought experiment (2.2)

We can calculate the standard deviation of the coefficients across simulations:

Thought experiment (2.2)

We can calculate the standard deviation of the coefficients across simulations:

```
> apply(coef.mat, 2, sd)
 Intercept      Slope
0.01943883 0.03216419
```

Thought experiment (2.2)

We can calculate the standard deviation of the coefficients across simulations:

```
> apply(coef.mat, 2, sd)
 Intercept      Slope
0.01943883 0.03216419
```

I stored the estimates in a matrix called `coef.mat`. This command says "calculate the standard deviation of the columns".

Thought experiment (2.2)

We can calculate the standard deviation of the coefficients across simulations:

```
> apply(coef.mat, 2, sd)
 Intercept      Slope
0.01943883 0.03216419
```

I stored the estimates in a matrix called `coef.mat`. This command says "calculate the standard deviation of the columns".

Again, we call these **standard errors.**

Thought experiment (2.2)

We can calculate the standard deviation of the coefficients across simulations:

```
> apply(coef.mat, 2, sd)
 Intercept      Slope
0.01943883 0.03216419
```

I stored the estimates in a matrix called `coef.mat`. This command says "calculate the standard deviation of the columns".

Again, we call these **standard errors**.

As with the simple polling case, we can also use some statistical theory to estimate the standard errors given a sample (i.e. without doing a simulation).

Standard errors in regression output

Standard errors in regression output

Output from a regression for one sample:

Standard errors in regression output

Output from a regression for one sample:

```
> summary(lm(support.leave[indices.to.sample] ~ attended.uni[indices.to.sample]))

Call:
lm(formula = support.leave[indices.to.sample] ~ attended.uni[indices.to.sample])

Residuals:
    Min      1Q  Median      3Q     Max 
-0.5987 -0.5987  0.4013  0.4013  0.5550 

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)    
(Intercept)                      0.59874    0.01960  30.544  < 2e-16 ***
attended.uni[indices.to.sample] -0.15370    0.03219  -4.774 2.07e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4932 on 1004 degrees of freedom
Multiple R-squared:  0.0222,Adjusted R-squared:  0.02123 
F-statistic: 22.79 on 1 and 1004 DF,  p-value: 2.071e-06
```

Standard errors in regression output

Output from a regression for one sample:

```
> summary(lm(support.leave[indices.to.sample] ~ attended.uni[indices.to.sample]))

Call:
lm(formula = support.leave[indices.to.sample] ~ attended.uni[indices.to.sample])

Residuals:
    Min      1Q  Median      3Q     Max 
-0.5987 -0.5987  0.4013  0.4013  0.5550 

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)    
(Intercept)                      0.59874    0.01960  30.544  < 2e-16 ***
attended.uni[indices.to.sample] -0.15370    0.03219  -4.774 2.07e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4932 on 1004 degrees of freedom
Multiple R-squared:  0.0222,Adjusted R-squared:  0.02123 
F-statistic: 22.79 on 1 and 1004 DF,  p-value: 2.071e-06
```

Very close to our estimates of the standard errors from simulation.

The most important use of regression standard errors: hypothesis testing

The most important use of regression standard errors: hypothesis testing

Suppose the coefficient on *AttendedUniversity* in your sample is -0.154, as in this regression output.

The most important use of regression standard errors: hypothesis testing

Suppose the coefficient on *AttendedUniversity* in your sample is -0.154, as in this regression output.

We want to know: have you proven conclusively that university attendance is related to Brexit support *in the population*, or might it just be a fluke *in your sample*?

The most important use of regression standard errors: hypothesis testing

Suppose the coefficient on *AttendedUniversity* in your sample is -0.154, as in this regression output.

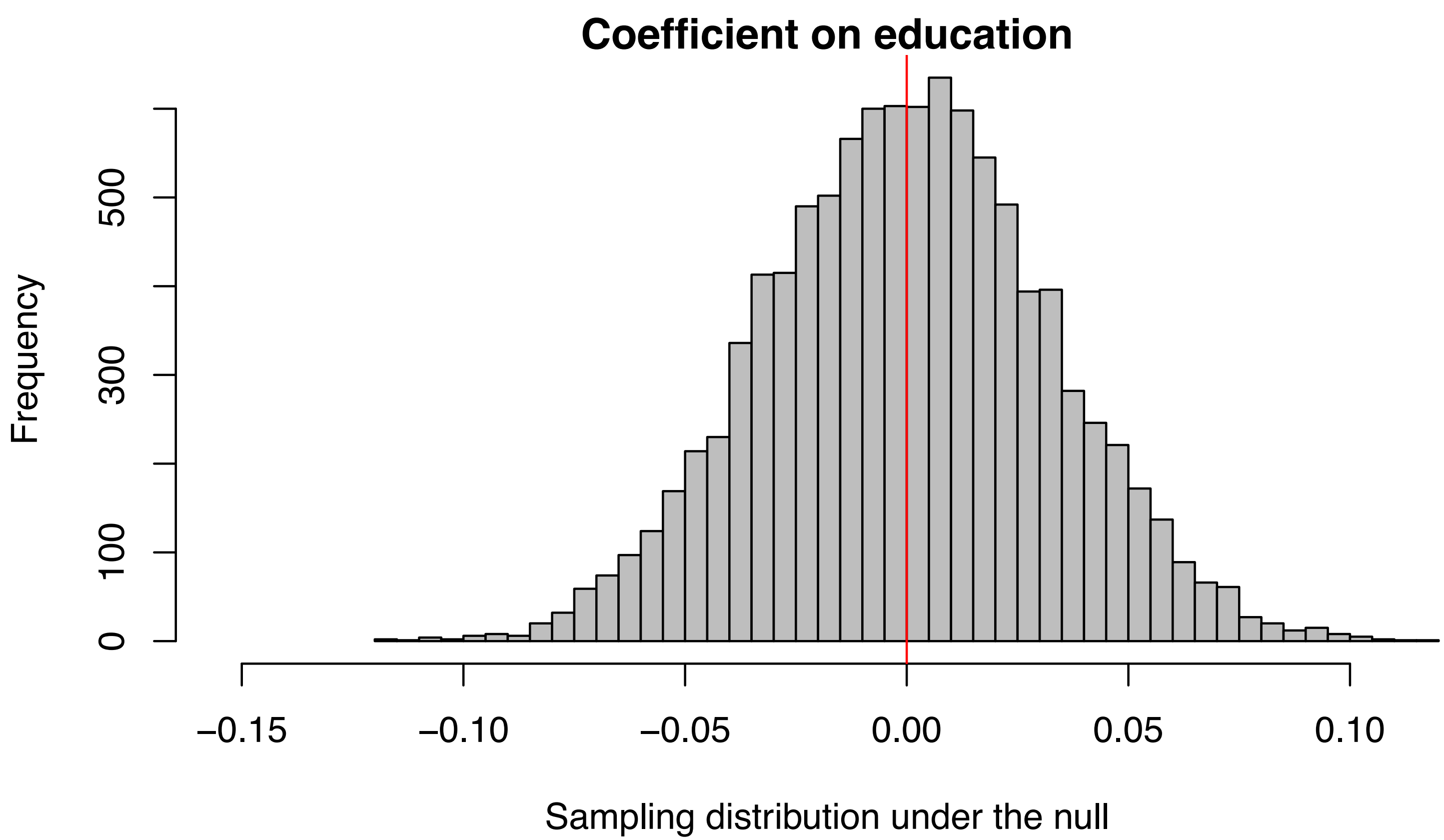
We want to know: have you proven conclusively that university attendance is related to Brexit support *in the population*, or might it just be a fluke *in your sample*?

Put differently: How likely is it that you would get a coefficient that far from 0 in your sample if the true coefficient were in fact 0?

Hypothesis testing for regression coefficients (2)

Hypothesis testing for regression coefficients (2)

The sampling distribution of the coefficient on AttendedUni, **if the true coefficient were 0**, would be something like



Hypothesis testing for regression coefficients (3)

Hypothesis testing for regression coefficients (3)

So what's the probability of getting a sample that yields a coefficient as far from zero as your estimate? (p-value)

Hypothesis testing for regression coefficients (3)

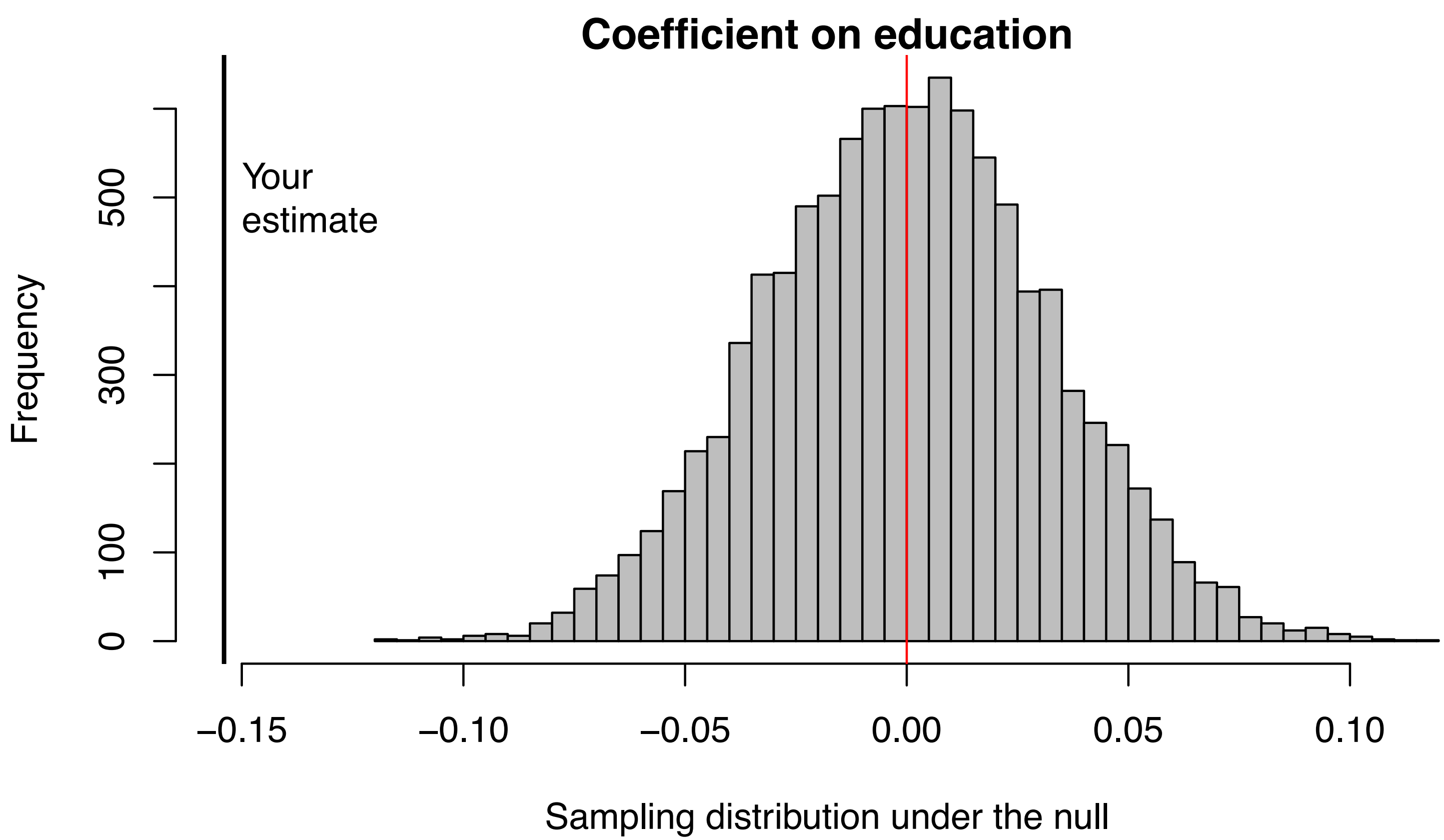
So what's the probability of getting a sample that yields a coefficient as far from zero as your estimate? (p-value)

Basically zero in this case!

Hypothesis testing for regression coefficients (3)

So what's the probability of getting a sample that yields a coefficient as far from zero as your estimate? (p-value)

Basically zero in this case!



Hypothesis testing in regression output

Hypothesis testing in regression output

Output from a regression for one sample:

Hypothesis testing in regression output

Output from a regression for one sample:

```
> summary(lm(support.leave[indices.to.sample] ~ attended.uni[indices.to.sample]))

Call:
lm(formula = support.leave[indices.to.sample] ~ attended.uni[indices.to.sample])

Residuals:
    Min      1Q  Median      3Q     Max 
-0.5987 -0.5987  0.4013  0.4013  0.5550 

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)    
(Intercept)                      0.59874    0.01960  30.544  < 2e-16 ***
attended.uni[indices.to.sample] -0.15370    0.03219  -4.774 2.07e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4932 on 1004 degrees of freedom
Multiple R-squared:  0.0222,Adjusted R-squared:  0.02123 
F-statistic: 22.79 on 1 and 1004 DF,  p-value: 2.071e-06
```

Hypothesis testing in regression output

Output from a regression for one sample:

```
> summary(lm(support.leave[indices.to.sample] ~ attended.uni[indices.to.sample]))

Call:
lm(formula = support.leave[indices.to.sample] ~ attended.uni[indices.to.sample])

Residuals:
    Min      1Q  Median      3Q     Max 
-0.5987 -0.5987  0.4013  0.4013  0.5550 

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)    
(Intercept)                      0.59874    0.01960  30.544  < 2e-16 ***
attended.uni[indices.to.sample] -0.15370    0.03219  -4.774 2.07e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4932 on 1004 degrees of freedom
Multiple R-squared:  0.0222,Adjusted R-squared:  0.02123 
F-statistic: 22.79 on 1 and 1004 DF,  p-value: 2.071e-06
```

Basically zero.

Now you should understand:

Dependent variable: Nobel Prizes awarded per capita (in log scale)

	(1)	(2)	(3)
Intercept	-1.629* (0.509)	-3.166* (0.511)	-2.982* (0.527)
Chocolate consumption per capita (log scale)	2.092* (0.298)	1.026* (0.326)	0.709 (0.415)
GDP/capita (thousands of USD)		0.105* (0.024)	0.106* (0.024)
NW Europe			0.549 (0.452)
R^2	0.70	0.85	0.86
N	34	34	34

Standard errors in parentheses. * Indicates p<0.05

- what a dependent variable is
- what an independent variable is
- what the coefficients mean (intercept, slopes)
- what the stars mean (i.e. what $p<0.05$ means)
- what the standard errors mean

To consider

To consider

In the thought experiments above, there are standard errors because of **sampling**. But what about when the sample **is** the population? (e.g. Lijphart's analysis of all countries that have been democratic since 1988)

To consider

In the thought experiments above, there are standard errors because of **sampling**. But what about when the sample **is** the population? (e.g. Lijphart's analysis of all countries that have been democratic since 1988)

The broader view is that history offers one sample, but if "re-run" it might have produced another. Less philosophically satisfying!