Bivariate relationships: introduction to regression

Andy Eggers
Assoc. Professor
Department of Politics and International Relations

We want you to understand:

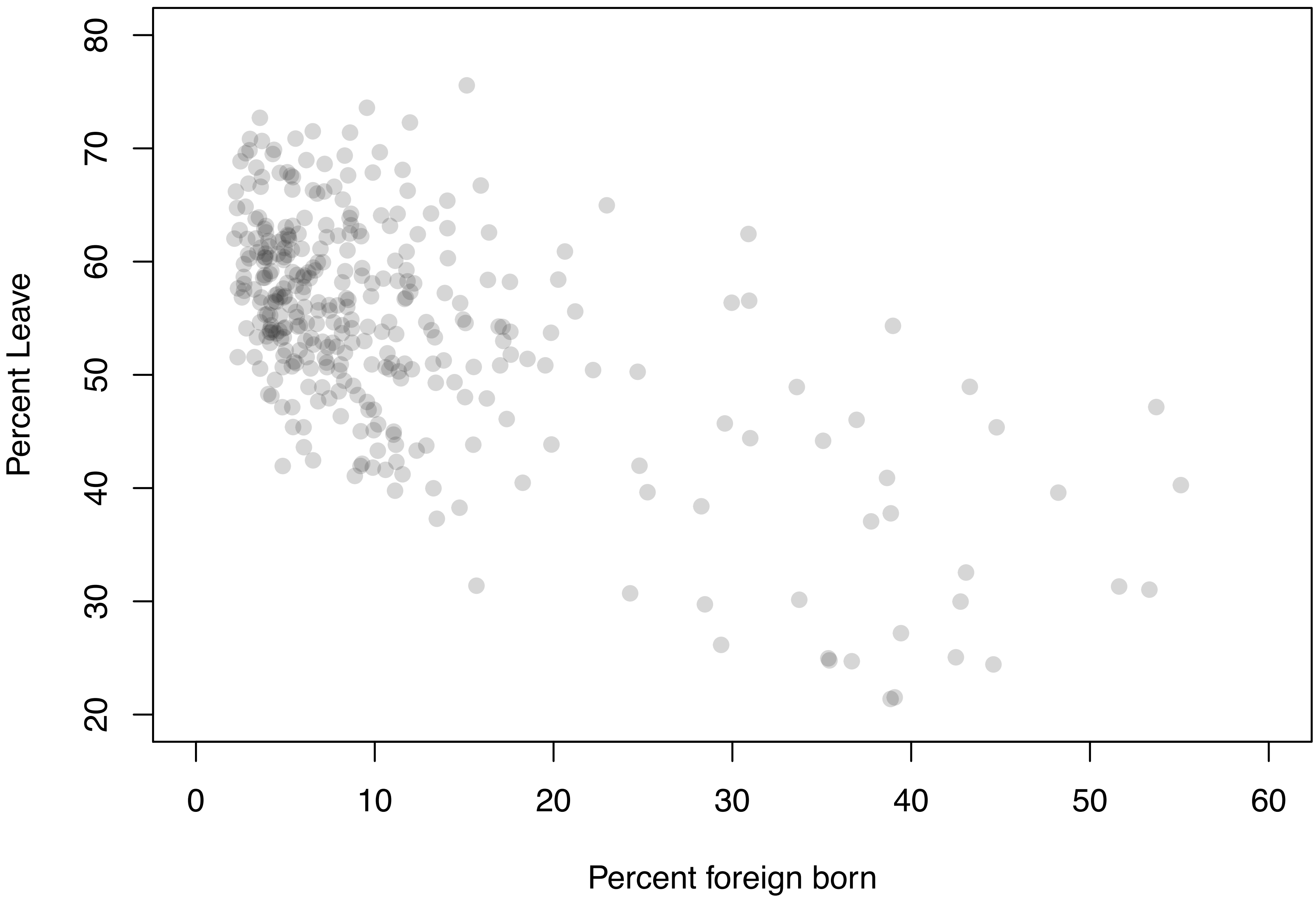
Dependent variable: Nobel Prizes awarded per capita (in log scale)

	(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

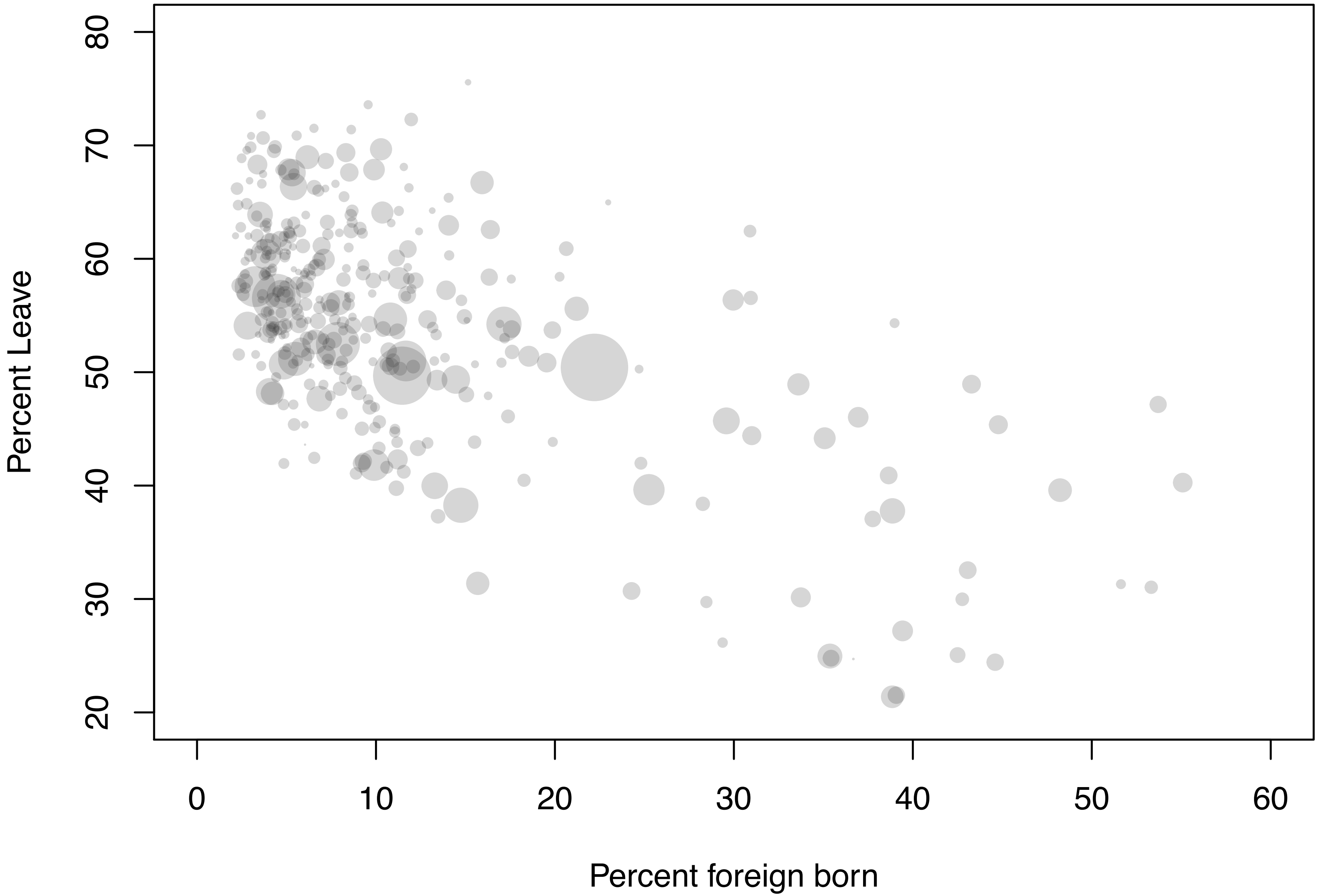
Standard errors in parentheses. * Indicates $p<0.05$

- 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

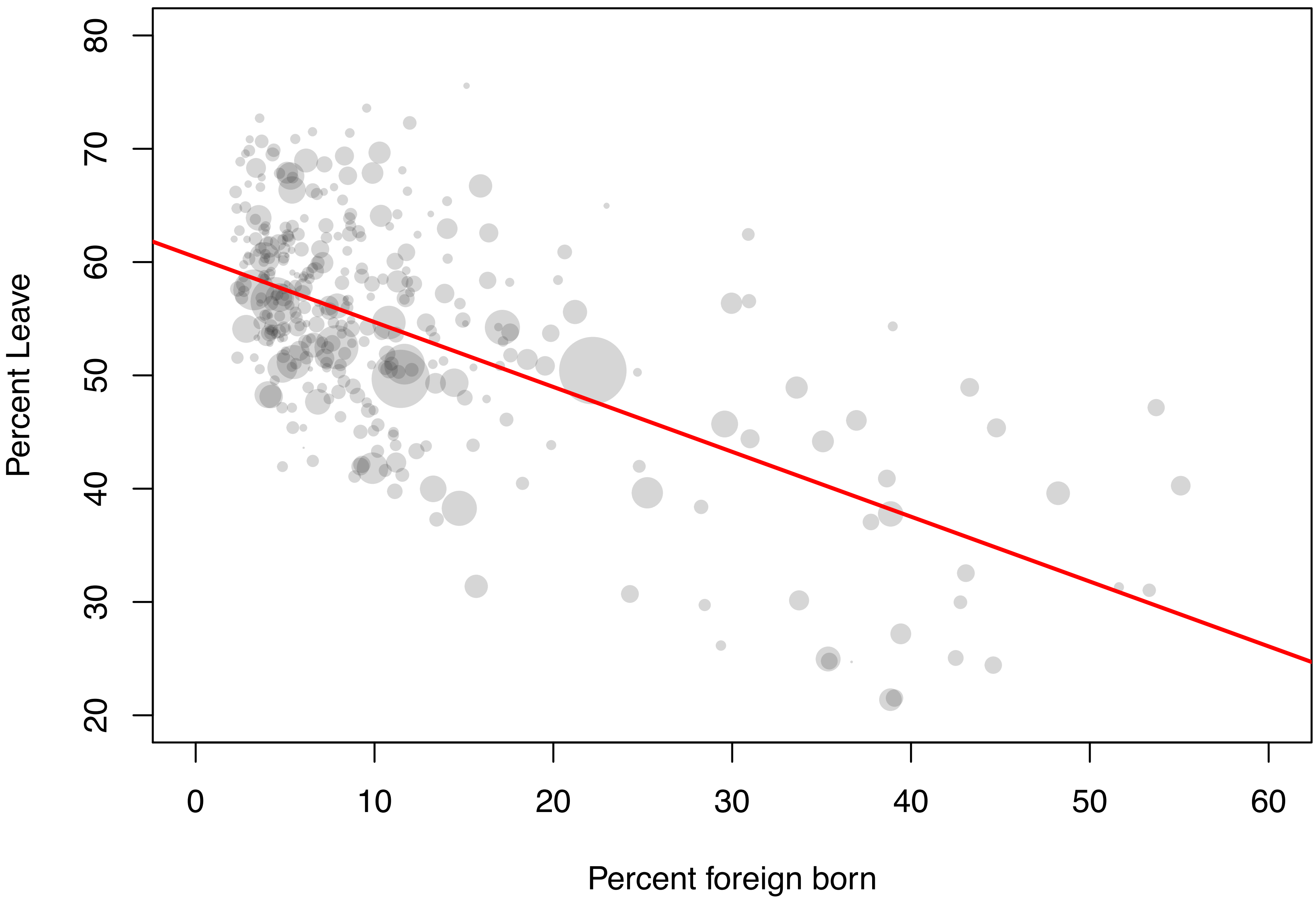
Local authorities with more foreign-born residents were less supportive of Brexit



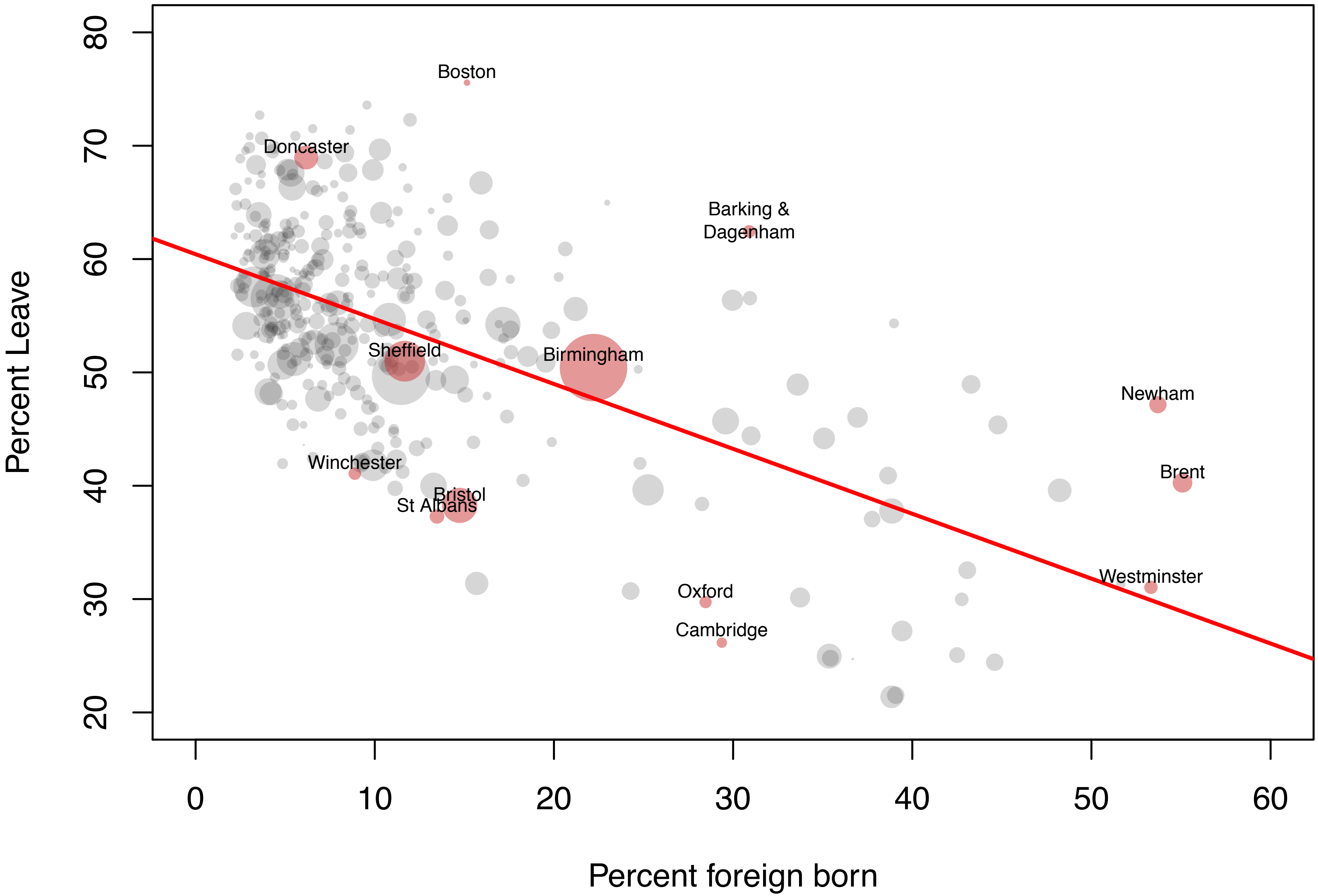
Local authorities with more foreign-born residents were less supportive of Brexit



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)

My lectures in weeks 5, 6, and 7

My lectures in weeks 5, 6, and 7

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

My lectures in weeks 5, 6, and 7

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

Plan:

My lectures in weeks 5, 6, and 7

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

My lectures in weeks 5, 6, and 7

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

My lectures in weeks 5, 6, and 7

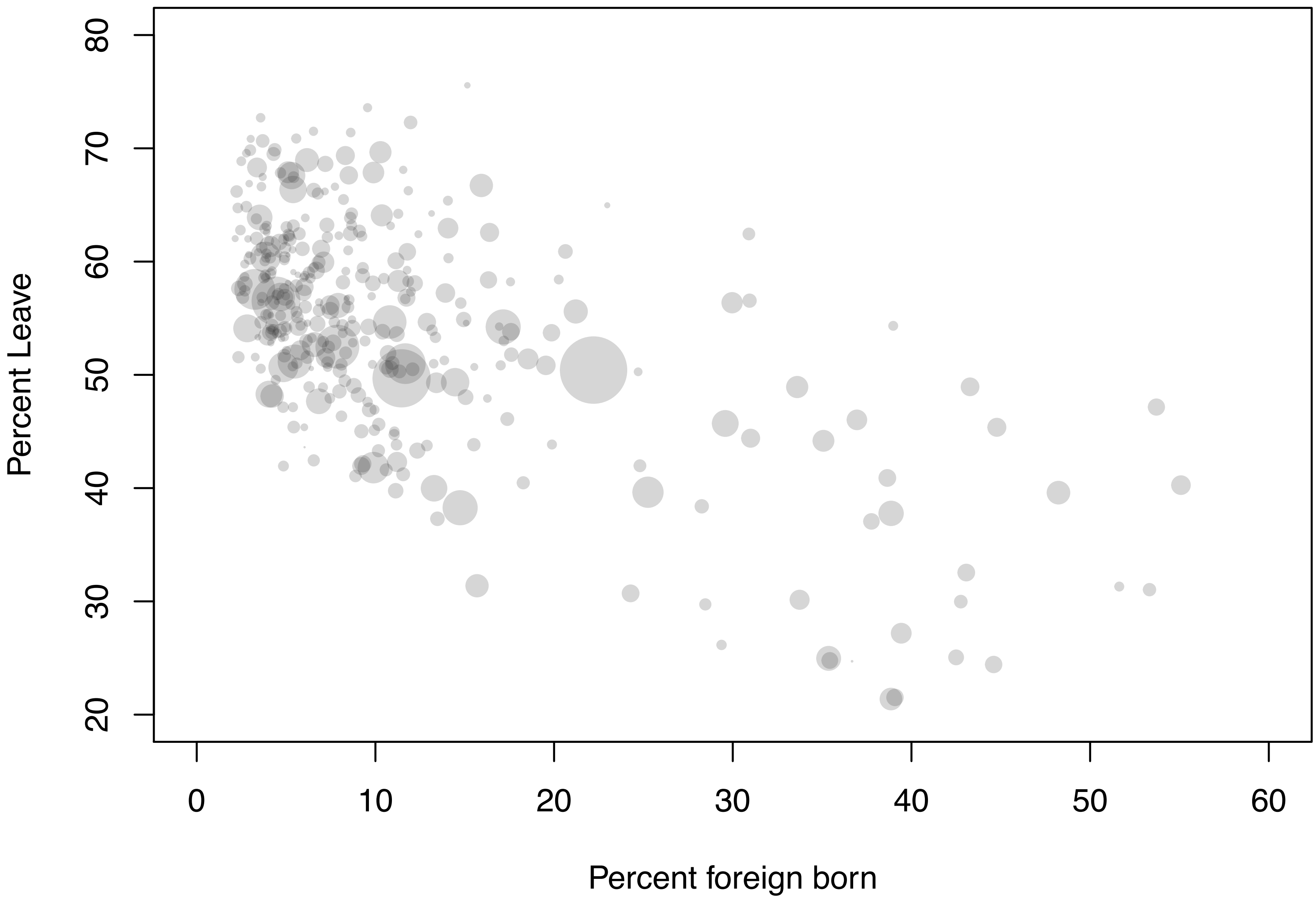
Running question: Why is there such a strong relationship between % foreign born and opposition to Brexit?

Plan:

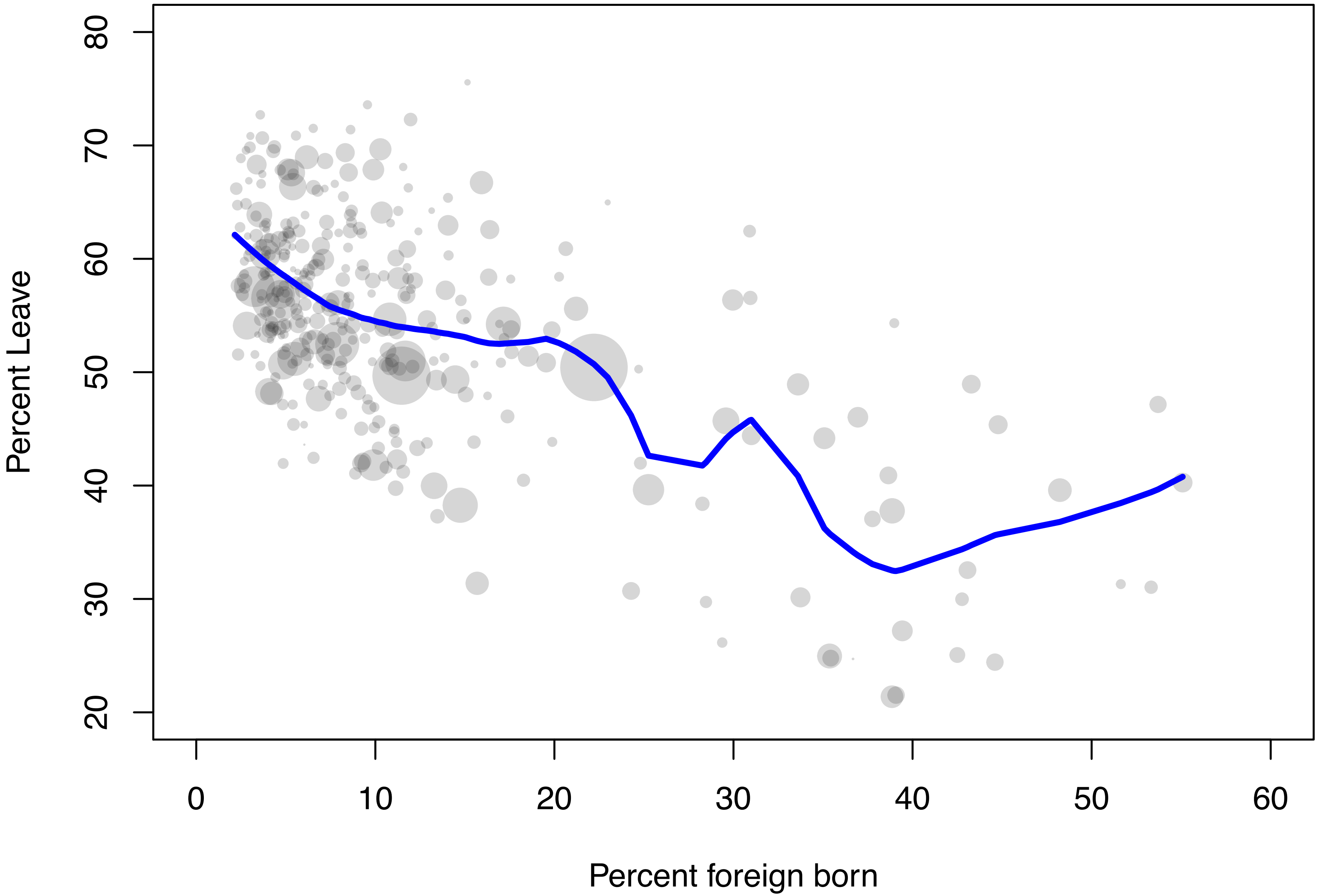
- How do we summarize the relationship between two variables?
  - bivariate OLS regression as main focus
- How do we summarize the relationship between two variables controlling for a third variable?
  - multivariate OLS regression as main focus
- How do we summarize our uncertainty about our conclusions?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

Summarizing bivariate relationships: options other than OLS regression

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



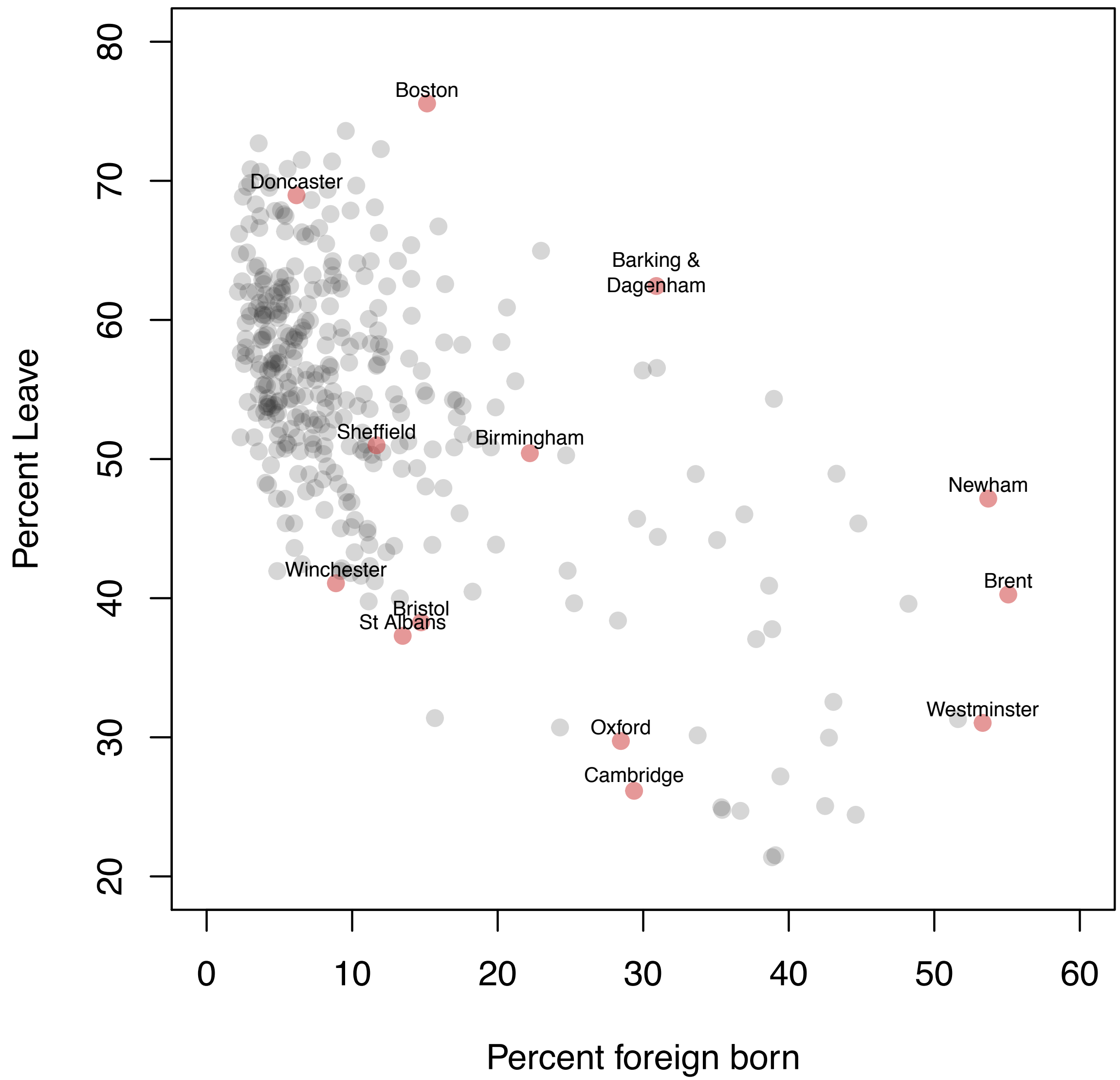
Kernel smoother (lokern function in R)



Single-number summaries: covariance

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

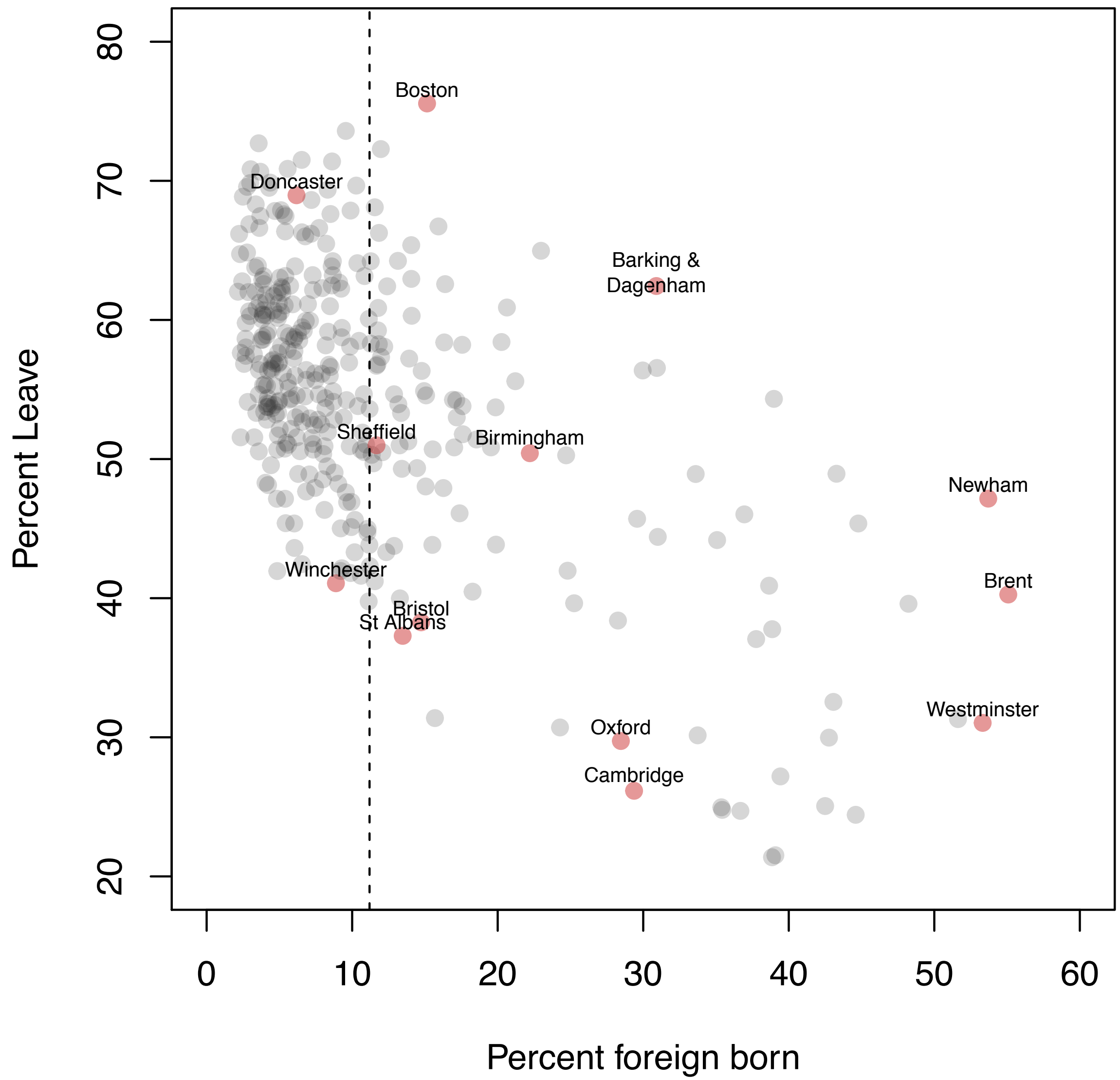
When x is above its mean, is y also above its mean? By how much?



Single-number summaries: covariance

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

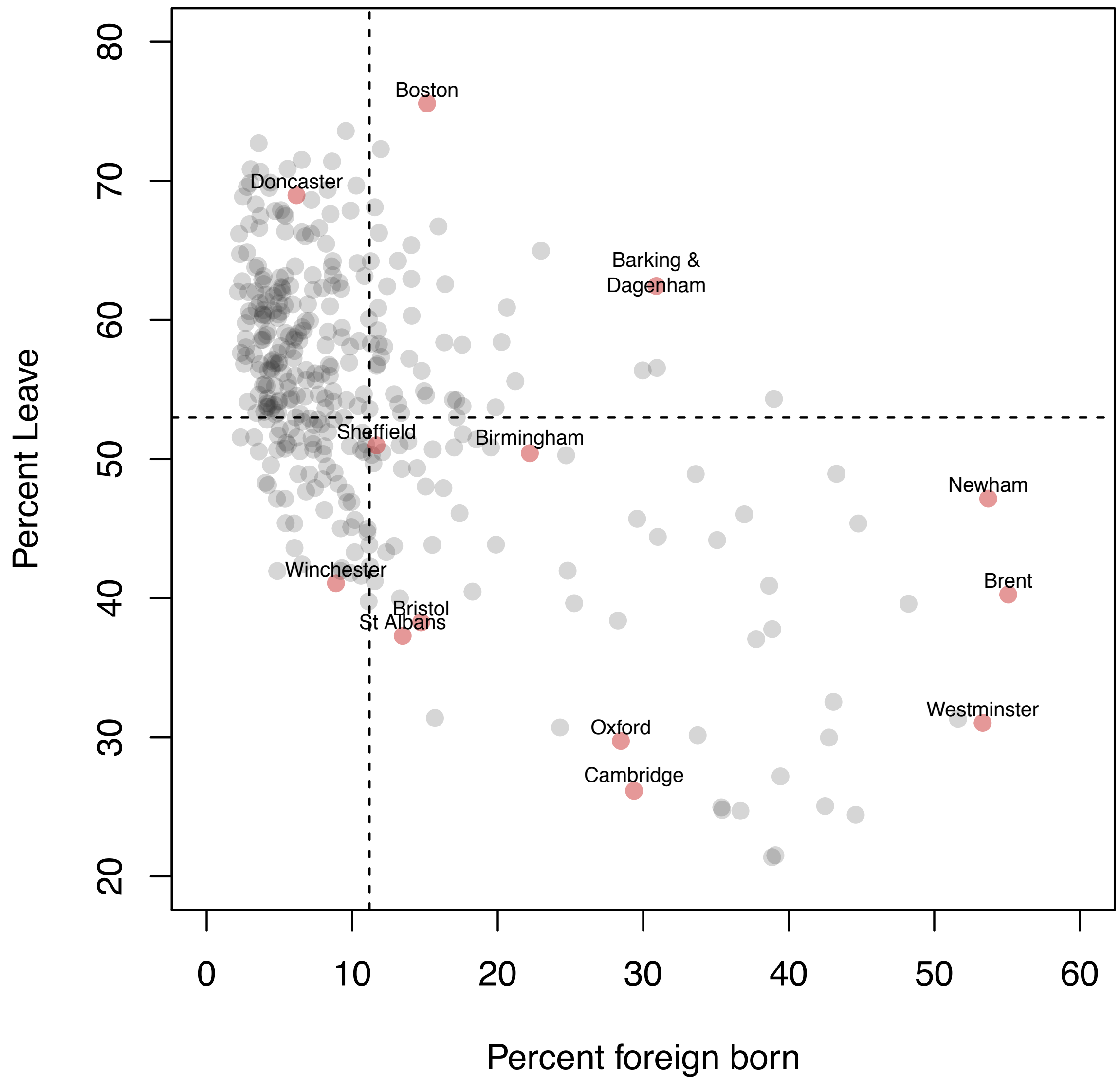
When x is above its mean, is y also above its mean? By how much?



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

Single-number summaries: covariance

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

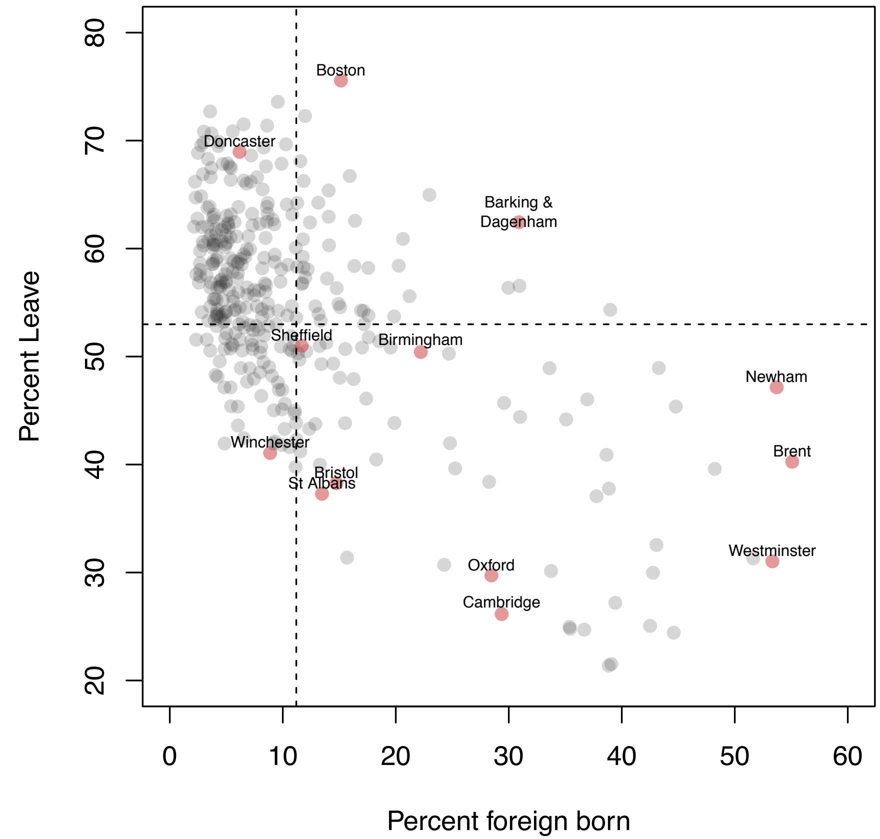
Single-number summaries: covariance

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

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)?

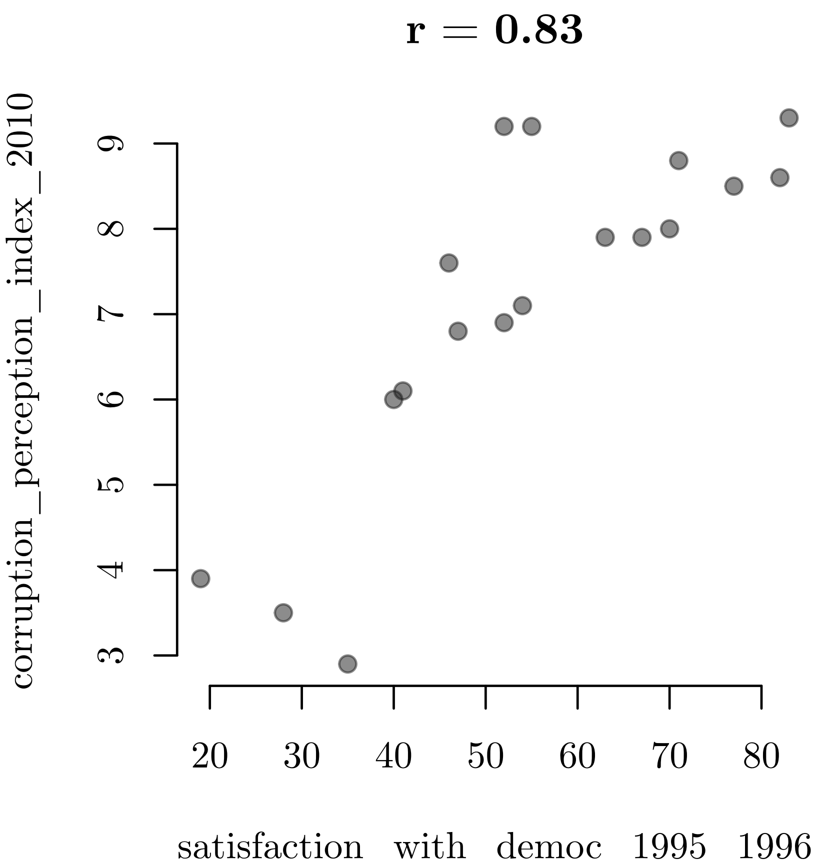
$$\mathrm{Cor}(x,y) = \frac{\mathrm{Cov}(x,y)}{\mathrm{sd}(x)\mathrm{sd}(y)}$$

"standard deviation of y"

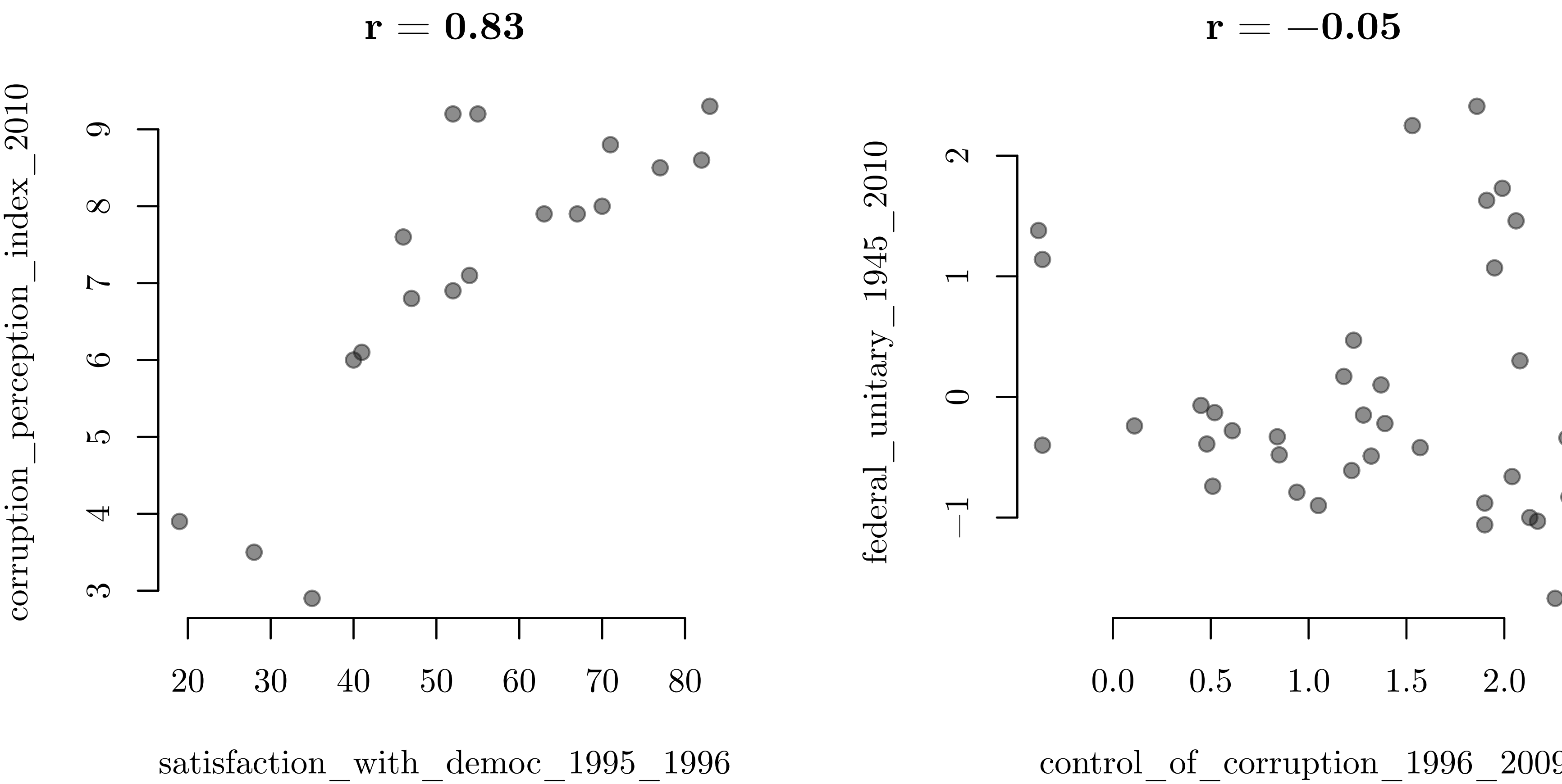
```
> cor(d$Percent_foreign_born, d$Percent_Leave, use = "complete")
[1] -0.6125353
```

Correlation examples from Lijphart's data

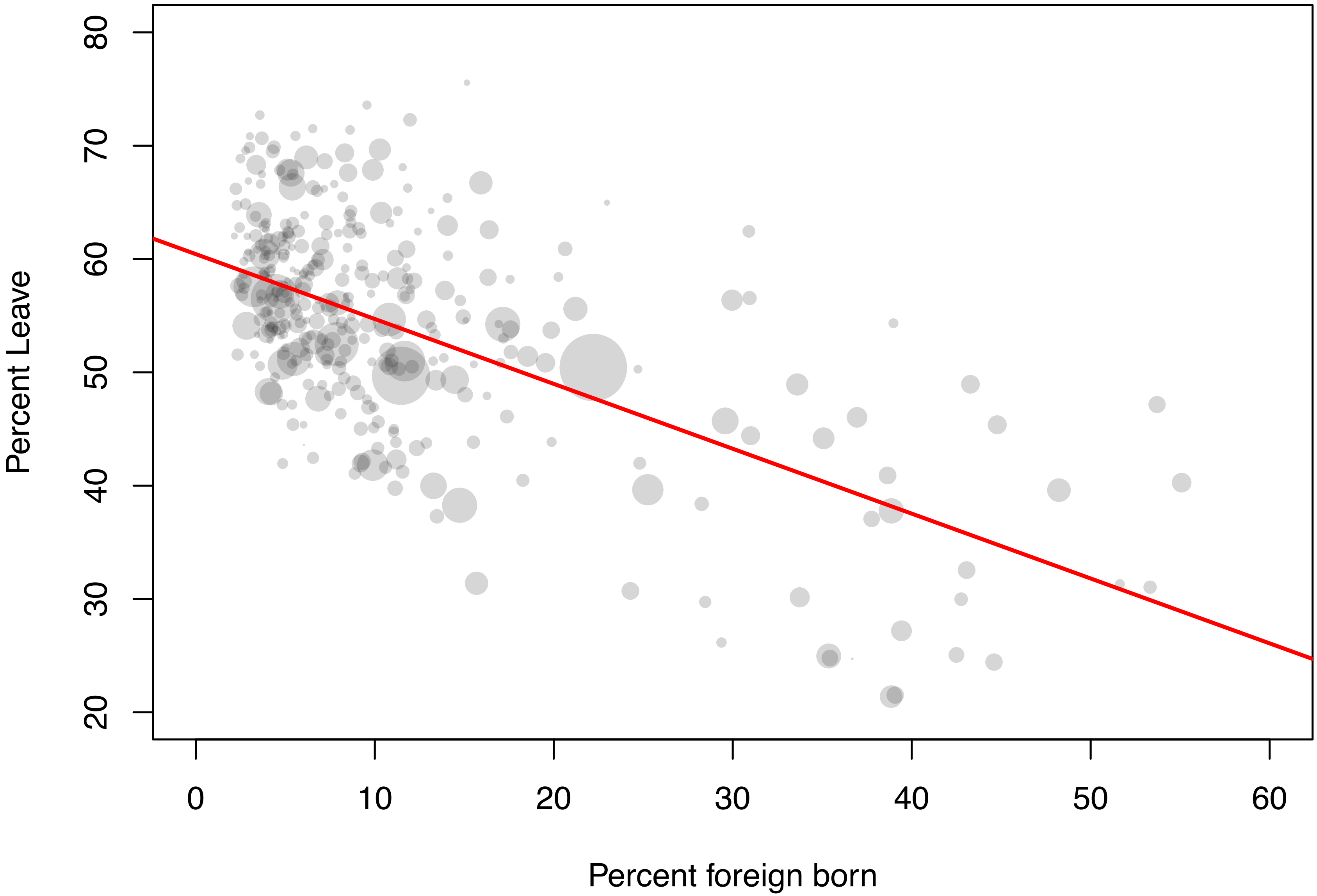
Correlation examples from Lijphart's data



Correlation examples from Lijphart's data

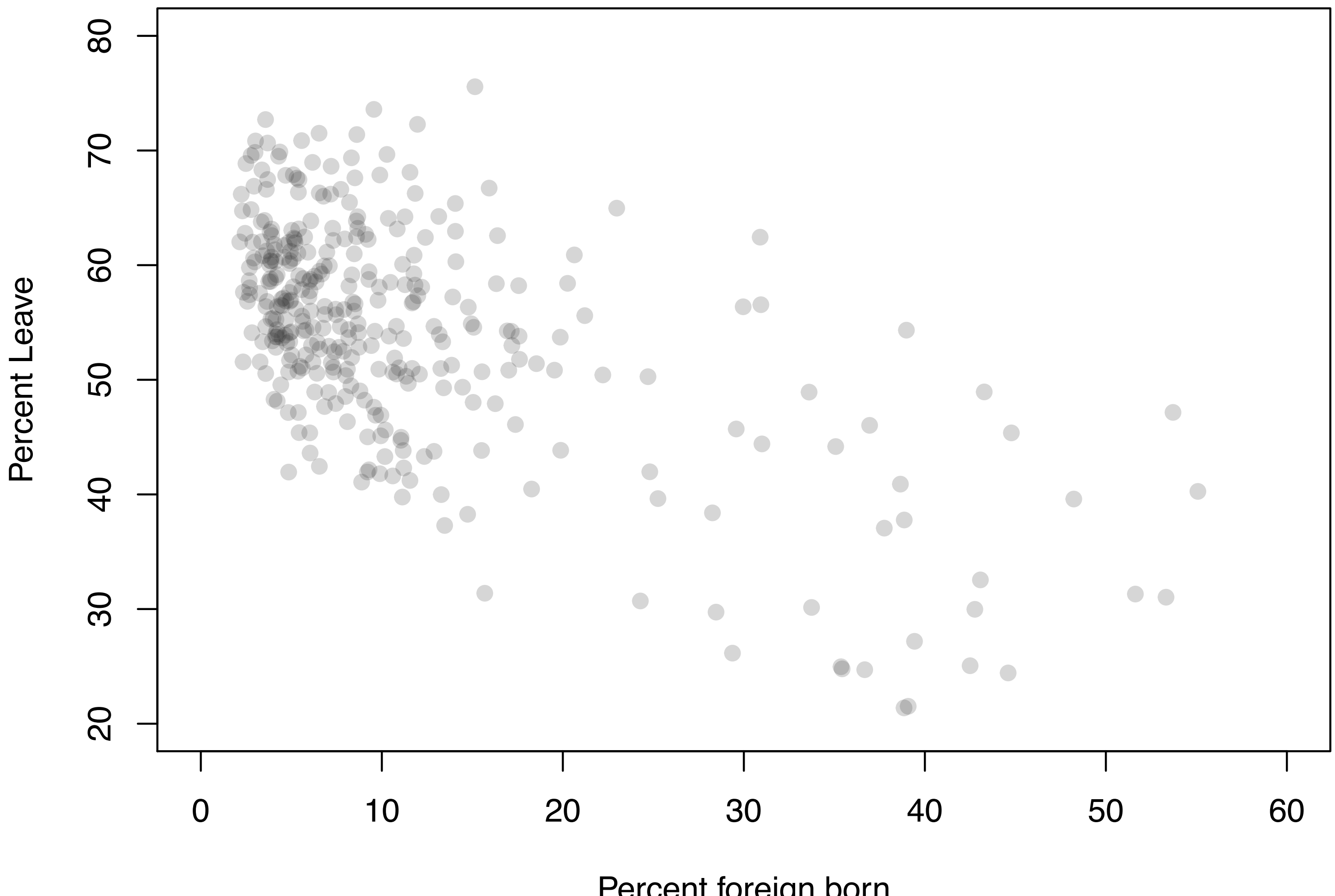


The most important summary: OLS regression



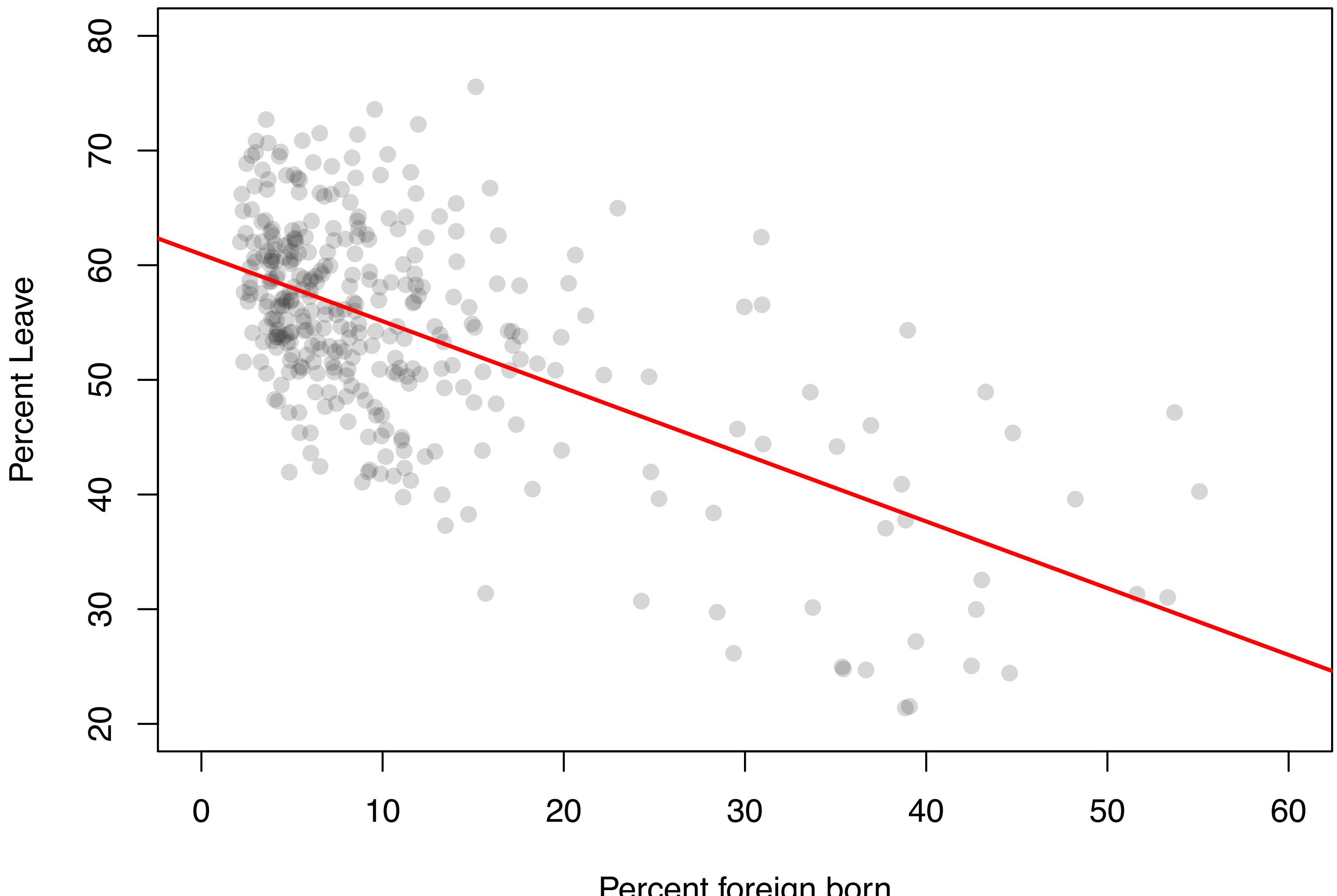
Step 1 for understanding OLS: residuals

A **residual** is the difference between the *actual* y-value and the *predicted* y-value.



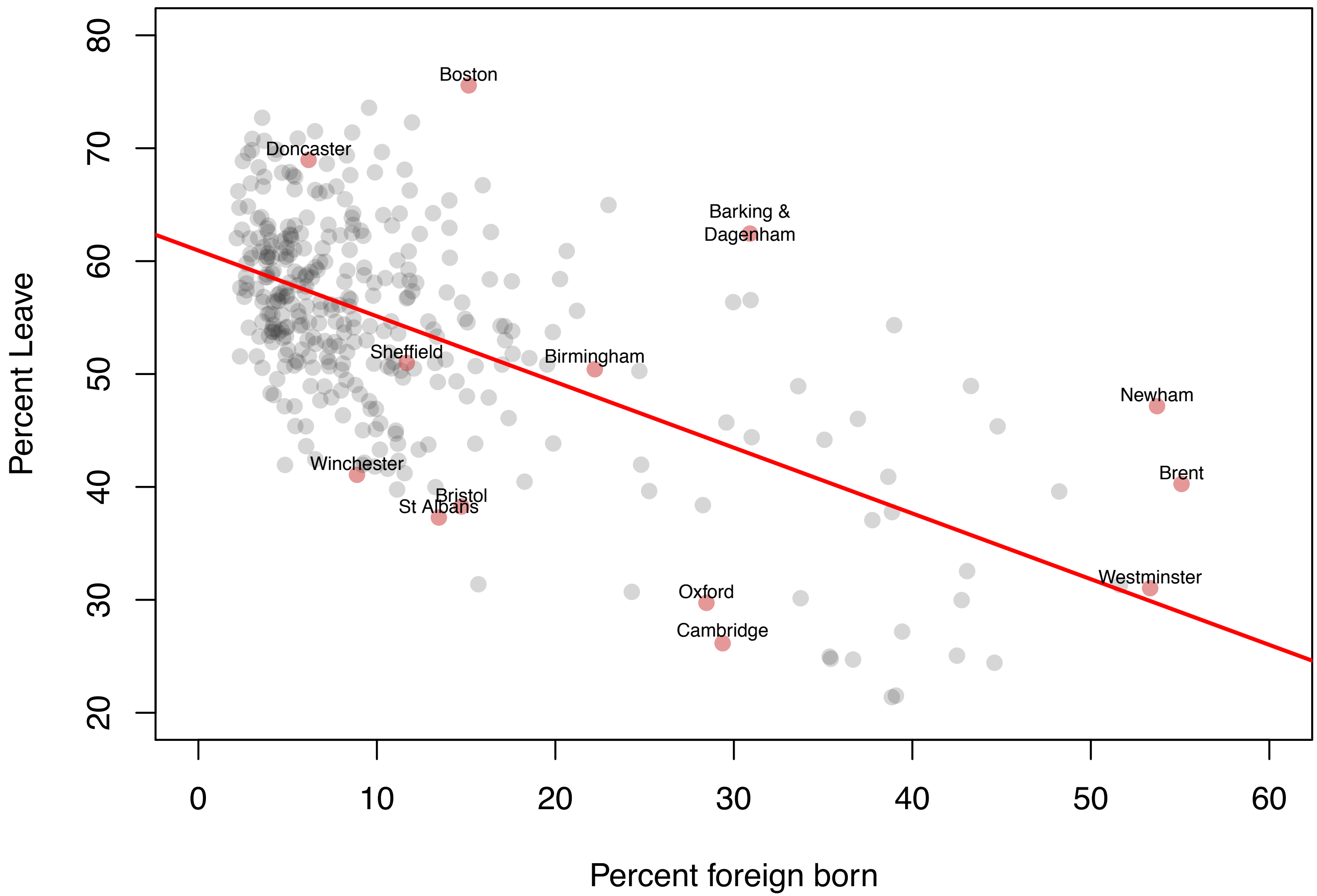
Step 1 for understanding OLS: residuals

A **residual** is the difference between the *actual* y-value and the *predicted* y-value.



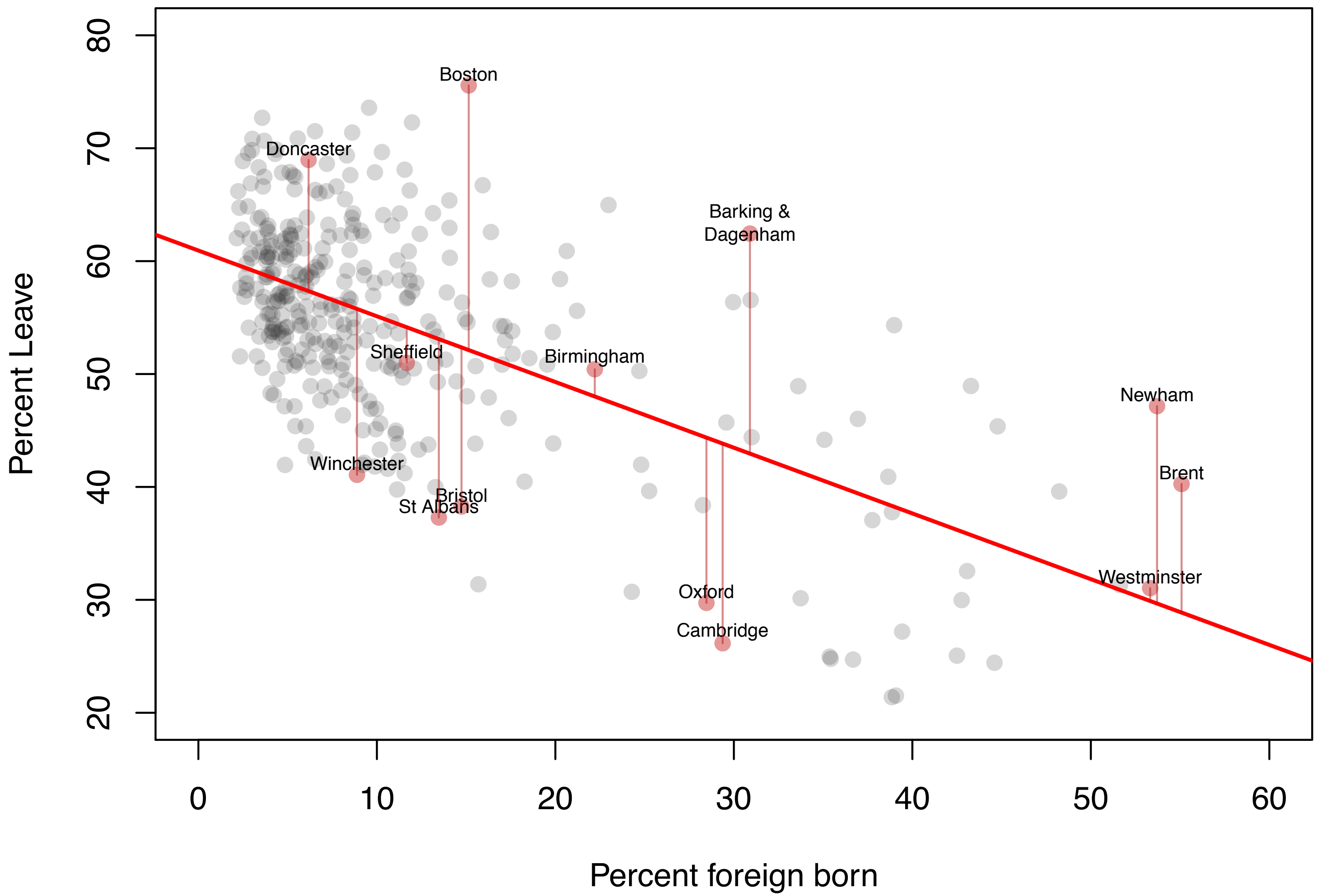
Step 1 for understanding OLS: residuals

A **residual** is the difference between the *actual* y-value and the *predicted* y-value.



Step 1 for understanding OLS: residuals

A **residual** is the difference between the *actual* y-value and the *predicted* y-value.

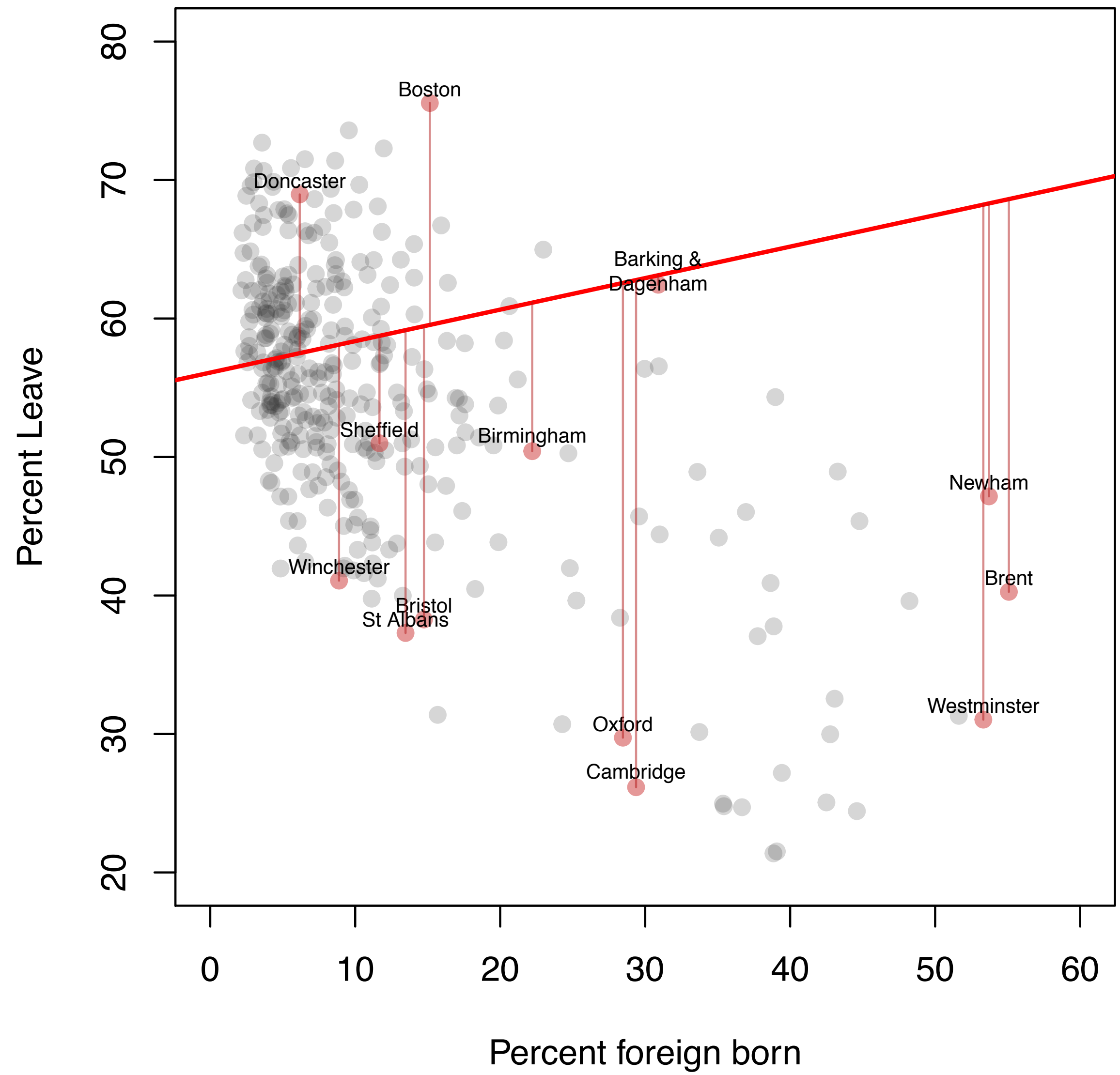


Step 2 for understanding OLS: sum of squared residuals

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

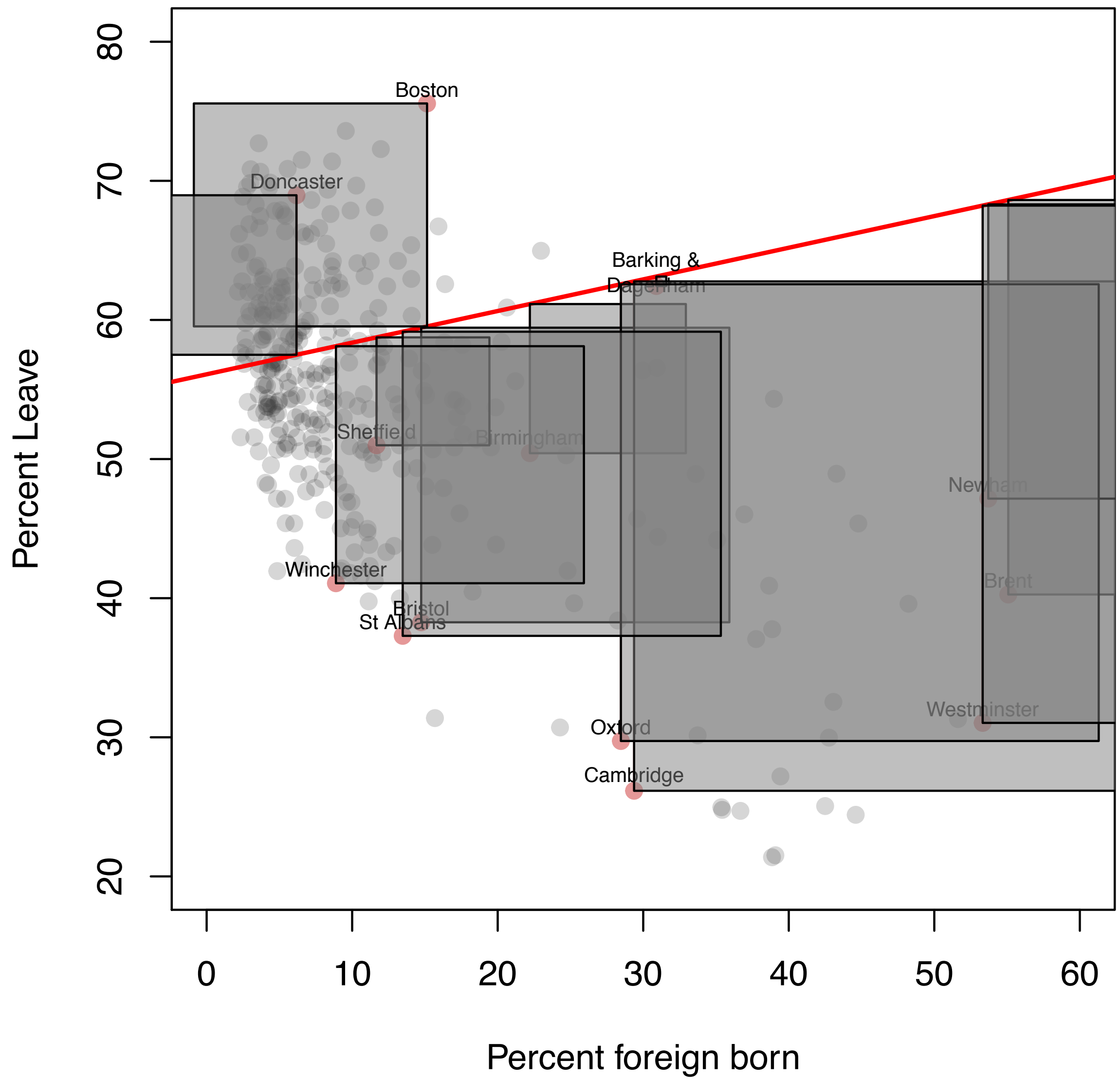
Step 2 for understanding OLS: sum of squared residuals

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



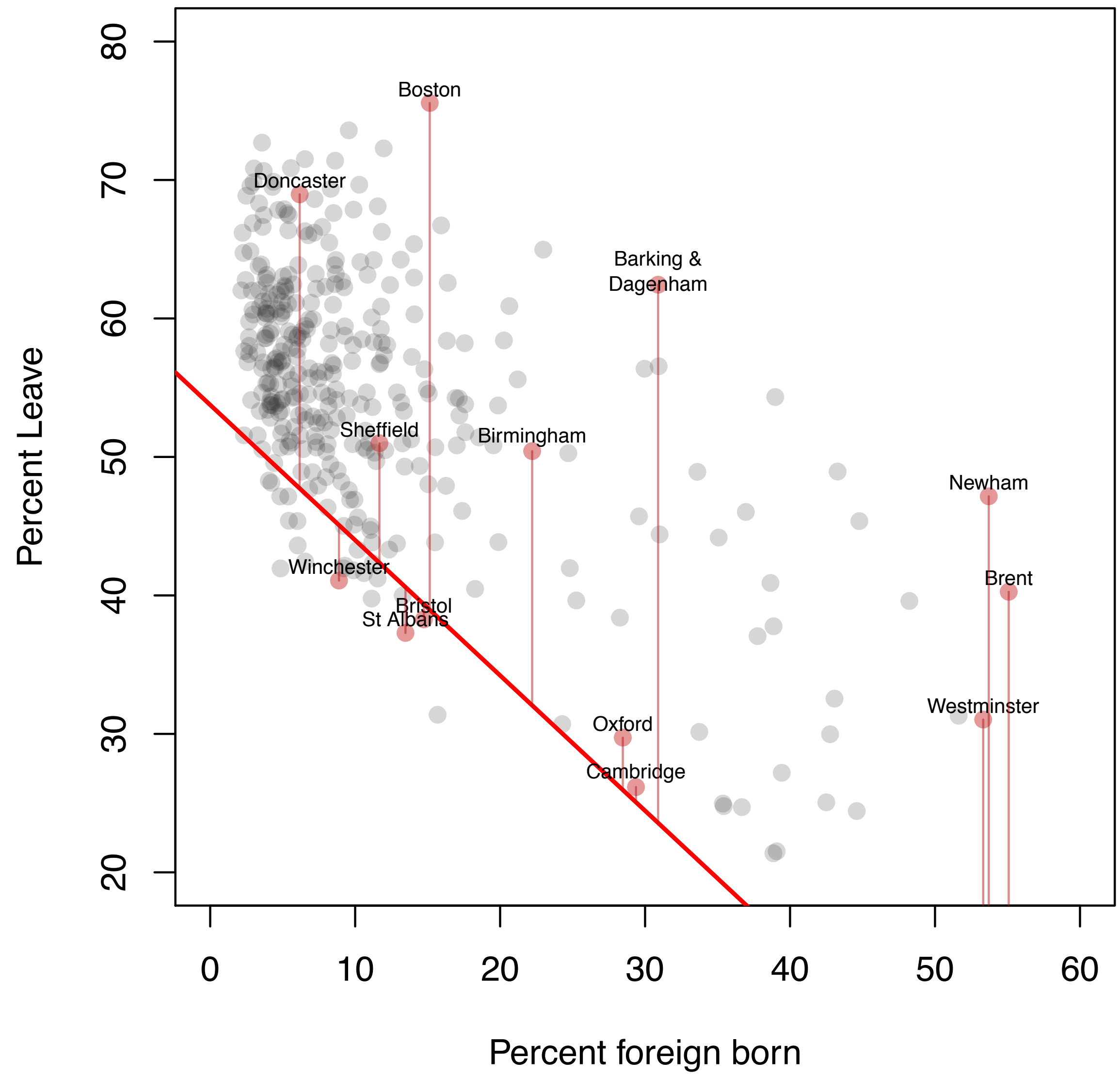
Step 2 for understanding OLS: sum of squared residuals

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



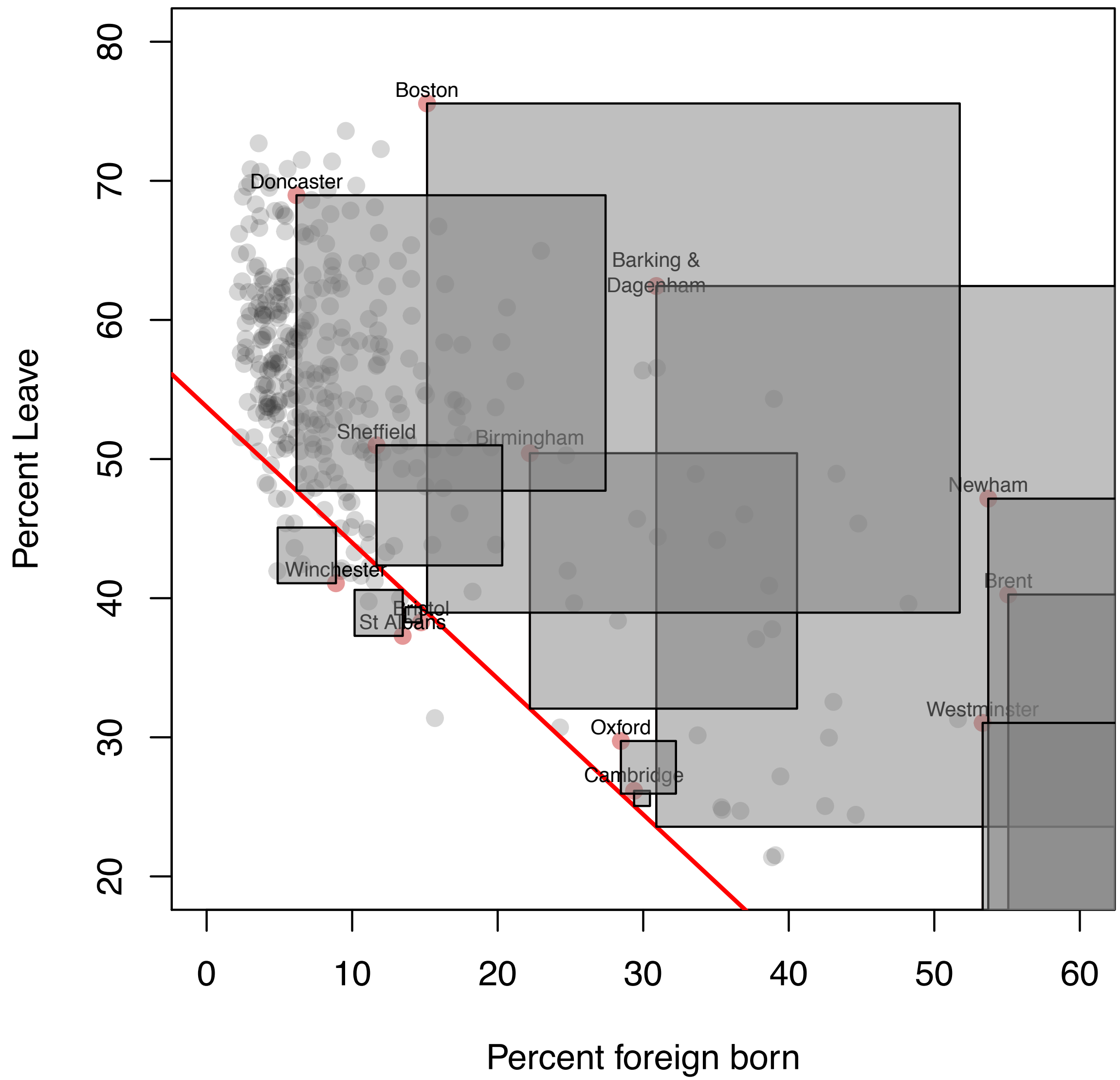
Step 2 for understanding OLS: sum of squared residuals

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



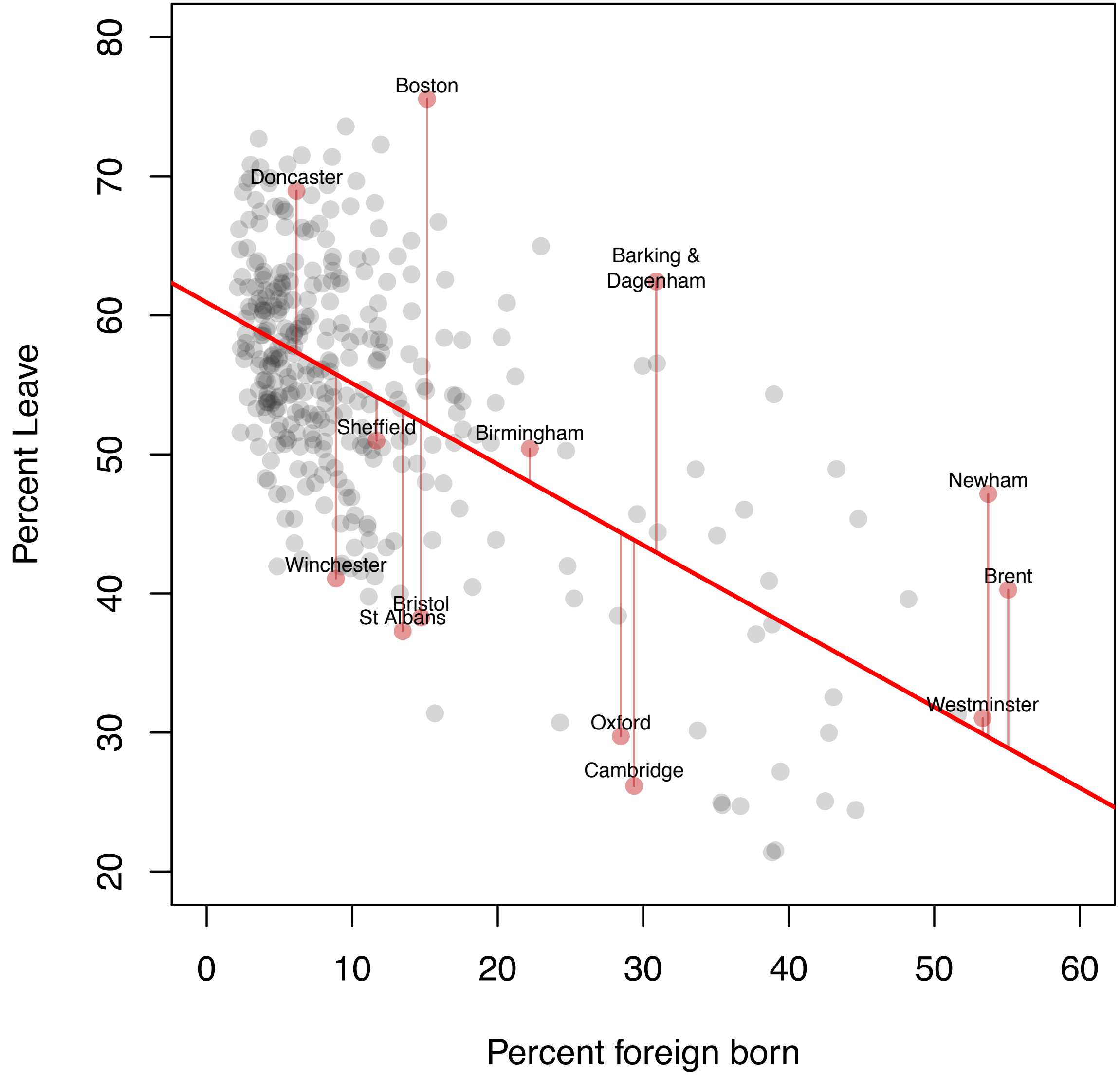
Step 2 for understanding OLS: sum of squared residuals

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



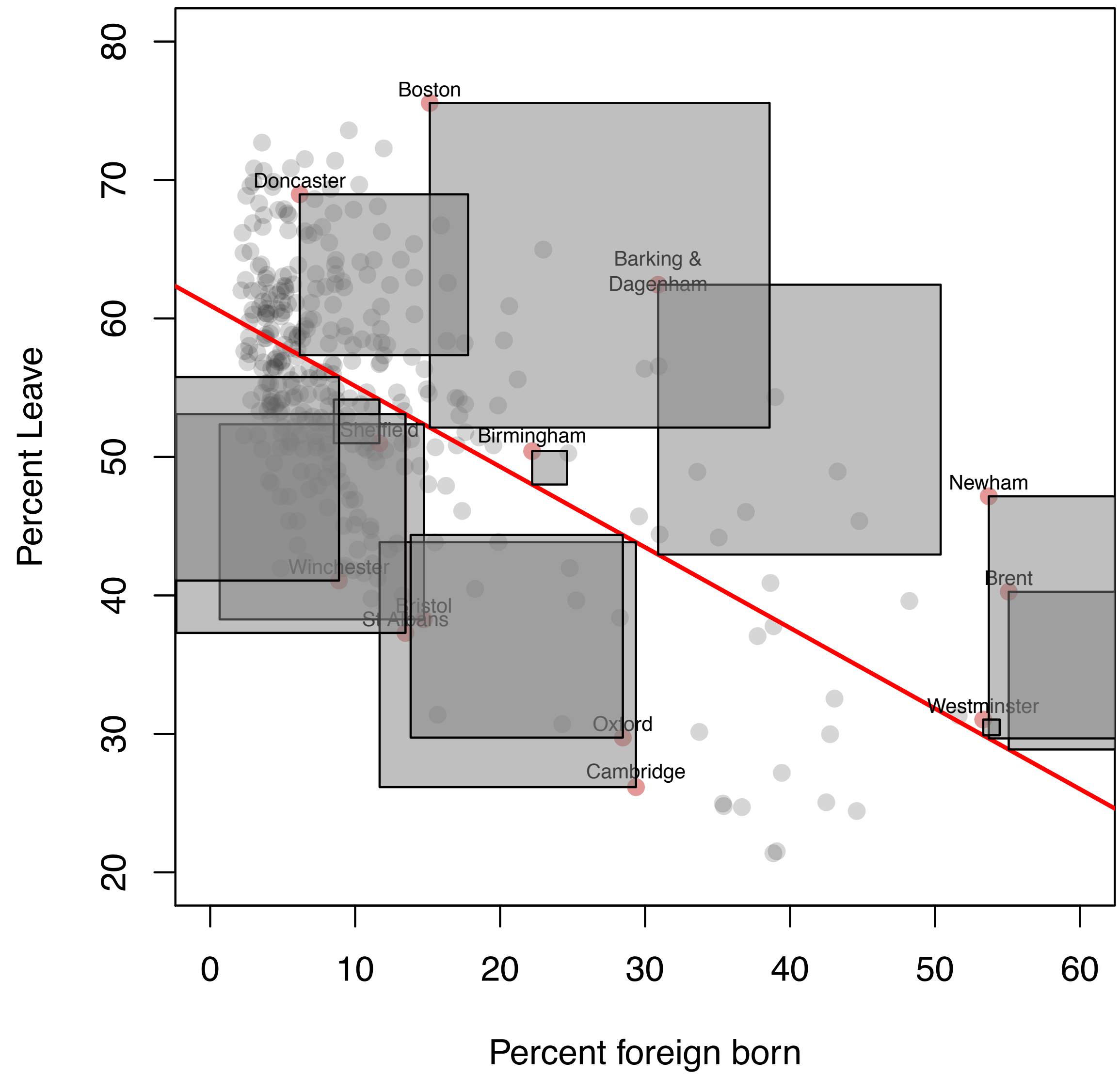
Step 2 for understanding OLS: sum of squared residuals

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



Step 2 for understanding OLS: sum of squared residuals

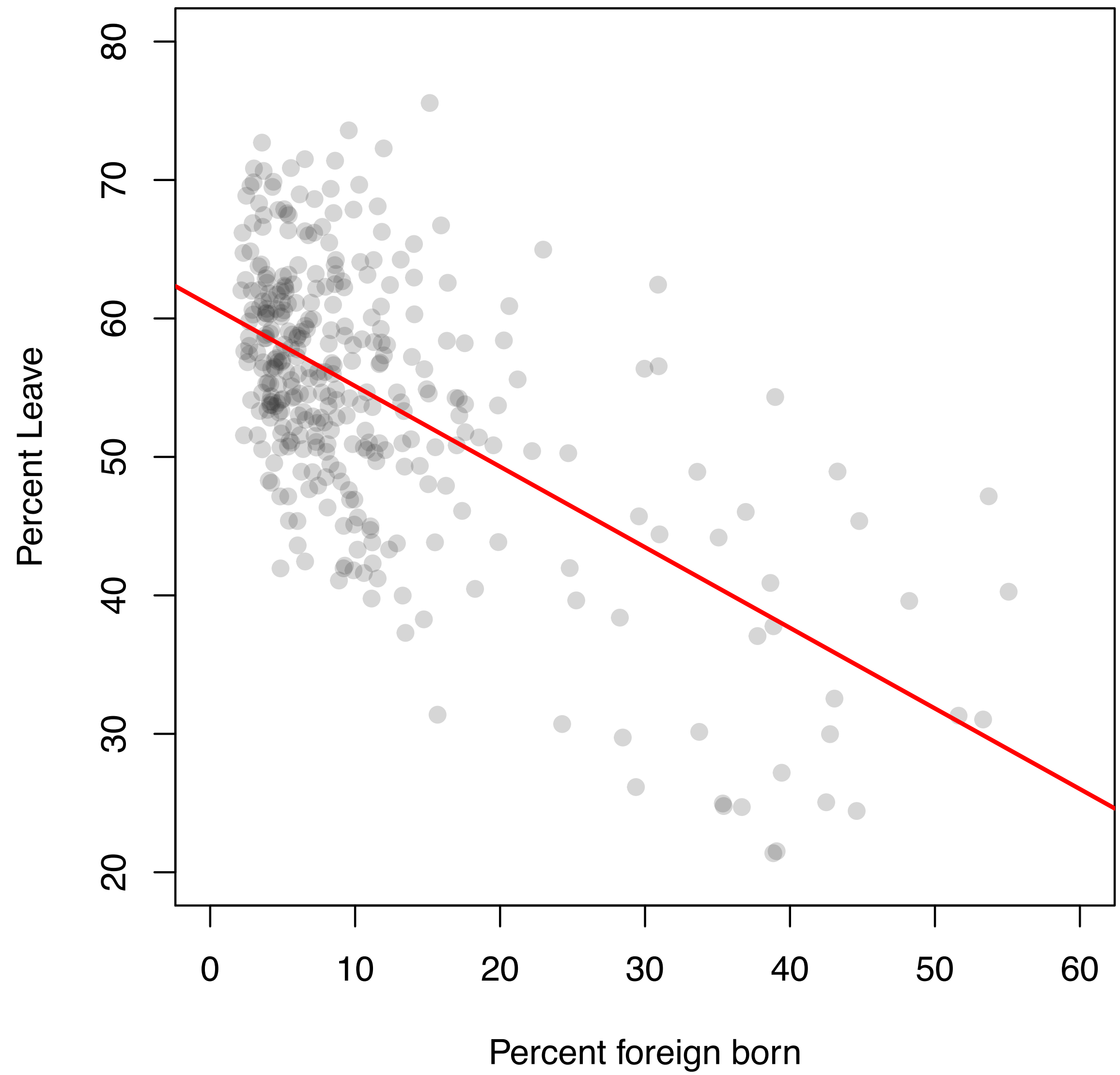
For any prediction line you draw, you can calculate residuals, square them, and sum them.



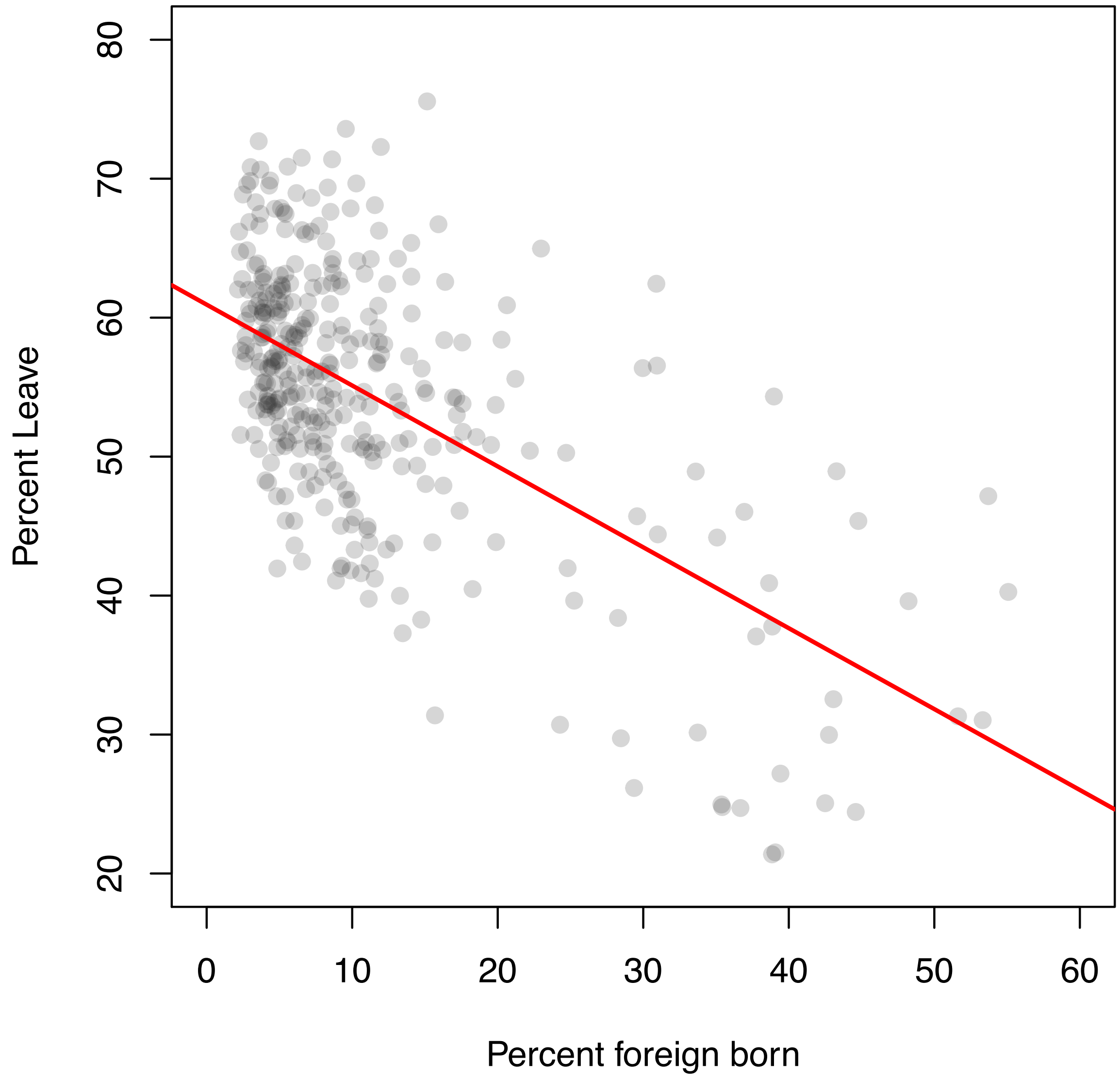
Step 3 for understanding OLS: minimizing the sum of squared residuals

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

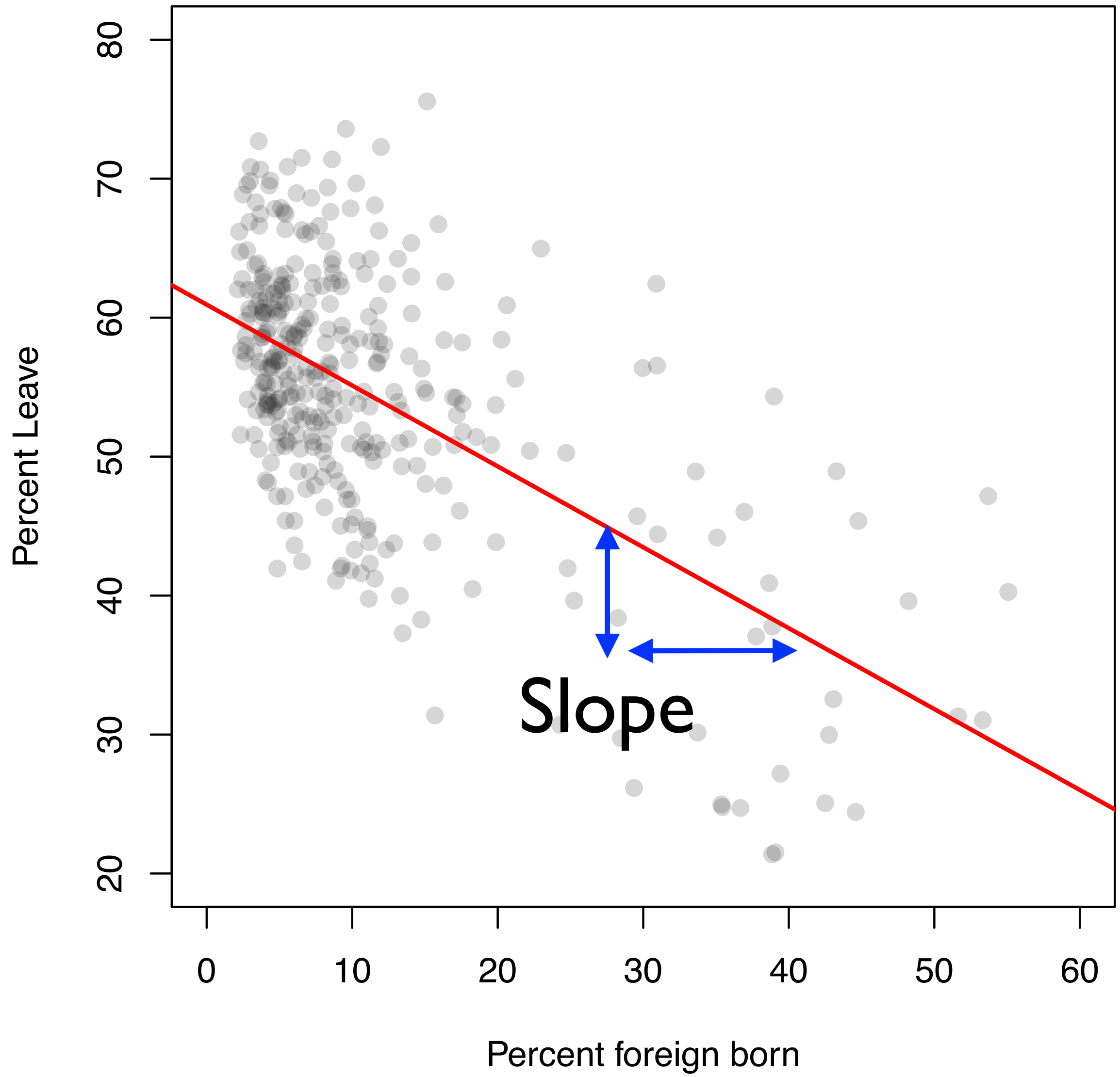
Hence ordinary **least squares**.



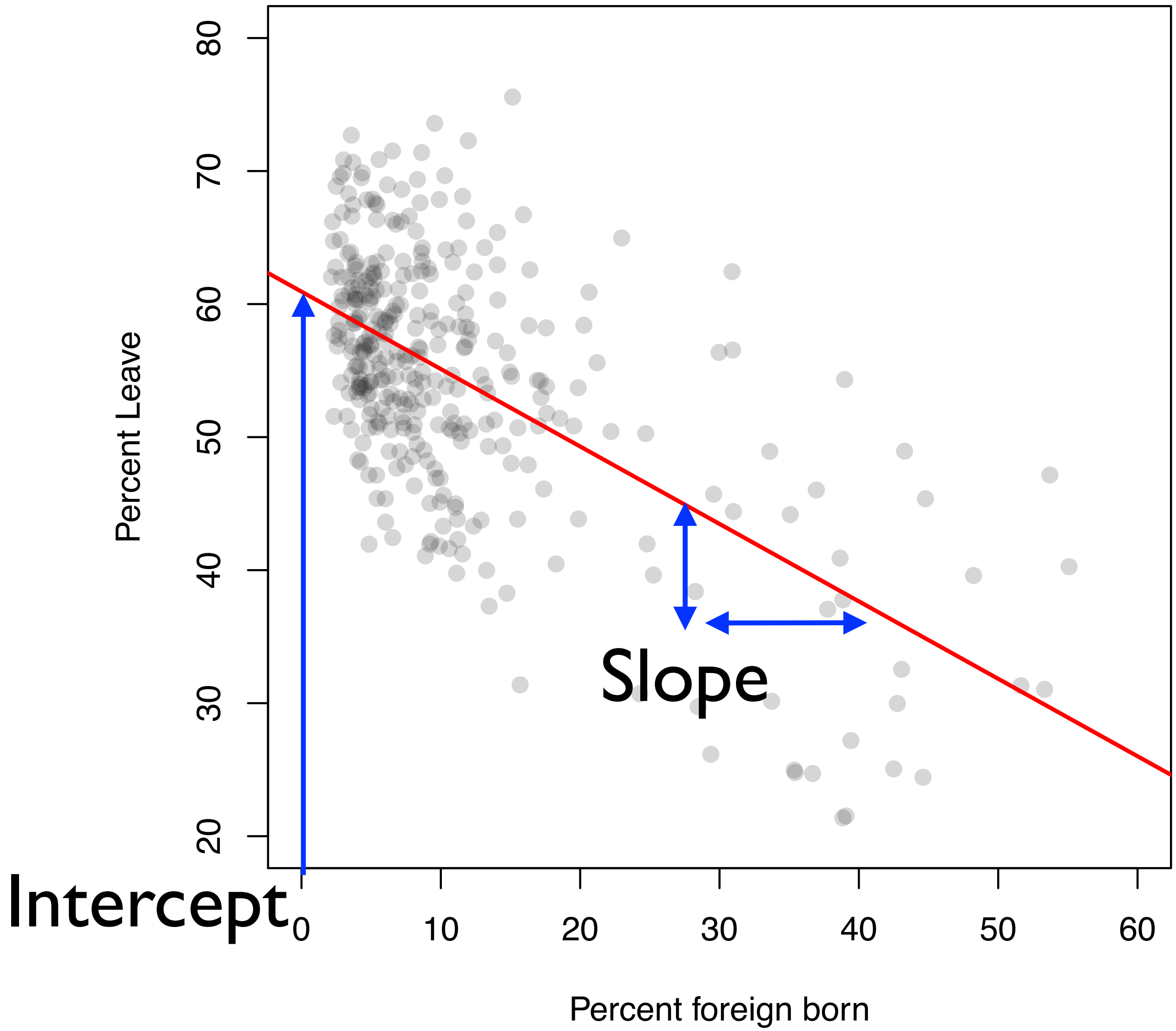
Coefficients: the two variables in a bivariate regression



Coefficients: the two variables in a bivariate regression



Coefficients: the two variables in a bivariate regression



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:

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

Variance of x:

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

Slope from OLS regression of y on x:

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

A (surprising?) fact about the slope coefficient

Covariance of x and y: $$\mathrm{Cov}(x,y) = \frac{\Sigma_i\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{n-1}$$

Variance of x: $$\mathrm{Var}(x) = \frac{\Sigma_i\left(x_i - \overline{x}\right)^2}{n-1}$$

Slope from OLS regression of y on x: $$\hat{\beta} = \frac{\mathrm{Cov}(x,y)}{\mathrm{Var}(x)}$$

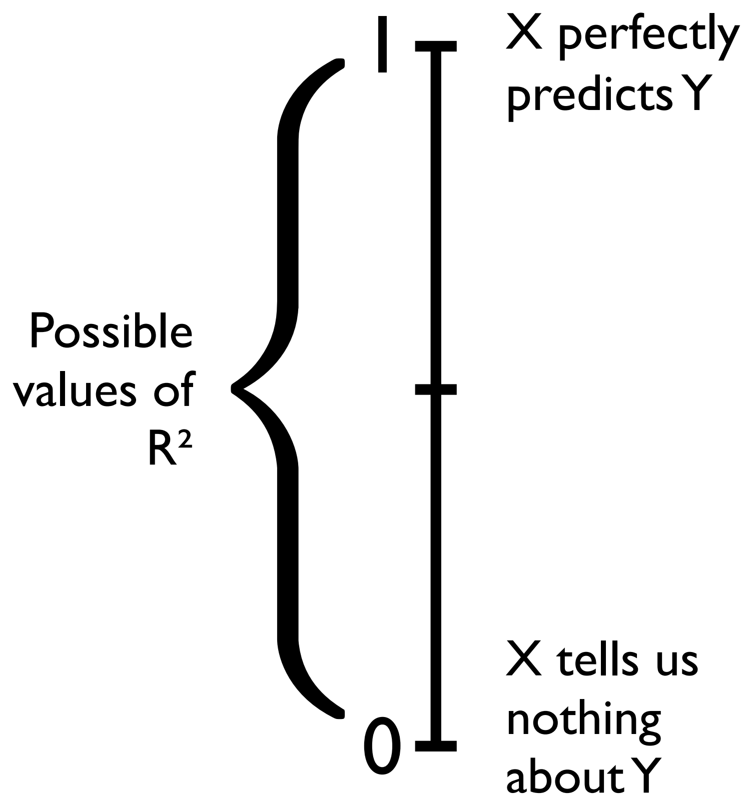
```
> cov(d$Percent_Leave, d$Percent_foreign_born, use = "complete")/var(d
$Percent_foreign_born, na.rm = T)
[1] -0.582101
> lm(d$Percent_Leave ~ d$Percent_foreign_born)

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

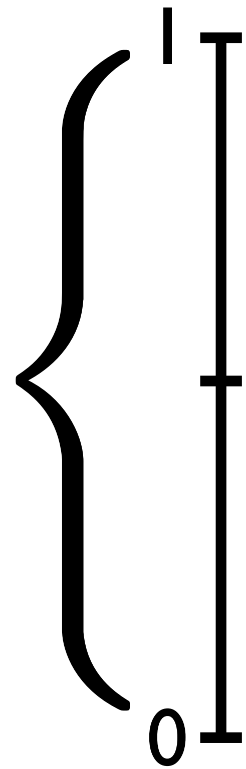
How much better are the predictions from our OLS regression line than the predictions from a flat line (i.e. not using X at all)?



R^2: intuition

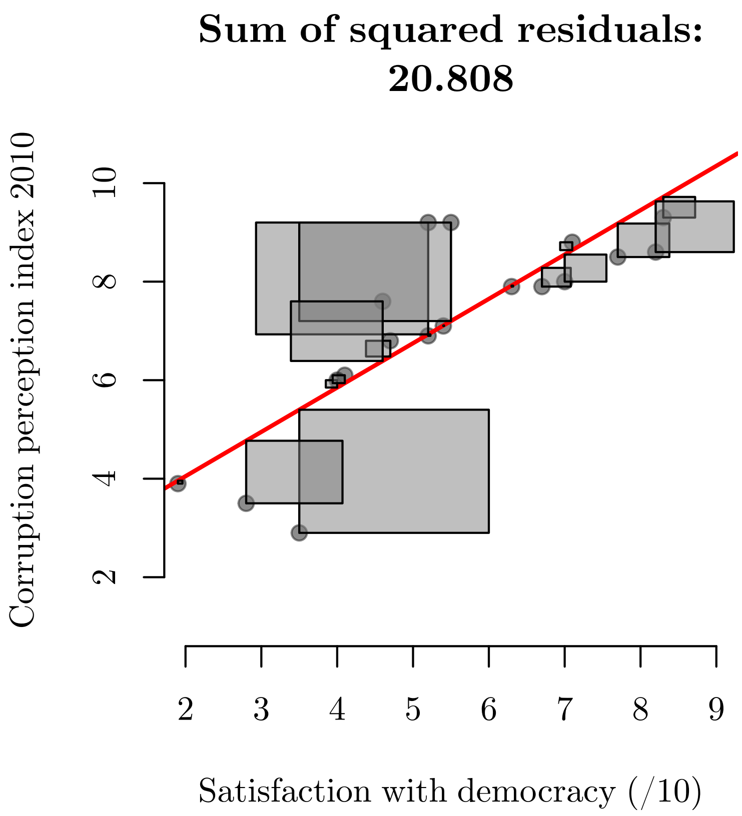
How much better are the predictions from our OLS regression line than the predictions from a flat line (i.e. not using X at all)?

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



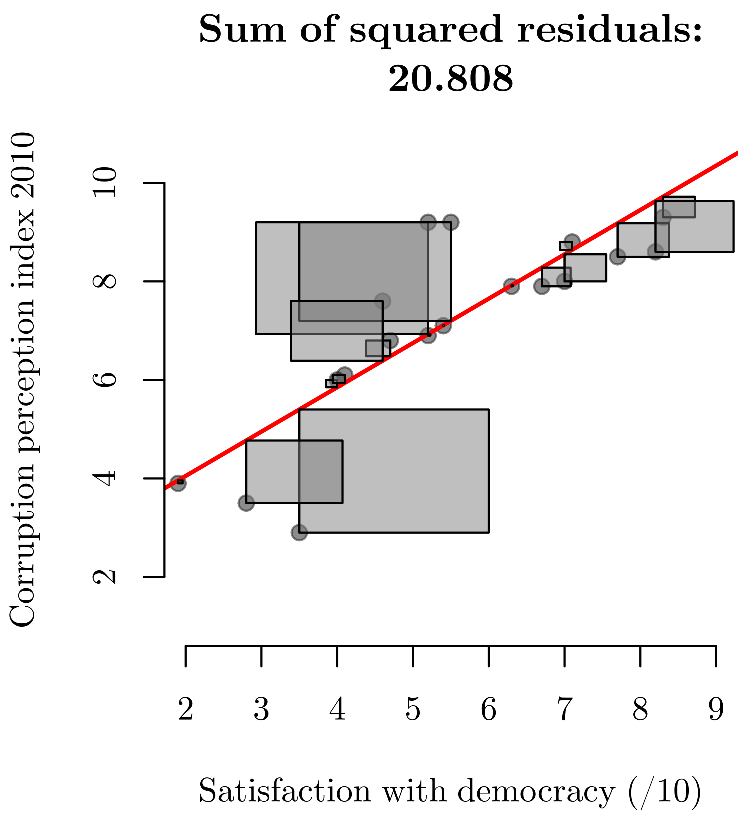
R^2: calculation

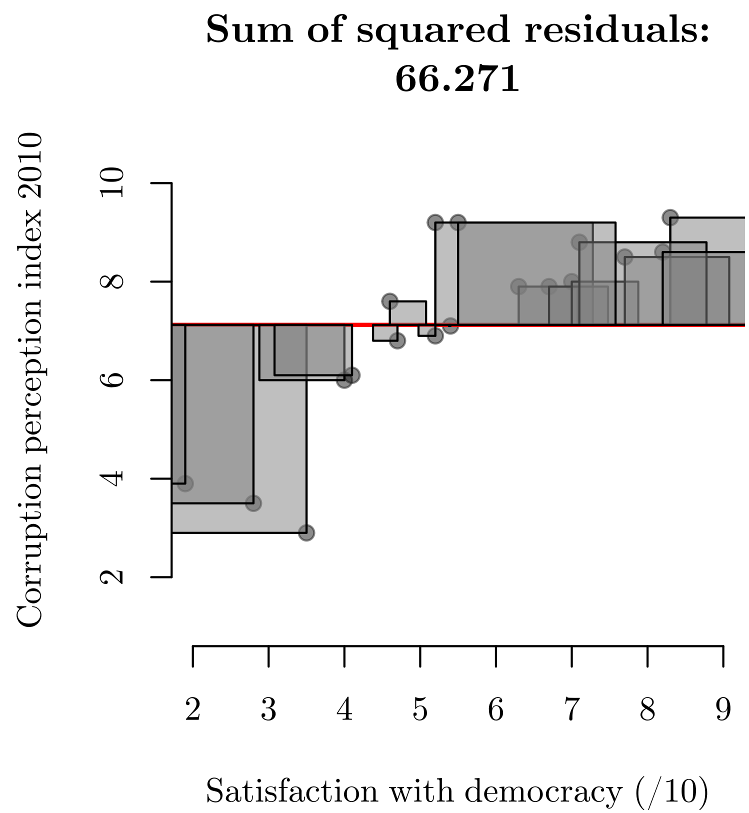
R^2: calculation



R^2: calculation

“Total sum of squares”

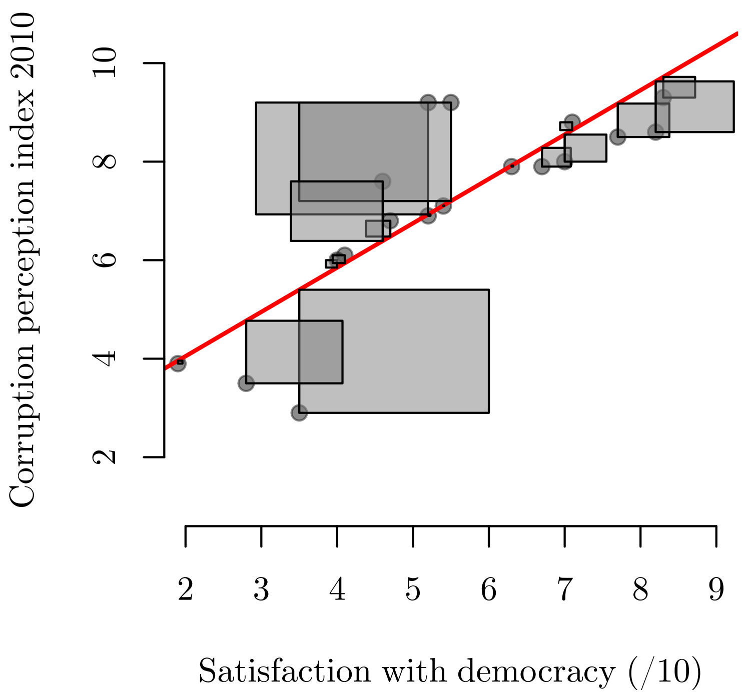




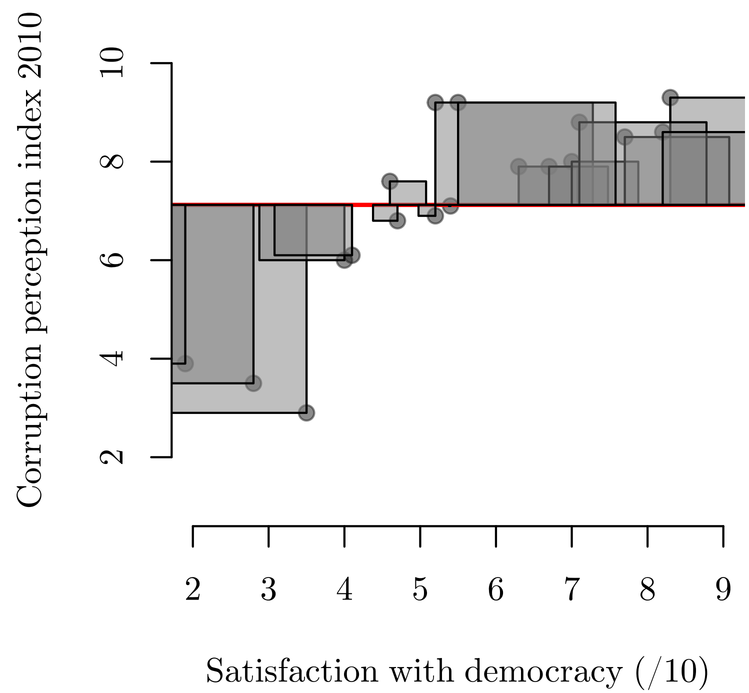
R^2: calculation

"Total sum of squares"

Sum of squared residuals: 20.808



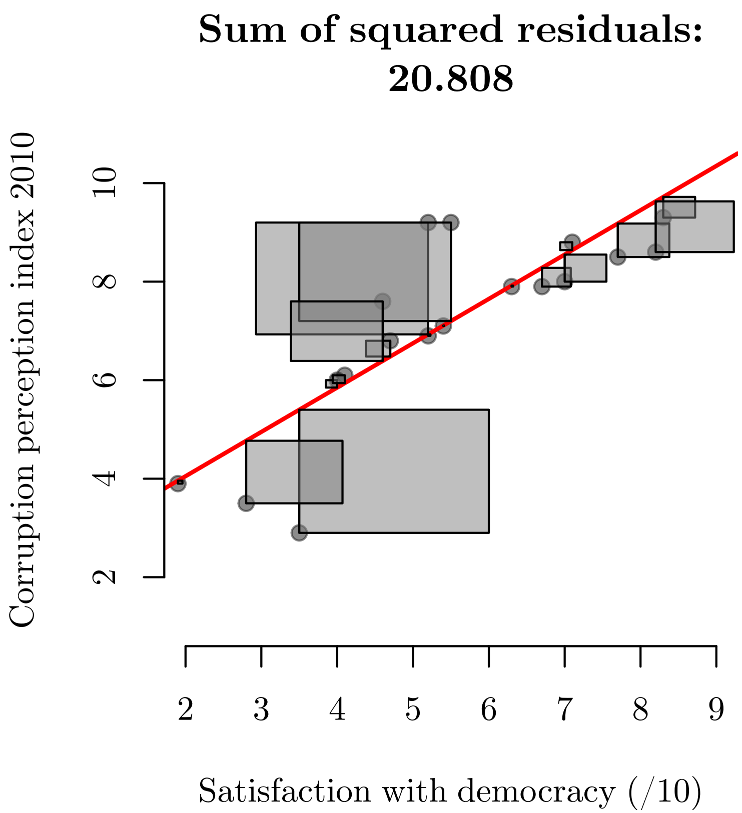
Sum of squared residuals: 66.271

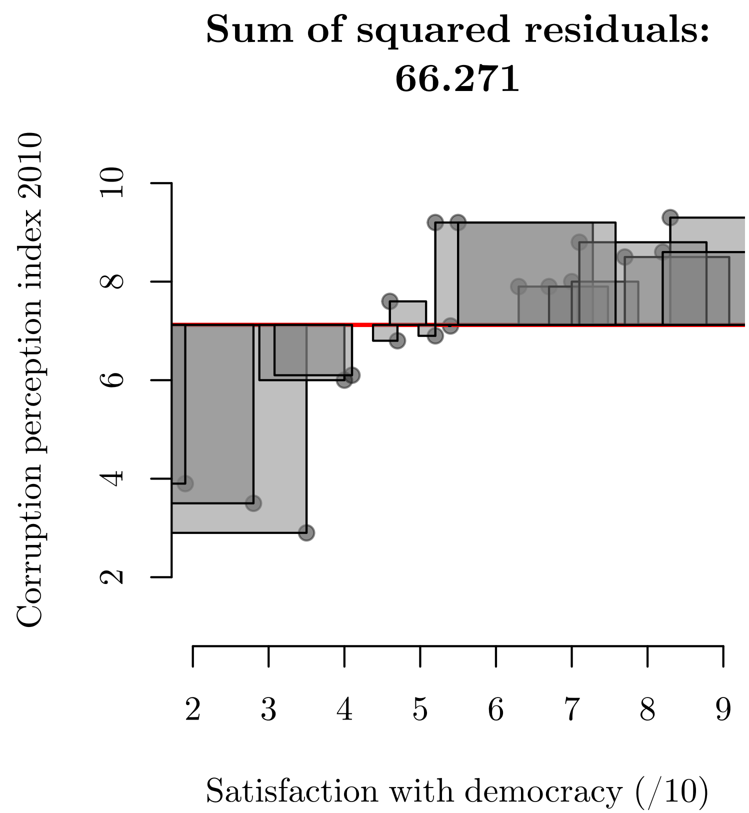


$$1 - \frac{20.808}{66.271} =$$

R^2: calculation

“Total sum of squares”





$$1 - \frac{20.808}{66.271} = 0.6864$$

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

Why are we minimizing squared residuals?

There are other ways to draw a predictive line.

But OLS (minimizing squared residuals)

- produces nice analytical solutions
- recovers the mean
- among unbiased estimators, minimizes variance