Oxford Q-Step Data Day: How we learn from data

Andy Eggers

Assoc. Professor

Department of Politics and International Relations

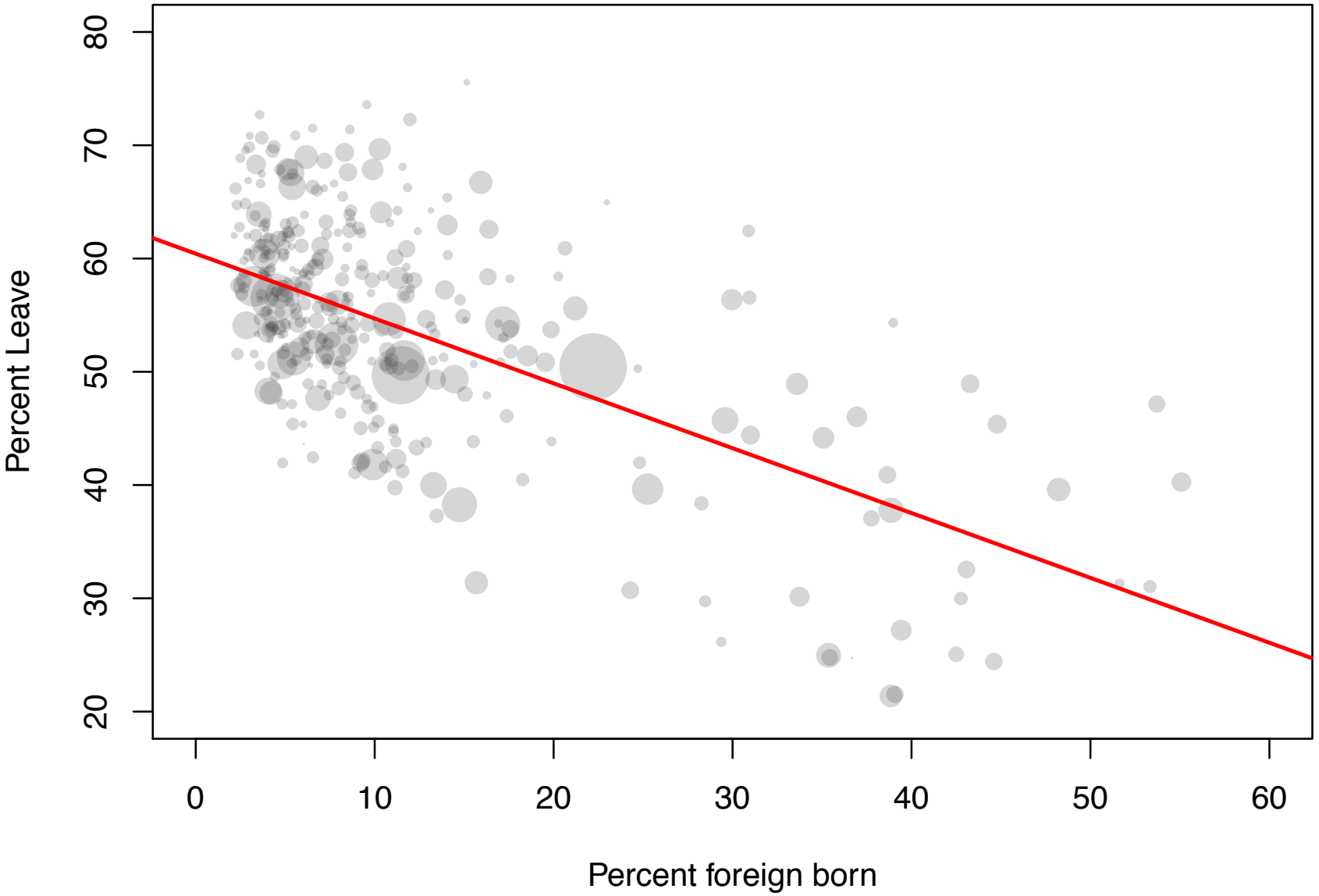
Data literacy: the three stages

1) Learn facts about data

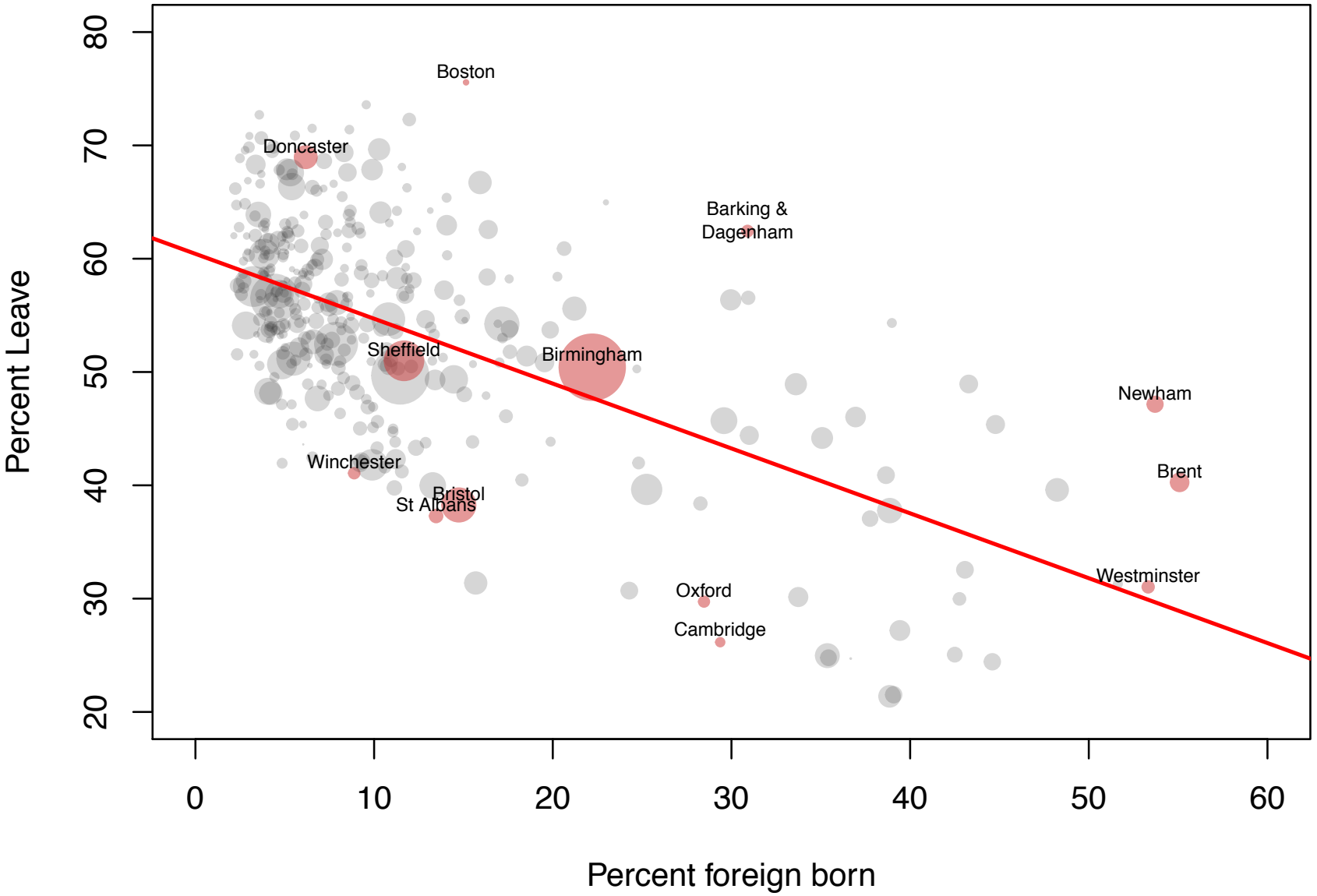
2) Understand and question how others collect and use data

3) Collect and use data to produce new knowledge

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



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

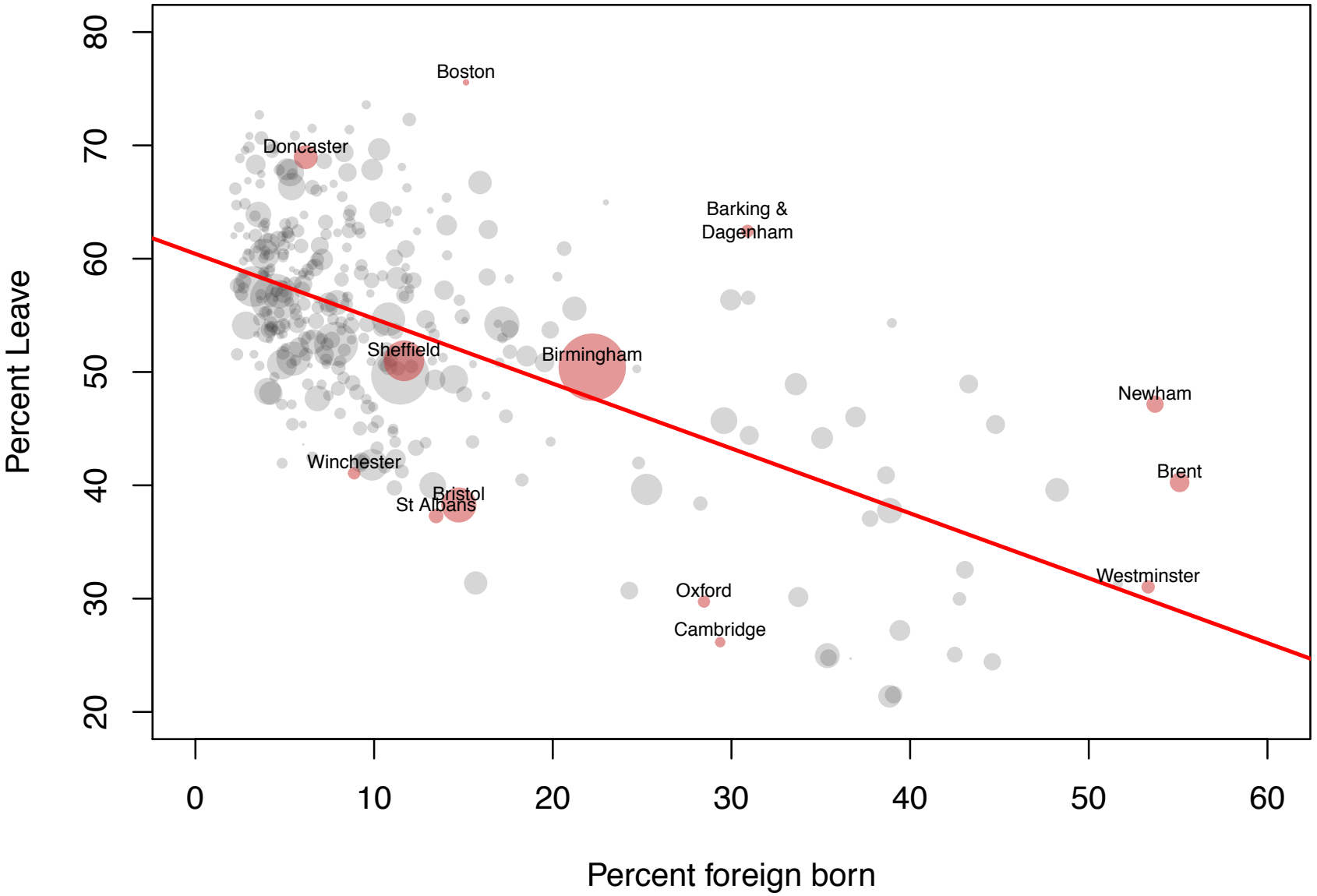


Contact hypothesis

"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*

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



Group activity

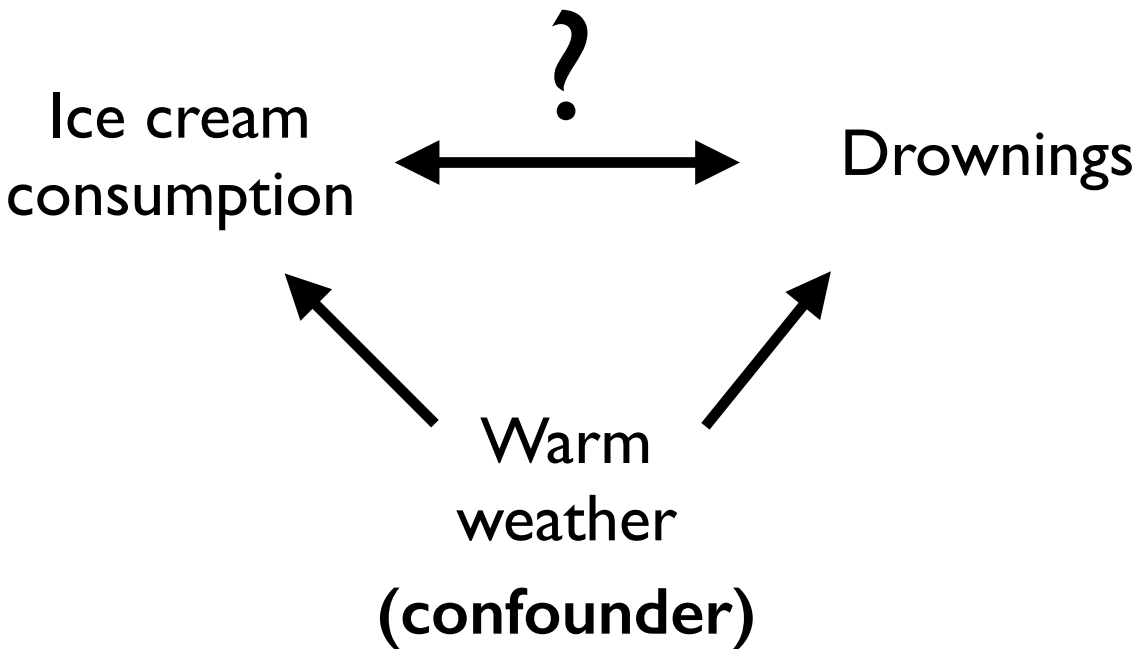
Brexit support is higher in places with fewer foreign-born residents. Is this evidence in favor of the contact hypothesis?

Be ready to tell the rest of us

- why it might be evidence of the contact hypothesis
- other reasons why this pattern might arise

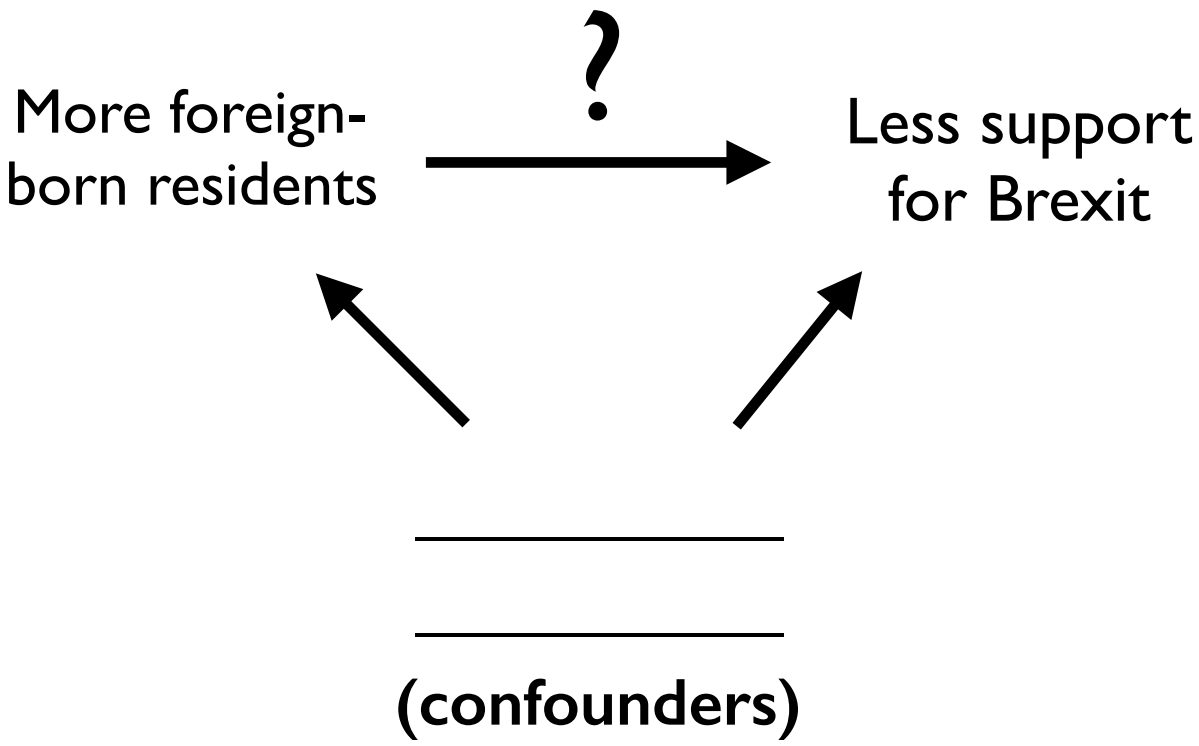
Confounders

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



Confounders (2)

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



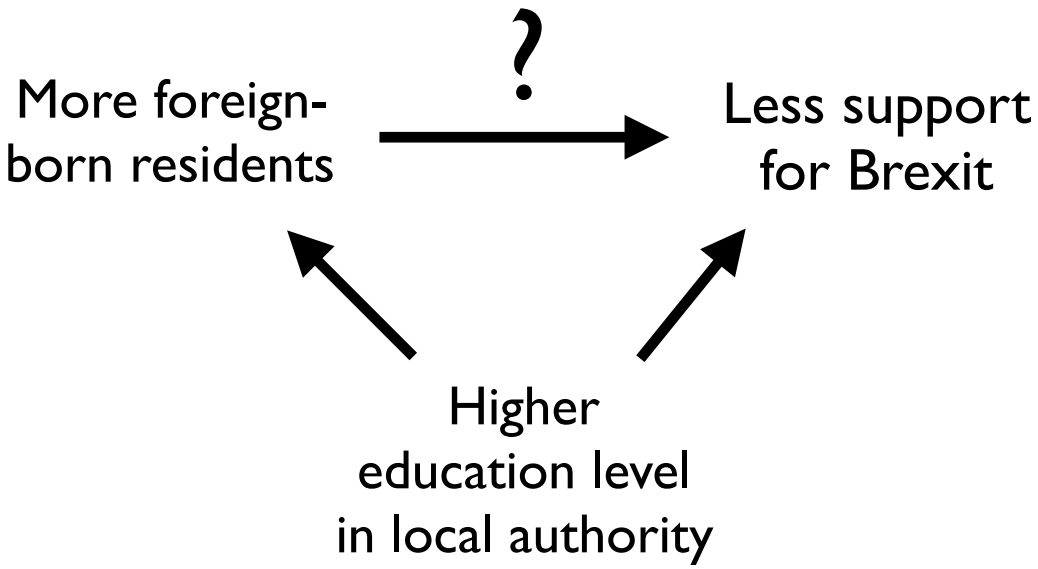
Controlling for confounders

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 _____?

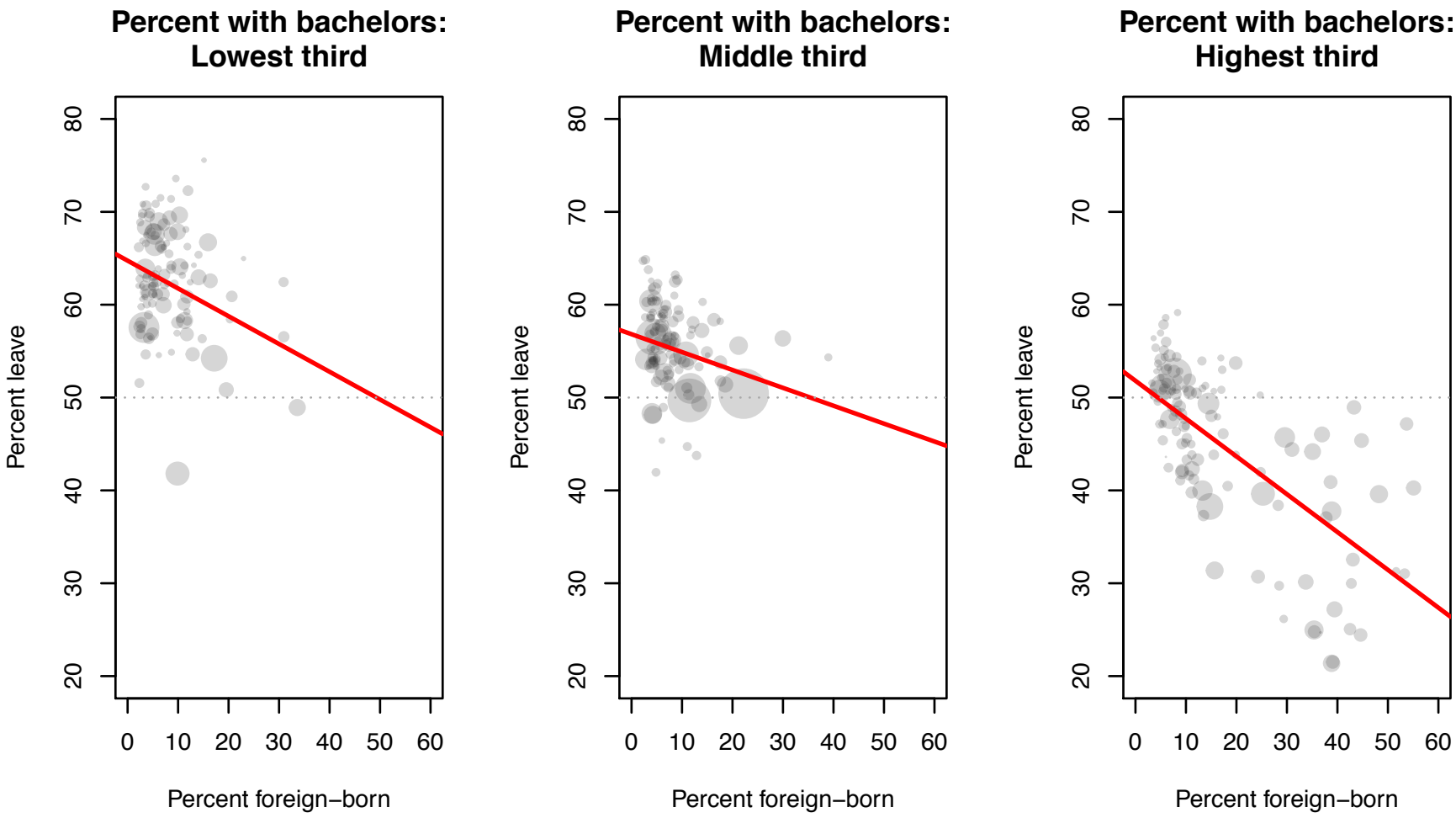
How do we control for confounders?

Let's focus on education level 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



A more general approach: multiple regression

Goal: measure relationship between

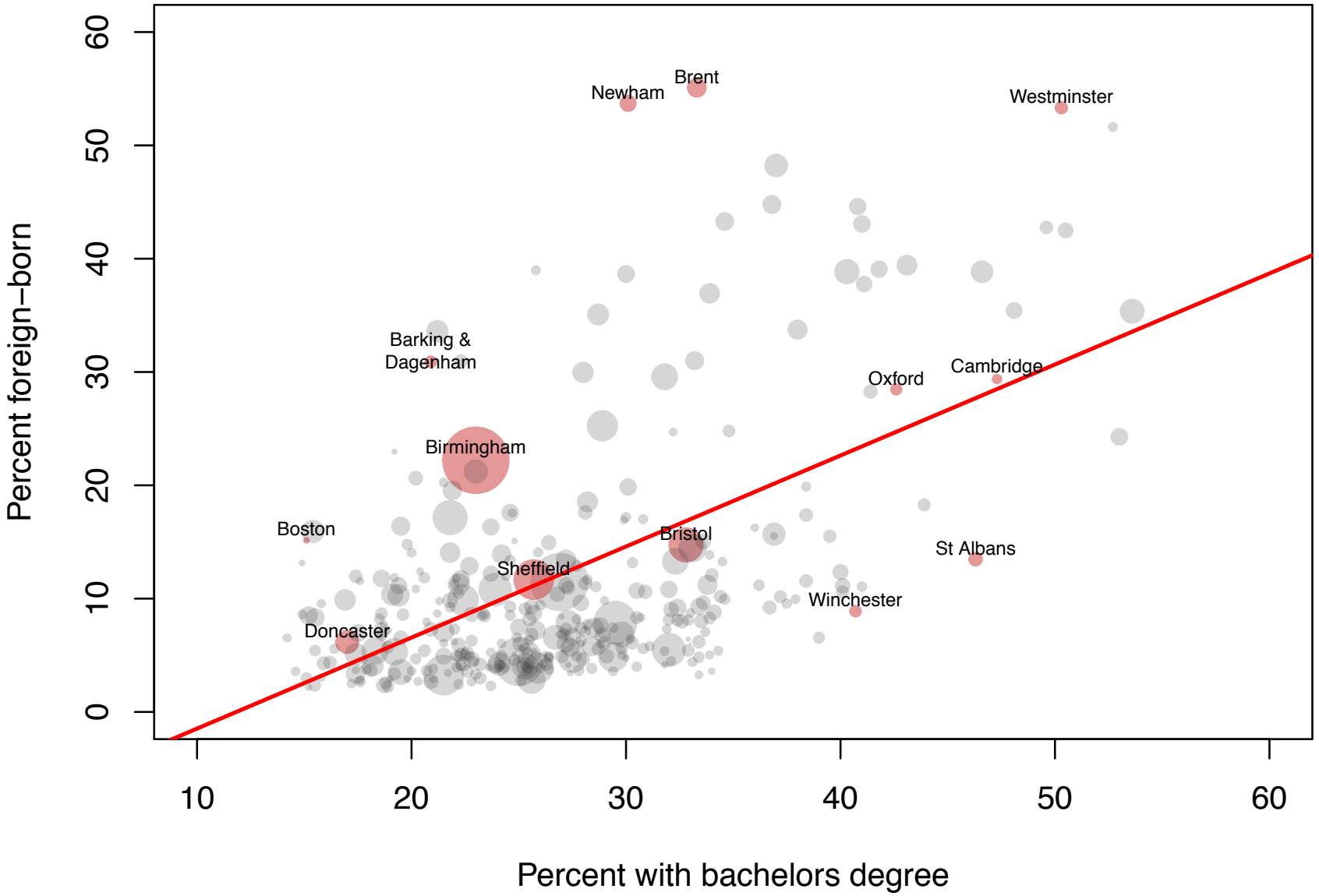
- "support for Leave" and
- "% foreign-born"

controlling for "% bachelors degree".

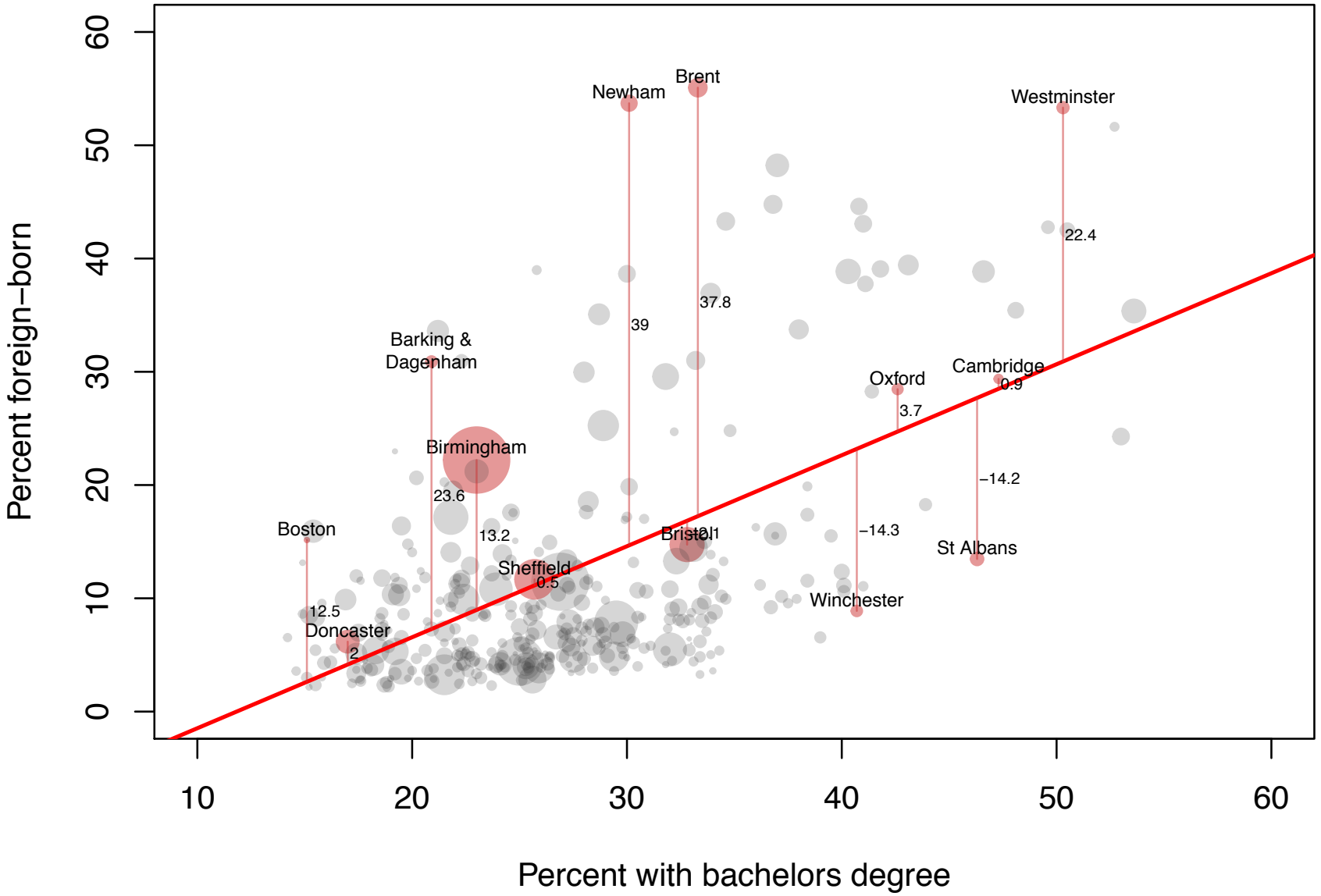
Basic idea: measure relationship between

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

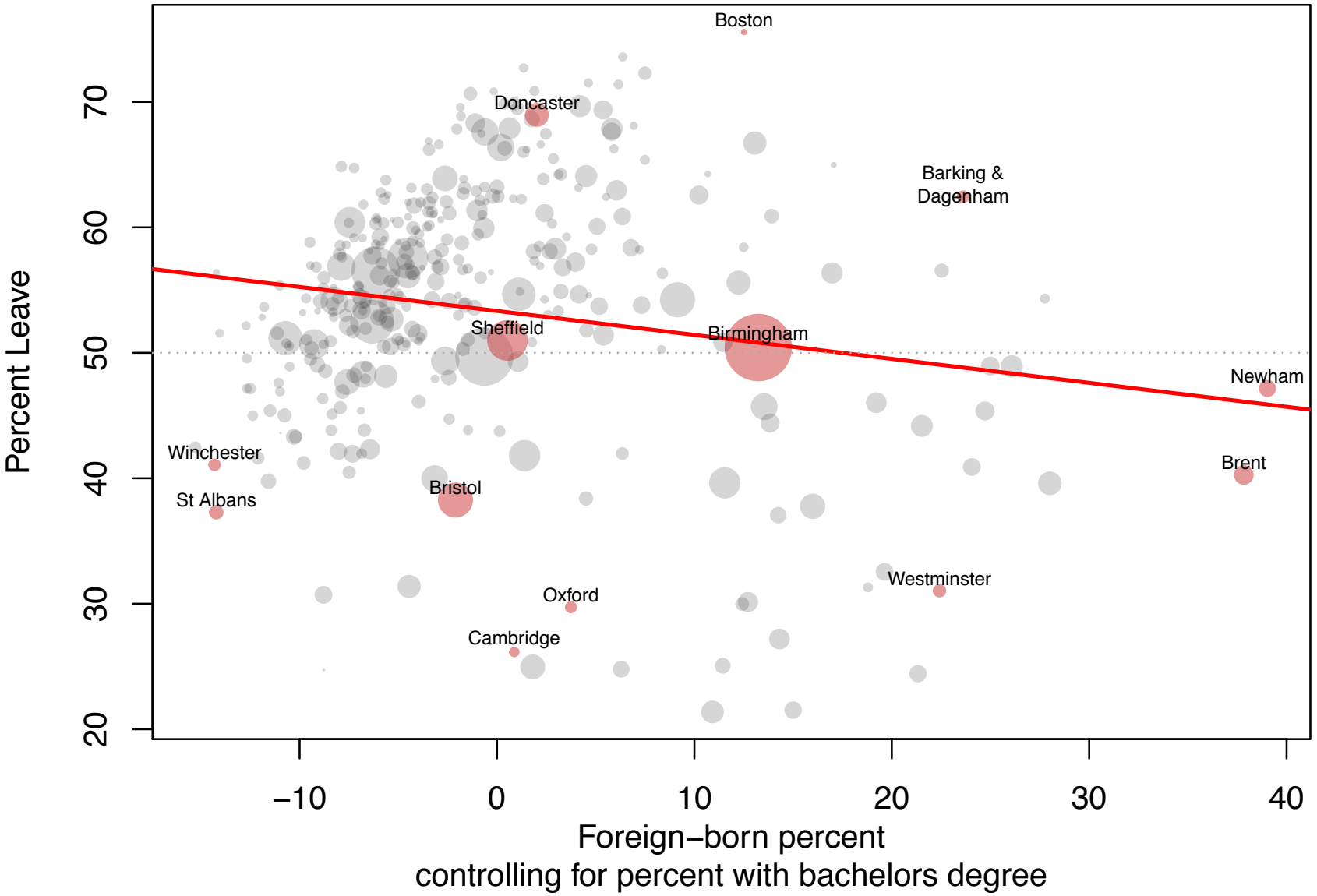
Step 1: measure relationship between explanatory variable and confounder



Step 2: calculate difference between explanatory variable and prediction line



Step 3: measure relationship between those differences and outcome (% Leave)



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?

Want to use multiple regression?

Today was about understanding the intuition.

How do you actually do multiple regression?

- In R,
  - load data
  - `lm(y ~ x1 + x2 + x3)` tells you how `x1` is related to `y`, controlling for `x2` & `x3`.
- In Google Spreadsheets with (free!) Statistics Add-on installed,
  - load data
  - "Add-ons" → "Statistics" → "Regression . . ." and choose `y` as "Response variable" and `x1`, `x2` & `x3` as "Predictor variables"