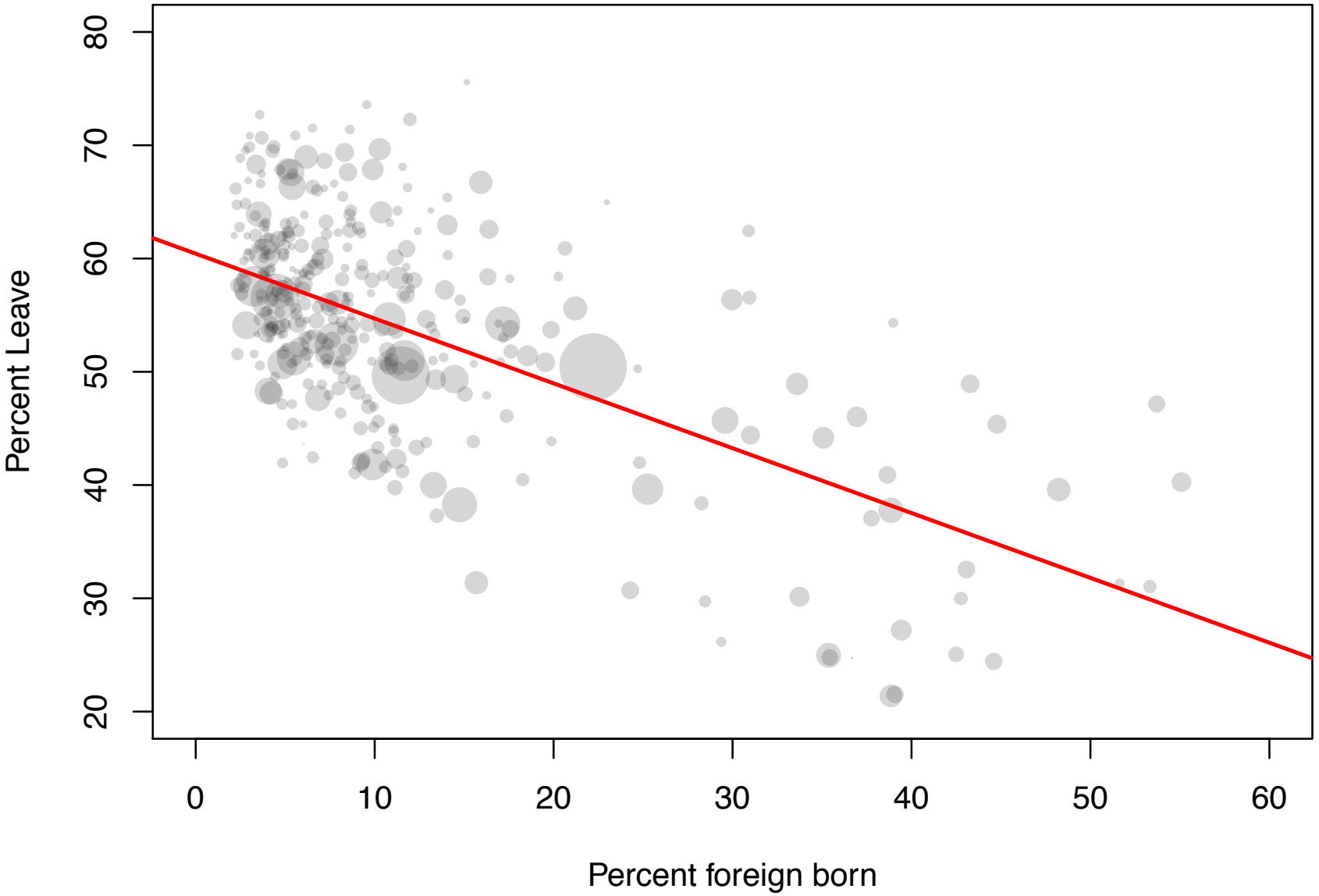
A brief, non-technical introduction to regression

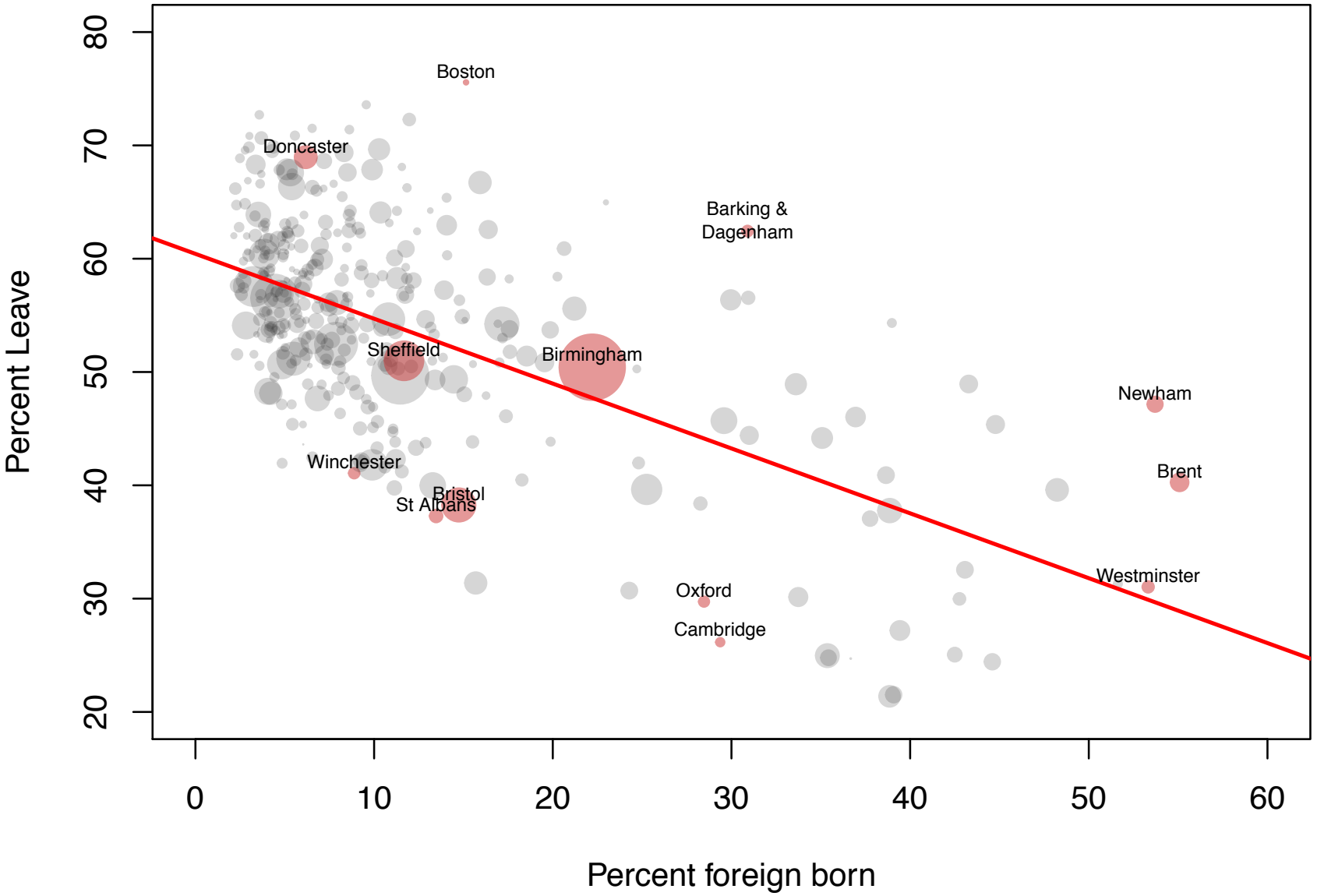
Andy Eggers

Oxford Q-Step Centre

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

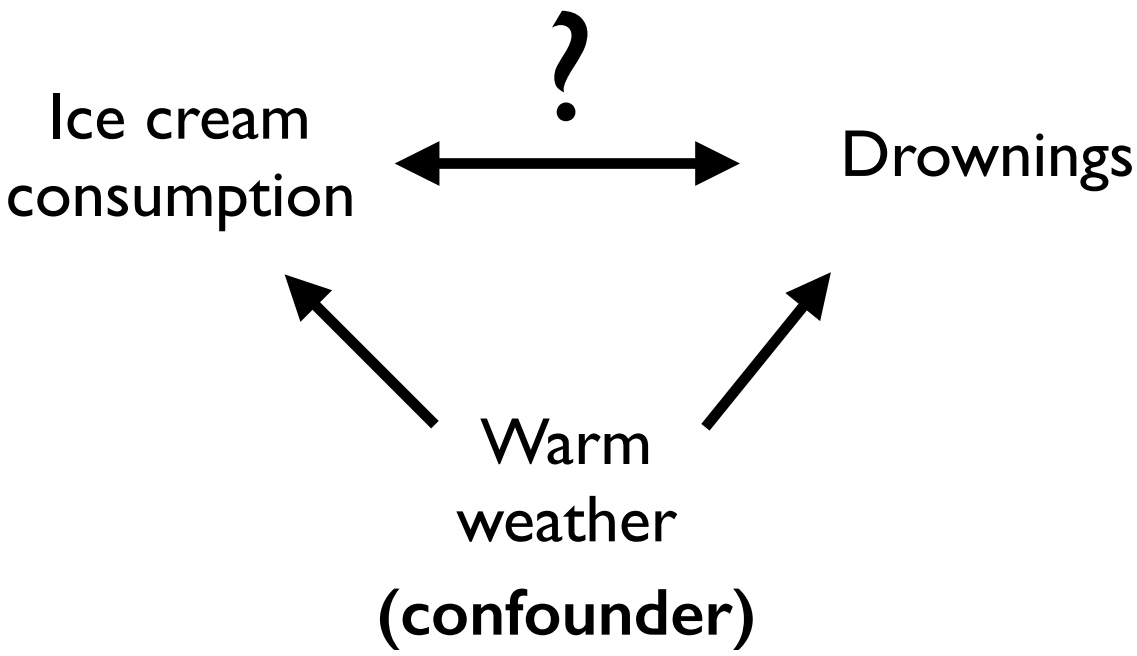


Did this pattern arise because contact with immigrants makes people less opposed to immigration?



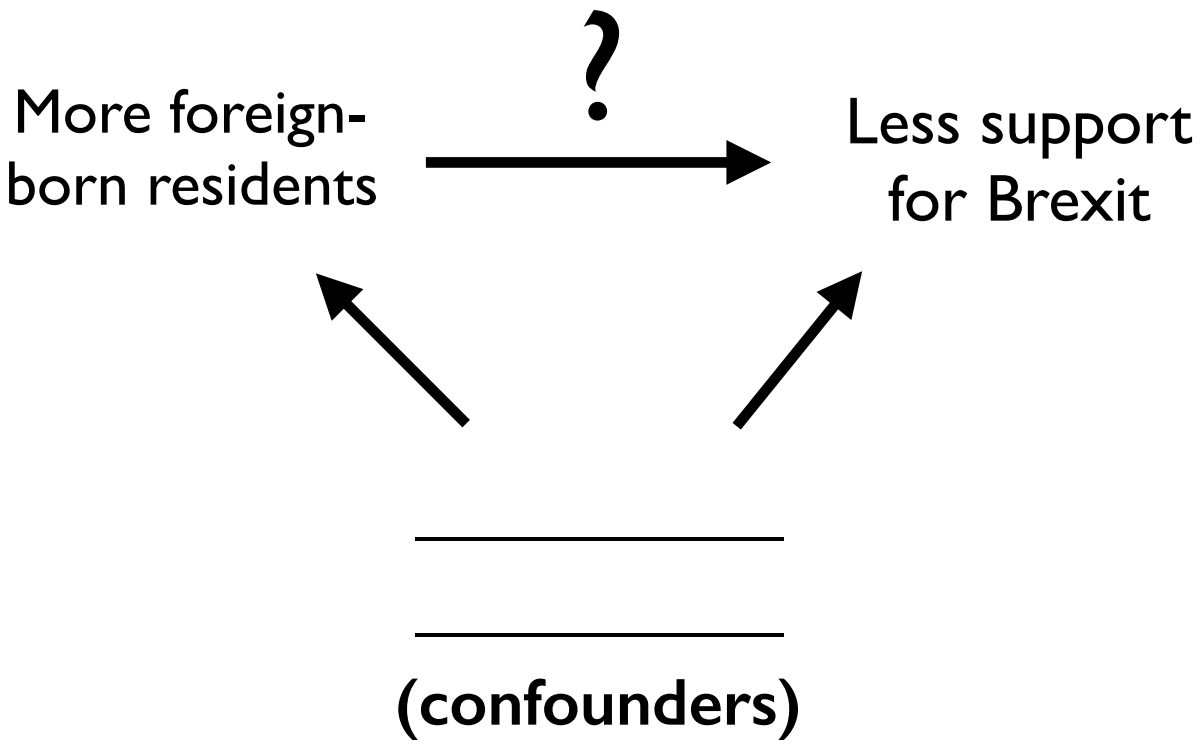
Confounders

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?



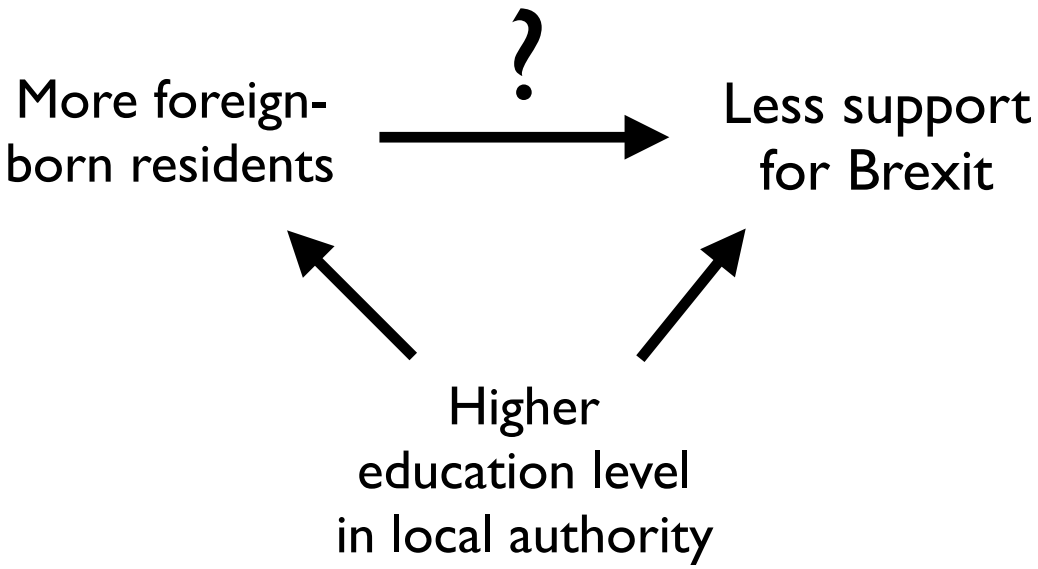
Controlling for confounders

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 _____?

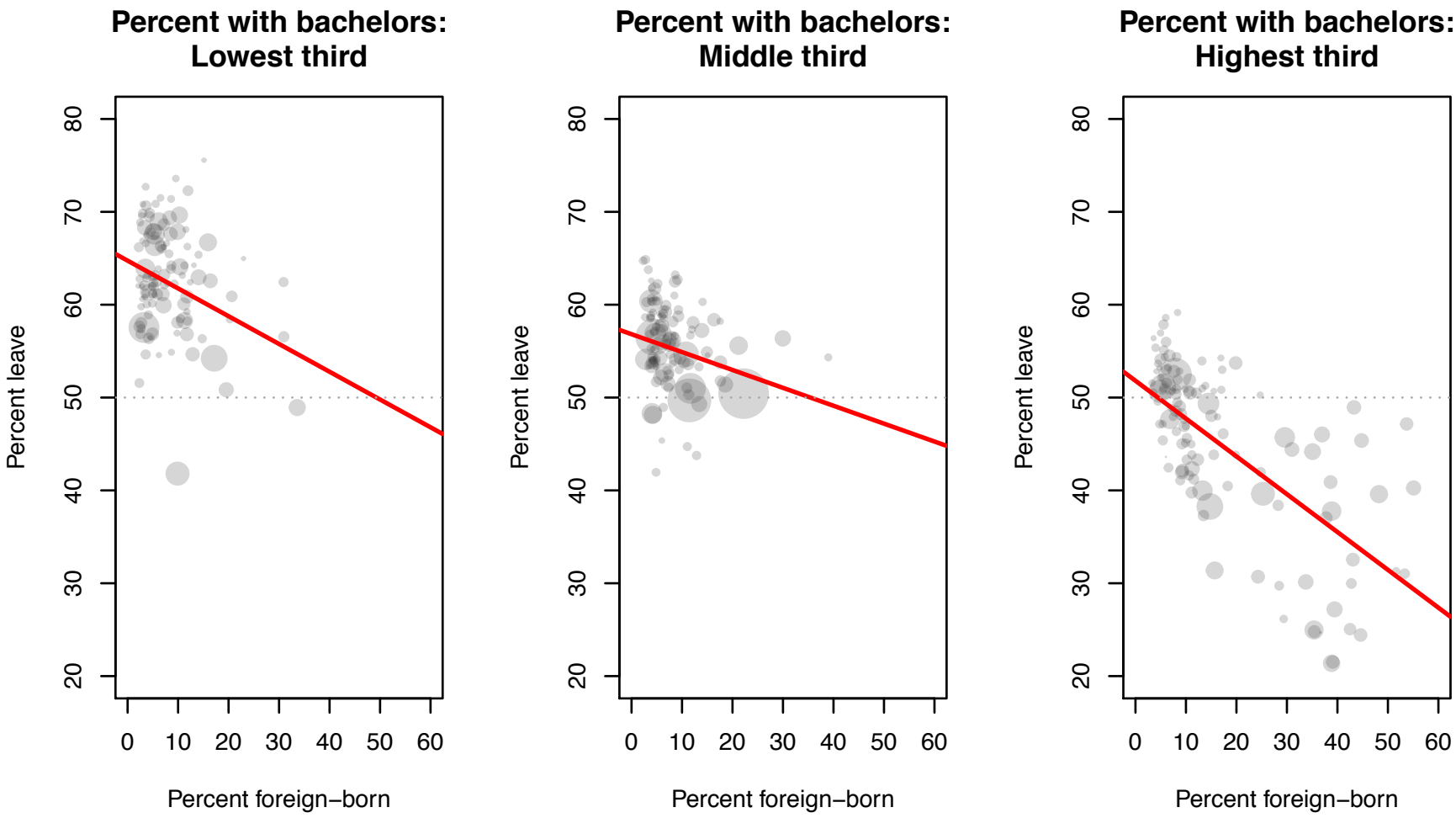
How do we control for confounders?

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

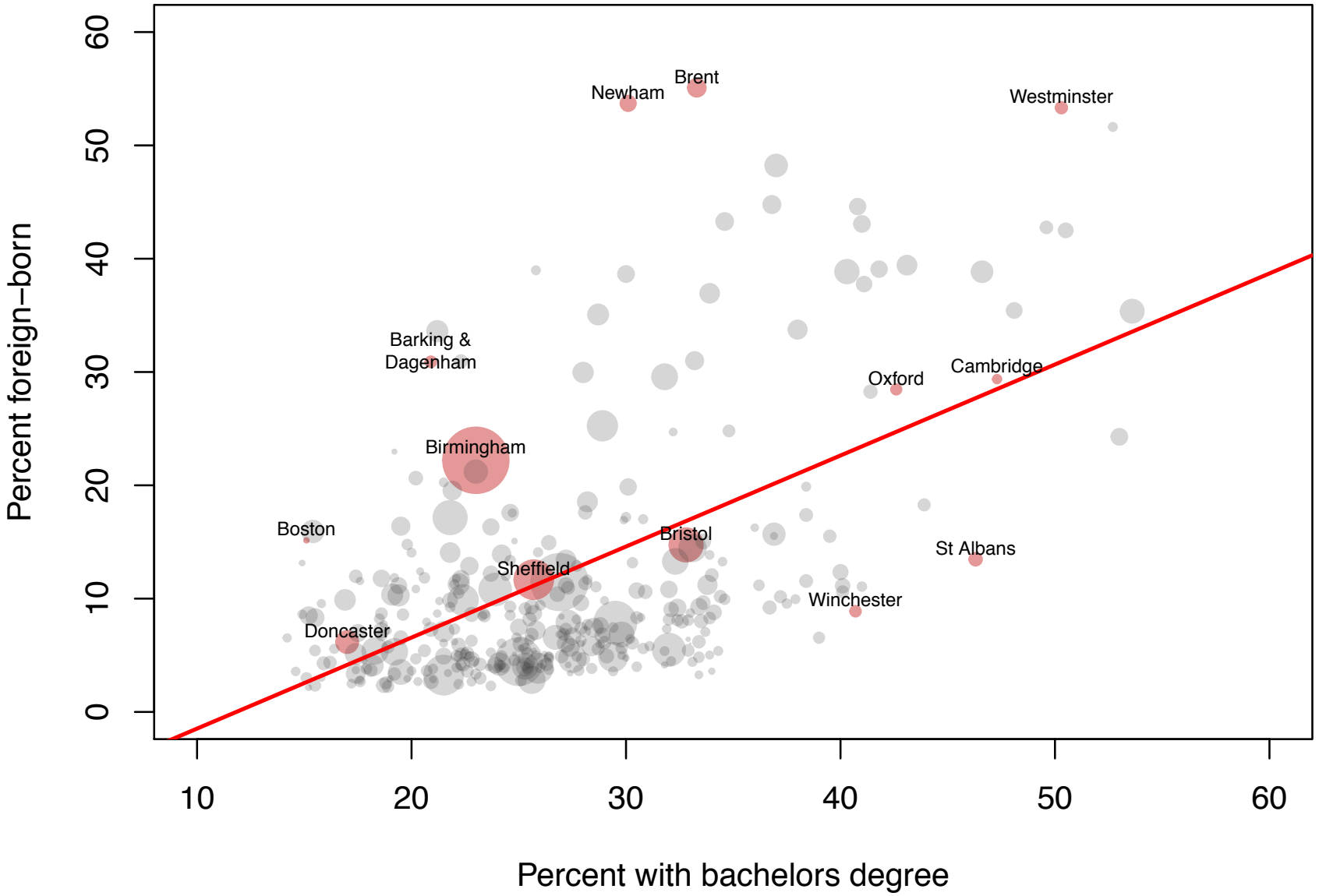
- "support for Leave" and
- "% foreign-born"

controlling for "% bachelors degree".

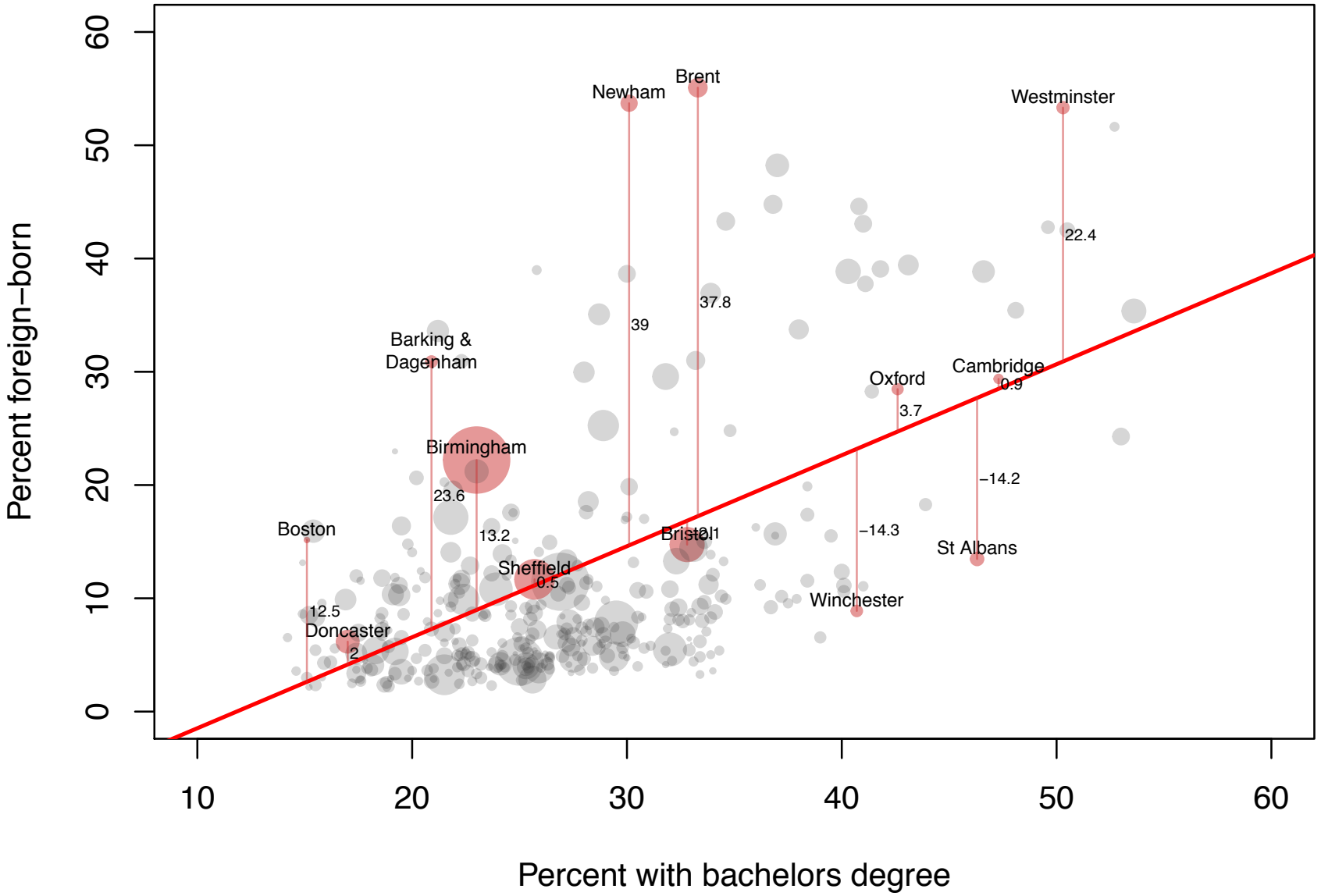
Basic idea: measure relationship between

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

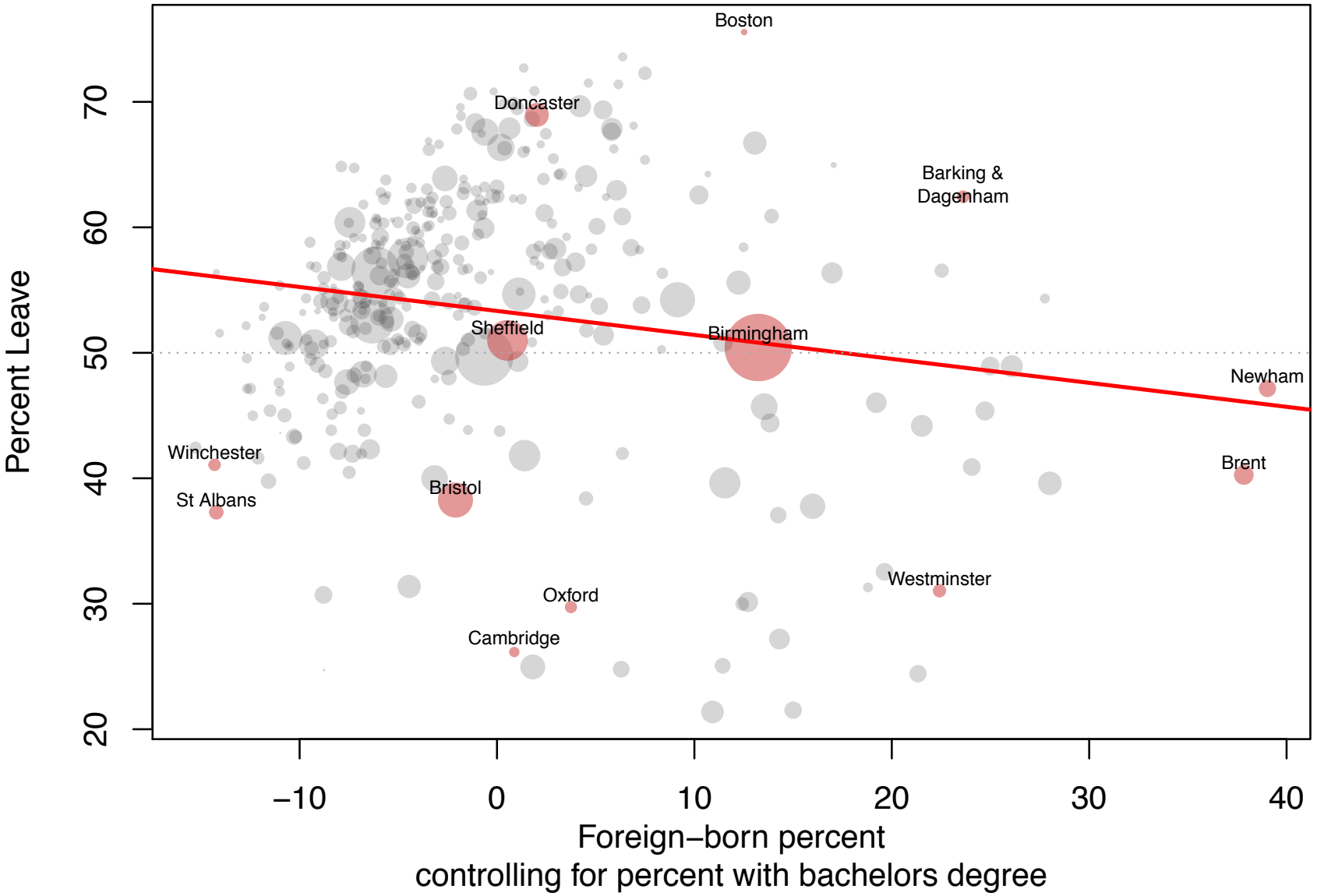
Step 1: measure relationship between explanatory variable and confounder



Step 2: calculate difference between explanatory variable and prediction line



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



How to do your own multiple regression

To regress `y` on `x1`, controlling for `x2` & `x3`:

- Google Spreadsheets (with Statistics Add-on):
  - load data ("File" → "Open")
  - "Add-ons" → "Statistics" → "Regression . . ." and choose `y` as "Response variable" and `x1`, `x2` & `x3` as "Predictor variables"
- R:
  - load dataset: `D = read.csv("filename.csv")`
  - run regression: `lm(y ~ x1 + x2 + x3, data = D)`