Multivariate relationships: more regression

Andy Eggers

Assoc. Professor

Department of Politics and International Relations

OCCASIONAL NOTES

Chocolate Consumption, Cognitive Function, and Nobel Laureates

Franz H. Messerli, M.D.

Dietary flavonoids, abundant in plant-based foods, have been shown to improve cognitive function. Specifically, a reduction in the risk of dementia, enhanced performance on some cognitive tests, and improved cognitive function in elderly patients with mild impairment have been associated with a regular intake of flavonoids.[1,2] A subclass of flavonoids called flavanols, which are widely present in cocoa, green tea, red wine, and some fruits, seems to be effective in slowing down or even reversing the reductions in cognitive performance that occur with aging. Dietary flavanols have also been shown to improve endothelial

cause the population of a country is substantially higher than its number of Nobel laureates, the numbers had to be multiplied by 10 million. Thus, the numbers must be read as the number of Nobel laureates for every 10 million persons in a given country.

All Nobel Prizes that were awarded through October 10, 2011, were included. Data on per capita yearly chocolate consumption in 22 countries was obtained from Chocosuisse (www.chocosuisse.ch/web/chocosuisse/en/home), Theobroma-cacao (www.theobroma-cacao.de/wissen/wirtschaft/international/konsum), and

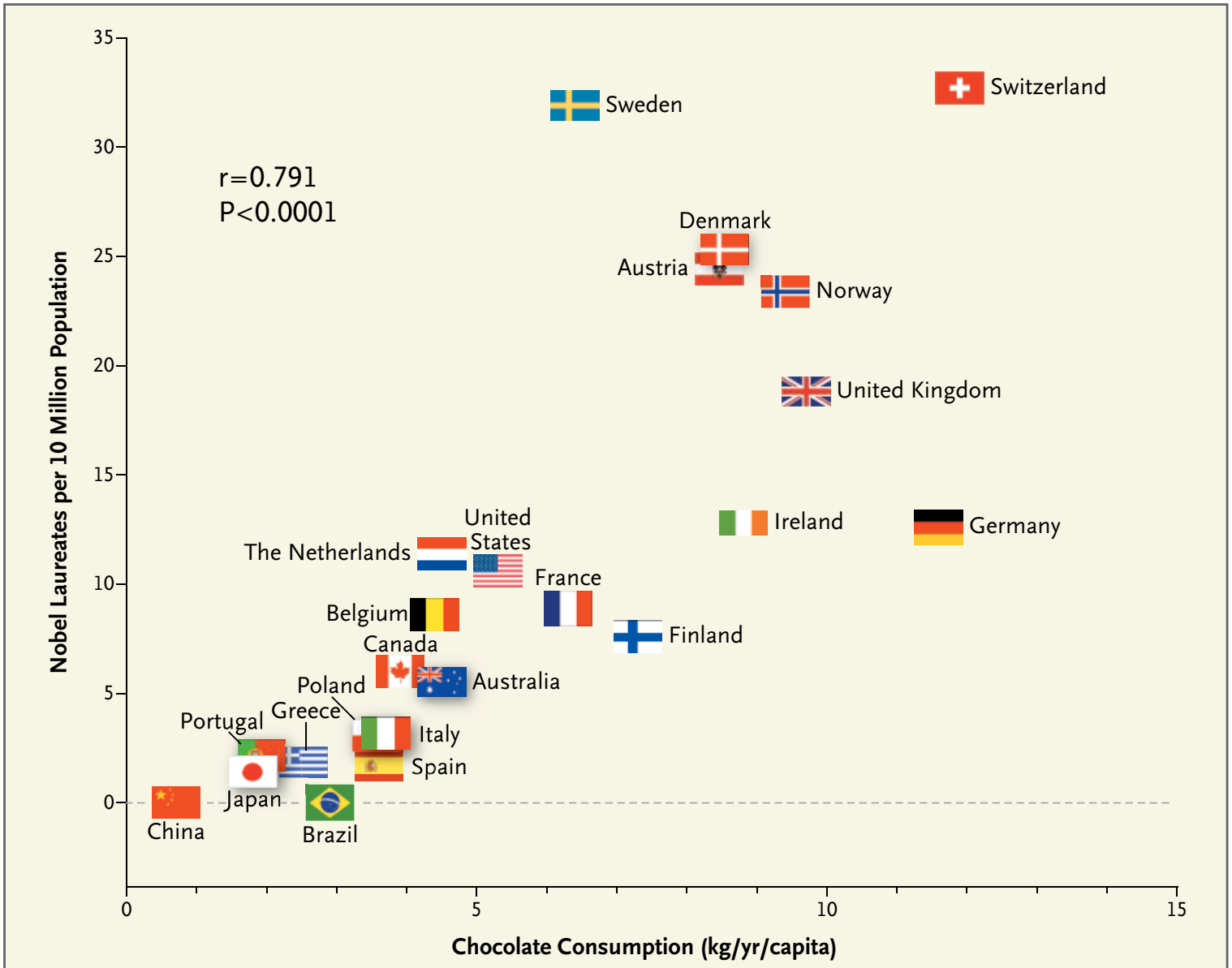


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

We want you to understand:

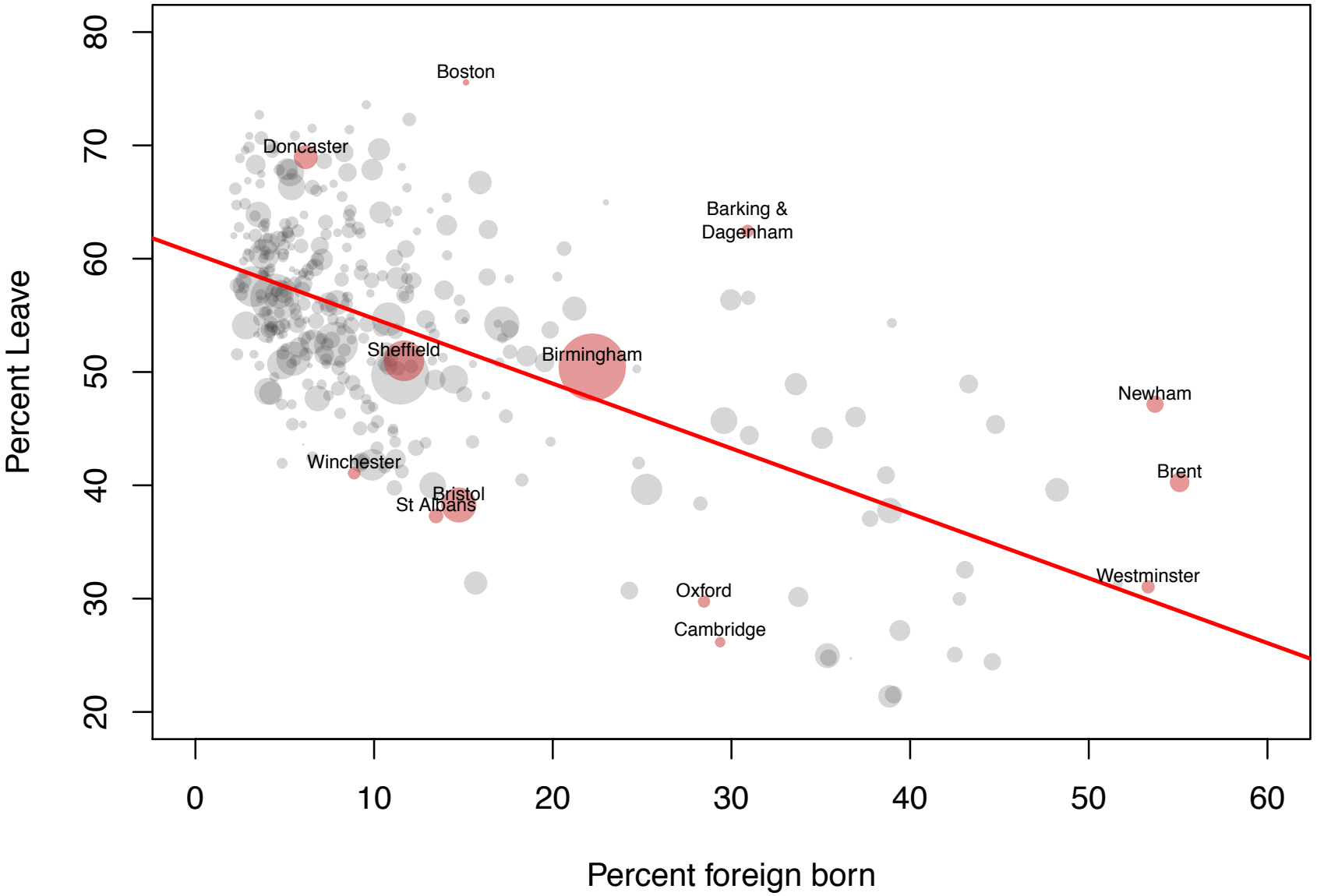
Dependent variable: Nobel Prizes awarded per capita (in log scale)

	(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

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

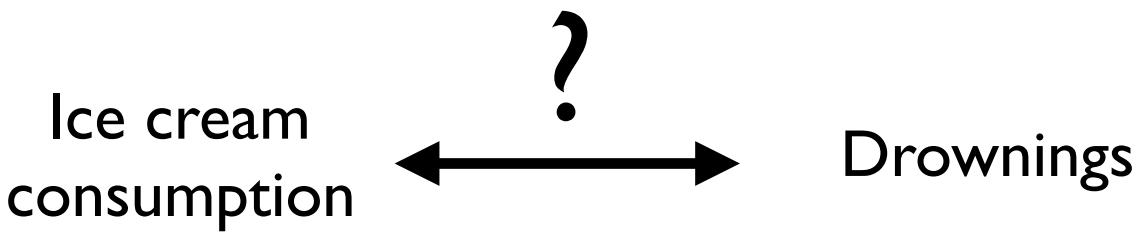
One reason why two phenomena can be correlated is the presence of a **confounder**.

Ice cream consumption

Drownings

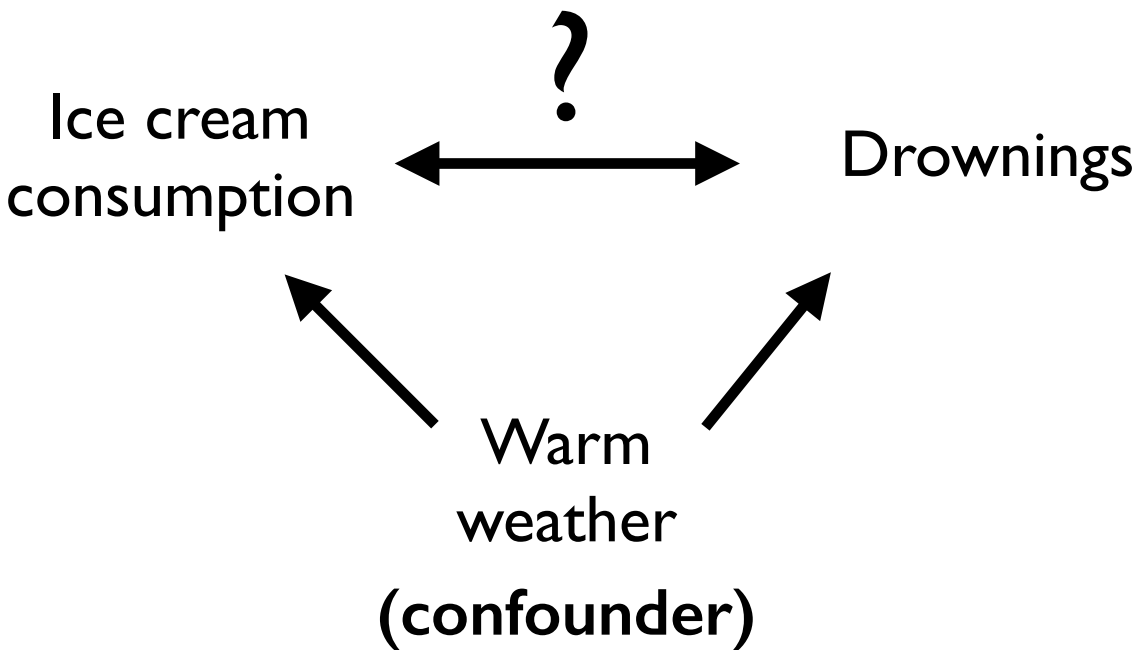
Confounders

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



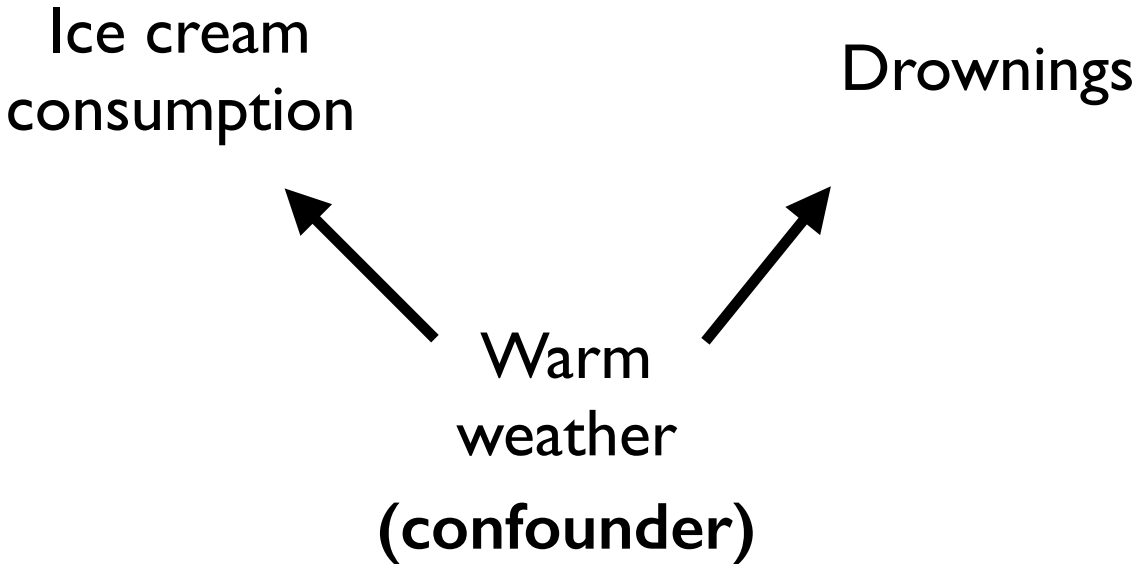
Confounders

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

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**.

Chocolate
consumption

Confounders

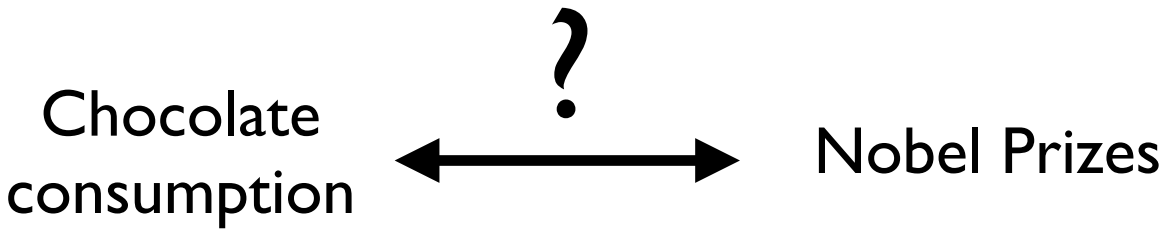
One reason why two phenomena can be correlated is the presence of a **confounder**.

Chocolate consumption

Nobel Prizes

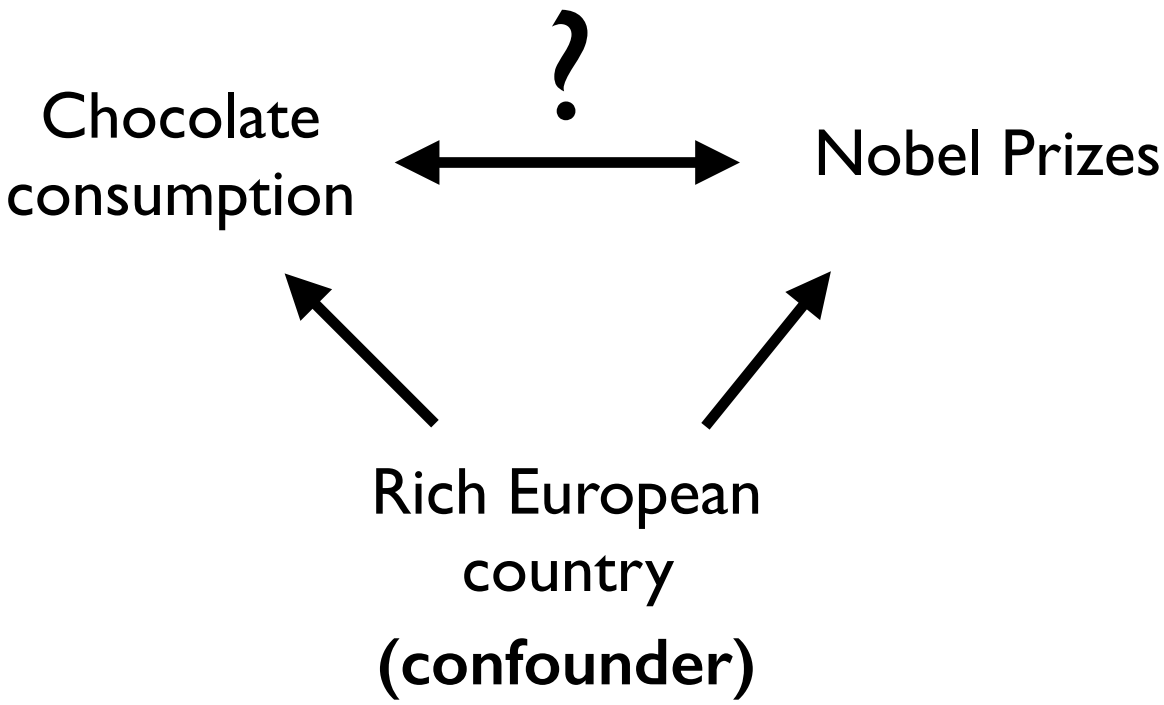
Confounders

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



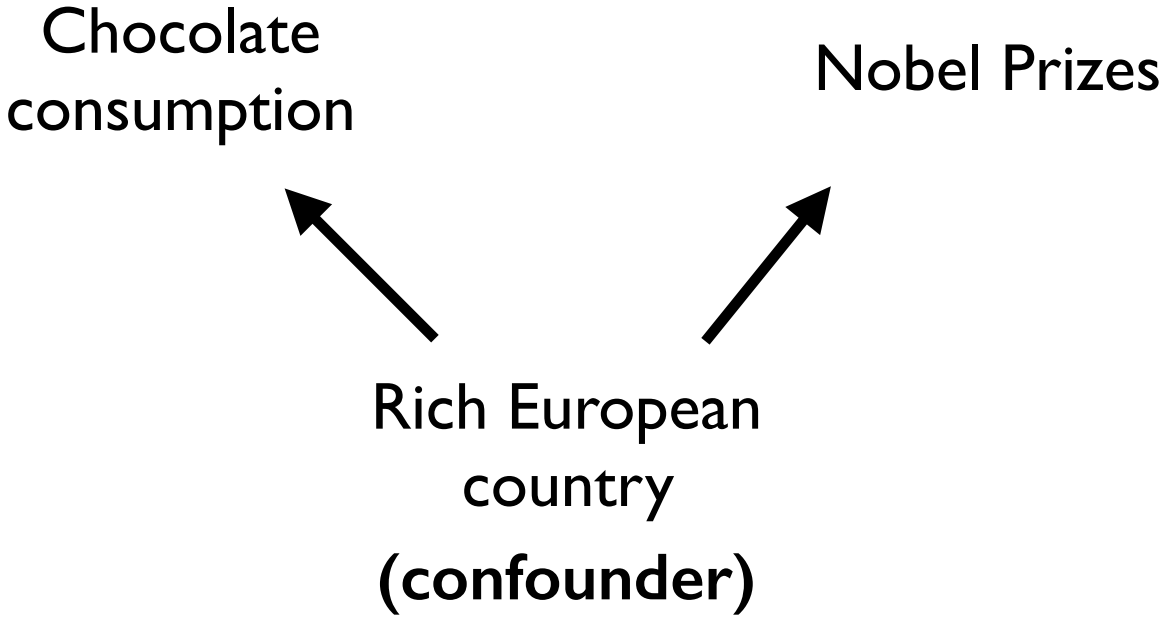
Confounders

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

Confounders

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

Confounders

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

Exercising in your 40s

Confounders

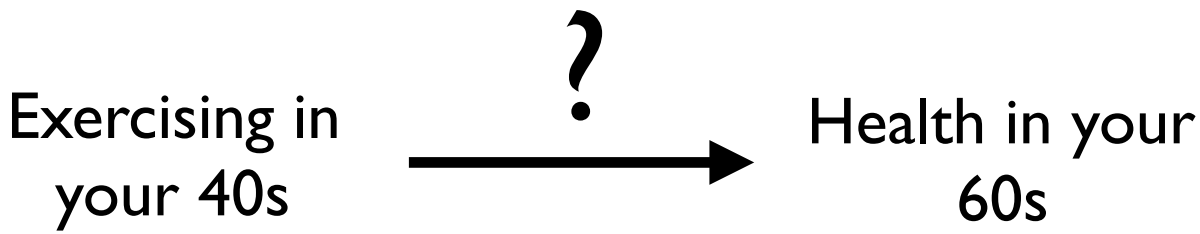
What are possible confounders in the relationship between exercise in your 40s and health in your 60s?

Exercising in your 40s

Health in your 60s

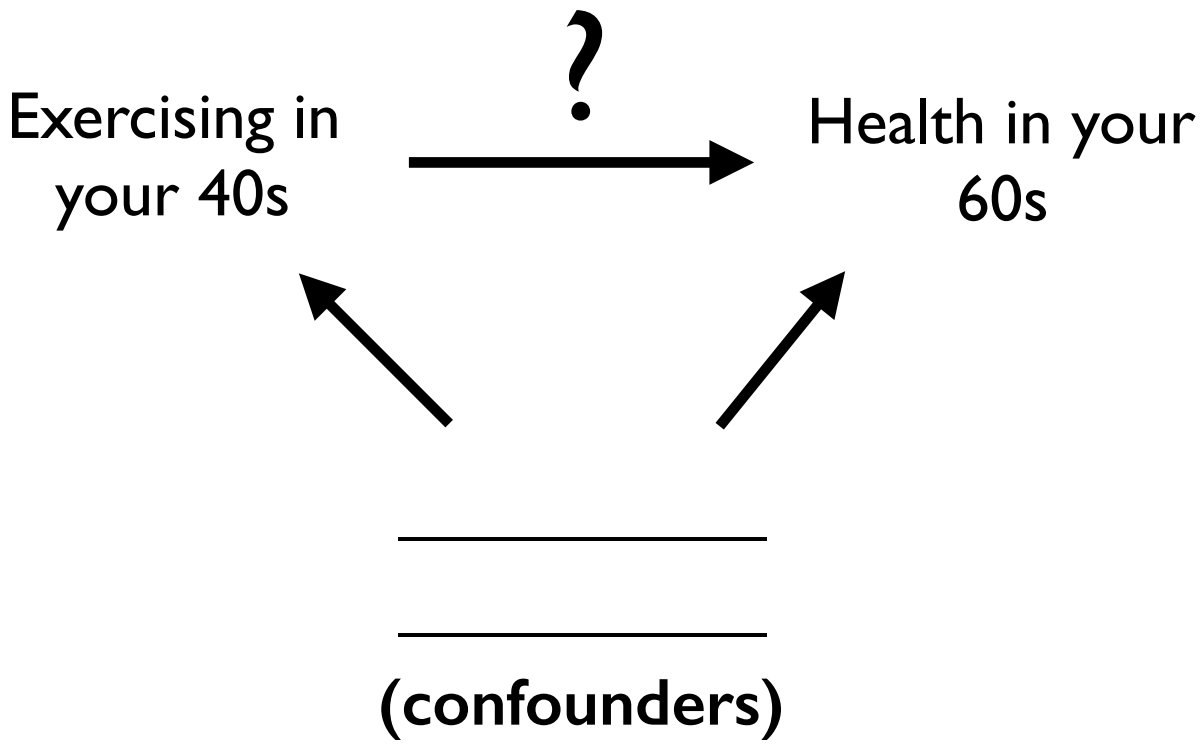
Confounders

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



Confounders

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



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)

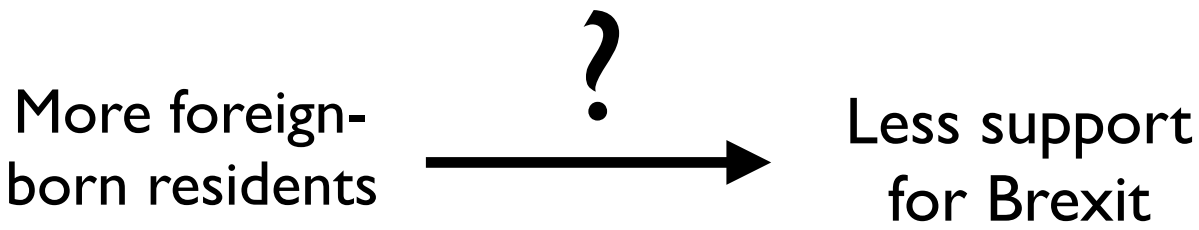
What are possible confounders in the relationship between percent of foreign-born residents and support for Brexit?

More foreign-born residents

Less support for Brexit

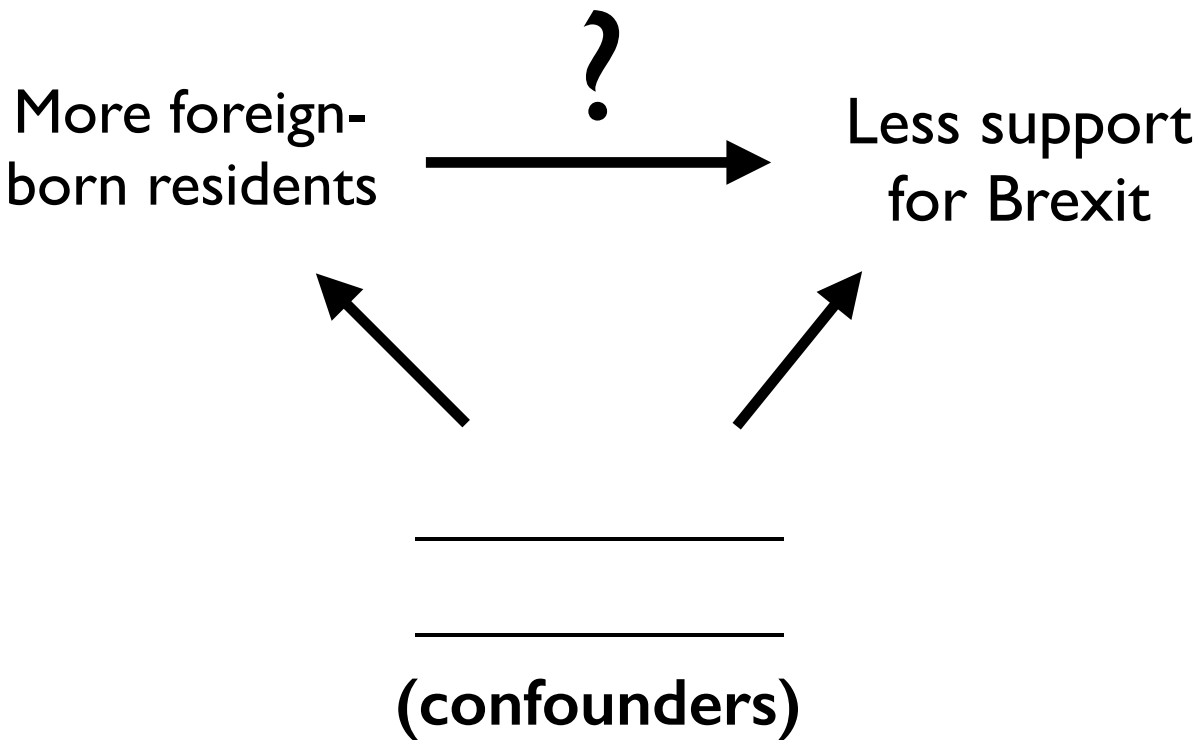
Confounders (2)

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?



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?

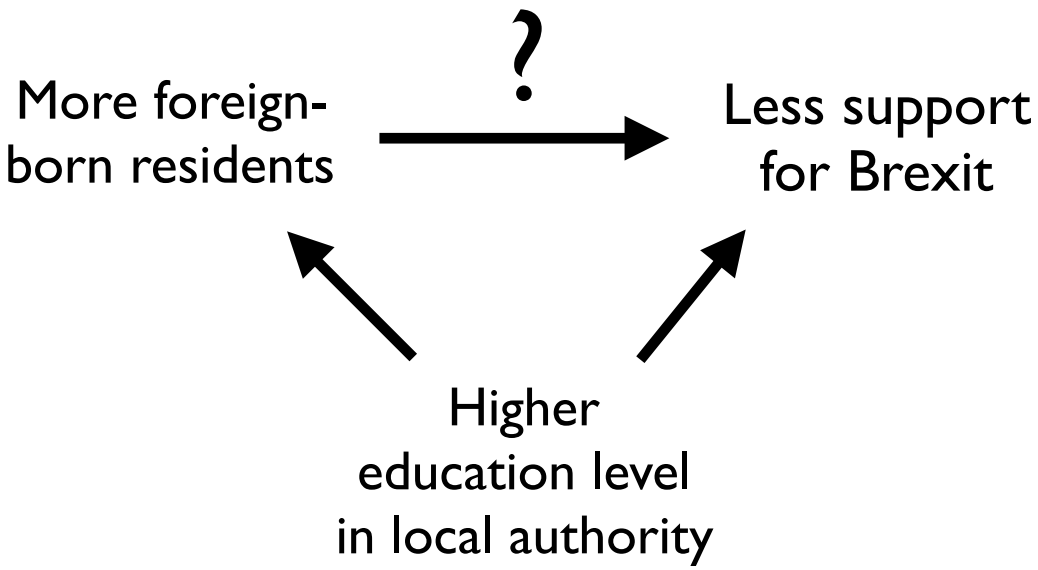
Controlling for confounders

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 _____?

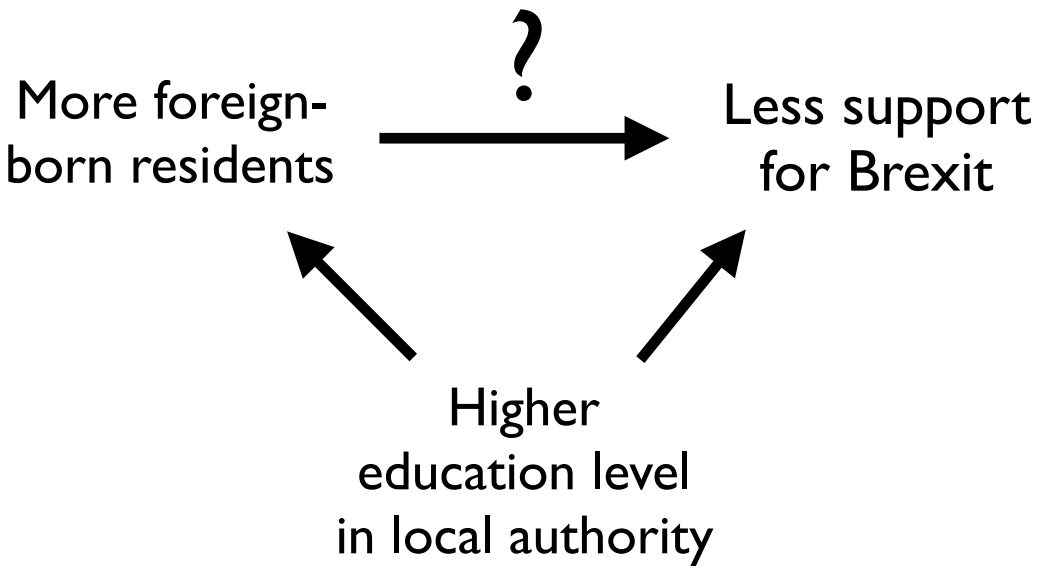
How do we control for confounders?

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



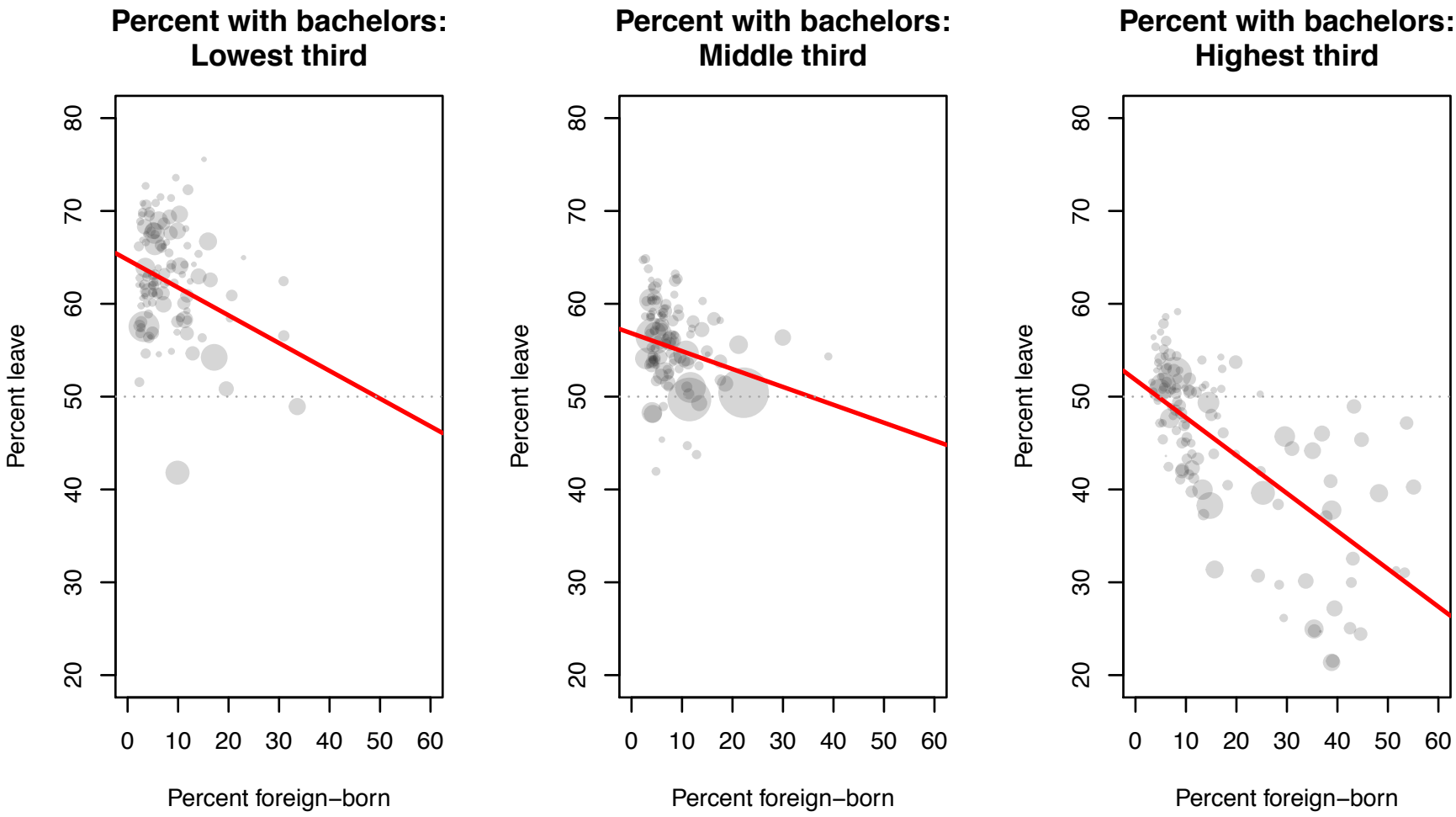
How do we control for confounders?

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: measure relationship between %foreign-born and %leave after stratifying by education level



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"

Thinking about multivariate regression in terms of derivatives (1)

Thinking about multivariate regression in terms of derivatives (1)

Last week, we chose an intercept (β_0) and slope (β_1) to minimize the sum of squared residuals produced by this equation:

Thinking about multivariate regression in terms of derivatives (1)

Last week, we chose an intercept (β_0) and slope (β_1) to minimize the sum of squared residuals produced by this equation:

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

β_1 was the **derivative** of predicted PercentLeave w.r.t. PercentForeignBorn, i.e. the slope.

Thinking about multivariate regression in terms of derivatives (2)

Thinking about multivariate regression in terms of derivatives (2)

Now, we choose an intercept (β_0) and **two** slopes (β_1 and β_2) to minimize the sum of squared residuals produced by this equation:

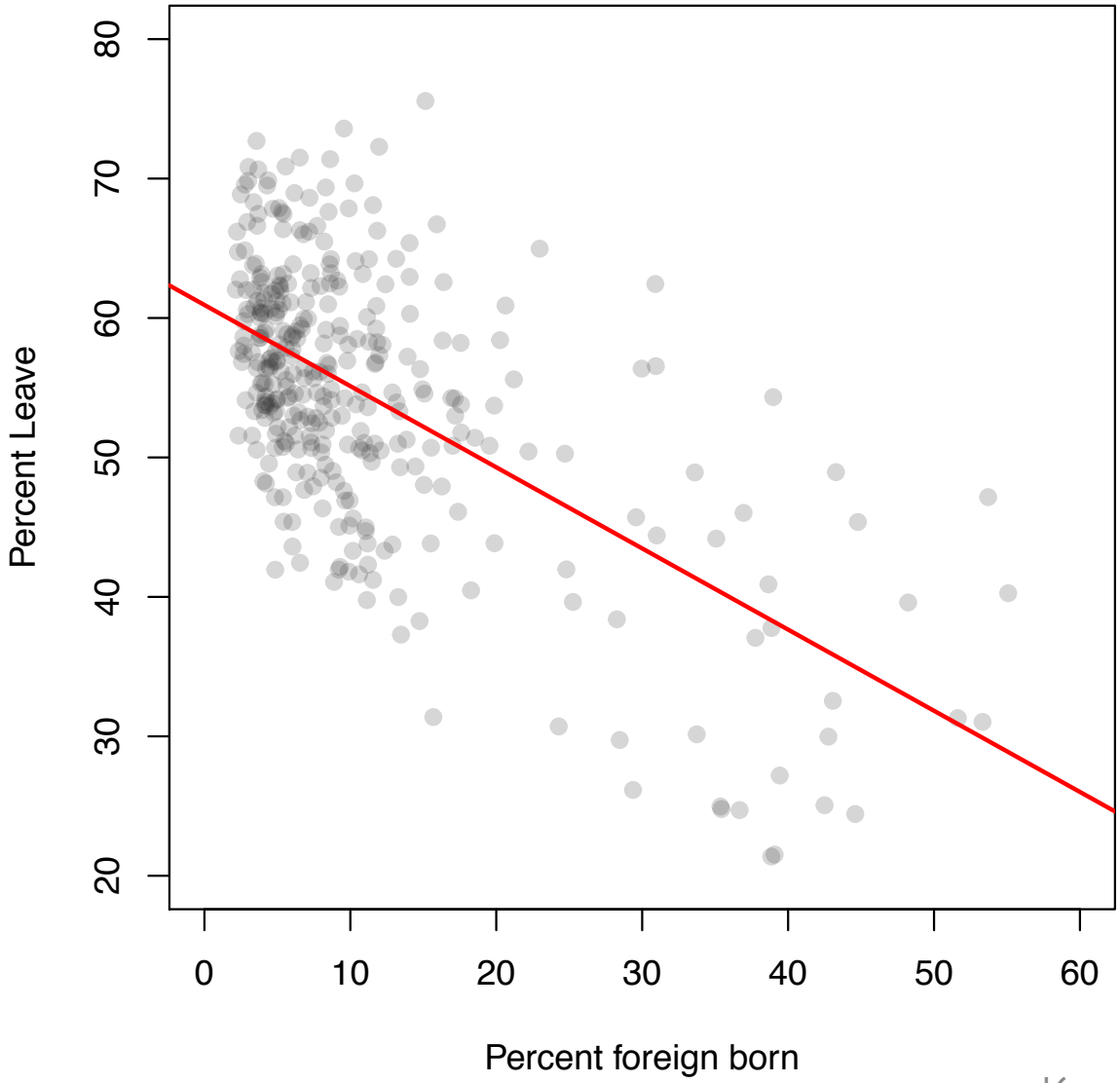
Thinking about multivariate regression in terms of derivatives (2)

Now, we choose an intercept (β_0) and **two** slopes (β_1 and β_2) to minimize the sum of squared residuals produced by this equation:

$$\text{PercentLeave} = \beta_0 + \beta_1 \text{PercentForeignBorn} + \beta_2 \text{Education}$$

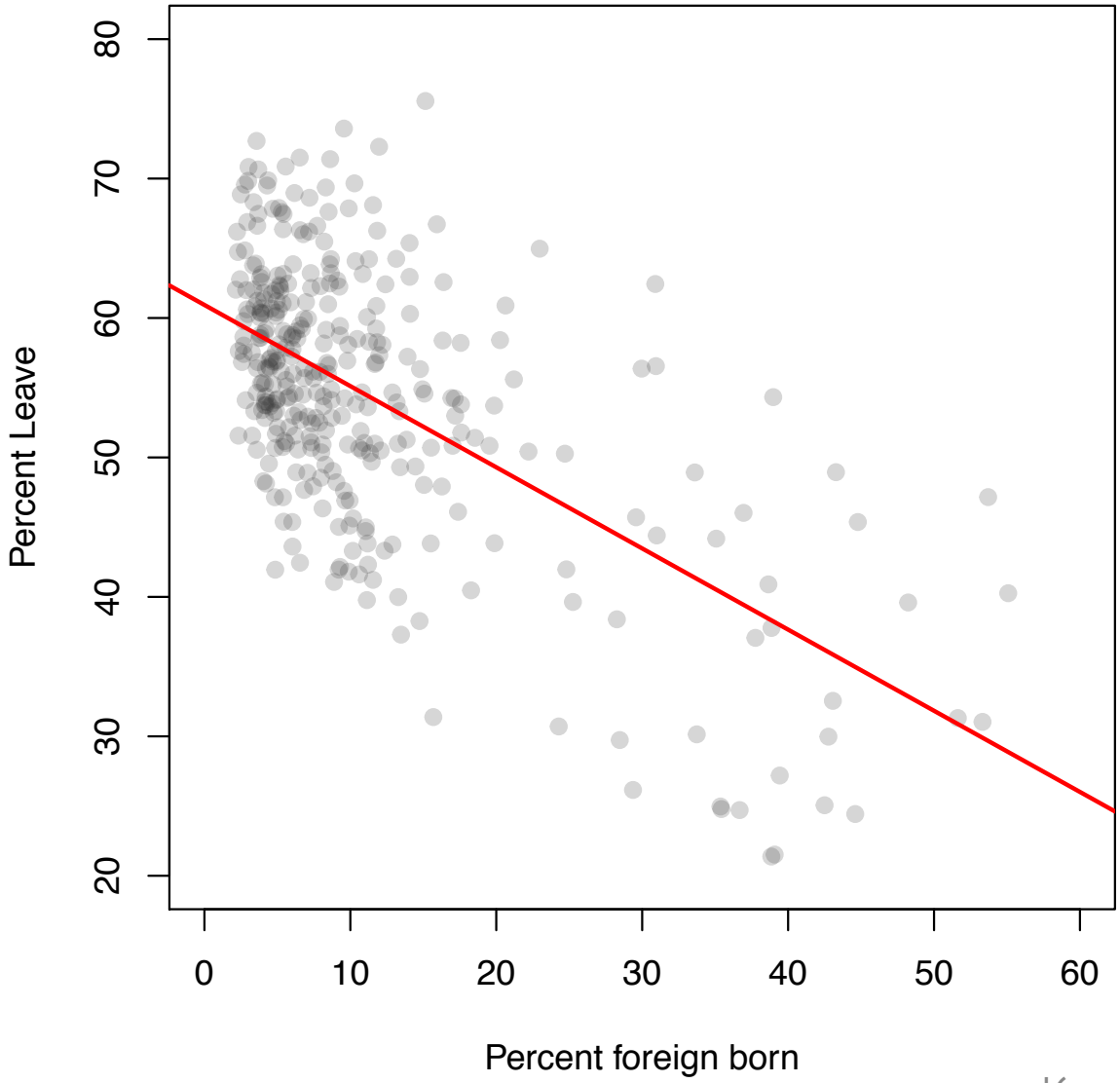
β_1 is now the **partial derivative** of predicted PercentLeave w.r.t. PercentForeignBorn, i.e. the slope holding constant Education.

Visualizing bivariate regression

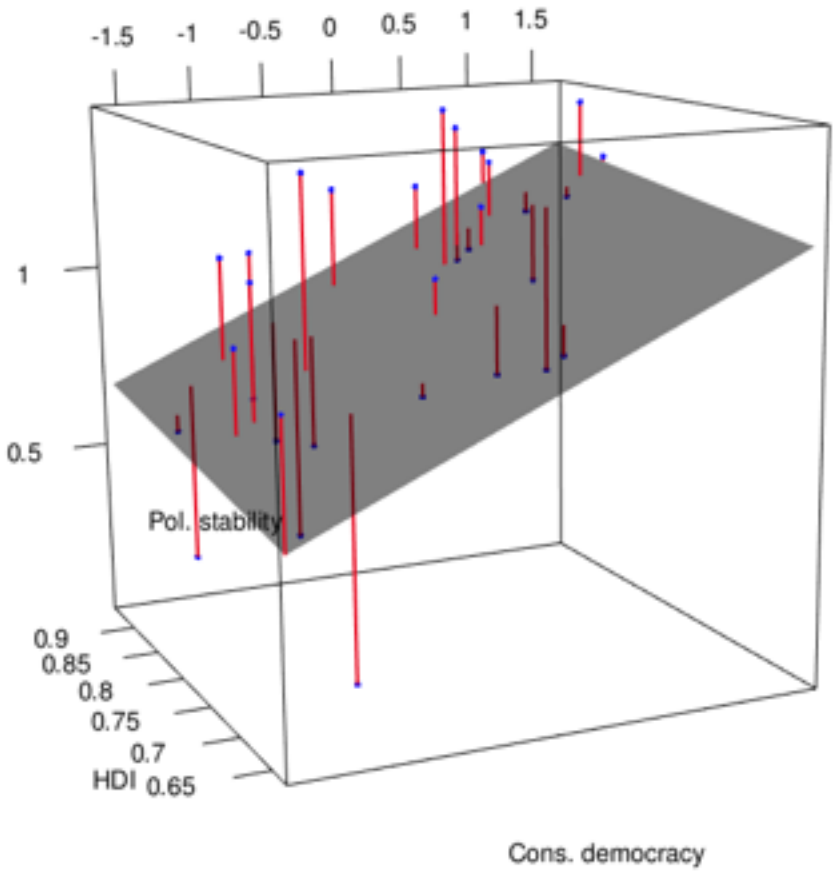


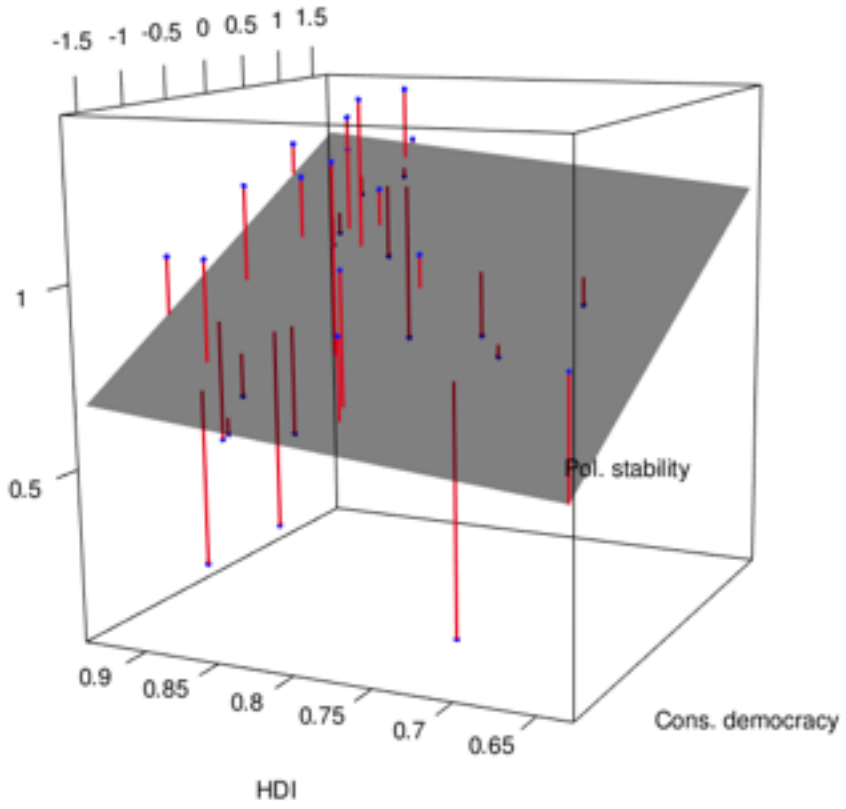
Visualizing bivariate regression

With just two variables (bivariate regression), the regression equation is a **line**.

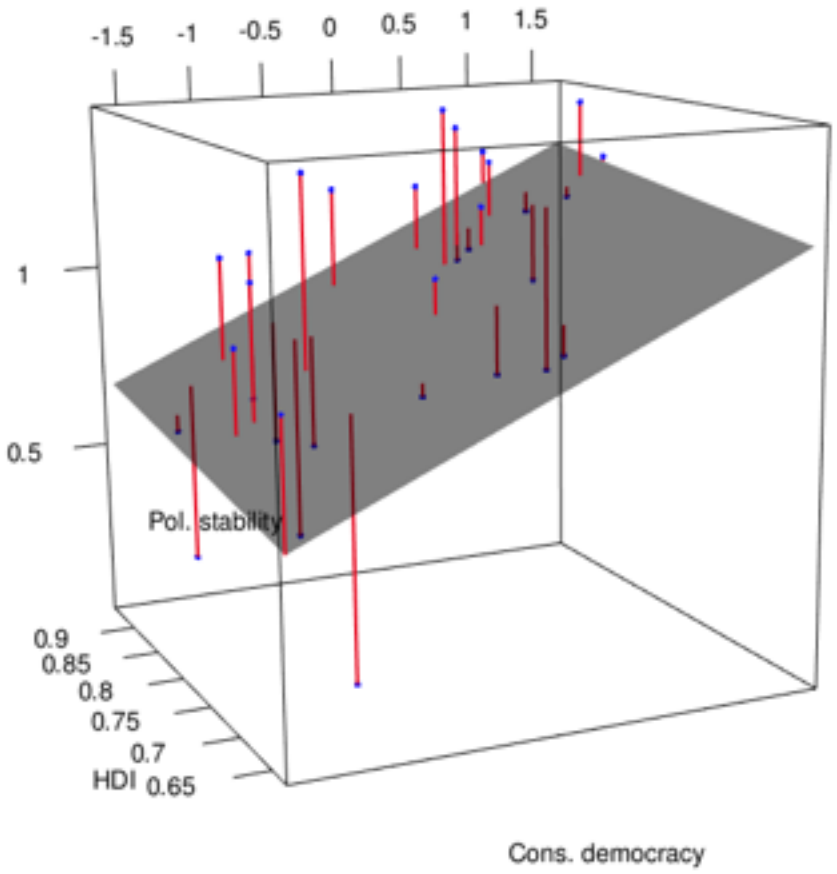


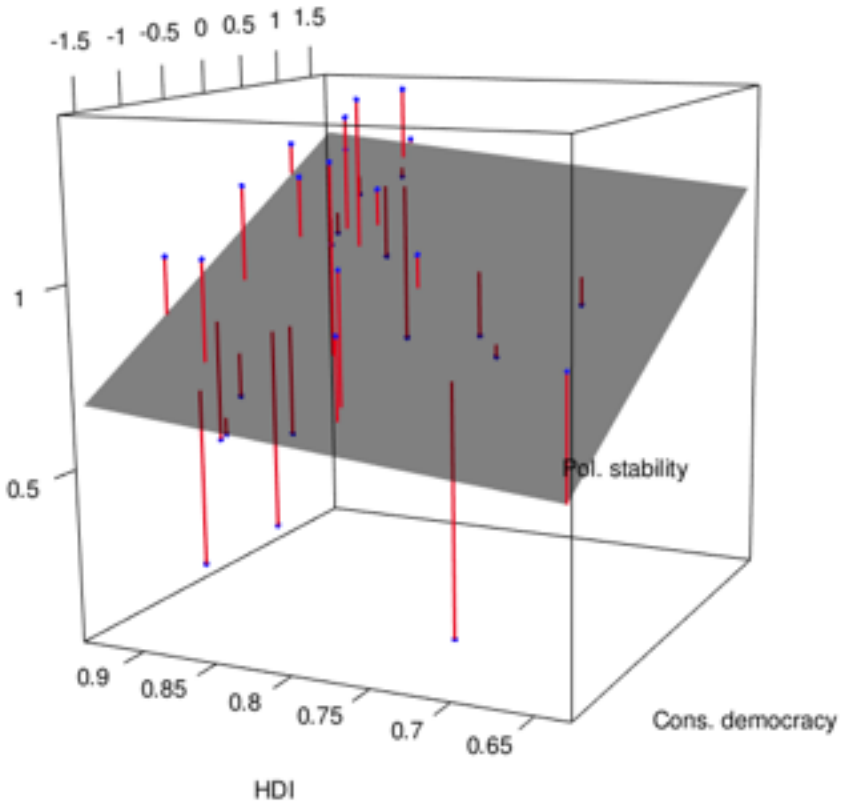
Visualizing multivariate regression





Visualizing multivariate regression





With three variables (multivariate regression), the regression equation is a **plane**.

Visualizing multivariate regression

Visualizing multivariate regression

We can still visualize multivariate regression coefficients with a scatterplot — and it clarifies what it means to **control**.

Visualizing multivariate regression

We can still visualize multivariate regression coefficients with a scatterplot — and it clarifies what it means to **control**.

Visualizing multivariate regression

We can still visualize multivariate regression coefficients with a scatterplot — and it clarifies what it means to **control**.

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

Visualizing multivariate regression

We can still visualize multivariate regression coefficients with a scatterplot — and it clarifies what it means to **control**.

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

Step 2: calculate residuals (the part of % foreign-born not explained by education)

Visualizing multivariate regression

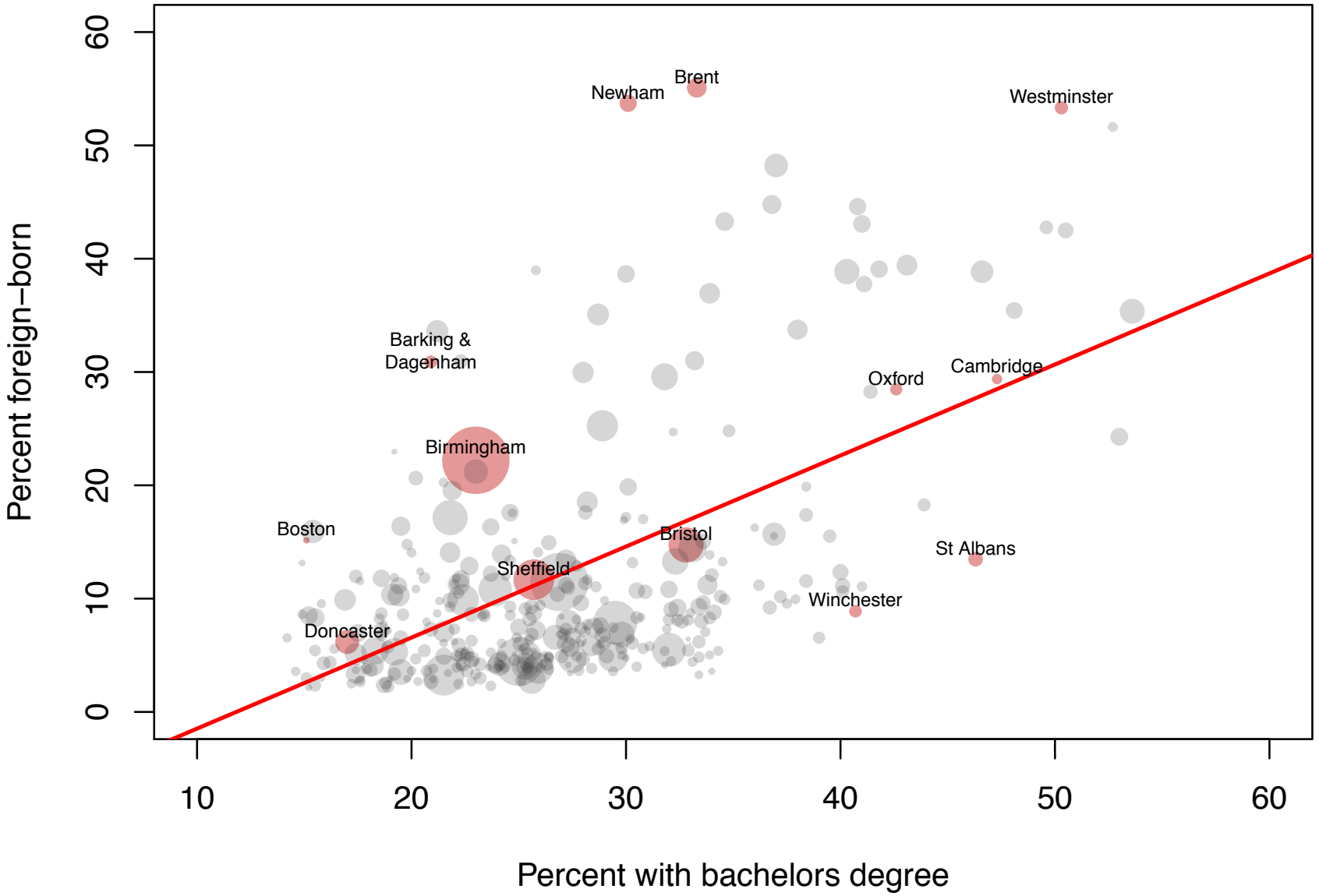
We can still visualize multivariate regression coefficients with a scatterplot — and it clarifies what it means to **control**.

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

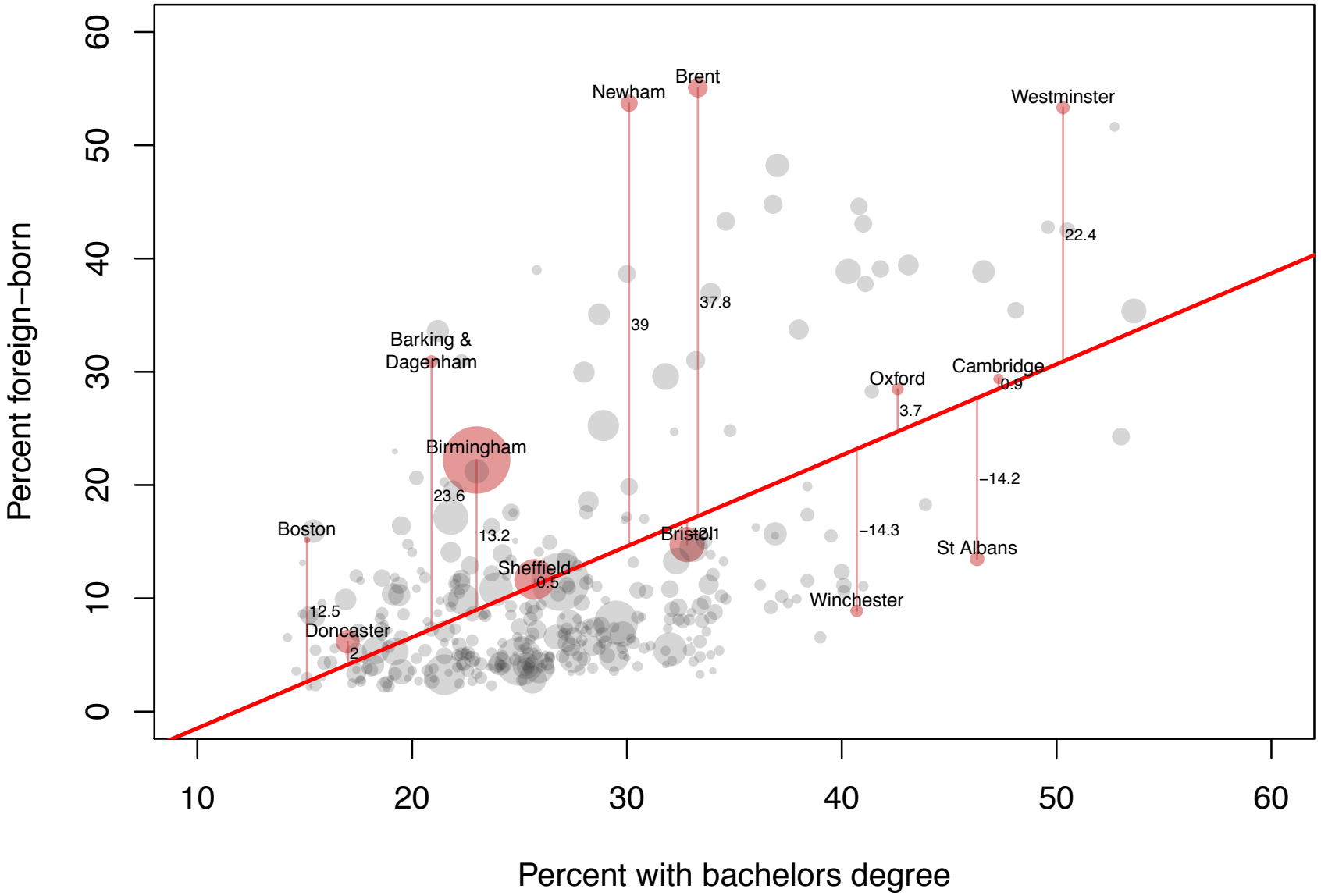
Step 2: calculate residuals (the part of % foreign-born not explained by education)

Step 3: regress outcome (% leave) on residuals from step 2

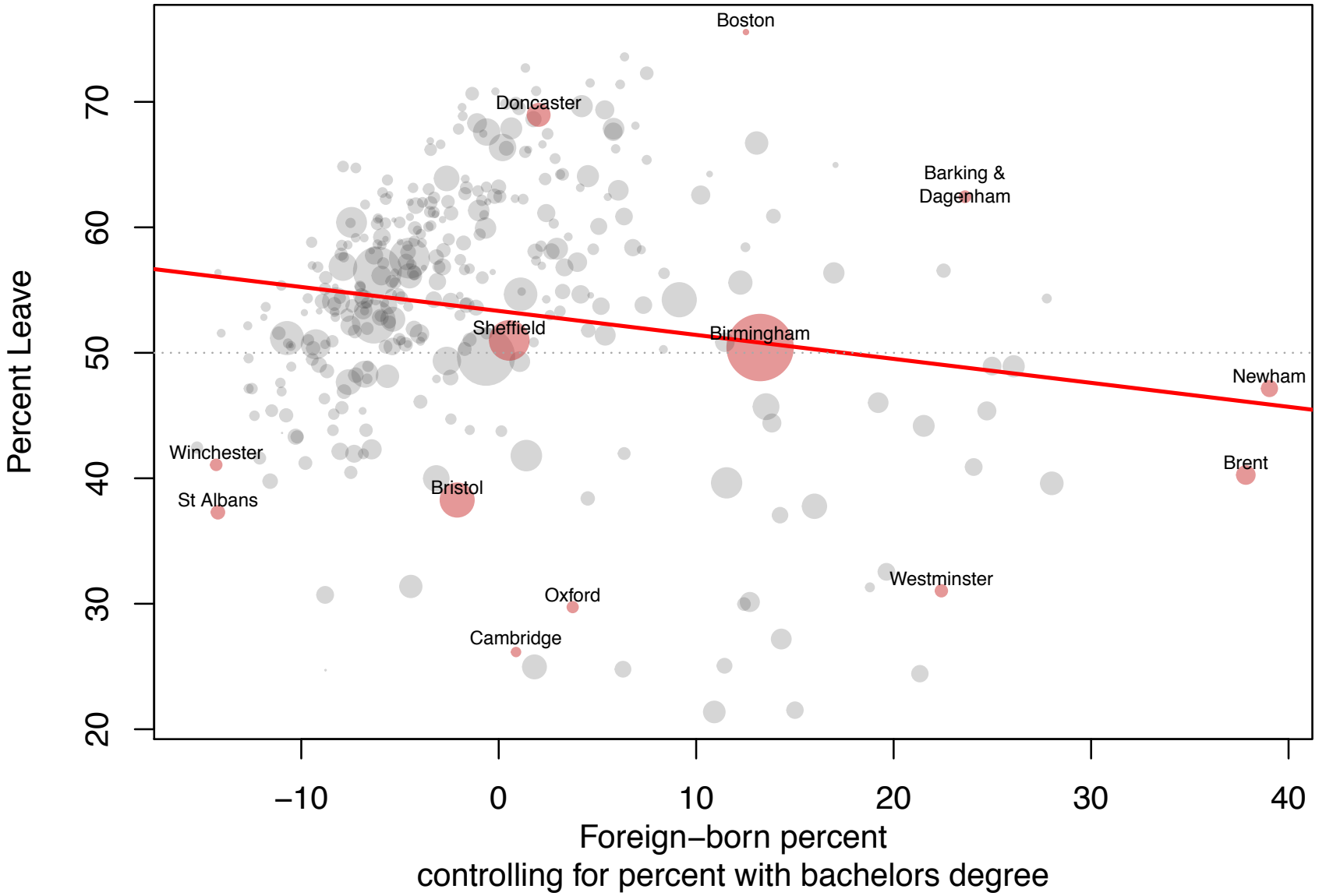
Step 1: regress explanatory variable (%foreign-born) on confounder (education)



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



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



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
```