

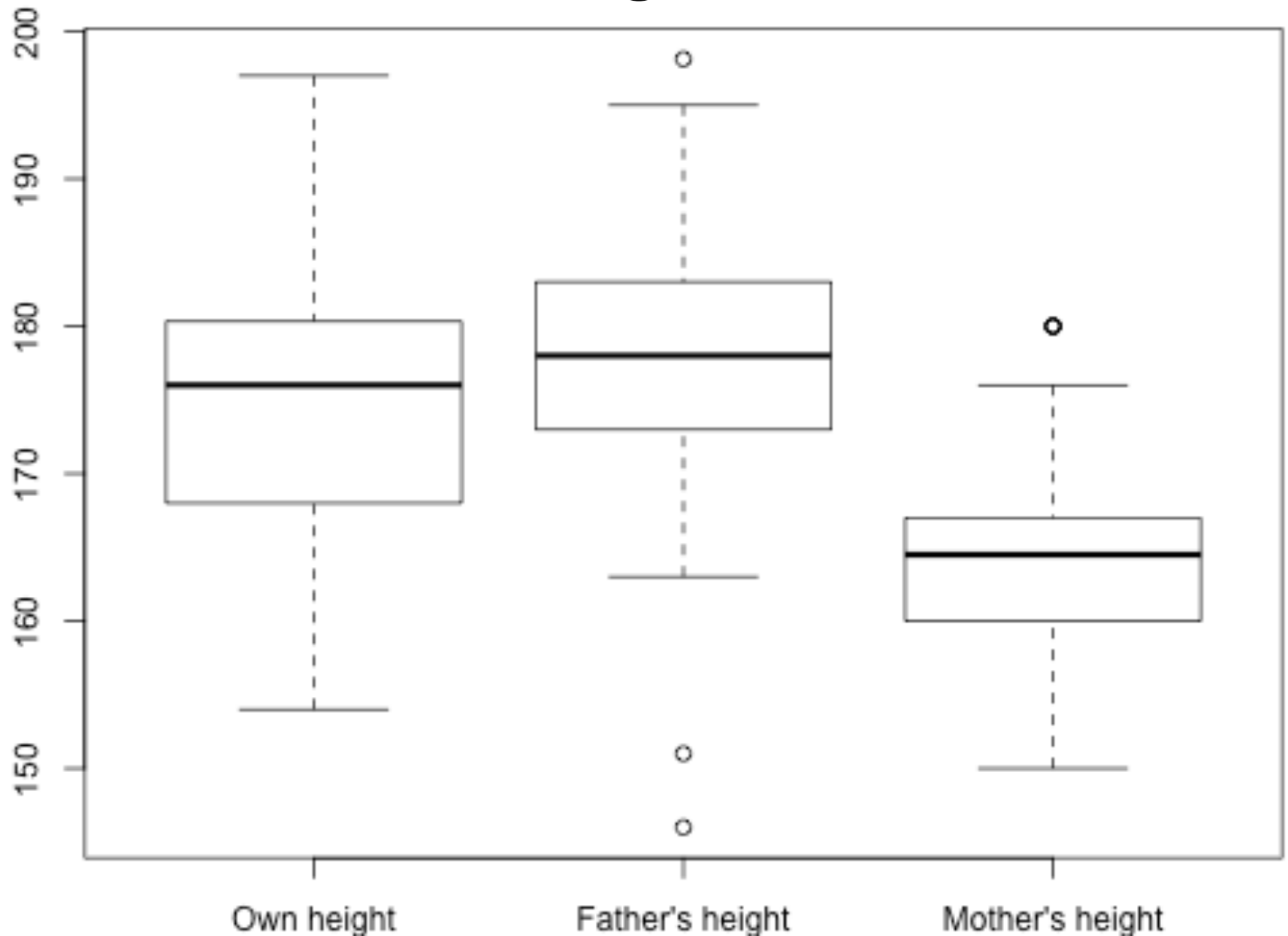
Multivariate relationships

Week 7

1 March, 2015

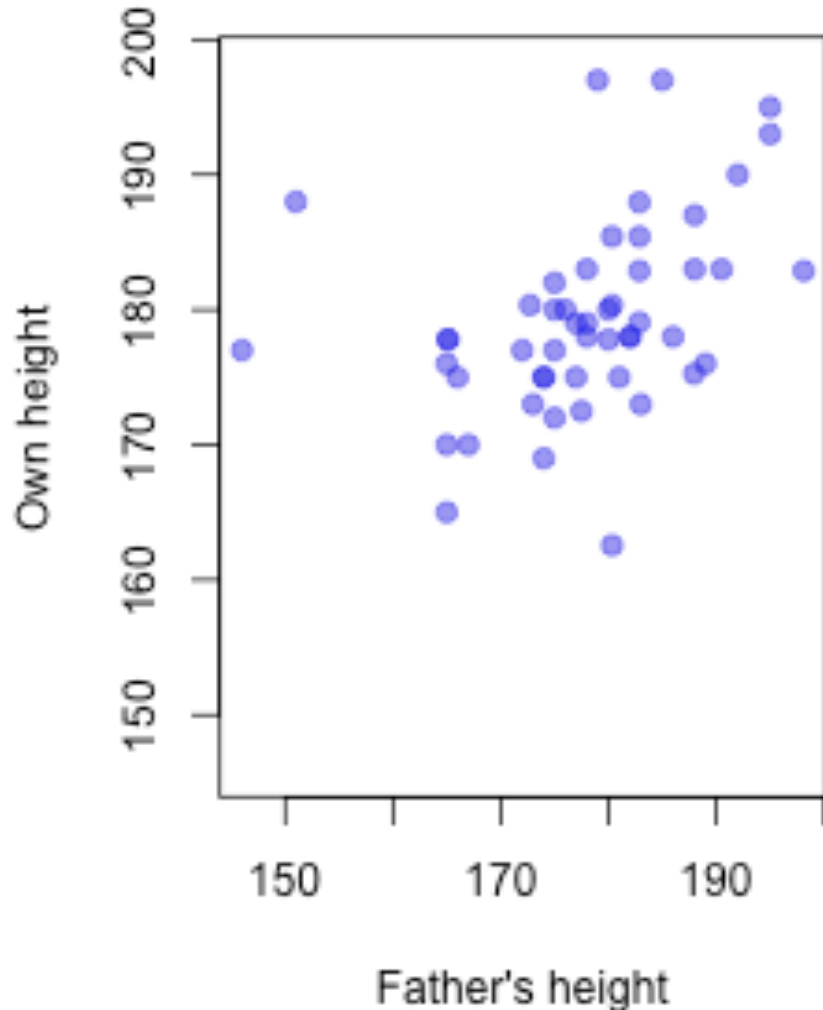
Prof. Andrew Eggers

Your height data

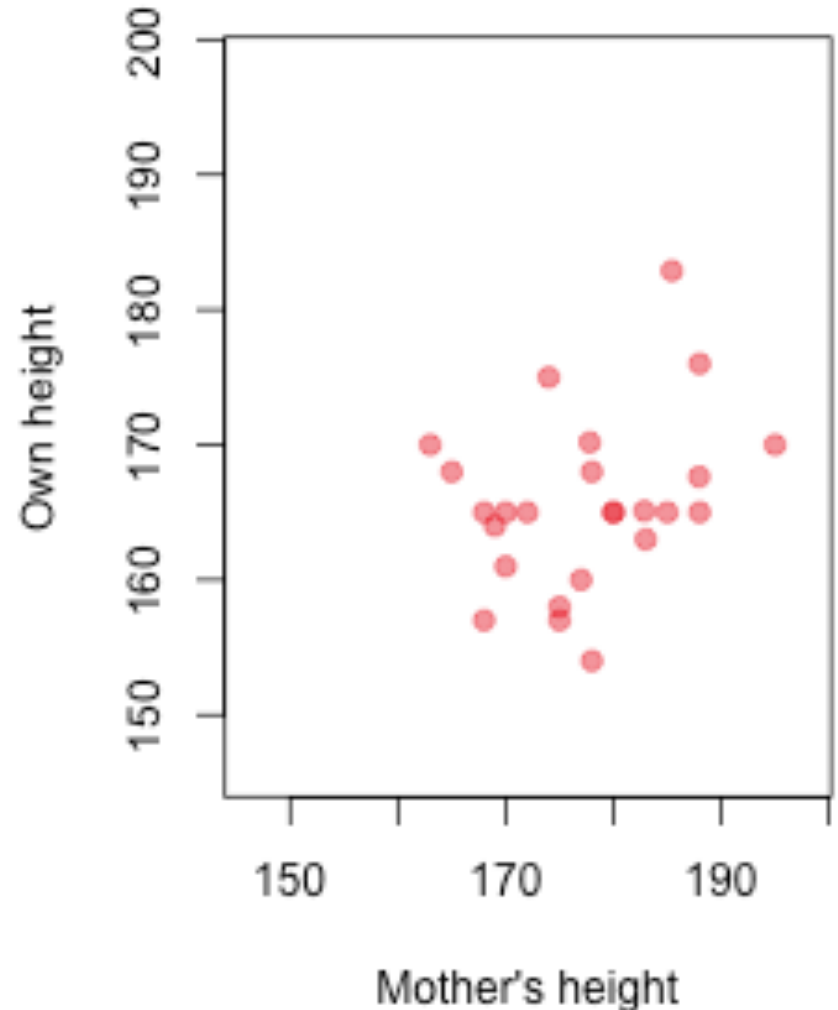


Your height data, by gender

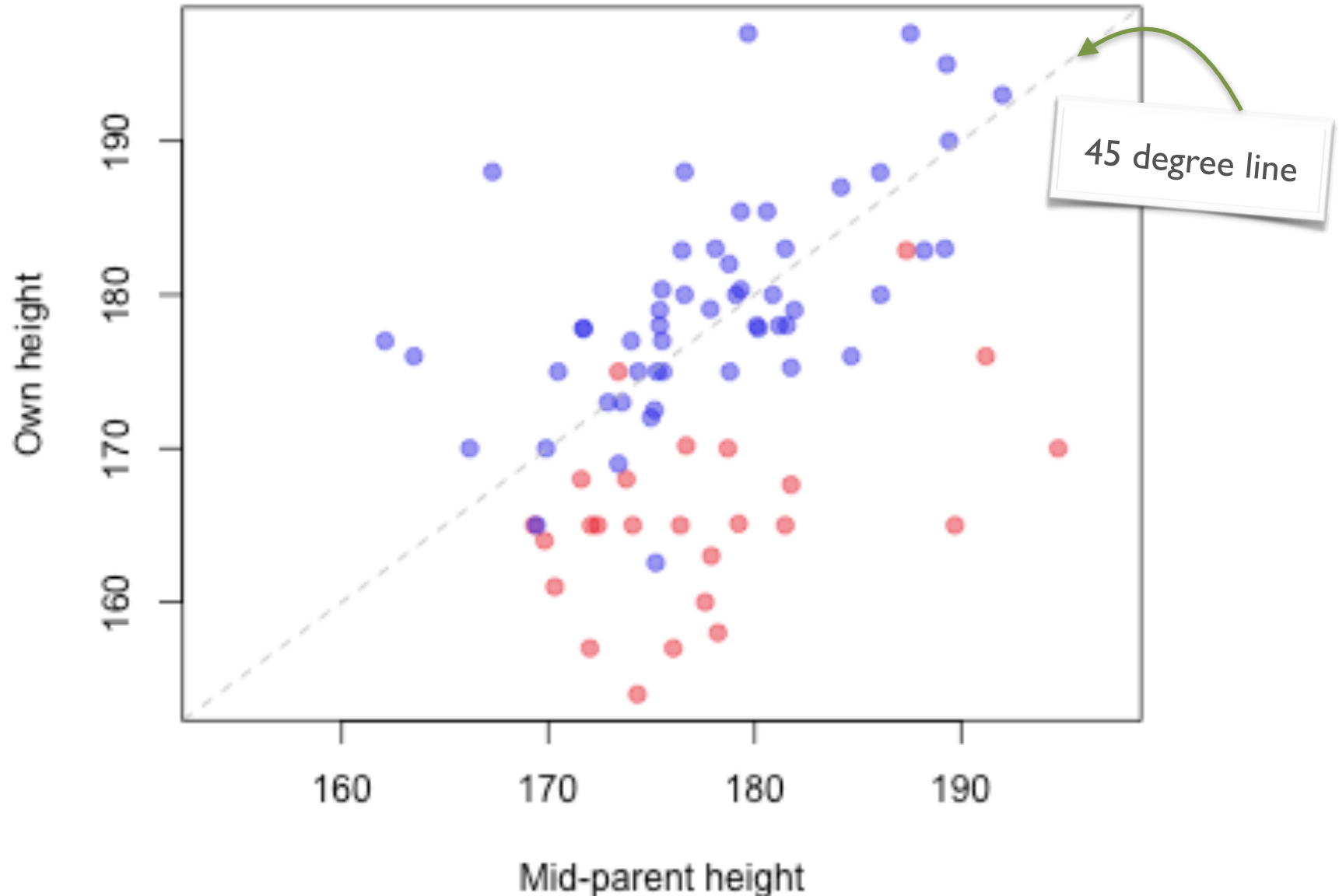
Male students



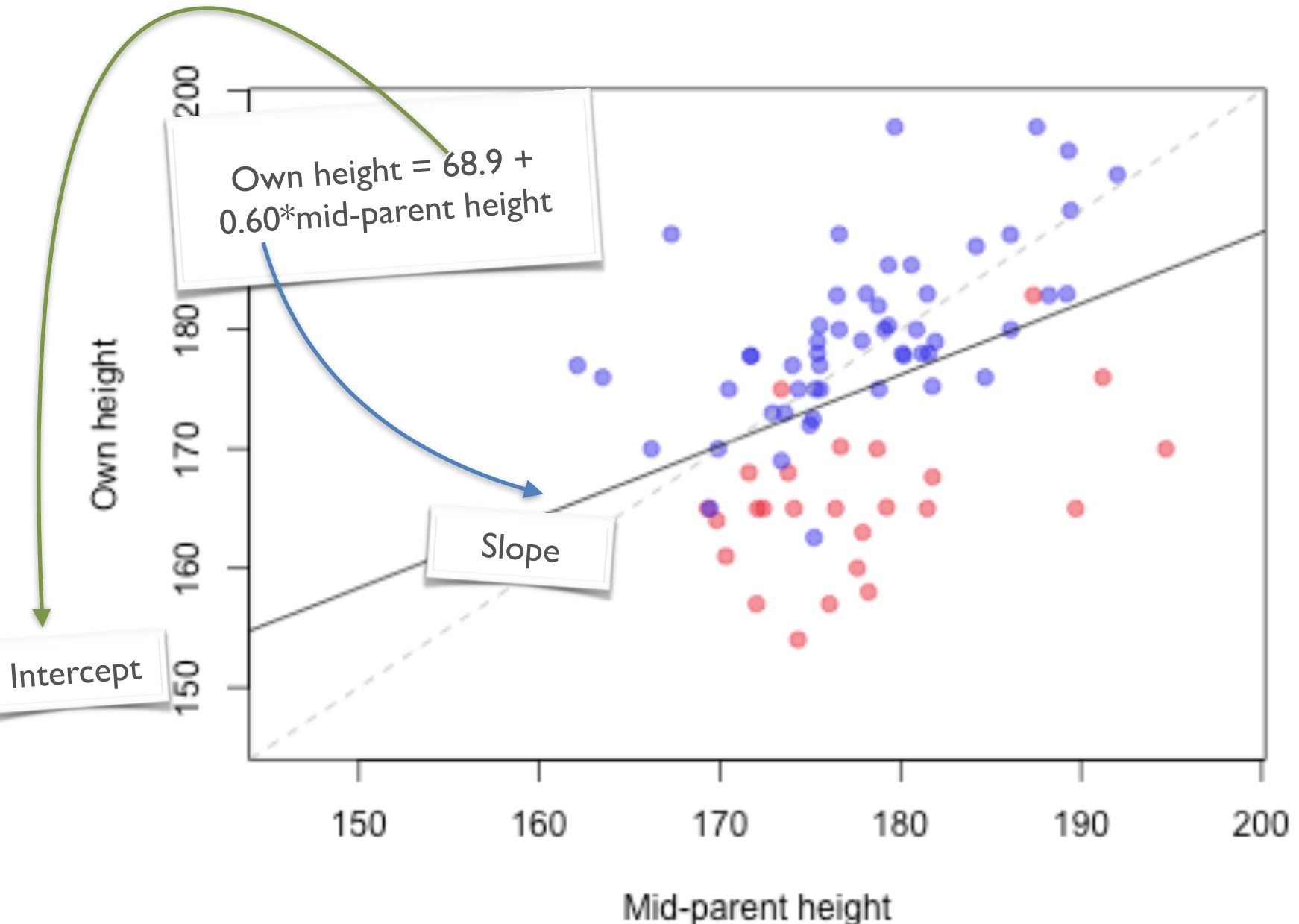
Female students



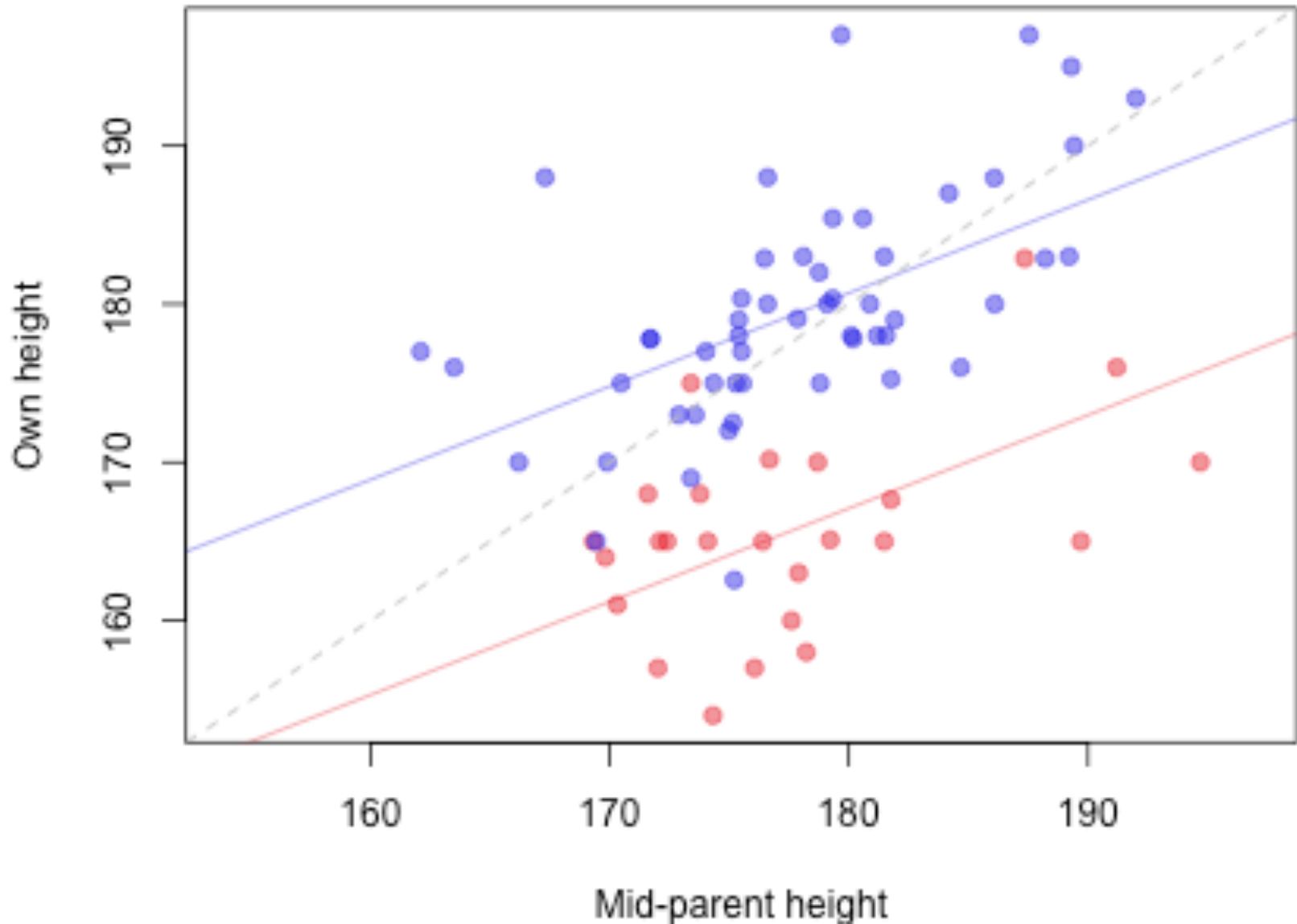
Own height and “mid-parent” height



Regression line shows “regression to the mean”!

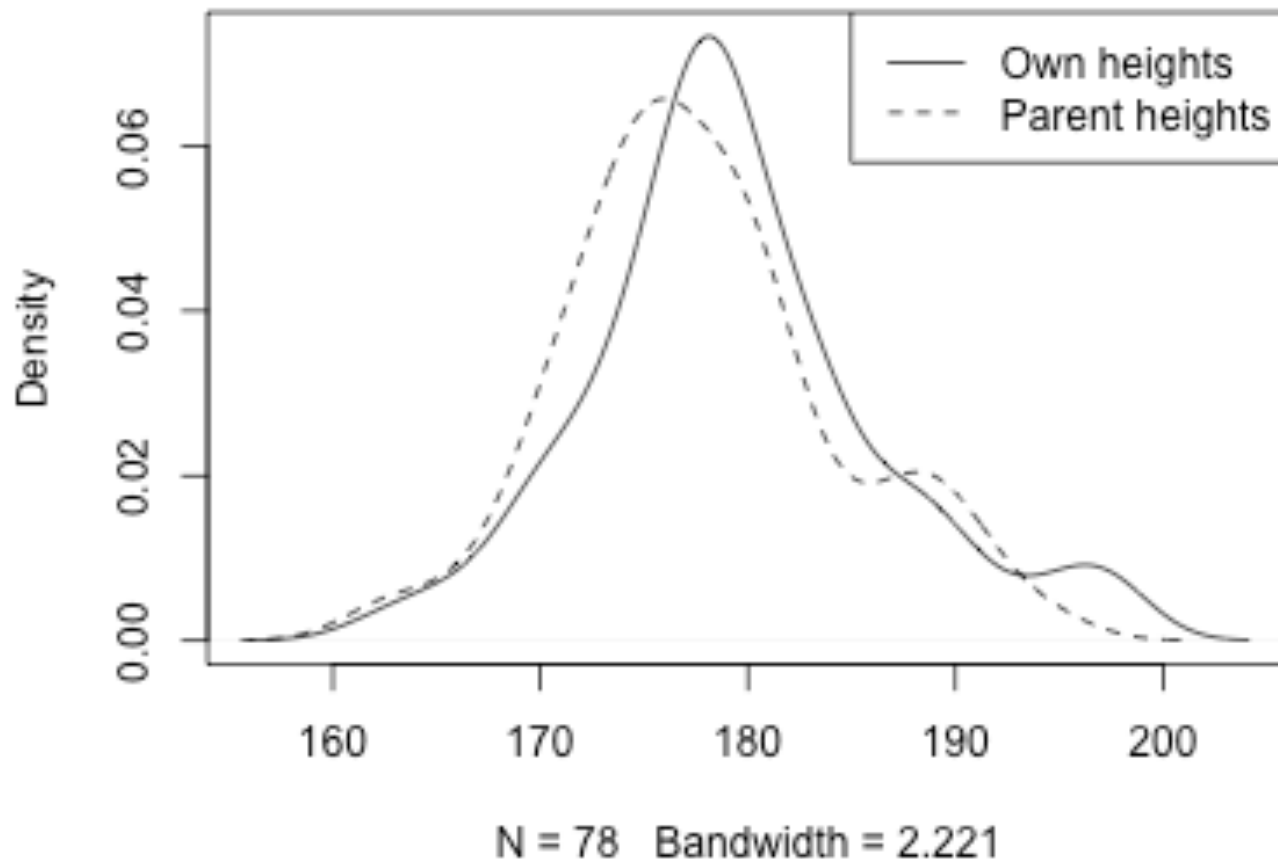


Same thing, controlling for own gender



(So is each generation more average than the last? No.)

Distribution of own heights and parent heights
(Both adjusted for gender)



Regression to the mean: because height is not entirely heritable (due to e.g. randomness of genetics), very tall/short people will be taller/shorter than their children **and their parents**. This happens even when the distribution stays same from generation to the next.

19 November 2012 Last updated at 18:19



Does chocolate make you clever?

By Charlotte Pritchard

BBC News



Eating more chocolate improves a nation's chances of producing Nobel Prize winners - or at least that's what a recent study appears to suggest. But how much chocolate do Nobel laureates eat, and how could any such link be explained?

**In today's
Magazine**

The Swiss children

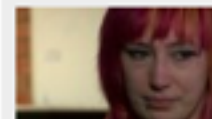
Top Stories

**Penalties 'do not stop'
drug use****Sickness benefit cuts 'considered'****Care plan 'to ease hospital pressure'****Child sex exploitation 'social norm'****MP Jim Murphy joins Labour contest**

Features

**Magical masterpiece**

The Leonardo hidden from Hitler in case it gave him special powers

**'GamerGate'**

The developer forced to leave her home due to threats

**Wake up**

Is eating sage better for your alertness than coffee?

**Rumble revisited**

Forty years since Ali took on Foreman in Kinshasa **BBC SPORT**

**Armageddon file**

The nuclear attack on the UK that...

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

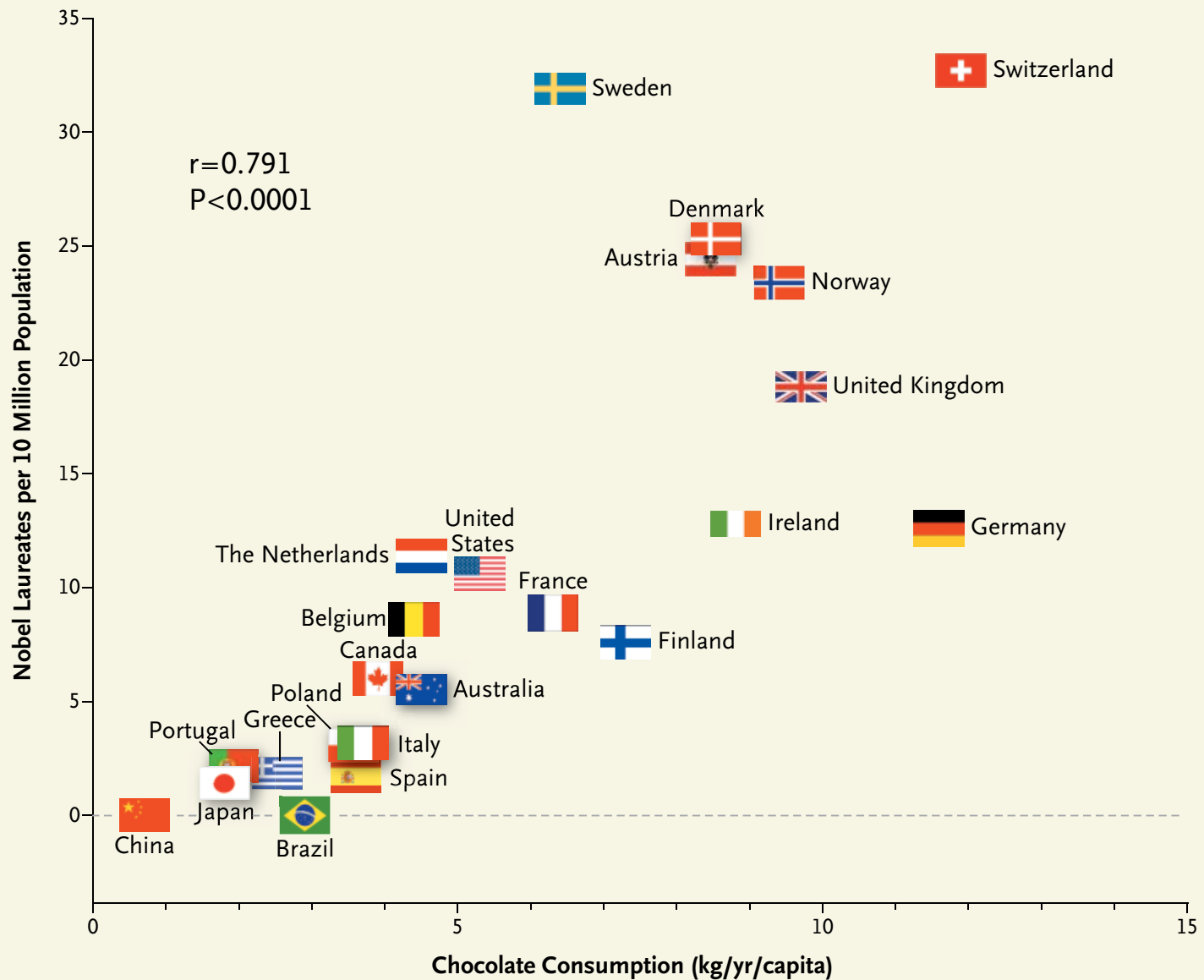
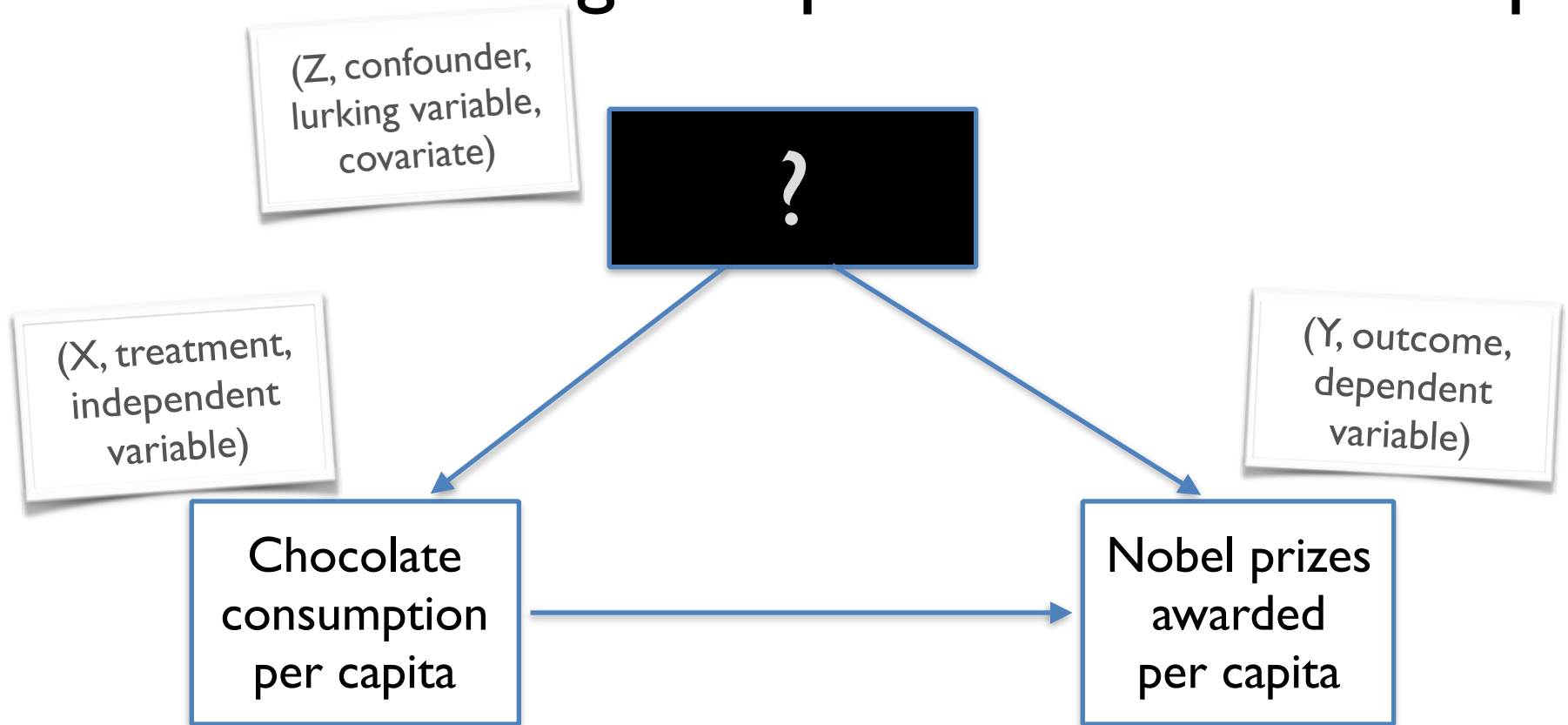


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

What else might explain this relationship?



Spurious?

How do we identify confounders?

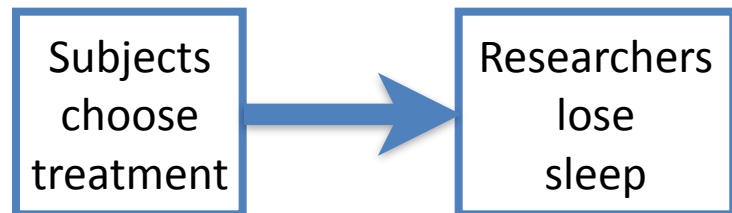
How do we **control** for them?

Best case: randomize treatment

In a randomized experiment, there should be no confounding variables.



In many social science settings, RCT is impossible: subjects (e.g. countries, individuals) choose their own treatment.



Next best: statistical control

TABLE 15.2

Multivariate regression analyses of the effect of consensus democracy (executives-parties dimension) on five indicators of violence, with controls for the effects of the level of economic development, logged population size, and degree of societal division, and with extreme outliers removed

Performance variables	Estimated regression coefficient	Absolute t-value	Countries (N)
Political stability and absence of violence (1996–2009)	0.189***	3.360	34
Internal conflict risk (1990–2004)	0.346**	2.097	32
Weighted domestic conflict index (1981–2009)	–105.0*	1.611	30
Weighted domestic conflict index (1990–2009)	–119.7**	2.177	33
Deaths from domestic terrorism (1985–2010)	–2.357**	1.728	33

* Statistically significant at the 10 percent level (one-tailed test)

** Statistically significant at the 5 percent level (one-tailed test)

*** Statistically significant at the 1 percent level (one-tailed test)

Source: Based on data in Kaufmann, Kraay, and Mastruzzi 2010; PRS Group 2004; Banks, 2010; and GTD Team 2010

Source: Lijphart (2012)

META-ANALYSIS

Sedentary time in adults and the association with diabetes, cardiovascular disease and death: systematic review and meta-analysis

E. G. Wilmot • C. L. Edwardson • F. A. Achana •
M. J. Davies • T. Gorely • L. J. Gray • K. Khunti •
T. Yates • S. J. H. Biddle

Diabetologia (2012) 55:2895–2905

2899

Table 1 Characteristics of cross-sectional and prospective cohort studies included in meta-analysis

Author [ref.]	Design, sample size	Outcome, no. cases	Sedentary measure used in meta-analysis	Confounders measured	Quality
Dunstan et al 2004 [21]	Cross-sectional 8,299 Australian men and women	Diabetes 252 cases (3%)	TV viewing >14 vs <14 h/week	Adjusted for age, education, FHx DM, smoking, diet and PA	5
Dunstan et al 2010 [32]	Prospective 6.6 year f/u 8,800 Australian men and women	Cardiovascular mortality 87 cases (1%) All-cause mortality 284 cases (3.2%)	TV viewing ≥4 vs <2 h/day	Adjusted for age, sex, smoking, education, diet	6
Ford et al 2010 [24]	Prospective 7.8 year f/u 23,855 German men and women	Diabetes 927 cases (3.9%)	TV viewing <1 vs >4 h/day	Adjusted for age, sex, education, occupational activity, smoking, alcohol, PA, diet, systolic BP	3
Hawkes et al 2011 [29]	Prospective 3 year f/u 1,966 Australian	Diabetes 247 cases (12.6%) ^a Cardiovascular	TV viewing <2 vs >4 h/day	Sex, age, education, marital status Diabetes outcome ^b	4

When statistical control really matters

Classic example: Cochran (1968) on risk of pipe smoking vs cigarette smoking



THE EFFECTIVENESS OF ADJUSTMENT BY SUBCLASSIFICATION IN REMOVING BIAS IN OBSERVATIONAL STUDIES

W. G. COCHRAN

Harvard University, Cambridge, Mass., U. S. A.

SUMMARY

In some investigations, comparison of the means of a variate y in two study groups may be biased because y is related to a variable z whose distribution differs in the two groups. A frequently used device for trying to remove this bias is adjust-

Are pipes worse than cigarettes?



Outcome (y): Death rate

Study groups (x): pipe smokers and cigarette smokers

Death rate is higher among pipe smokers.

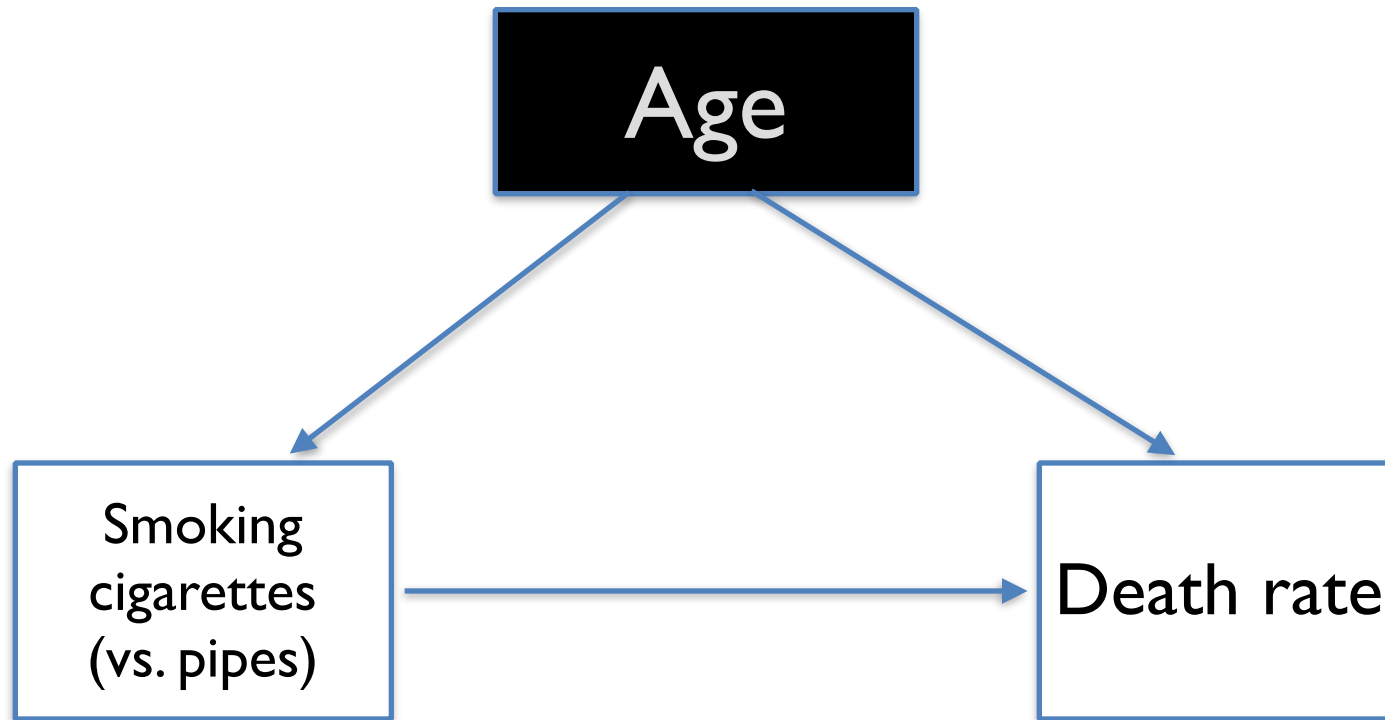
But: pipe smokers are older (z).



	US		UK	
	<i>Raw death rates</i>	<i>Adjusted for age</i>	<i>Raw death rates</i>	<i>Adjusted for age</i>
<i>Pipe smokers</i>	17.4	13.7	20.7	11.0
<i>Cigarette smokers</i>	13.5	21.2	11.0	14.8

Source: Cochran (1968)

Why does controlling for age reverse the conclusion?



When would controlling for a confounder **strengthen** the conclusion?

Is consensus democracy better than majoritarian democracy?

Outcome (y): e.g. unemployment

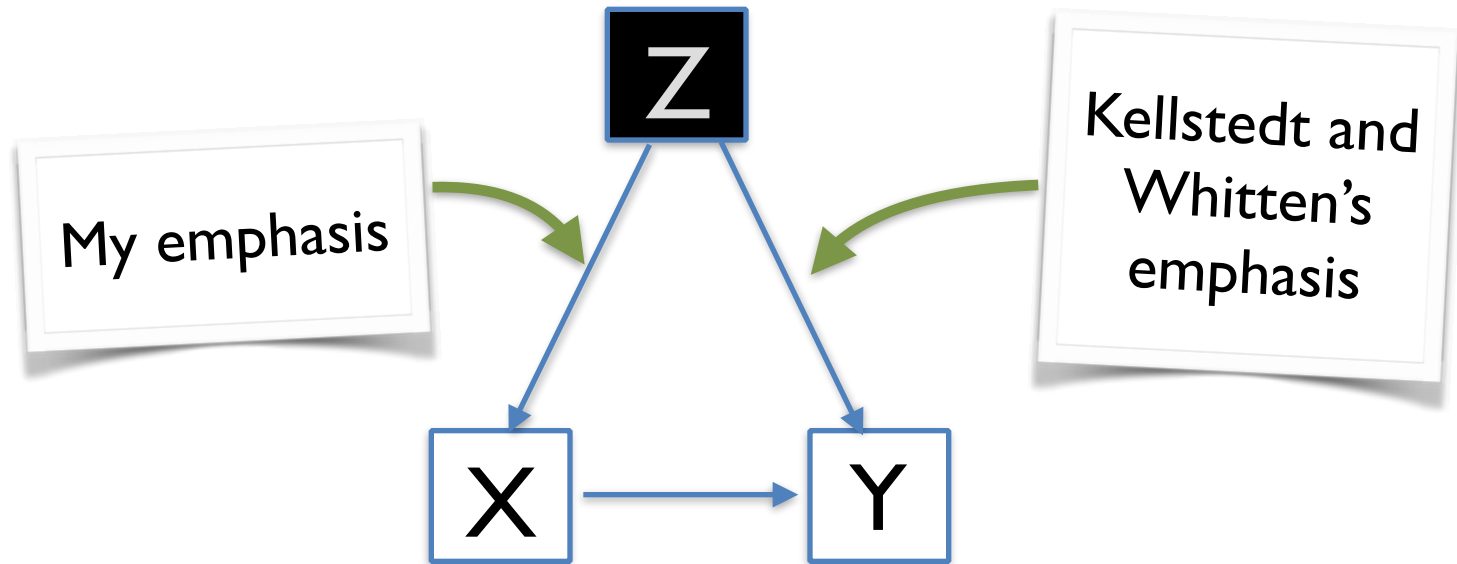
Study groups (x): countries with consensus forms of democracy (e.g. Finland, Netherlands) and majoritarian forms (e.g. UK, Bahamas)

Unemployment is lower in consensus democracies.

But: These countries differ in many other ways! (z). Which differences should we control for?



What do we need to control for?



Kellstedt and Whitten: control for “as many causes of the dependent variable as possible” (60) (“technically ... only those related to X”, fn p. 87)

Me: control for *determinants** of X that might also affect Y

*Not effects.



Example: Smoking and mortality



What determines whether someone smokes a pipe or a cigarette?

- Age
-
-
-
-

What determines whether someone dies?

- Age
-
-
-
-

What about when thinking about consensus democracy and unemployment?

How do we control?

Two intuitive approaches

Subclassification: compare outcomes for subjects within intervals of a covariate.

What about multiple confounders?

	<i>Mortality</i>	
	<i>Cig. smoker</i>	<i>Pipe smoker</i>
<i>Age 55-60</i>	8.2	6.1
<i>Age 60-65</i>	10.4	8.7

etc.




Matching: for every “treated” unit, find a similar “untreated” unit. Compare the two groups.

The regression approach

The bivariate regression

$$\text{Mortality}_i = \beta_0 + \beta_1 \text{PipeSmoker}_i$$


says “pipe smoking is worse than cigarettes”, but


$$(\beta_1 > 0)$$

the multivariate regression

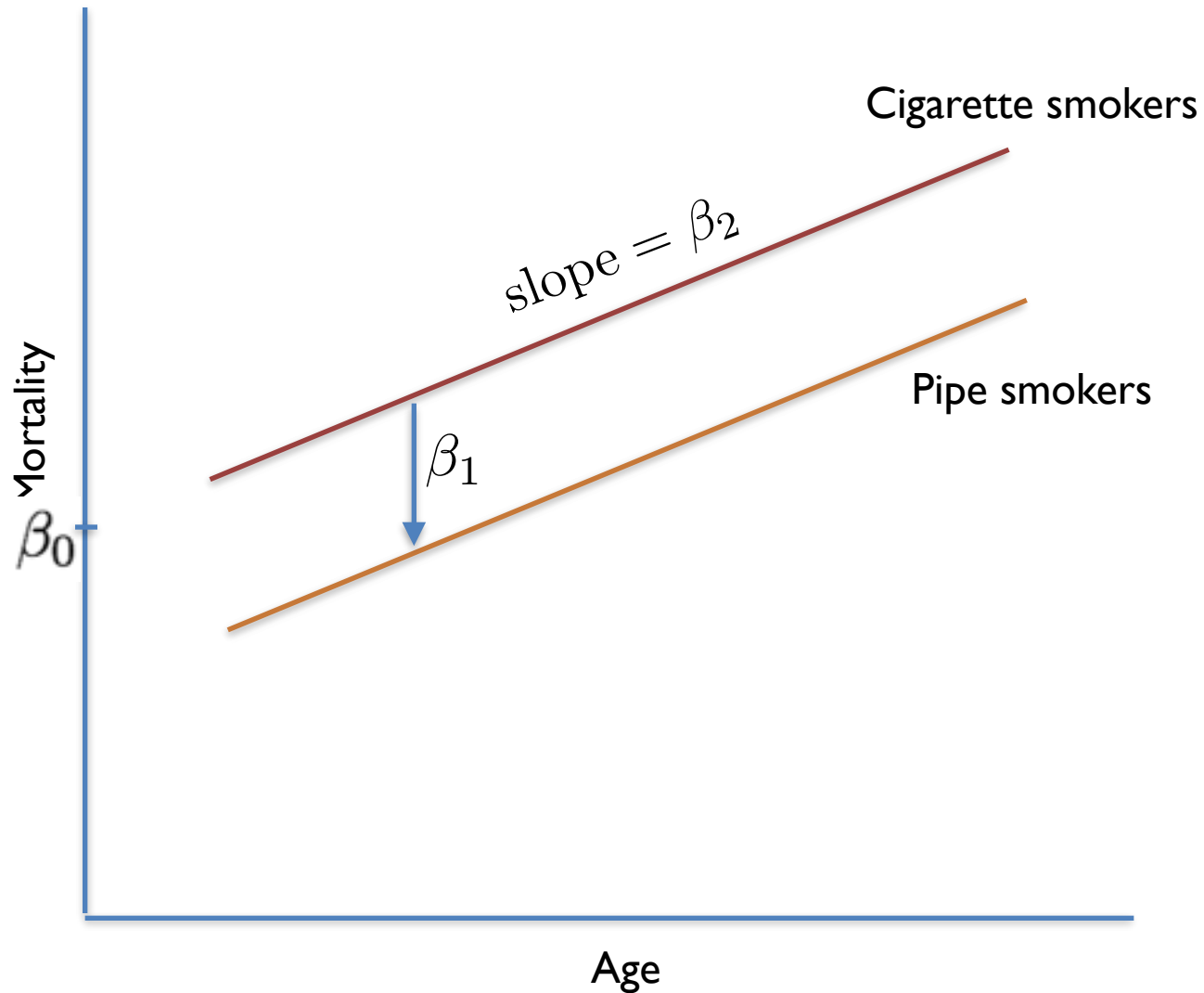
$$\text{Mortality}_i = \beta_0 + \beta_1 \text{PipeSmoker}_i + \beta_2 \text{Age}_i$$

says “cigarettes are worse”.


$$(\beta_1 < 0)$$

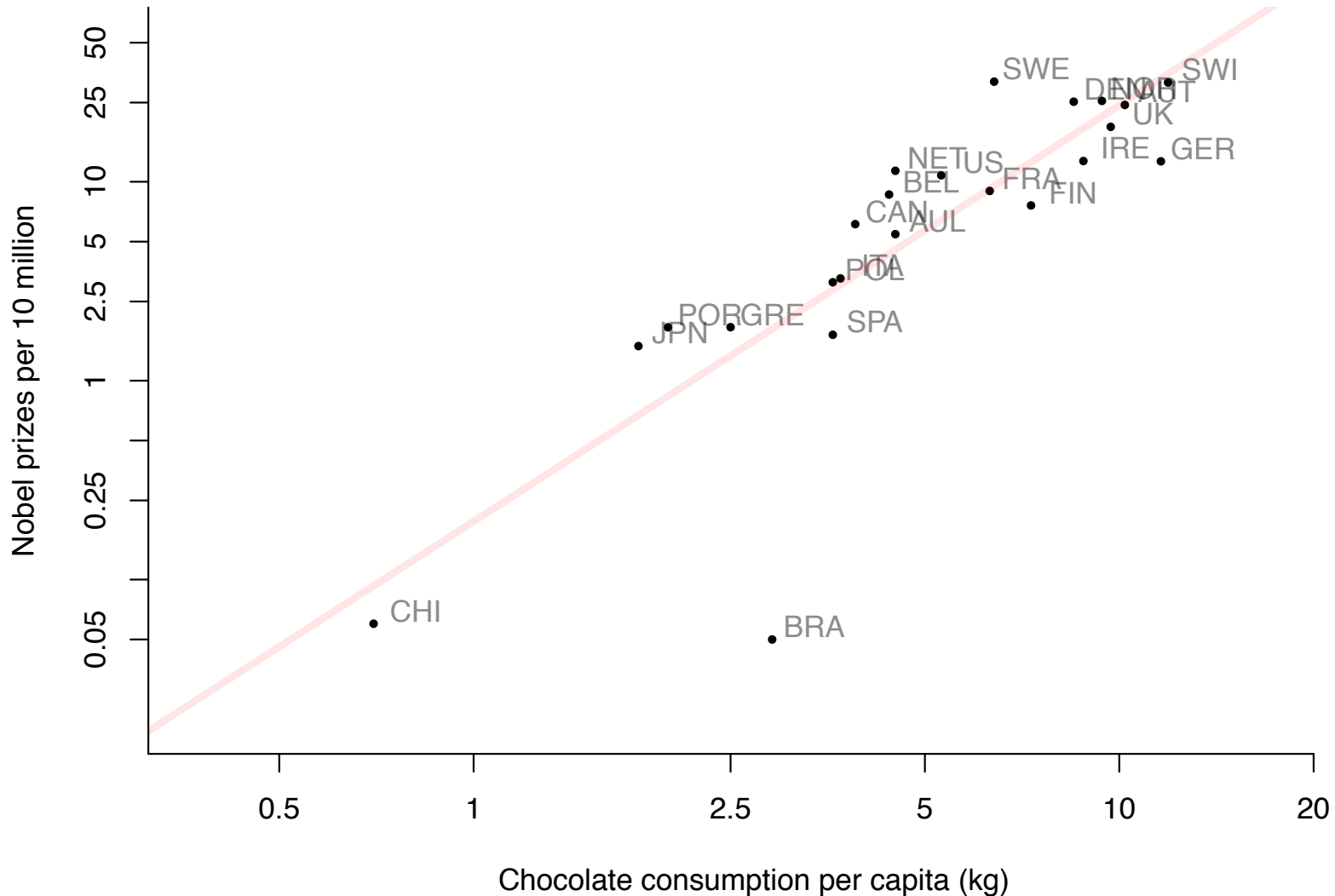
- What is the difference in predicted mortality between a 50-year-old and a 51-year-old?
- What is the difference in predicted mortality between a 60-year-old pipe smoker and a 60-year-old cigarette smoker?
- What is the difference in predicted mortality between a 50-year-old pipe smoker and a 60-year-old cigarette smoker?

The regression approach (2)



Back to chocolate

Nobel Prizes and chocolate consumption
(slope = 2.09)



Controlling for GDP/capita

$$\log(\text{NobelRate})_i = \beta_0 + \beta_1 \log(\text{ChocolateRate})_i$$

$$\log(\text{NobelRate})_i = \beta_0 + \beta_1 \log(\text{ChocolateRate})_i + \beta_2 \text{GDPperCapita}_i$$

```
> summary(lm(log(nobel_rate) ~ log(chocolate), data = cc))
```

Call:

```
lm(formula = log(nobel_rate) ~ log(chocolate), data = cc)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.5941	-0.2827	0.0883	0.5552	1.2067

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-1.6291	0.5093	-3.199	0.00432 **
log(chocolate)	2.0921	0.2982	7.015	6.32e-07 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9691 on 21 degrees of freedom
Multiple R-squared: 0.7009, Adjusted R-squared: 0.6867
F-statistic: 49.22 on 1 and 21 DF, p-value: 6.321e-07

```
> summary(lm(log(nobel_rate) ~ log(chocolate) + GDP_capk, data = cc))
```

Call:

```
lm(formula = log(nobel_rate) ~ log(chocolate) + GDP_capk, data = cc)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.99834	-0.35473	-0.05404	0.33435	1.10757

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-3.16640	0.51082	-6.199	4.69e-06 ***
log(chocolate)	1.02616	0.32611	3.147	0.00508 **
GDP_capk	0.10488	0.02386	4.395	0.00028 ***

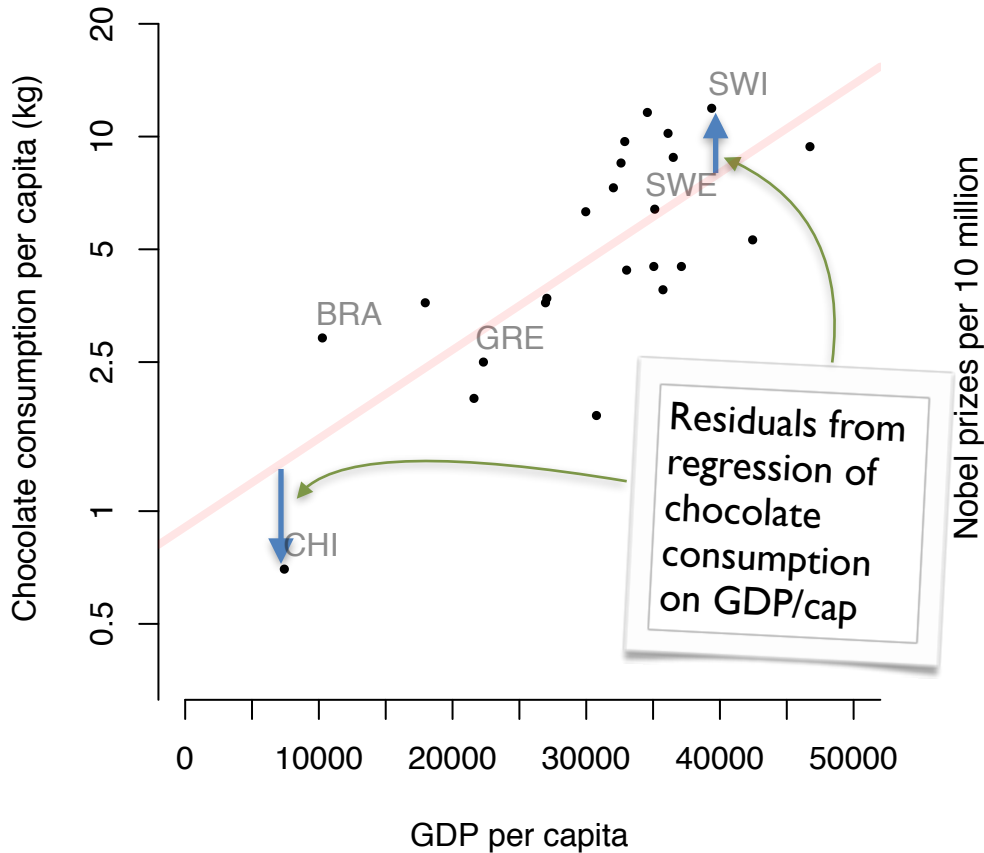
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7083 on 20 degrees of freedom
Multiple R-squared: 0.8478, Adjusted R-squared: 0.8326
F-statistic: 55.72 on 2 and 20 DF, p-value: 6.65e-09

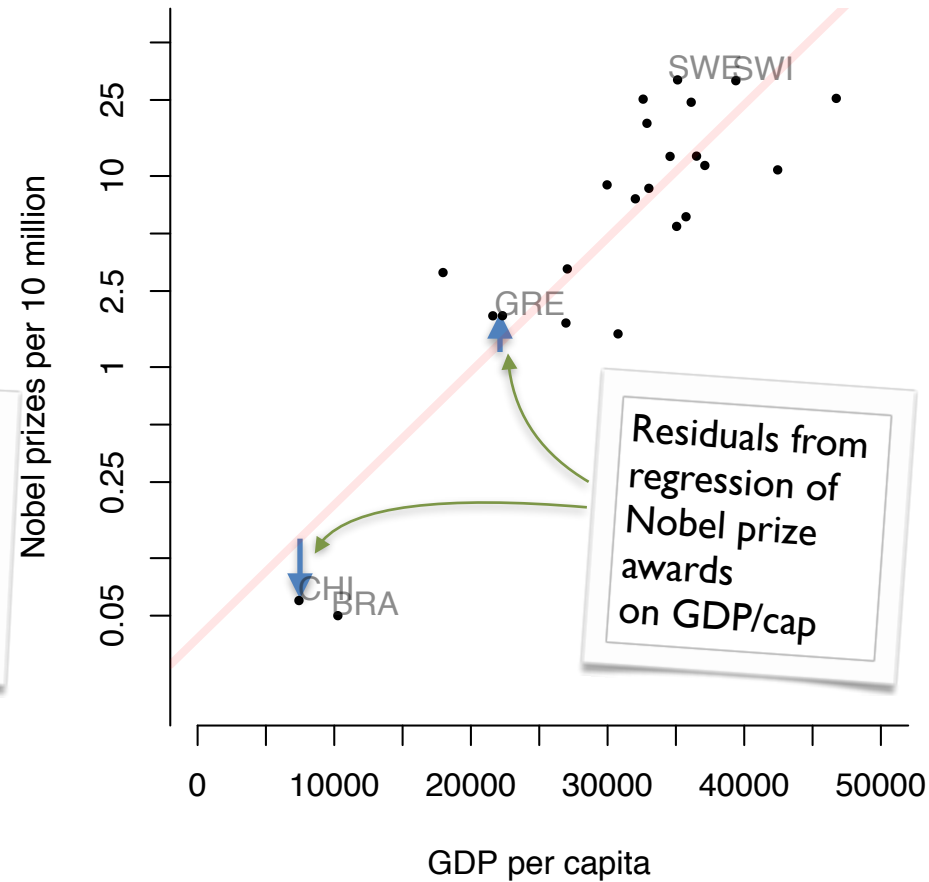
How do we interpret the coefficients from the regression of y on x and z ?

- Geometrically:
 - Picture cloud of data in 3D, with outcome (y) in vertical direction
 - Regression: choose a plane cutting through the points such that the sum of squared residuals (vertical distance from points to plane) is minimized
 - Intercept of regression is the value of y where plane intersects $x=0$ and $z=0$
 - Coefficient on x is slope of the plane at a given value of z
- In terms of **predictions**: the fitted regression says
$$\log(\text{nobel_rate}) = -3.16 + 1.02 \cdot \log(\text{chocolate}) + 0.104 \cdot \text{GDP/cap}$$
so holding fixed the GDP/cap, an increase of 1.0 in $\log(\text{chocolate})$ implies a predicted increase of 1.02 in $\log(\text{nobel_rate})$
- In terms of the **partial regression plot** (next 2 slides)

Chocolate consumption and GDP per capita



Nobel prizes and GDP per capita

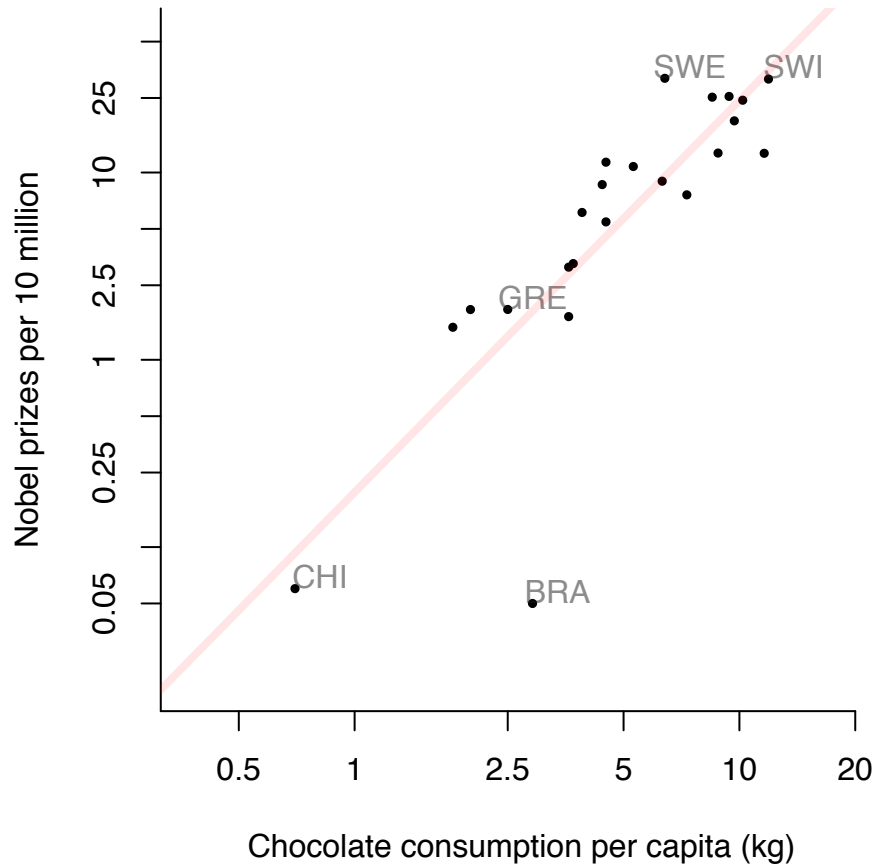


Which countries eat more/less chocolate than predicted by their GDP?

Which win more/fewer Nobel Prizes than predicted by GDP?

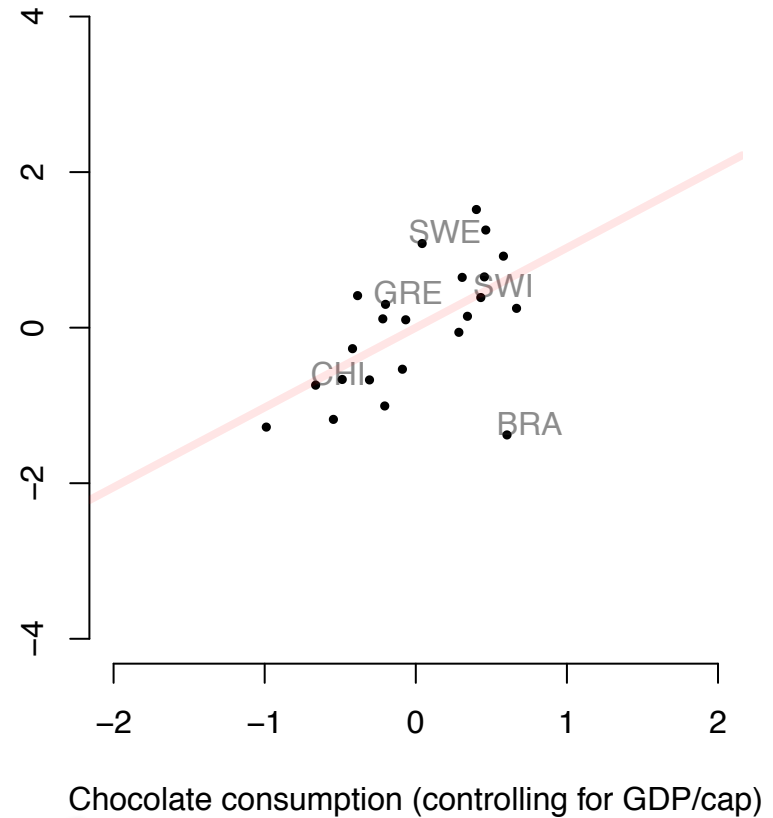
“Partial regression plot”!

**Nobel Prizes and chocolate consumption
(slope = 2.09)**



**Nobel prizes and chocolate consumption
controlling for GDP/cap
(slope = 1.03)**

Nobel prizes (controlling for GDP/cap)

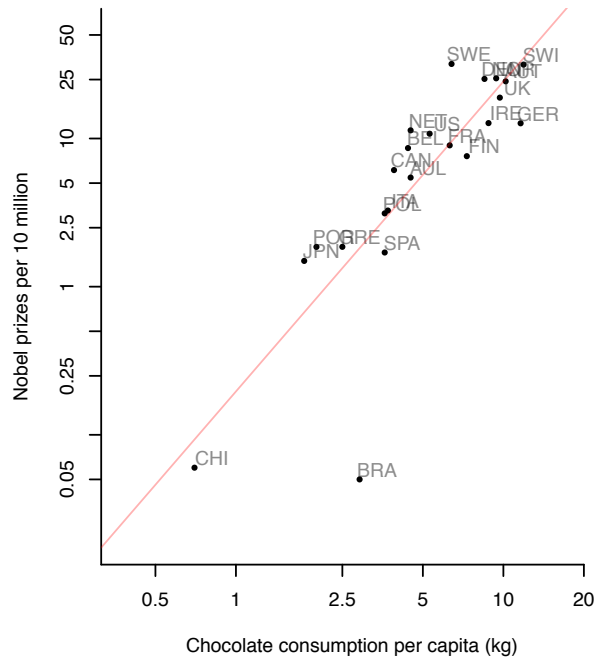


Residuals from regression of
Nobel prizes on GDP/cap

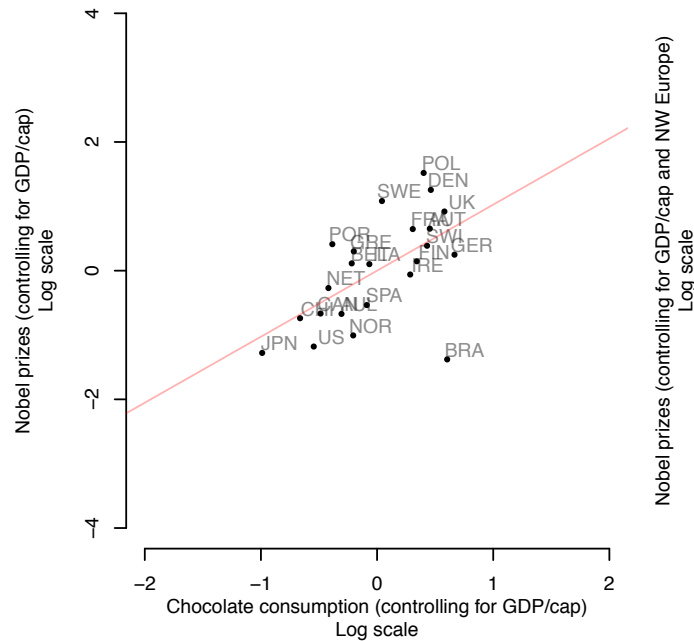
Residuals from regression of
chocolate consumption on GDP/cap

Relationship (mostly!) goes away when controlling for GDP per capita and region (NW Europe).

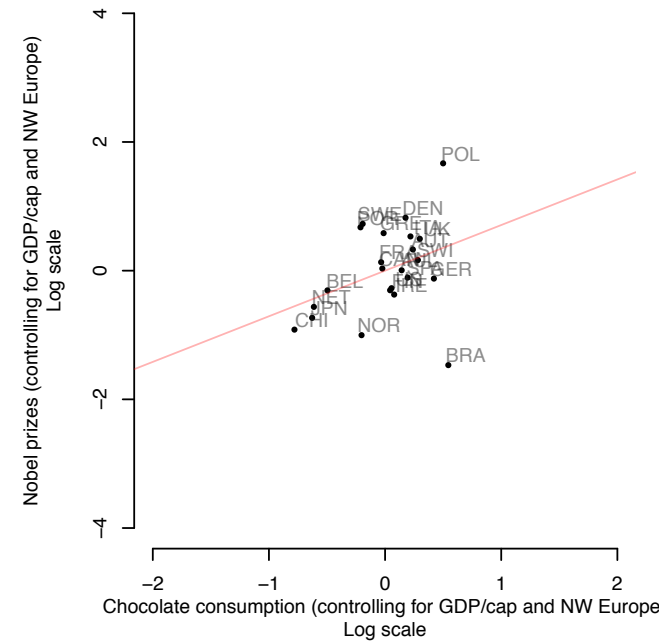
Nobel Prizes and chocolate consumption
(slope = 2.09)



Nobel prizes and chocolate consumption
controlling for GDP/cap
(slope = 1.03)



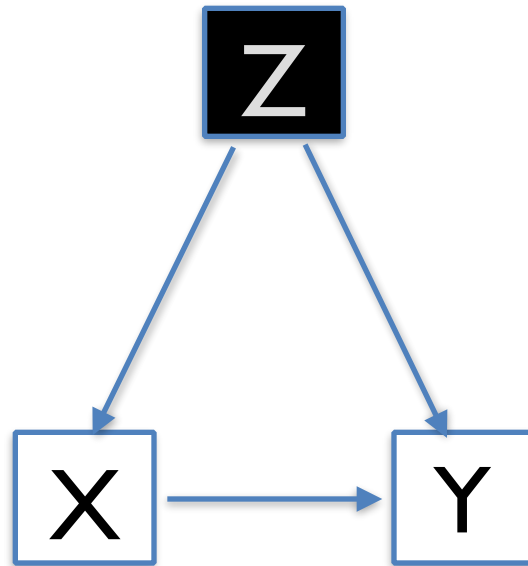
Nobel prizes and chocolate consumption
controlling for GDP/cap and NW Europe
(slope = 0.71)



How this relates to your essay

Some questions you can ask about one of Lijphart's findings:

- What are some difference between consensus democracies and majoritarian democracies other than the aspects Lijphart controls for?
- What should Lijphart control for, given his questions and claims?
- Are the regression results the same when you control for an additional variable?
- Are the regression results the same when you include or exclude outliers?



Next week:

Lecture: Inference — i.e. assessing our confidence in an estimate.

Labs: the regression commands you'll need for your essay.