SOC 3811/5811:
BASIC SOCIAL STATISTICS

Two-Way ANOVA and Three-Variable Relationships

Quick Review of Causality

We’ve talked about three basic criteria that must be met in order to infer that X causes Y:

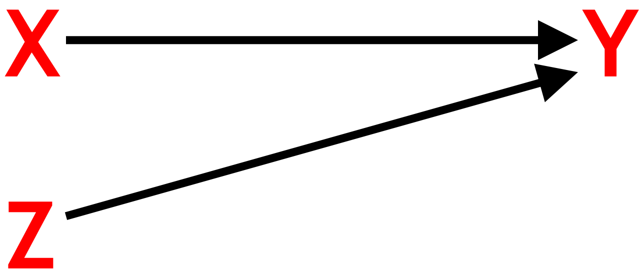
1. X and Y must be associated

2. X must precede Y in time … this is a matter of research design, and is often fairly easily to establish (especially with longitudinal data)

3. There must be no third variable(s), Z, that induce spuriousness in the association between X and Y

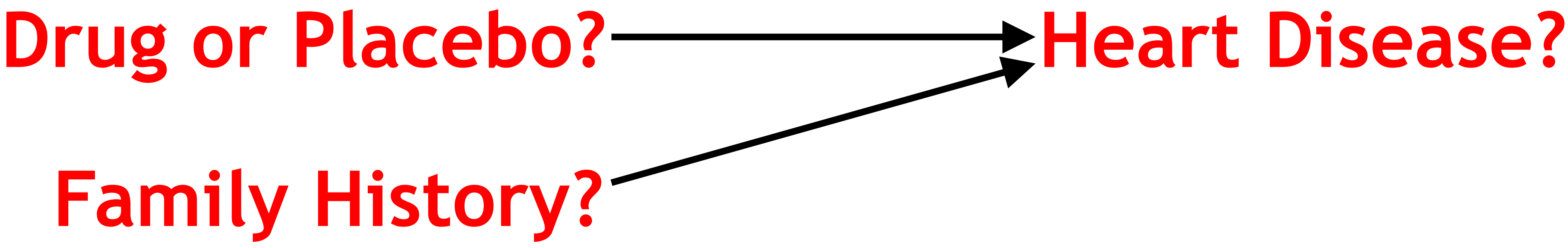
Does X Causally Affect Y?

In experimental research, X precedes Y in time and spuriousness (aka confounding) is not possible. So, any association between X and Y is entirely causal in nature



Does X Causally Affect Y?

In experimental research, X precedes Y in time and spuriousness (aka confounding) is not possible. So, any association between X and Y is entirely causal in nature



Drug or Placebo?

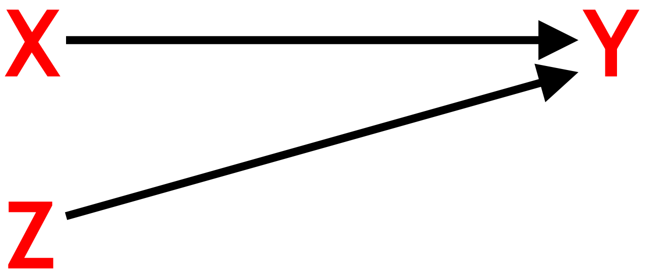
Heart Disease?

Family History?

Does X Causally Affect Y?

In observational research, if X and Y are associated with one another *and* X precedes Y in time, one of three things could be true:

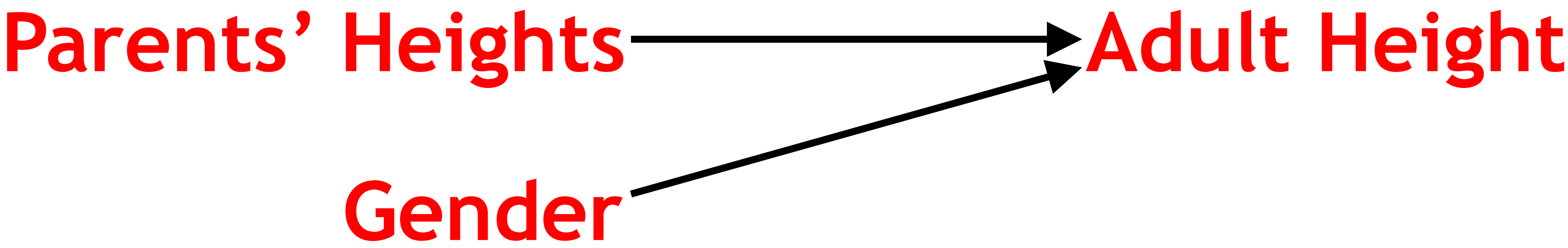
1. The association is entirely causal in nature



Does X Causally Affect Y?

In observational research, if X and Y are associated with one another *and* X precedes Y in time, one of three things could be true:

1. The association is entirely causal in nature



Parents' Heights

Adult Height

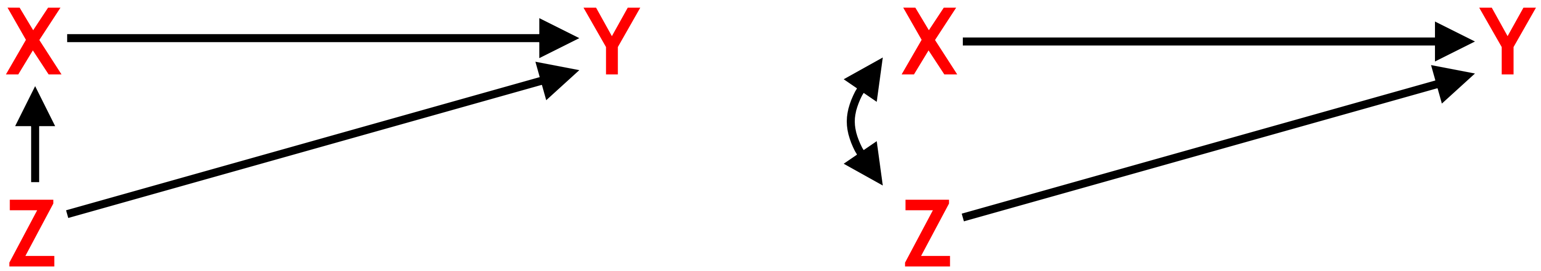
Gender

Does X Causally Affect Y?

In observational research, if X and Y are associated with one another *and* X precedes Y in time, one of three things could be true:

2. The association is partly causal and partly spurious

    owing to confounder Z

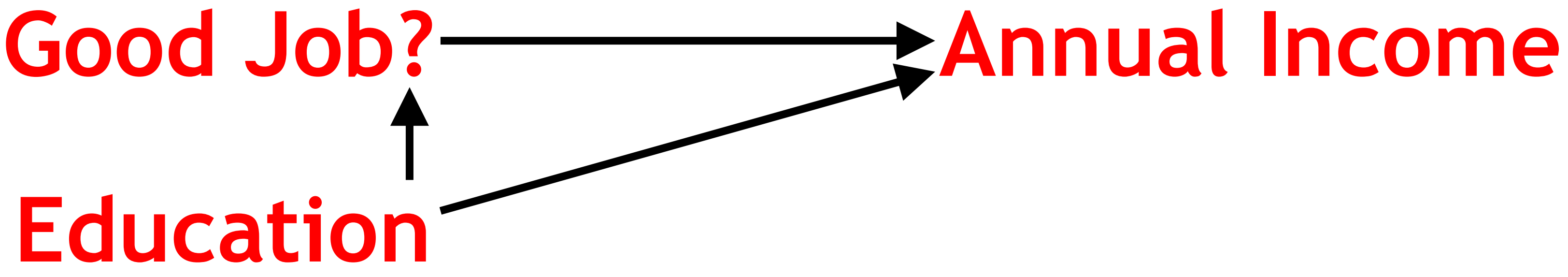


Does X Causally Affect Y?

In observational research, if X and Y are associated with one another *and* X precedes Y in time, one of three things could be true:

2. The association is partly causal and partly spurious

   owing to confounder Z

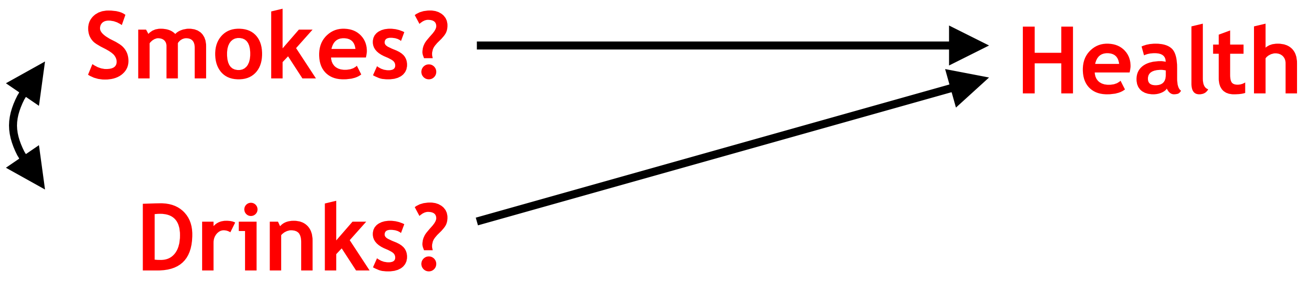


Does X Causally Affect Y?

In observational research, if X and Y are associated with one another *and* X precedes Y in time, one of three things could be true:

2. The association is partly causal and partly spurious

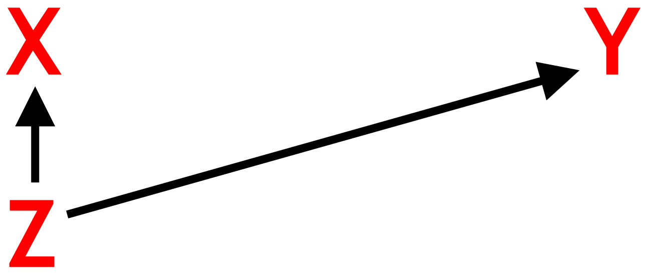
   owing to confounder Z

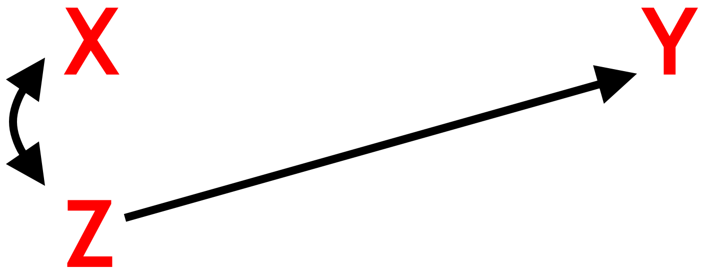


Does X Causally Affect Y?

In observational research, if X and Y are associated with one another *and* X precedes Y in time, one of three things could be true:

3. The association is entirely spurious owning to Z

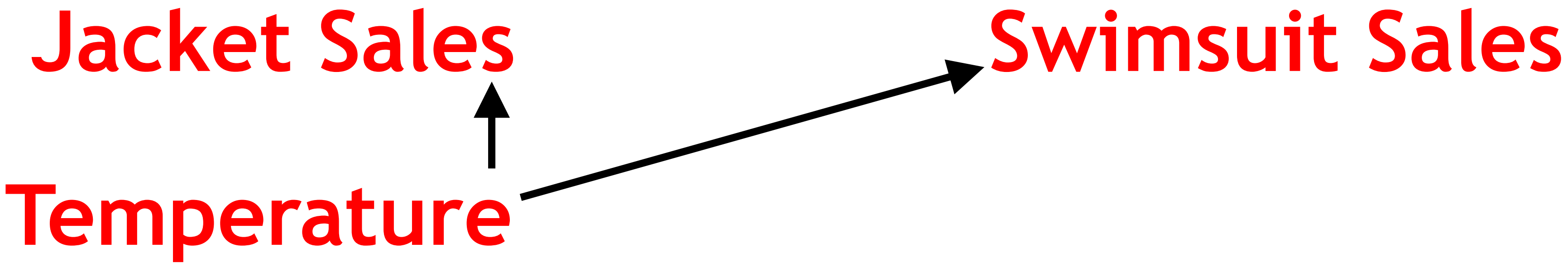




Does X Causally Affect Y?

In observational research, if X and Y are associated with one another *and* X precedes Y in time, one of three things could be true:

3. The association is entirely spurious owning to Z



Jacket Sales

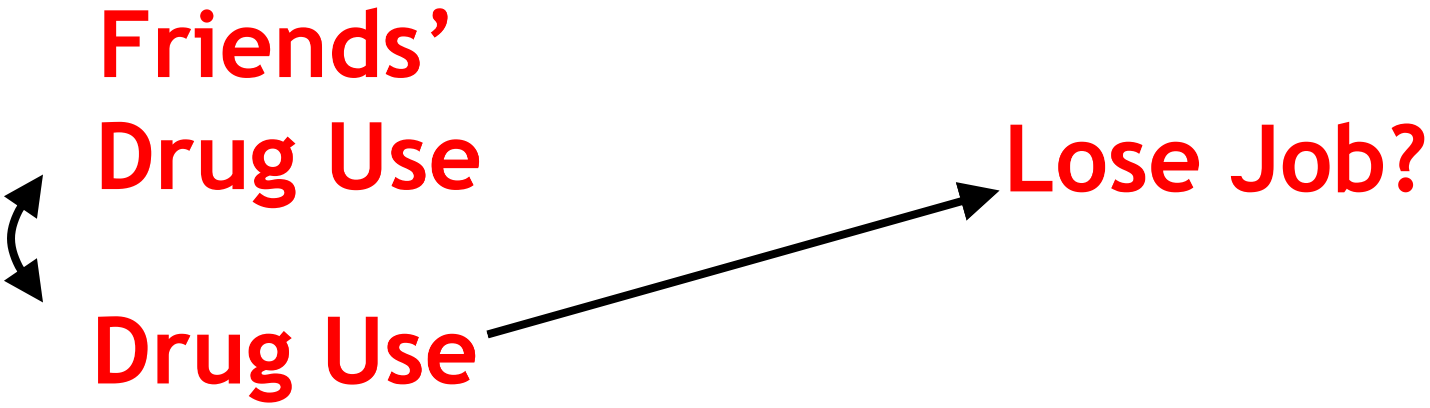
Swimsuit Sales

Temperature

Does X Causally Affect Y?

In observational research, if X and Y are associated with one another *and* X precedes Y in time, one of three things could be true:

3. The association is entirely spurious owning to Z



Friends' Drug Use

Lose Job?

Drug Use

“Control” for Z?

In observational research, **should we “control” for Z**?

“Controlling for” Z means focusing on the association between X and Y for cases with the same level of Z

Imagine that X and Y are associated, and that X precedes Y in time

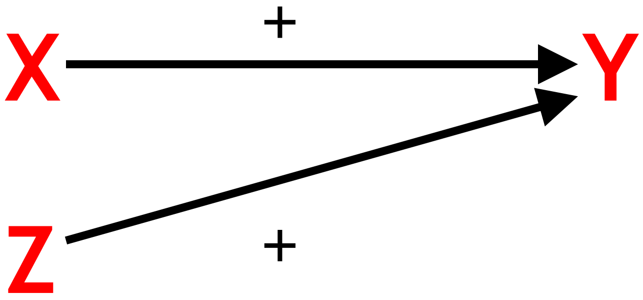
1. What are the consequences of failing to control for Z in each of the three situations reviewed above?
2. What are the consequences of controlling for Z in each situation?

“Control” for Z?

The true causal effect of X on Y is β

If we *fail to control for Z*, we would infer that the effect of X on Y is β;

If we *control for Z*, we would infer that the effect of X on Y is β ... so, controlling for Z makes no difference

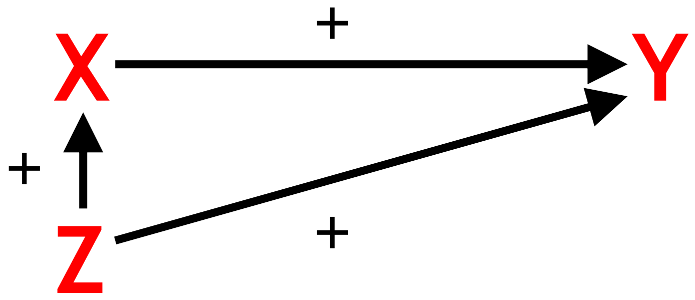


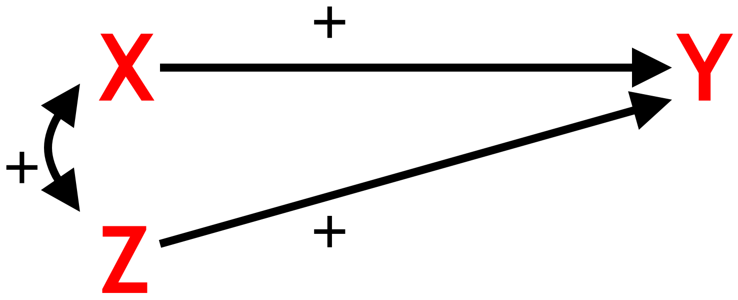
"Control" for Z?

The true causal effect of X on Y is β

If we *fail to control for Z*, we would infer that the effect of X on Y is **larger** than β;

If we *control for Z*, we would infer that the effect of X on Y is β ... so, controlling for Z is a good thing!



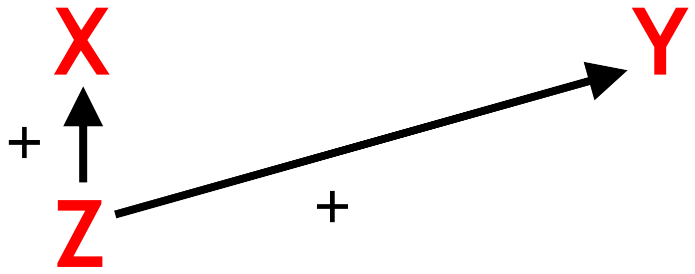


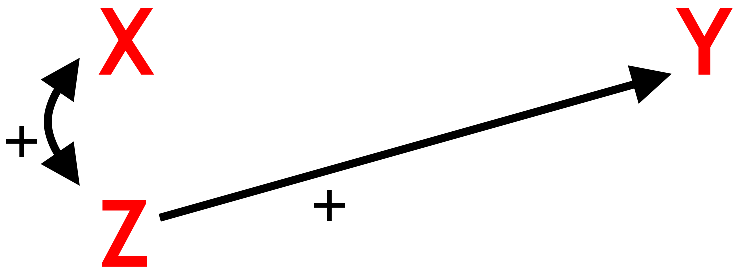
“Control” for Z?

The true causal effect of X on Y is 0

If we *fail to control for Z*, we would infer that the effect of X on Y is **larger** than 0;

If we *control for Z*, we would infer that the effect of X on Y is 0 ... so, controlling for Z is a good thing!





Three Variable Relationships

So why not just control for Z every time we can think of any variables (Z) that might be associated with X and Y?

In establishing the causal impact of X on Y we have—to this point—only thought of third variable(s) Z as inducing spuriousness … a bad thing

This suggests that to estimate the true, causal impact of X on Y we should always “control” (or “adjust”) for Z

Not so fast…

Three Variable Relationships

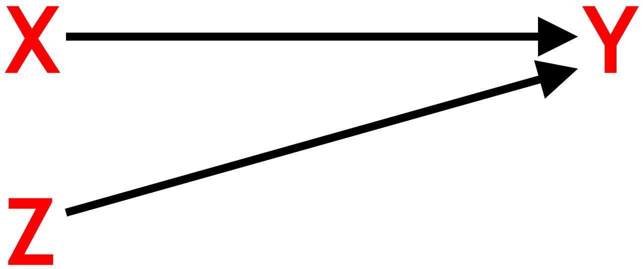
In fact there are a variety of ways in which X, Y, and Z might be related to one another

Theory and **prior evidence** should guide our decision about how Z plays into the relationship between X and Y

Depending on our theoretical understanding of how Z plays into the relationship between X and Y, we might or might not want to statistically control for Z

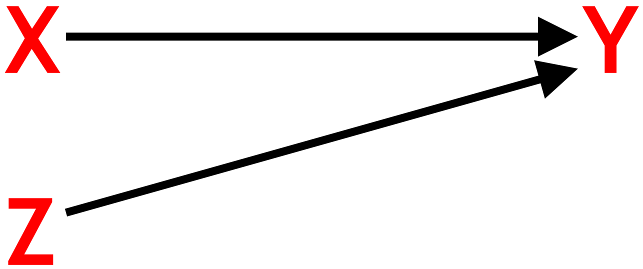
Three Variable Relationships

Under the scenario depicted below, the association between X and Y is unaffected by the presence of third variable(s) Z



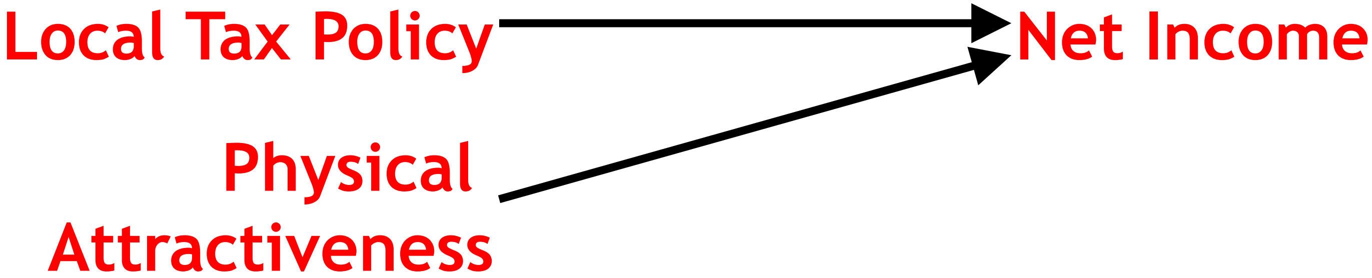
Three Variable Relationships

Statistically controlling for Z has no bearing on our assessment of the association between X and Y … so there is no need to do so (but it doesn't hurt anything)



Three Variable Relationships

Example



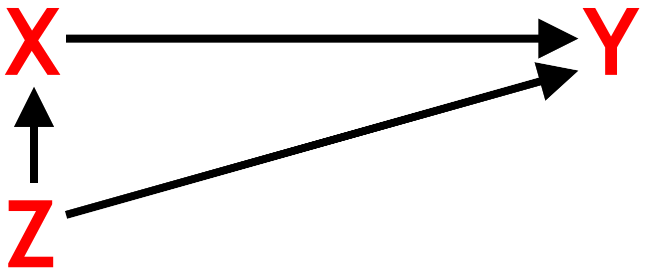
Local Tax Policy

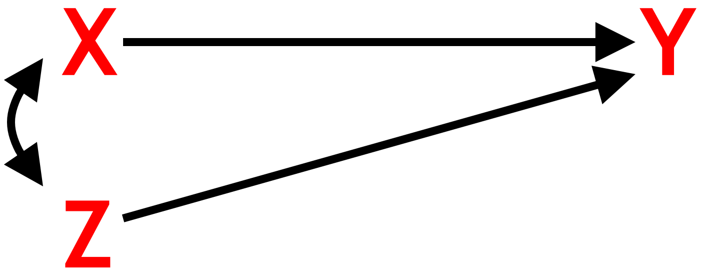
Net Income

Physical Attractiveness

Three Variable Relationships

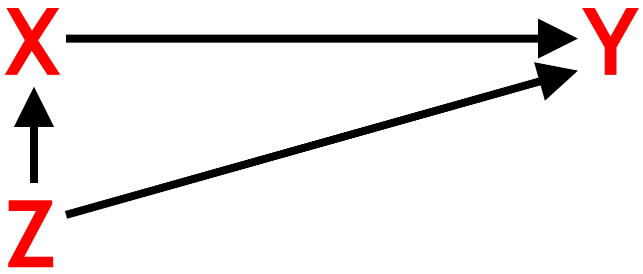
Under both scenarios depicted below, the association between X and Y is—at least partly—spurious owing to the influence of confounding variable(s) Z

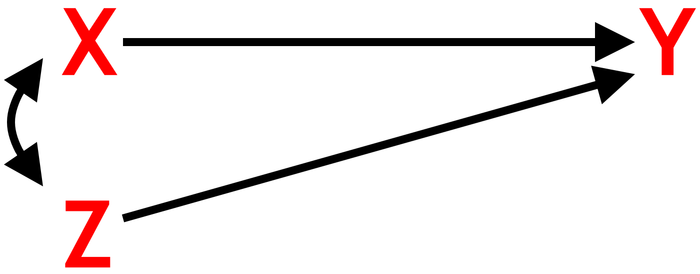




Three Variable Relationships

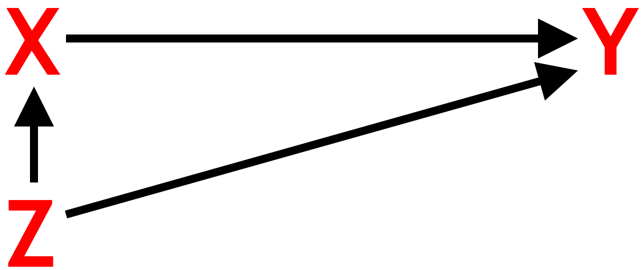
Under each of these scenarios, statistically controlling for Z is designed to estimate the *independent* effect of X on Y … the effect "net of" Z

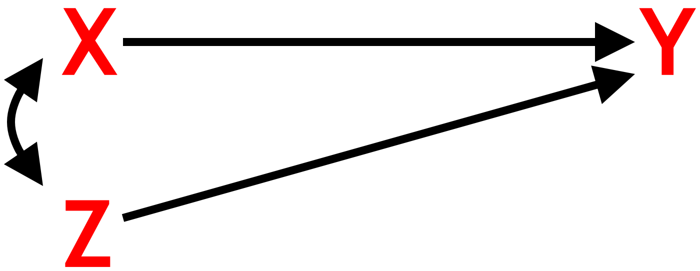




Three Variable Relationships

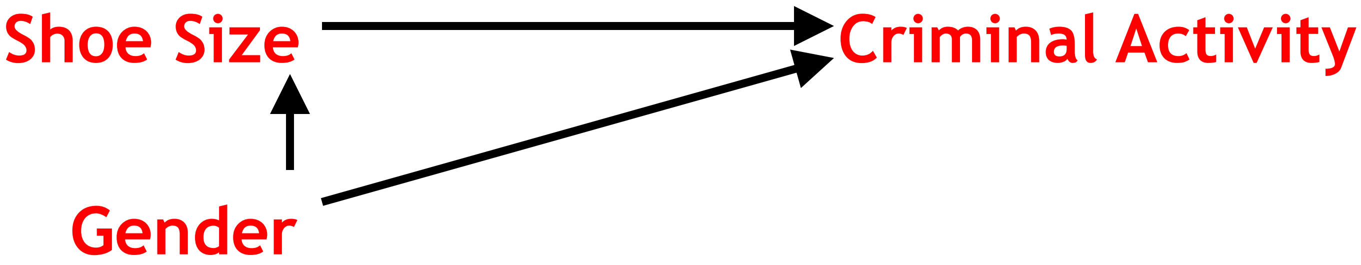
Here, the independent effect of X on Y … the effect “net of” Z … represents the direct effect of X on Y **(but don’t forget about the other criteria for establishing causality)**





Three Variable Relationships

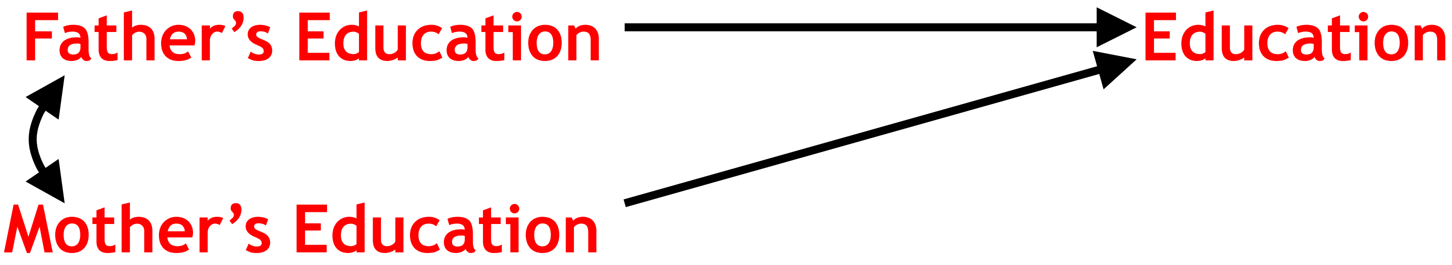
Example



Shoe Size

Criminal Activity

Gender



Father's Education

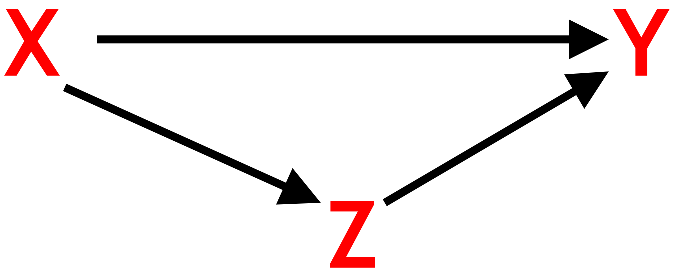
Education

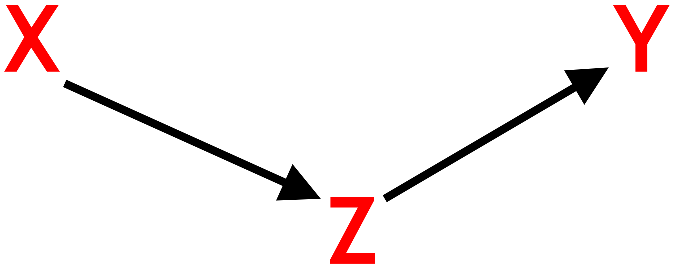
Mother's Education

Three Variable Relationships

Under the scenarios depicted below, Z is associated with both X and Y but does **not** induce spuriousness

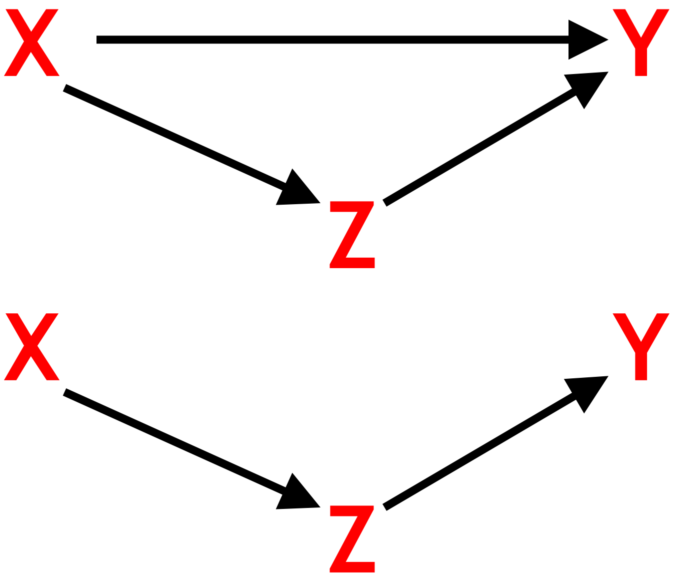
Z is a mechanism through which X affects Y … Z is known as a mediating variable(s)





Three Variable Relationships

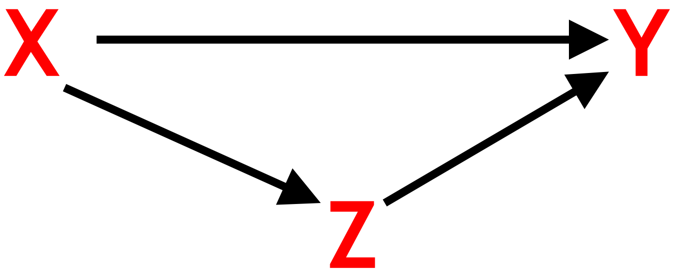
Statistically controlling for Z under these scenarios is also designed to estimate the independent effect of X on Y … the effect “net of” Z

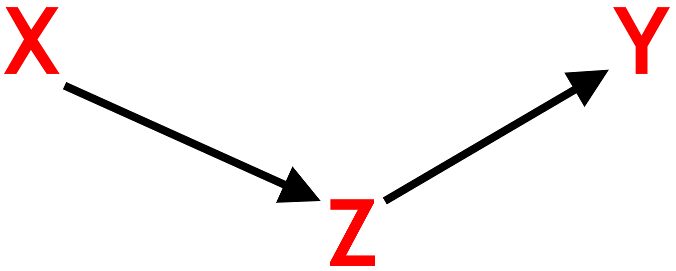


Three Variable Relationships

Conceptually, in these cases the independent effect of X on Y … the effect "net of" Z … represents the direct effect of X on Y

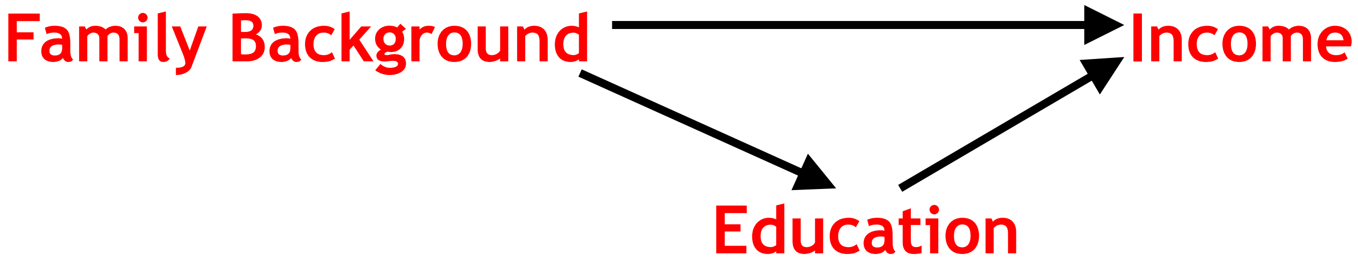
Is that really what we want to estimate? **It depends…**





Three Variable Relationships

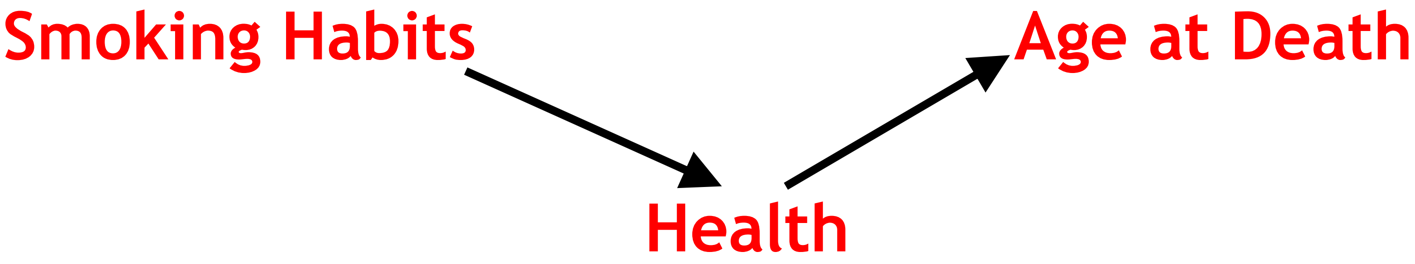
Examples:



Family Background

Income

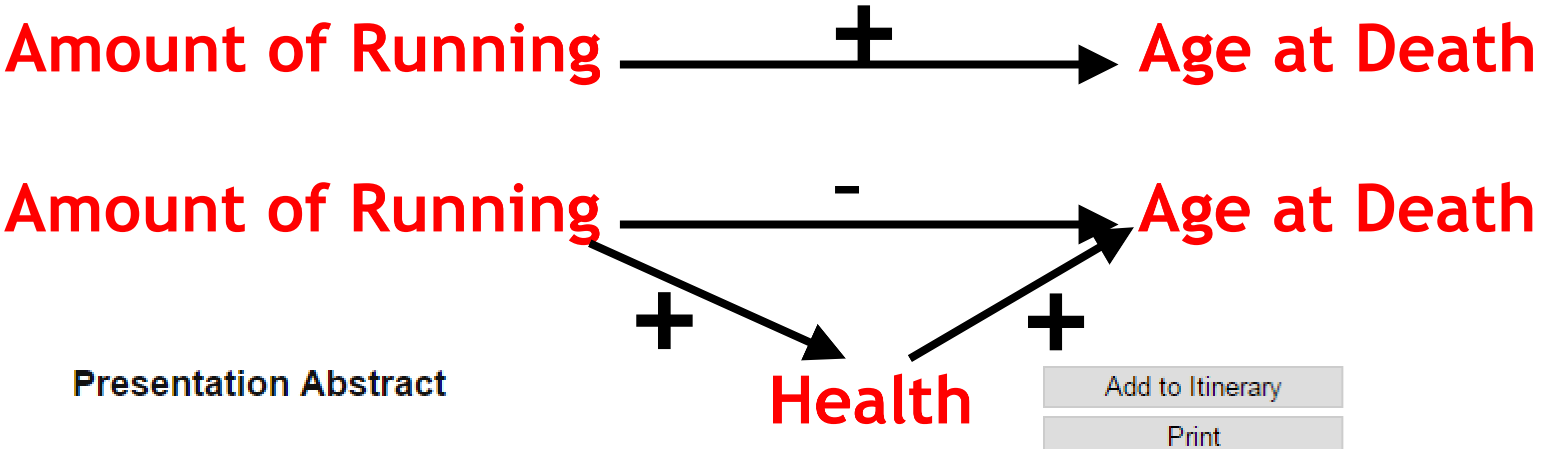
Education

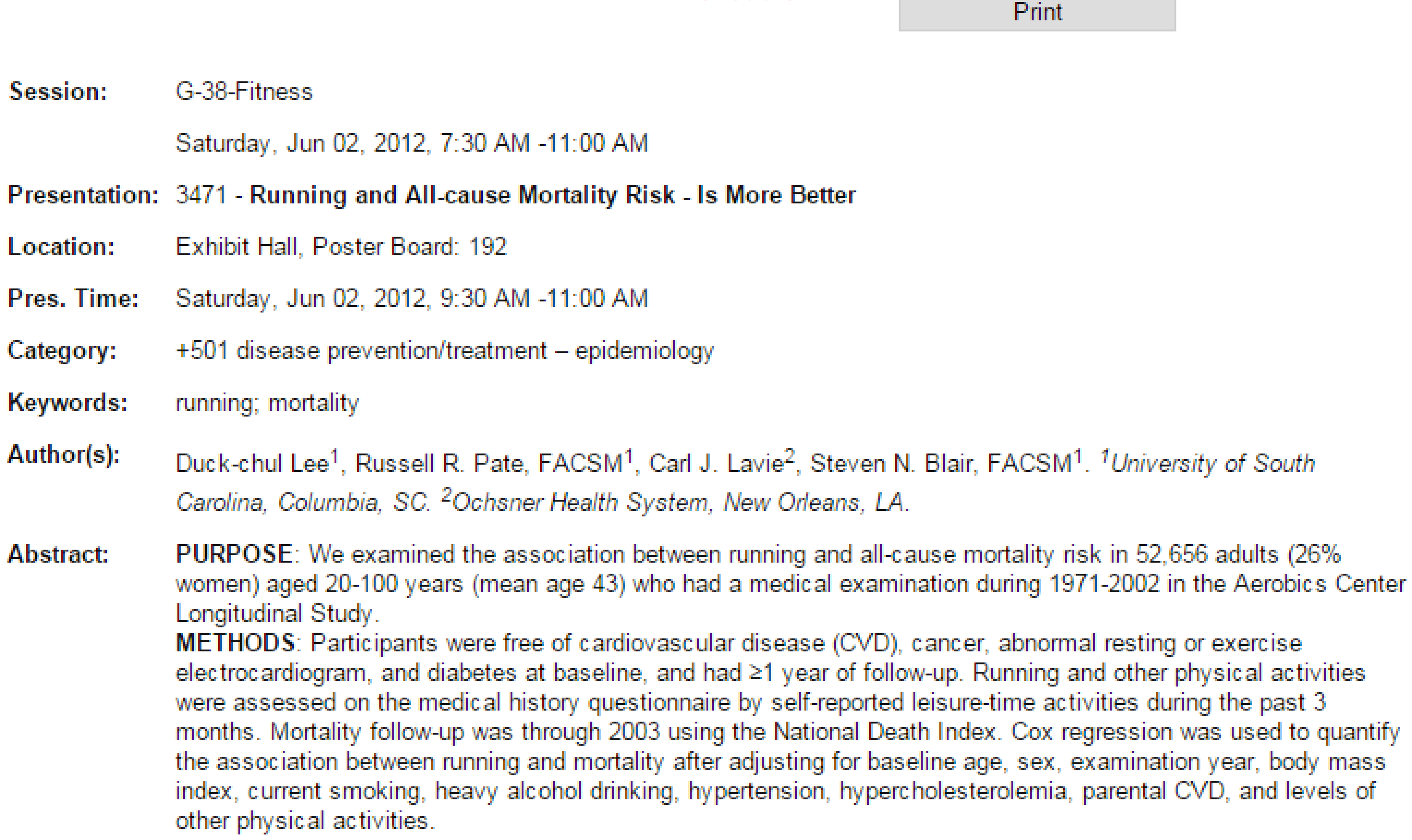


Smoking Habits

Age at Death

Health





Presentation Abstract

Add to Itinerary

Print

Session: G-38-Fitness

Saturday, Jun 02, 2012, 7:30 AM -11:00 AM

Presentation: 3471 - **Running and All-cause Mortality Risk - Is More Better**

Location: Exhibit Hall, Poster Board: 192

Pres. Time: Saturday, Jun 02, 2012, 9:30 AM -11:00 AM

Category: +501 disease prevention/treatment – epidemiology

Keywords: running; mortality

Author(s): Duck-chul Lee[1], Russell R. Pate, FACSM[1], Carl J. Lavie[2], Steven N. Blair, FACSM[1]. [1]*University of South Carolina, Columbia, SC.* [2]*Ochsner Health System, New Orleans, LA.*

Abstract: **PURPOSE**: We examined the association between running and all-cause mortality risk in 52,656 adults (26% women) aged 20-100 years (mean age 43) who had a medical examination during 1971-2002 in the Aerobics Center Longitudinal Study.
METHODS: Participants were free of cardiovascular disease (CVD), cancer, abnormal resting or exercise electrocardiogram, and diabetes at baseline, and had ≥1 year of follow-up. Running and other physical activities were assessed on the medical history questionnaire by self-reported leisure-time activities during the past 3 months. Mortality follow-up was through 2003 using the National Death Index. Cox regression was used to quantify the association between running and mortality after adjusting for baseline age, sex, examination year, body mass index, current smoking, heavy alcohol drinking, hypertension, hypercholesterolemia, parental CVD, and levels of other physical activities.

Worksheet

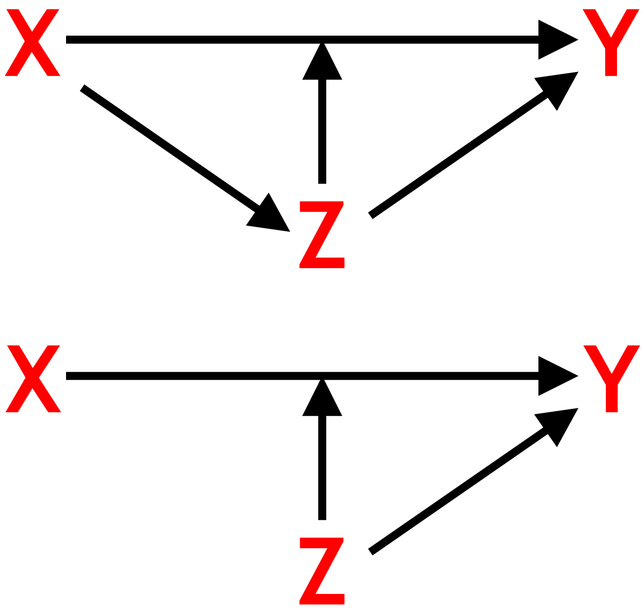
What are some potential **confounders** and some potential **causal mechanisms** linking X to Y in the following examples?

1. Kids who frequently misbehave in high school (X) are less likely to go to college (Y)
2. Men who are married (X) live longer (Y) than men who are not married
3. Eating breakfast (X) improves employees’ productivity (Y) during the day

Three Variable Relationships

In some cases the association between X and Y may be different across levels of Z

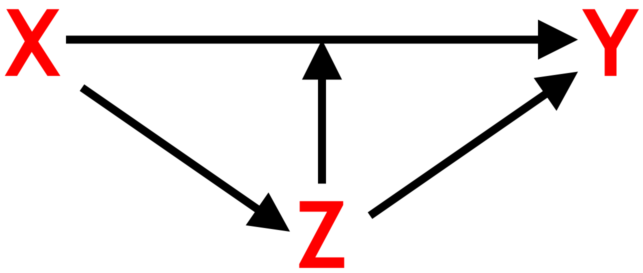
In these cases we say that there is an interaction between X and Z … Z is known as a moderating variable

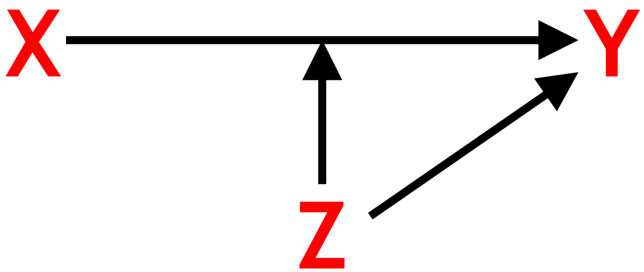


Three Variable Relationships

If we simply statistically adjust for Z in these situations, then the resulting “net” effect of X on Y isn’t very meaningful

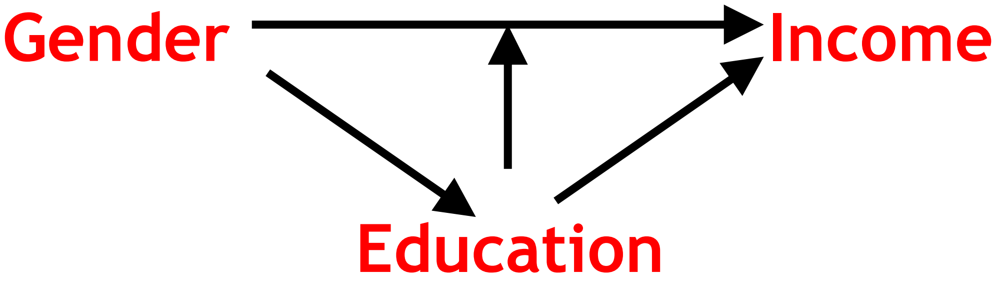
We must allow for different effects by level of Z

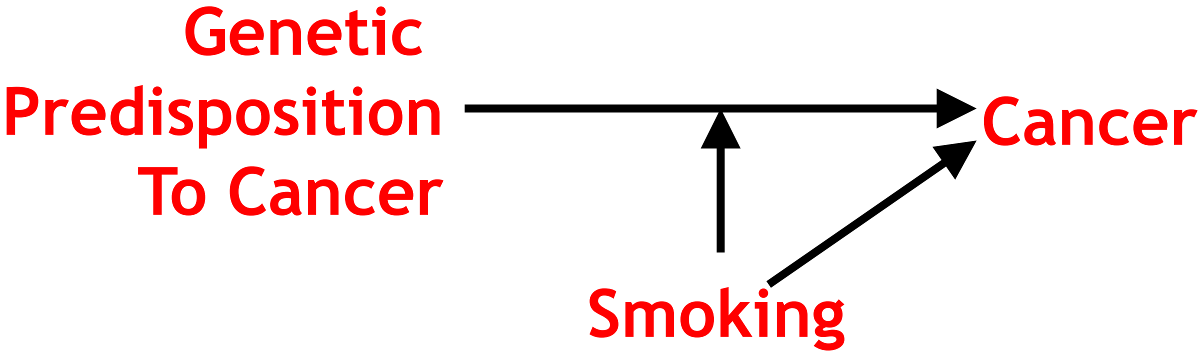




Three Variable Relationships

Example

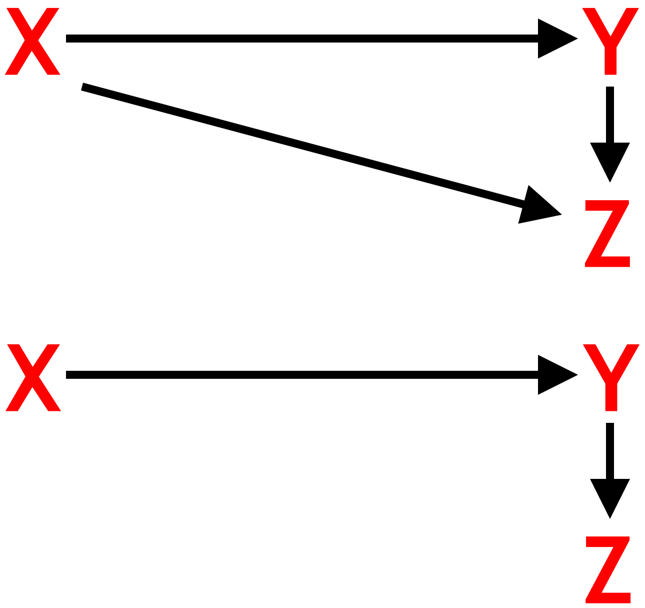




Three Variable Relationships

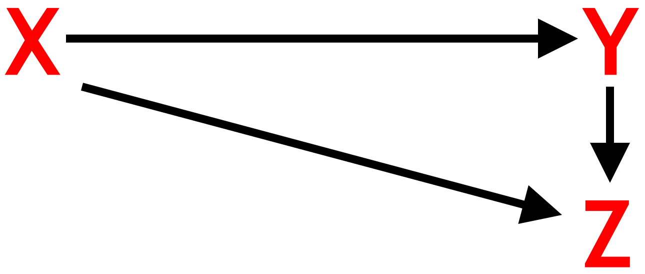
Under the scenarios depicted below, Z is affected by Y (and also possibly X)

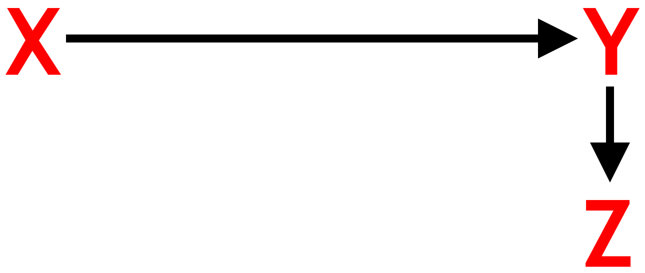
Here, Y is a mechanism through which X affects Z



Three Variable Relationships

If we are trying to assess the independent effect of X on Y, then in these situations statistically controlling for Z is a **terrible idea**

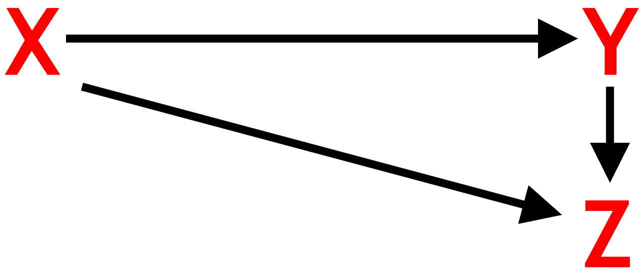


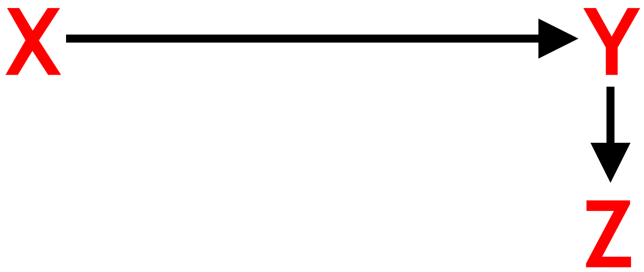


Three Variable Relationships

In effect, controlling for Z in these situations is like selecting on the dependent variable

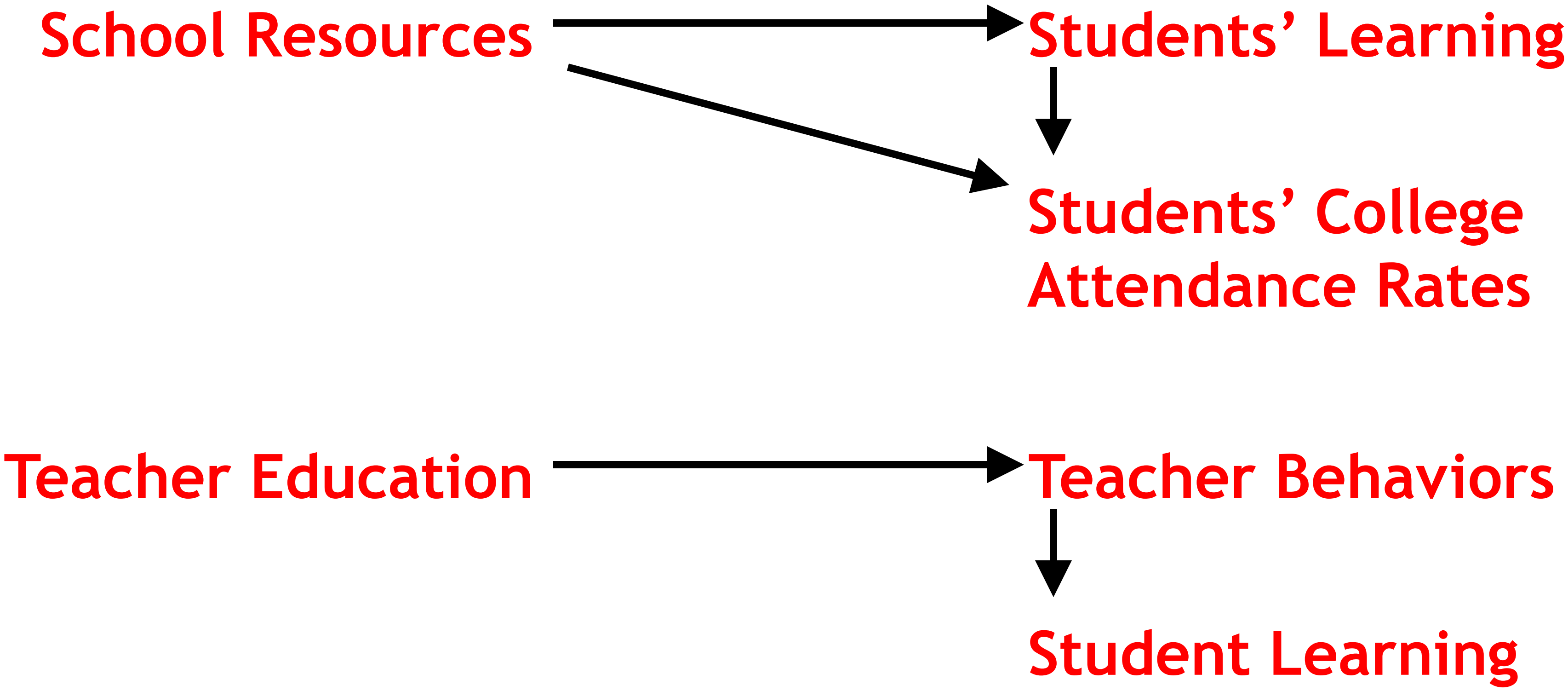
This is known as over-controlling





Three Variable Relationships

Example



School Resources

Students' Learning

Students' College Attendance Rates

Teacher Education

Teacher Behaviors

Student Learning

Three Variable Relationships

In assessing the causal effect of X on Y, is Z…

…totally orthogonal to the association between X and Y?

…a confounding variable which must be controlled?

…a mediating variable which we might want to control?

…a moderating variable which we must treat with care?

…a variable that it would be a mistake to control for?

It depends on the situation … so carry out analyses on the basis of **theory** and on evidence from **prior research**

How Do We Control for “Z?”

Statistical Control in Cross-Tabulations

What is the effect of shoe size (X) on whether someone has committed a felony (Y)?

Start with a zero-order table … a cross-table in which zero third variables (Z) have been controlled

χ^2= 4.8 (p<0.05)

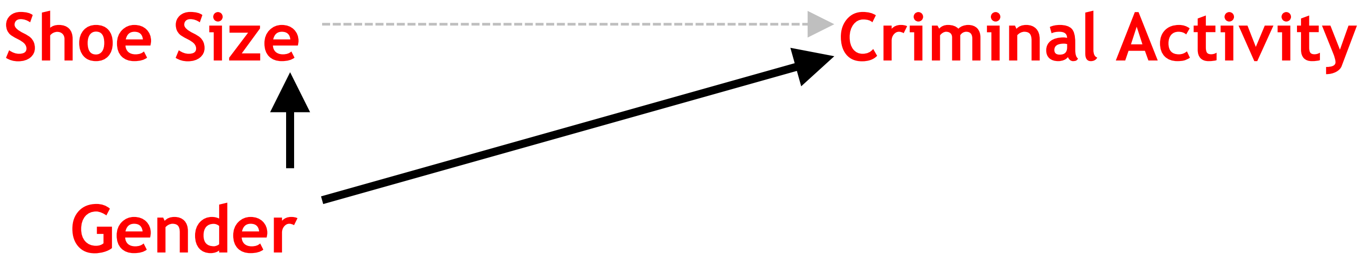
Gamma= 0.31

		Shoe Size		
		1=Small	*2=Large*	Row
Felony?	*2=Yes*	55	70	125
	1=No	45	30	75
	Column	100	100	N=200

Statistical Control in Cross-Tabulations

Shoe size is unlikely to causally affect people’s chances of committing a felony

The association is probably spurious owing to gender



Controlling for gender in this case would be appropriate … it would allow us to estimate the effect of shoe size on crime net of confounding variable gender

Statistical Control in Cross-Tabulations

What is the effect of shoe size (X) on whether someone has committed a felony (Y) net of gender (Z)?

Here we produce first-order tables … cross-tables in which one third variables (Z) has been controlled

χ^2 = 0.0

MEN

Shoe Size

Gamma = 0.0

Felony?		*1=Small*	*2=Large*	Row
	2=Yes	15	60	75
	1=No	5	20	25
	Column	20	80	N=100

χ^2 = 0.0

WOMEN

Shoe Size

Gamma = 0.0

Felony?		*1=Small*	*2=Large*	Row
	2=Yes	40	10	50
	1=No	40	10	50
	Column	80	20	N=100

Statistical Control in Cross-Tabulations

Note that the measures of association in each sub-table of this three-way cross-tabulation are known as conditional associations

The "conditional association" between X and Y is the association after controlling for Z

Note also that controlling for Z may…

…entirely account for the observed (zero-order) association between X and Y (as in our example) ,

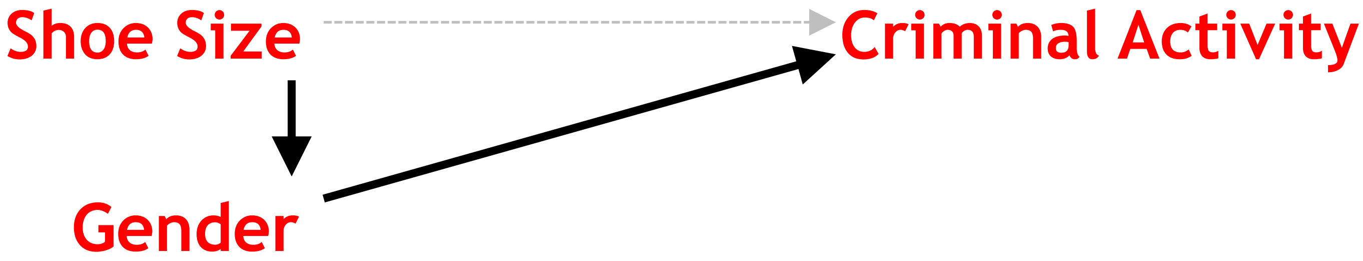
…partially account for that association, or

…not account for that association at all

Statistical Control in Cross-Tabulations

In the previous example, gender was a confounder

What if we thought—from a theoretical point of view—that gender was a mediating variable, such that:



Shoe Size

Criminal Activity

Gender

Here we (stupidly) conceive of gender as a mechanism through which shoe size affects criminality

How do we know whether gender is a confounder or a mediator? **Theory**! (The analysis looks exactly the same! How we make sense of the results is different, though)

Statistical Control in Cross-Tabulations

Question: “What is the effect of getting a physical exam (X) on whether people die within the next year (Y)?”

Start with a zero-order table … a cross-table in which zero third variables (Z) have been controlled

χ^2= 54.8 (p<0.01)

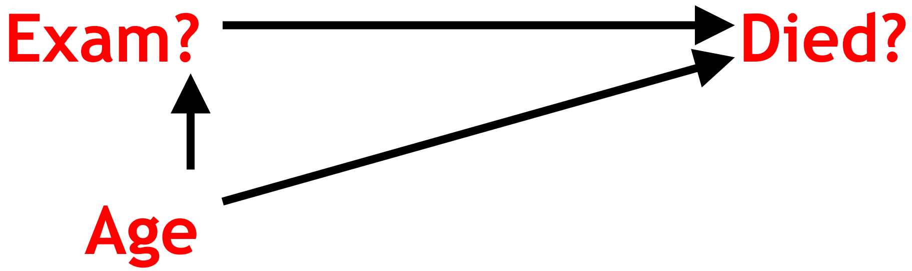
Gamma= 0.35

		Exam?		
		1=No	*2=Yes*	Row
Died?	*2=Yes*	270	330	600
	1=No	880	520	1400
	Column	1,150	850	N=2,000

Statistical Control in Cross-Tabulations

In the zero-order table, getting a physical exam is associated with a higher chance of dying!

Perhaps in this example age is a confounder, such that:



Controlling for age in this case would allow us to better estimate the effect of getting a physical exam on whether people die

Statistical Control in Cross-Tabulations

Whereas the effect in the zero-order table was positive, in the first-order table the effect is zero for the young and negative for the old

This is an example of an interaction effect

χ^2 = 0.0

Gamma = 0.0

YOUNG

		Exam?		
		1=No	*2=Yes*	Row
Died?	*2=Yes*	95	5	100
	1=No	855	45	900
	Column	950	50	N=1000

χ^2 = 140.6 ($p<0.01$)

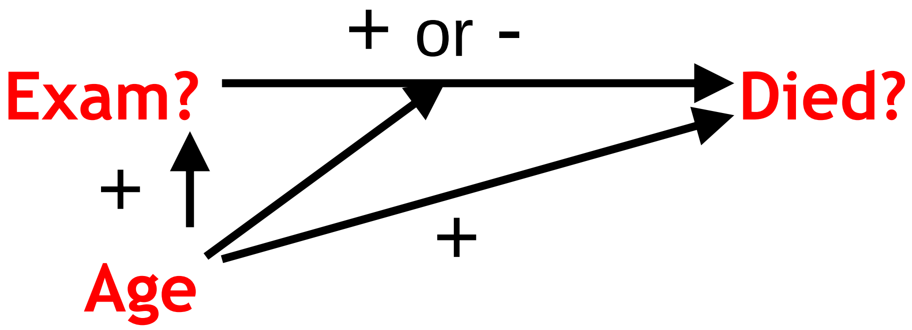
Gamma = -0.82

OLD

		Exam?		
		1=No	*2=Yes*	Row
Died?	*2=Yes*	175	325	500
	1=No	25	475	500
	Column	200	800	N=1000

Statistical Control in Cross-Tabulations

For the “old,” getting an exam is negatively associated with the chances of dying; for the “young,” there is no association between getting an exam and the chances of dying



Because the association is not the same across the first-order tables, we say that there is an interaction between X and Z, such that the effect of X varies by Z (and the effect of Z varies by X)

Statistical Control in Cross-Tabulations

Question: “What is the effect of educational credentials (X) on income (Y)?”

Start with a zero-order table … a cross-table in which zero third variables (Z) have been controlled

χ^2= 7.94 $(p<0.01)$

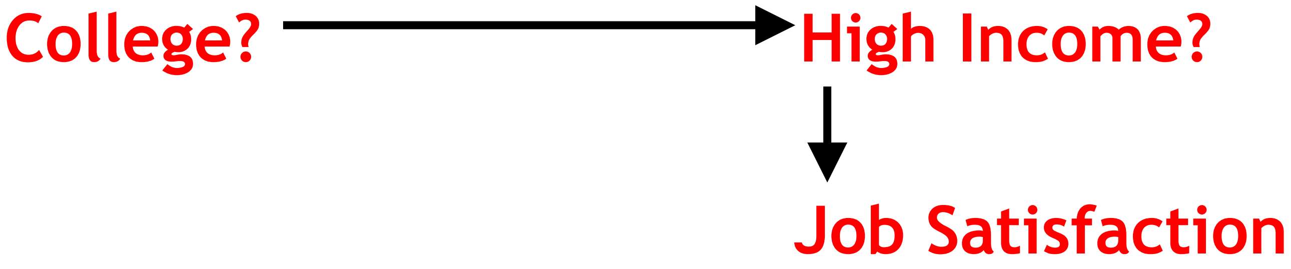
Gamma= 0.20

		College Grad?		
		1=No	*2=Yes*	Row
Income?	*2=High*	100	200	300
	1=Low	300	400	700
	Column	400	600	N=1,000

Statistical Control in Cross-Tabulations

In the zero-order table, completing college is associated with a greater chance high income

Imagine the world works this way:



College?

High Income?

Job Satisfaction

Controlling for job satisfaction in this case would take us **away** from estimating the causal effect of X on Y

Statistical Control in Cross-Tabulations

Whereas the "effect" in the zero-order table was 0.2, in the first-order tables the effect is different

This is an example of over-controlling

χ^2 = 0.0

Gamma = 0.0

Dissatisfied w/ Job

College Grad?

		1=No	*2=Yes*	Row
Income?	*2=High*	60	90	150
	1=Low	140	210	350
	Column	200	300	N=500

χ^2 = 15.9 ($p<0.01$)

Gamma =0.40

Satisfied w/ Job

College Grad?

		1=No	*2=Yes*	Row
Income?	*2=High*	40	110	150
	1=Low	160	190	350
	Column	200	300	N=500

Worksheet

How do you interpret the following results?

In the association between whether high school students work at paid jobs during the school year (X) and whether they drop out of high school (Y), gamma is 0.25

After statistically controlling for children’s family income and wealth (Z), gamma is reduced to 0.20

Worksheet

How do you interpret the following results?

The correlation between parents’ wealth (X) and their children’s wealth (Y) is 0.40

After statistically controlling for children’s education (Z), the correlation between X and Y is 0.10

Worksheet

How do you interpret the following results?

In a regression of final exam scores (Y) on number of hours people studied for the exam (X), the slope is estimated to be 5.0

After statistically controlling for students’ year in college (Z), the slope is estimated to be 2.0 among freshman and sophomores, 5.0 among juniors, and 8.0 among seniors.