Causality and Counterfactuals

When we talk about cause and effect in the social sciences, we usually talk about how one variable affects another variable

The explanatory variable (or independent variable) is said to affect or cause change in the response variable (or the dependent variable or the outcome variable)

Example: "Does poverty cause kids to commit crimes?"

Explanatory Variable: Whether a child is poor

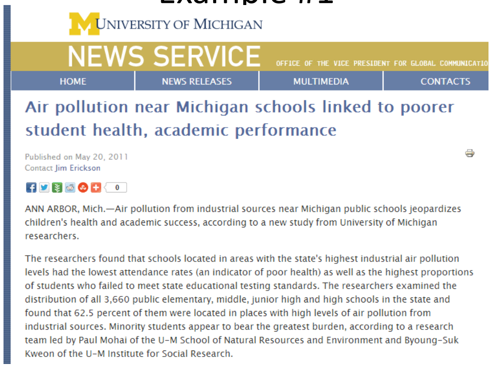
Response Variable: Whether a child commits crimes

Causality and Counterfactuals

$X \longrightarrow Y$

Explanatory Variable	Response Variable
Independent Variable	Dependent Variable
Predictor Variable	Outcome Variable
Upstream Variable	Downstream Variable
Exogenous Variable	Endogenous Variable
Regressor Variable	Regressand Variable
Cause	Effect

Example #1



Air pollution near Michigan schools linked to poorer student health, academic performance

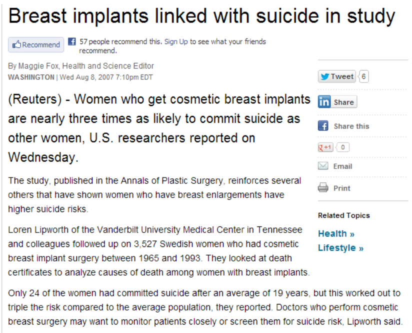
Published on May 20, 2011
Contact Jim Erickson

ANN ARBOR, Mich.—Air pollution from industrial sources near Michigan public schools jeopardizes children's health and academic success, according to a new study from University of Michigan researchers.

The researchers found that schools located in areas with the state's highest industrial air pollution levels had the lowest attendance rates (an indicator of poor health) as well as the highest proportions of students who failed to meet state educational testing standards. The researchers examined the distribution of all 3,660 public elementary, middle, junior high and high schools in the state and found that 62.5 percent of them were located in places with high levels of air pollution from industrial sources. Minority students appear to bear the greatest burden, according to a research team led by Paul Mohai of the U-M School of Natural Resources and Environment and Byoung-Suk Kweon of the U-M Institute for Social Research.

http://ns.umich.edu/new/releases/8398

Example #2



Breast implants linked with suicide in study

Recommend
57 people recommend this. Sign Up to see what your friends recommend.

By Maggie Fox, Health and Science Editor
WASHINGTON | Wed Aug 8, 2007 7:10pm EDT

Tweet 6
Share
Share this
+1 0
Email
Print

(Reuters) - Women who get cosmetic breast implants are nearly three times as likely to commit suicide as other women, U.S. researchers reported on Wednesday.

The study, published in the Annals of Plastic Surgery, reinforces several others that have shown women who have breast enlargements have higher suicide risks.

Loren Lipworth of the Vanderbilt University Medical Center in Tennessee and colleagues followed up on 3,527 Swedish women who had cosmetic breast implant surgery between 1965 and 1993. They looked at death certificates to analyze causes of death among women with breast implants.

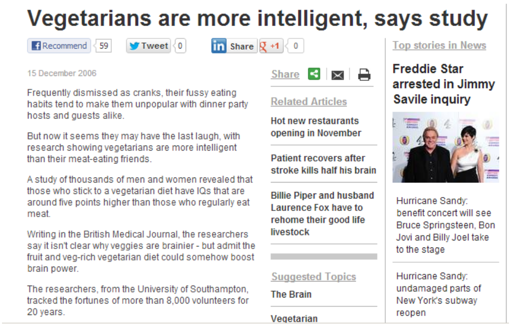
Only 24 of the women had committed suicide after an average of 19 years, but this worked out to triple the risk compared to the average population, they reported. Doctors who perform cosmetic breast surgery may want to monitor patients closely or screen them for suicide risk, Lipworth said.

Related Topics

Health »
Lifestyle »

http://www.reuters.com/article/2007/08/08/us-implants-suicide-idUSN0836919020070808?feedType=RSS&rpc=22&sp=true

Example #3



Vegetarians are more intelligent, says study

15 December 2006

Frequently dismissed as cranks, their fussy eating habits tend to make them unpopular with dinner party hosts and guests alike.

But now it seems they may have the last laugh, with research showing vegetarians are more intelligent than their meat-eating friends.

A study of thousands of men and women revealed that those who stick to a vegetarian diet have IQs that are around five points higher than those who regularly eat meat.

Writing in the British Medical Journal, the researchers say it isn't clear why veggies are brainier - but admit the fruit and veg-rich vegetarian diet could somehow boost brain power.

The researchers, from the University of Southampton, tracked the fortunes of more than 8,000 volunteers for 20 years.

Related Articles

Hot new restaurants opening in November

Patient recovers after stroke kills half his brain

Billie Piper and husband Laurence Fox have to rehome their good life livestock

Suggested Topics

The Brain

Vegetarian

Top stories in News

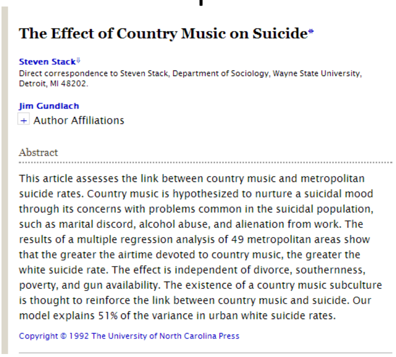
Freddie Star arrested in Jimmy Savile inquiry

Hurricane Sandy: benefit concert will see Bruce Springsteen, Bon Jovi and Billy Joel take to the stage

Hurricane Sandy: undamaged parts of New York's subway reopen

http://www.standard.co.uk/news/vegetarians-are-more-intelligent-says-study-7082629.html

Example #4



The Effect of Country Music on Suicide*

Steven Stack
Direct correspondence to Steven Stack, Department of Sociology, Wayne State University, Detroit, MI 48202.

Jim Gundlach
+ Author Affiliations

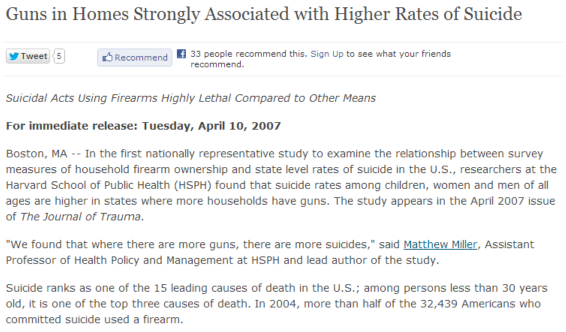
Abstract

This article assesses the link between country music and metropolitan suicide rates. Country music is hypothesized to nurture a suicidal mood through its concerns with problems common in the suicidal population, such as marital discord, alcohol abuse, and alienation from work. The results of a multiple regression analysis of 49 metropolitan areas show that the greater the airtime devoted to country music, the greater the white suicide rate. The effect is independent of divorce, southernness, poverty, and gun availability. The existence of a country music subculture is thought to reinforce the link between country music and suicide. Our model explains 51% of the variance in urban white suicide rates.

Copyright © 1992 The University of North Carolina Press

http://sf.oxfordjournals.org/content/71/1/211.short

Example #5



Guns in Homes Strongly Associated with Higher Rates of Suicide

33 people recommend this. Sign Up to see what your friends recommend.

Suicidal Acts Using Firearms Highly Lethal Compared to Other Means

For immediate release: Tuesday, April 10, 2007

Boston, MA -- In the first nationally representative study to examine the relationship between survey measures of household firearm ownership and state level rates of suicide in the U.S., researchers at the Harvard School of Public Health (HSPH) found that suicide rates among children, women and men of all ages are higher in states where more households have guns. The study appears in the April 2007 issue of *The Journal of Trauma*.

"We found that where there are more guns, there are more suicides," said Matthew Miller, Assistant Professor of Health Policy and Management at HSPH and lead author of the study.

Suicide ranks as one of the 15 leading causes of death in the U.S.; among persons less than 30 years old, it is one of the top three causes of death. In 2004, more than half of the 32,439 Americans who committed suicide used a firearm.

http://www.hsph.harvard.edu/news/press-releases/2007-releases/press04102007.html

Causality and Counterfactuals

At the heart of all case and effect statements is a counterfactual—*the situation that would have existed had the explanatory variable not changed*

For example, by making the causal claim:

“Living in poverty caused kids to commit crimes”

we are simultaneously making the counterfactual claim:

“If the kids had **not** lived in poverty, then they would **not** have committed crimes"

A fundamental problem in explanatory research is that we never actually observe the counterfactual

Worksheet

What are the counterfactual claims that are implied by these causal statements?

1. Air pollution affects kids’ school performance
2. Breast implants affect women’s suicide rates
3. Eating a vegetarian diet makes you smarter
4. Listening to country music leads to suicide
5. Guns in homes lead to more suicides

Causality and Counterfactuals

The fact that we do not observe the counterfactual presents a major logical problem

How do we know what the value of the response variable would have been had the explanatory variable taken on a different value?

We can never get around this fundamental logical problem … but we can design research projects in such a way that we can make sound inferences about the nature of the counterfactual

Causality and Counterfactuals

Experimental research designs are much stronger with respect to their ability to yield valid causal claims

However, we are typically only able to carry out non-experimental or observational research

Multivariate statistical techniques are valuable for use in non-experimental, observational research precisely because they help us approximate counterfactual situations

Criteria for Establishing Causality

Momentarily putting aside this logical issue associated with our inability to observe the counterfactual…

…and recognizing that there are many philosophical schools of thought on our ability to make inferences about cause-and-effect relationships…

…and regardless of our research design (experimental or non-experimental)…

…there are some general criteria that must be satisfied in order to claim that change in explanatory variable X causes change in response variable Y

Criteria for Establishing Causality

Three conditions that must be met in order to establish that X causes Y:

1. X and Y must be empirically associated (criteria of association)
2. X must precede Y in time (criteria of temporal ordering)
3. There must be no third variable, Z, which acts as a "confounder"—or which induces "spuriousness"—in the association between X and Y (criteria of nonspuriousness)

Different research designs have their own ways of meeting these criteria, of facilitating inferences about the nature of the counterfactual, and thus of allowing us to make defensible causal claims

Criteria for Establishing Causality

Criteria of Association

In order to establish that X causes Y—or that change in the explanatory variable X causally leads to change in the response variable Y—we must observe an empirical association between X and Y

We have spent the past several weeks learning how to measure the association between variables (with our technique for doing so dependent upon whether X and Y are discrete, continuous, or a combination of the two)

Criteria for Establishing Causality

How was the association between X and Y established in these examples?

1. Air pollution affects kids' school performance
2. Breast implants affect women's suicide rates
3. Eating a vegetarian diet makes you smarter
4. Listening to country music leads to suicide
5. Guns in homes lead to more suicides

Criteria for Establishing Causality

Criteria of Temporal Ordering

In order to establish that X causes Y—or that change in the explanatory variable X causally leads to change in the response variable Y—it must be the case that the change in X occurred before the change in Y

As we will discuss, this can be difficult in observational research

Criteria for Establishing Causality

Is the criteria of causal ordering clearly met in these examples?

1. Air pollution affects kids' school performance
2. Breast implants affect women's suicide rates
3. Eating a vegetarian diet makes you smarter
4. Listening to country music leads to suicide
5. Guns in homes lead to more suicides

Criteria for Establishing Causality

Criteria of Nonspuriousness

Suppose we observe that shoe size at age 18 is associated with frequency of criminal behavior in adulthood, such that people with bigger shoes at age 18 commit more crimes later on

We've met the first two criteria required to make causal statement— but can we say that shoe size at age 18 causally affects criminal behavior in adulthood?

Criteria for Establishing Causality

Criteria of Nonspuriousness

If some third variable(s) affects both shoe size at age 18 and criminal behavior in adulthood, then the association between shoe size and criminal behavior is a spurious relationship, not a causal relationship

Can you think of any?

The "third variables" are called confounding variables

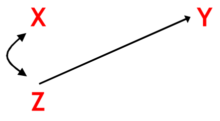
Criteria for Establishing Causality

Criteria of Nonspuriousness

What is the effect of X on Y?

In this example, X and Y are associated, but the association is *entirely* spurious owning to Z

Z is a counfounder



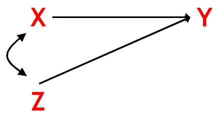
Criteria for Establishing Causality

Criteria of Nonspuriousness

What is the effect of X on Y?

Here X and Y are associated

That association is *partly* spurious (owing to confounder Z) and *partly* causal



Criteria for Establishing Causality

Could spuriousness threaten the validity of these causal claim?

1. Air pollution affects kids' school performance
2. Breast implants affect women's suicide rates
3. Eating a vegetarian diet makes you smarter
4. Listening to country music leads to suicide
5. Guns in homes lead to more suicides

Worksheet

1. Having a TV set in the bedroom (A) correlated with couples having lower frequency of sexual activity (B) (research finding).
2. Listening to certain types of sexual lyrics (A) correlated with teen sexual activity (B) (research finding).
3. Alcohol consumption (A) correlated with violent behavior (B) (research findings).
4. In his career, when Adrian Peterson rushes for 100 yards or more (A) the Vikings usually win, going 20-10 through November 1, 2012 (B)

Causality and Research Design

In experimental research cases are randomly assigned to two or more comparison groups … X is defined by the group to which cases are assigned

The response variable Y is measured before (pre-test) and after (post-test) the manipulation of X

If the change in Y between the pre-test and the post-test differs across levels of X, then X and Y are associated

Because the value of X is assigned at random, spuriousness is not possible

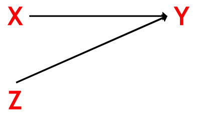
Causality and Research Design

Experimental Design

What is the effect of X on Y?

X and Y are associated; X precedes Y in time

Because the value of X is assigned at random, spuriousness is not possible



Causality and Research Design

Main weaknesses of experimental methods:

1. Experiments can be expensive or unethical to conduct in many circumstances
2. Experimental research does not typically yield any information about causal mechanisms
3. Randomly assigning cases to comparison groups is not the same thing as randomly selecting a sample … results from experiments are not always generalizable

Causality and Research Design

Observational research involves studying naturally-occurring variation in X and Y, with no intervention from the researcher

There are a number of techniques for assessing the magnitude of association between X and Y

However, in observational research…

…it is often difficult to establish temporal ordering

…it is usually extremely difficult to rule out the possibility that the observed relationship is spurious

Causality and Research Design

Despite these threats to making valid causal statements, most social science research is observational

If we want to study the effects of most important social it is usually not possible (or ethical) to randomly assign cases to treatment and control groups

To establish temporal ordering, researchers sometimes use longitudinal (as opposed to cross-sectional) designs

To avoid problems with spuriousness, researchers attempt to statistically control for confounding variables

Causality and Research Design

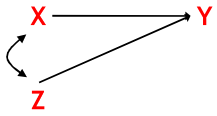
Observational Design

What is the effect of X on Y?

X and Y are associated; X may precede Y in time

The association between X and Y that remains after statistically controlling for Z is causal in nature

(but only if we control for all Z variables!)



Statistical Control

Statistical control is a technique used in observational research to reduce the risk of spuriousness

The confounding variables Z are “held constant” so that we can observe the independent association between X and Y “net of” Z

Conceptually, holding Z “constant” means observing the association between X and Y among people with equal values on the confounding Z variables

We presume that the association between X and Y that persists after statistically controlling for Z is causal in nature (as long as the other criteria have been met)

Statistical Control

Example: What is the effect of education on family income as an adult?

		Family Income as Adult		
		< Avg	Avg	> Avg
	< H.S.	2,141	2,959	440
Education	H.S.	2,346	5,252	1,403
	> H.S.	2,525	5,682	4,406

Source: GSS

Gamma = 0.37

Criteria of Association? … OK

Temporal Ordering Criteria? … OK

Nonspuriousness? … What about family background?

Statistical Control

Example: What is the effect of education on family income as an adult *net of father's education*?

Father: <H.S.

		Family Income as Adult		
		< Avg	Avg	> Avg
	< H.S.	1,811	2,445	341
Education	H.S.	1,461	3,106	731
	> H.S.	908	1,981	1,300

Source: GSS

Gamma = 0.34

Father: H.S.

		Family Income as Adult		
		< Avg	Avg	> Avg
	< H.S.	242	359	53
Education	H.S.	613	1,594	467
	> H.S.	684	1,776	1,190

Source: GSS

Gamma = 0.30

Father: >H.S.

		Family Income as Adult		
		< Avg	Avg	> Avg
	< H.S.	88	155	46
Education	H.S.	272	552	205
	> H.S.	933	1,925	1,916

Source: GSS

Gamma = 0.32

The association between X and Y is partially spurious

Causality: Myths Dispelled

Myth #1: "X isn't the only (or even the most important) cause of Y, so it's wrong to say that X causes Y"

- Mechanical vs. probabilistic explanations
- Whether you get lung cancer depends on a lot of things … only one of which is cigarette smoking
- Just because smoking isn't the only factor that affects whether you get lung cancer doesn't mean it isn't a causal factor

Causality: Myths Dispelled

Myth #2: “Exceptions Disprove the Rule”

- Mechanical vs. probabilistic explanations
- Remember that causal statements pertain to how one *variable* affects another
- Even if smoking does cause lung cancer, we should still expect to find many smokers who don’t get lung cancer and many people with lung cancer who never smoked

Causality: Myths Dispelled

Myth #3: “X doesn’t always—or even usually—lead to Y, so therefore X doesn’t cause Y”

- Does driving drunk cause people to crash their cars?
- The vast majority of drunk drivers get home safely
- Does this mean that drunk driving has no effect on the odds of crashing? No…
- As long as the probability of crashing is higher for drunk drivers than for sober drivers — even if the probability is still very low for drunk drivers — then it is still true that driving drunk is causally related to crashing

BREAK

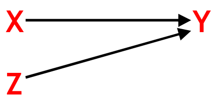
Quick Review of Causality

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 observed association between X and Y

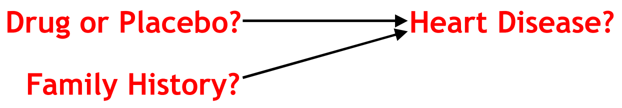
Does X Causally Affect Y?

In experimental research, X precedes Y in time and spurious is not possible. So, any association between X and Y is entirely causal in nature



Does X Causally Affect Y?

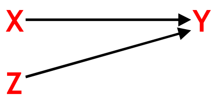
In experimental research, X precedes Y in time and spurious is not possible. So, any association between X and Y is entirely causal in nature



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:

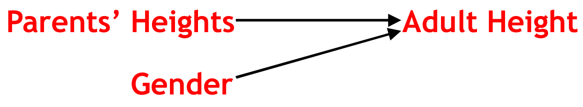
1. The association is entirely causal in nature



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:

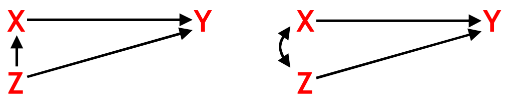
1. The association is entirely causal in nature



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:

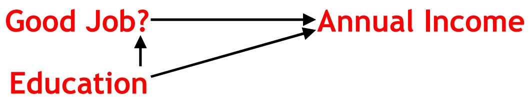
2. The association is partly causal and partly spurious owing to confounder Z



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:

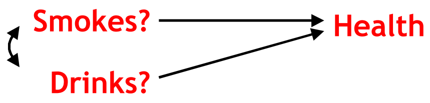
2. The association is partly causal and partly spurious owing to confounder Z



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:

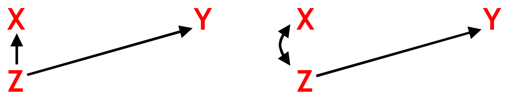
2. The association is partly causal and partly spurious owing to confounder Z



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:

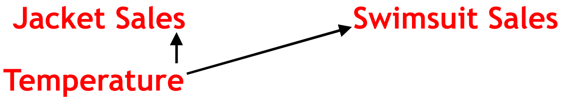
3. The association is entirely spurious owning to Z



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

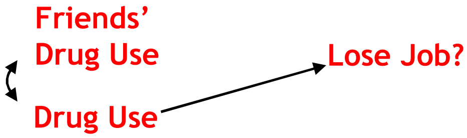
Swimsuit Sales

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

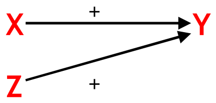
1. What are the consequences of *failing to control for Z* in each of the three situations reviewed above?
2. What are the consequences of *controlling for Z* in each situation?

“Control” for Z?

Here, 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

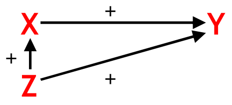


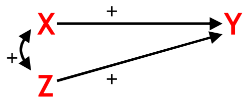
“Control” for Z?

Here, 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!



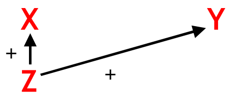


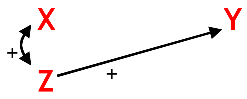
“Control” for Z?

Here, 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

Three Variable Relationships

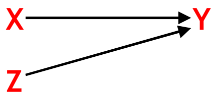
In fact there are a variety of ways in which X, Y, and Z might be related to one another

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 <u>might</u> or <u>might not</u> want to statistically control for Z

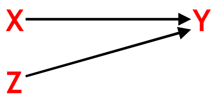
Three Variable Relationships

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



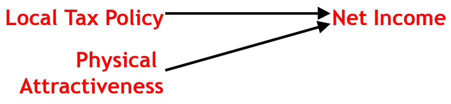
Three Variable Relationships

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)



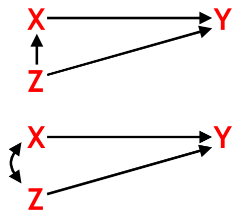
Three Variable Relationships

Example



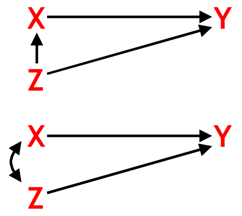
Three Variable Relationships

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



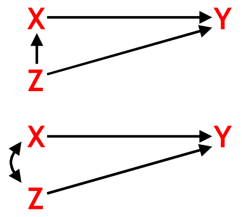
Three Variable Relationships

Under each of these scenarios, statistically controlling for Z is designed to estimate the *independent* association between X and Y … the association “net of” Z



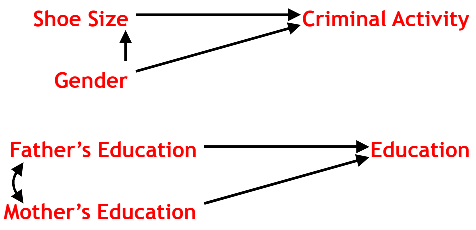
Three Variable Relationships

Here, the independent association between X and Y … the association “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

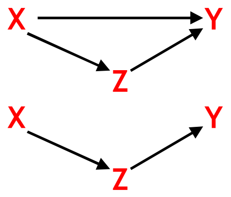
Example



Three Variable Relationships

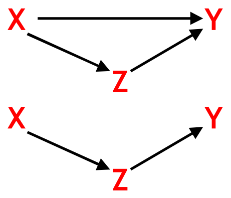
Under the scenarios depicted below, Z is associated with both X and Y but does **not** induce spuriousness

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



Three Variable Relationships

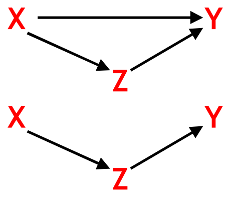
Statistically controlling for Z under these scenarios is also designed to estimate the independent association between X and Y … the association “net of” Z



Three Variable Relationships

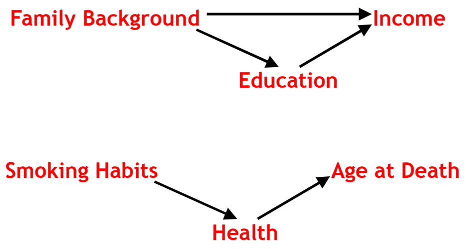
Conceptually, in these cases the independent association between X and Y … the association “net of” Z … represents the direct effect of X on Y

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



Three Variable Relationships

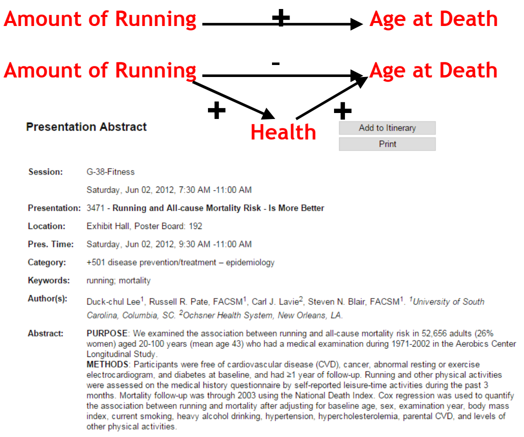
Examples:



Family Background → Income

Family Background → Education → Income

Smoking Habits → Health → Age at Death



Amount of Running + → Age at Death

Amount of Running - → Age at Death

Amount of Running + → Health + → Age at Death

Presentation Abstract

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

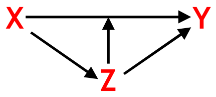
What are some potential confounders and some potential causal mechanisms linking X to Y in the following examples?

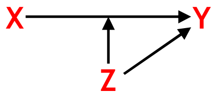
1. Kids who frequently misbehave in high school (X) are less likely to go to college (Y)
2. Men who are married (X) live longer (Y) than men who are not married
3. Eating breakfast (X) improves employees’ productivity (Y) during the day

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

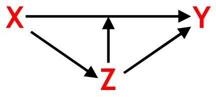

X
Y
Z

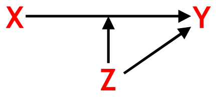

X
Y
Z

Three Variable Relationships

If we simply statistically adjust for Z in these situations, then the resulting "net" association between X and Y isn't very meaningful

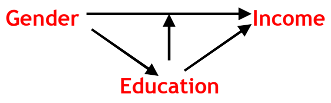
We must allow for different associations by level of Z

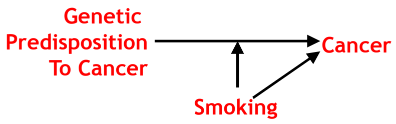

X
Y
Z


X
Y
Z

Three Variable Relationships

Example

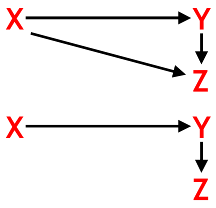

Gender
Income
Education


Genetic Predisposition To Cancer
Cancer
Smoking

Three Variable Relationships

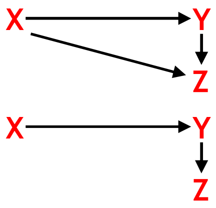
Under the scenarios depicted below, Z is affected by Y (and also possibly X)

Here, Y is a mechanism through which X affects Z



Three Variable Relationships

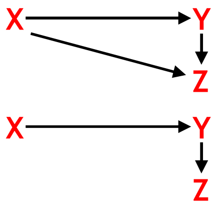
If we are trying to assess the independent association between X and Y, then in these situations statistically controlling for Z is a **terrible idea**



Three Variable Relationships

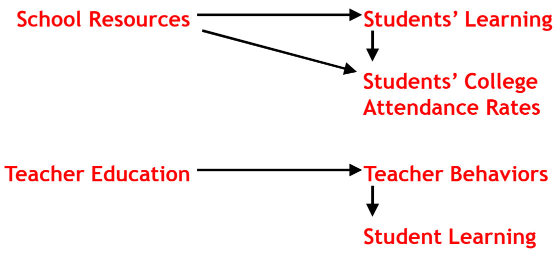
In effect, controlling for Z in these situations is like selecting on the dependent variable

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)

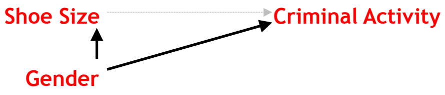
Gamma= 0.31

		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

Gamma = 0.0

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

χ^2 = 0.0

Gamma = 0.0

		WOMEN Shoe Size		
		1=Small	*2=Large*	Row
Felony?	*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

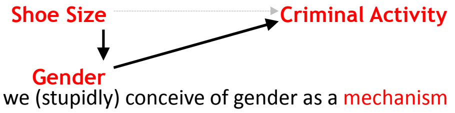
Note also that controlling for Z may…

- …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:



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)

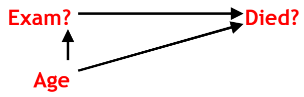
Gamma= 0.35

		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:



Exam?

Died?

Age

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 $(p<0.01)$

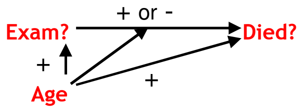
Gamma = -0.82

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



+ or -

Exam?

Died?

+

+

Age

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)

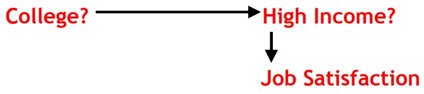
Gamma= 0.20

		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.