Statistical Control

- Statistical Control is about
  - underlying causation
  - under-specification
  - measurement validity
  - "Correcting" the bivariate relationship
- Regression & Residual formulas
- Correlations among y, x, y', & residuals
- Control for Design & Measurement problems
- "Controlling" proxy variables
- Limitations & advantages of statistical control

Variations of statistical control

Control of single variables

partial correlation -- correlation between two variables (x & y) controlling **both** for some 3rd variable (z)

-- $r_{yx.z}$

semi-partial correlation -- correlation between two variables (x & y) controlling **one** of the variables for some 3rd variable (z)
(part correlation)

-- $r_{y\,(x.z)}$ & $r_{x\,(y.z)}$

Control of multiple variables…

multiple partial correlation -- like partial, but with "multiple 3rd variables"

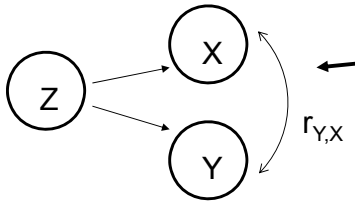
-- $r_{yx.zabc}$

multiple semi-partial correlation -- like semi-partial, but with "multiple 3rd variables"

-- $r_{y(x.zabc)}$ vs. $r_{x(y.zabc)}$

ANCOVA -- ANalysis of COVAriance -- a kind of semi-partial or multiple semi-partial corr
-- test of the DV mean difference between IV conditions that is independent of some 3rd variable (called the "covariate")
-- $r_{DV\,(IV.cov)}$

Statistical control is about the underlying causal model of the variables …

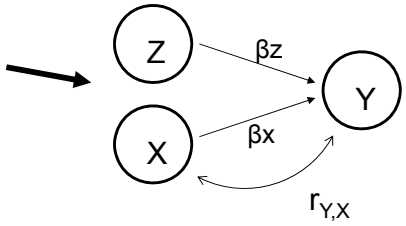


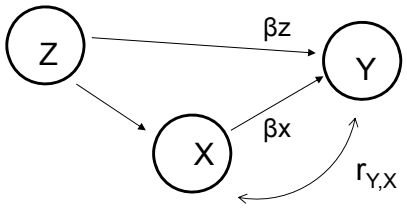
3rd variable problem …

- relationship between x & y exists because a 3rd variable is influencing both
- Z might be a causal or a psychometric influence

Misspecification problem …

- by ignoring Z, the X-Y relationship misses part of the story



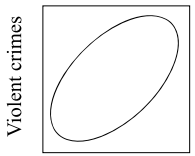


Combining the two…

Part of the story may be "indirect effects"

"3rd variable problems"

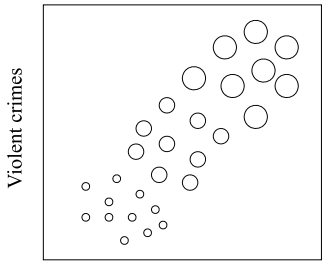
Here's a well-known example ...



When plotted by week or by month -- there is a +r between ice cream sales & amount of violent crime. Huh?

- Does eating ice cream make you violent ?
- Does being violent make you crave ice cream ?

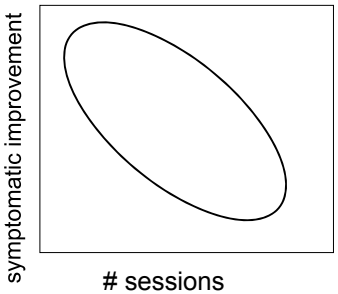
Is there some "3rd variable" variable that is "producing" the bivariate correlation? What might it be ???



Here's the same scatterplot, but with the size of each data point representing the average temperature of that period.

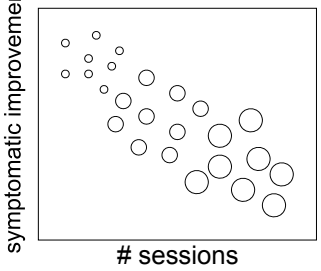
We can see that there is no relationship between violent crimes and ice cream sales after controlling both for temperature. Temperature might be that "3rd variable".

Another example ...



We found a -r between # therapy sessions and amount of symptomatic improvement! Huh?!? Let's think through this...

- The sample is heterogeneous with respect to initial level of depression
- Initial level of depression is likely to be related to # sessions they attend
- Initial level of depression is likely to be related to symptomatic improvement
- So, is the relationship each of these variables has with initial level of depression "producing" the bivariate correlation we found?



dotsize indicates initial depression…

Those who are more depressed come to more sessions and show less improvement.

Those who are less depressed come to fewer sessions and show more improvement.

Misspecification problems…

When we take a bivariate look at part of a multivariate picture …

- we'll underestimate how much we can know about the criterion
- we'll likely mis-estimate how that predictor relates to the criterion
  - leaving predictors out usually leads to over-estimation $r > \beta$
  - leaving predictors out changes the collinearity structure, and so, might cause us to miss suppressor effects $r < \beta$

Statistical control in an attempt to improve this with a "better r"…

Allman Brothers Band - No One To Run With.mp3

- what "would be" the bivariate relationship between these variables, in a population for which the control variable(s) is a constant (and so is not collinear with these variables) ?
- it is very much like looking at the β for that predictor in a multiple regression, but is in the form of a "corrected" simple correlation

$$r_{y(x.z)} \approx \beta_x \text{ from } y' = \beta_x X + \beta_z Z$$

"What is relationship between y & part of x independent of Z?

Statistical control is about "correcting" the bivariate correlation to take the control variable(s) "into account"

What do we get from this? Here's where opinions differ …

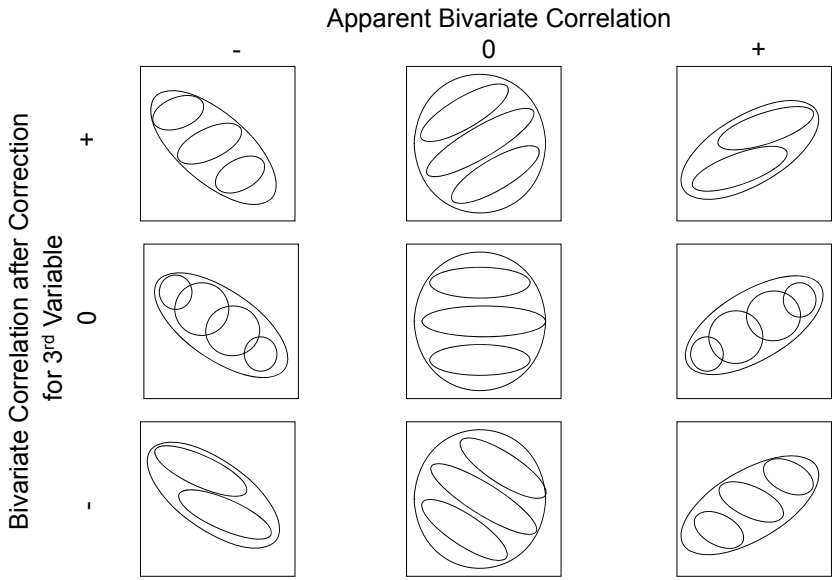
1. A substitute for experimantal control ?
2. A better estimate of the causal relationship of the 2 variables ?
3. A substitute for construct validity ?
4. Solves under-specification problem ?
5. Probably a better description of the relationship of these 2 variables than is the bivariate analysis ?

Rejection of 1-4 (probably for good reasons) has led some to reject the 5th as well …but then what are we to do?

- Only perform bivariate analyses (known to be flawed) ?
- Expect to construct the "full story" from convergent research? (which is already the answer for "what's the correct study"!!!)

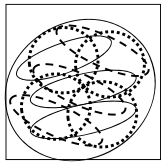
More complex models are, on average, more likely to be accurate!

Some examples of taking the 3rd variable into account

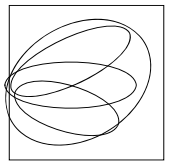


Of course, it can be uglier than that…

Here, what "correction" you get depends upon which variable you control for – remember, larger models are only more accurate "on average"



Here, the addition of the "Z" variable will help, but there is also a 3rd – the interaction of X & Z



Here are two formula that contain "all you need to know"

$y' = bx + a$ $\quad$ residual = $y - y'$

y = the criterion $\quad$ x = the predictor

y' = the predicted criterion value
-- that part of the criterion that is related to the predictor (Note: $r_{yx} = r_{yy'}$)

residual = difference between criterion and predicted criterion values
-- the part of the criterion not related to the predictor
(Note: $r_{x\,res} = 0$ or $r_{x(y-y')} = 0$)

Summary of Interrelationships among the key values…

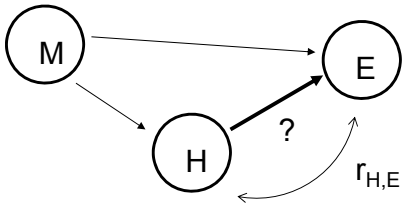
	X	Y	Y'	Y - Y'
X	---	y' "is" x		
Y	r_{yx}	---	so, y is related to y' as to x	
Y'	1.0	r_{yx}	---	
Y - Y'	0.0	$\sqrt{(1 - r_{yx}^2)}$	0.0	---

the residual is the part of the criterion not related to the y or y' (which "is" x)

Examples of Statistical Control A design problem …

We want to assess the correlation between the amount of homework completed and performance on exams. Since we can not RA participants to amount of homework, we are concerned that motivation will be a confound. Specifically, we think that motivation will influence both the number of homeworks each participant completes and how well they prepare for the exam.

So, the question we want to ask can be phrased as, "What is the correlation between amount of homework completed and test performance that is independent of (not related to) motivation?"



To do this we want to correlate the part of homeworks completed that is not related to motivation, with the part of test scores that is not related to motivation.

Design control using Partial Correlation by residualization

Step 1a predict # homeworks completed from motivation

#hwks' = b(motivation) + a

Step 1b find the residual (part of #hwks not related to motivation)

residual #hwks = #hwks - #hwks' (rem: $r_{x\,(y\text{-}y')} = 0$)

Step 2a predict test scores from motivation

test' = b(motivation) + a

Step 2b find the residual (part of test scores not related to motivation)

residual test = test - test' (again: $r_{x\,(y\text{-}y')} = 0$)

Step 3 correlate the residuals (the part of # hwks not related to motivation and the part of test scores not related to motivation) to find the relationship between # hwks and test scores that is independent of motivation

Design control using Semi-Partial Correlation by residualization

Consider a version of this question for which we want to control only homework scores for motivation

Step 1a predict # homeworks completed from motivation

#hwks' = b(motivation) + a

Step 1b find the residual (part of #hwks not related to motivation)

residual #hwks = #hwks - #hwks' (rem: $r_{x\,(y\text{-}y')} = 0$)

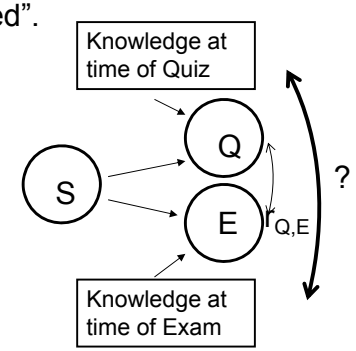
Step 2 correlate the residual of homework (the part of # hwks not related to motivation) and original test scores to find the relationship between test scores and that part of homework scores that is independent of motivation.

Notice that the difference between partial and semi-partial (part) correlations is that when doing the semi-partial we residualize only one of the variables we want to correlate.

Examples of Statistical Control A measurement problem …

We want to assess the correlation between performance on quizzes and performance on the final exam. But we know that both of these variables include "test taking speed". So, we are concerned that the correlation between the two types of performance will be "tainted" or "inflated".

So, the question we want to ask can be phrased as, "What is the correlation between quiz performance and final exam performance that is independent of (not related to) test taking speed?"



To do this we want to correlate that part of quiz scores that is not related to test taking speed, with that part of final exam scores that is not related to test taking speed.

Measurement control using Partial Correlation by residualization

Step 1a predict quiz scores from test taking speed

quiz' = b(speed) + a

Step 1b find the residual (part of quiz scores not related to test taking speed)

residual quiz = quiz - quiz'

Step 2a predict final exam scores from test taking speed

final' = b(speed) + a

Step 2b find the residual (part of final exam scores not related to test taking speed)

residual final = final - final'

Step 3 correlate the residuals (the part of quiz scores not related to speed and the part of final scores not related to speed) to find the relationship between quiz scores and final scores that is independent of test taking speed

Measurement control using Semi-Partial Correlation by residualization

Consider a version of this question that involves controlling only quiz scores for test taking speed.

Step 1a predict quiz scores from test taking speed

quiz' = b(speed) + a

Step 1b find the residual (part of quiz scores not related to test taking speed)

residual quiz = quiz - quiz' (rem: $r_{x\,(y\text{-}y')} = 0$)

Step 2 correlate the residual of quiz scores (the part of quiz scores not related to speed) and original final scores to find the relationship between final scores that part of quiz scores that is independent of test taking speed

Notice that the difference between partial and semi-partial (part) correlations is that when doing the semi-partial we residualize only one of the variables we want to correlate.

"Customized" Statistical Control…

It is possible to "control" the X and Y variables for different variables, for example…

controlling Final scores for test taking speed while controlling Homework scores for motivation & #practices.

Step 1

final' = b(speed) + a
residual final = final - final'

Step 2

#hwks' = b(motivation) b(#pract) + a
residual #hwks = #hwks - #hwks'

Step 3 Correlate residual final & residual #hwks

Please note: 1) There is no MReg equivalent,
2) There must be a theoretical reason for doing this!

Statsitical control is an obvious help with under-specification, but it can also be help with proxy variables …

Say we find $r_{experf,\ hwperf} = .4$ → we might ask for what is "homework performance" a proxy?

Consider other things that are related to hwperf & add those to the model, to see if "the part of hwperf that isn't them" is still related to experf… lots of possible results → each with a different interp!!!

$r_{experf,(hwperf.mot)} = .1$ → part of hwperf is "really" motivation (usual collinearity result)

$r_{experf,(hwperf.mot)} = .4$ → none of hwper is motivation (mot and hwperf not collinear)

$r_{experf,(hwperf.mot)} = .6$ → part of hwperf not related to mot is more corr with experf than hwperf (suppressor)

Should continue with other variables → can't control everything, so we should focus on most important variables to consider as confounds or measurement confounds.

More about "Problems" with statistical control

In analyses such as the last two examples, we have statistically changed the research question, for example.

1st e.g. -- Instead of asking, "What is the relationship between quiz scores and final exam scores that is independent of test taking speed in the population represented by the sample?" We are asking, "What would be the correlation between quiz scores and final exam scores in the population, *if all the members of that population had the same test taking speed*?"

2nd e.g. -- Instead of asking, "What is the relationship between the amount of homework completed and test performance that is independent of motivation?" We are asking, "What would be the correlation between amount of homework completed and test performance in the population, *if all members of that population had the same motivation*?"

Populations such as these are unlikely to be representative !!

Advantages of Statistical Control

1st one -- Sometimes experimental control is impossible. Sometimes “intrusion free” measures aren’t available

2nd one -- Statistical control is often less expensive (might even be possible with available data, if the control variables have been collected). Doing the analysis with statistical control can help us decide whether or not to commit the resources to perform the experimental control or create better measures

3rd one -- often the only form of experimental control available is post-hoc matching (because you’re studying natural or intact groups). Preliminary analyses that explore the most effective variables for “statistical control” can help you target the most appropriate variables to match on. Can also retain sample size (hoc matching of mis-matched groups can decrease sample size dramatically)