BIAS CORRECTION THROUGH INSTRUMENTAL VARIABLES

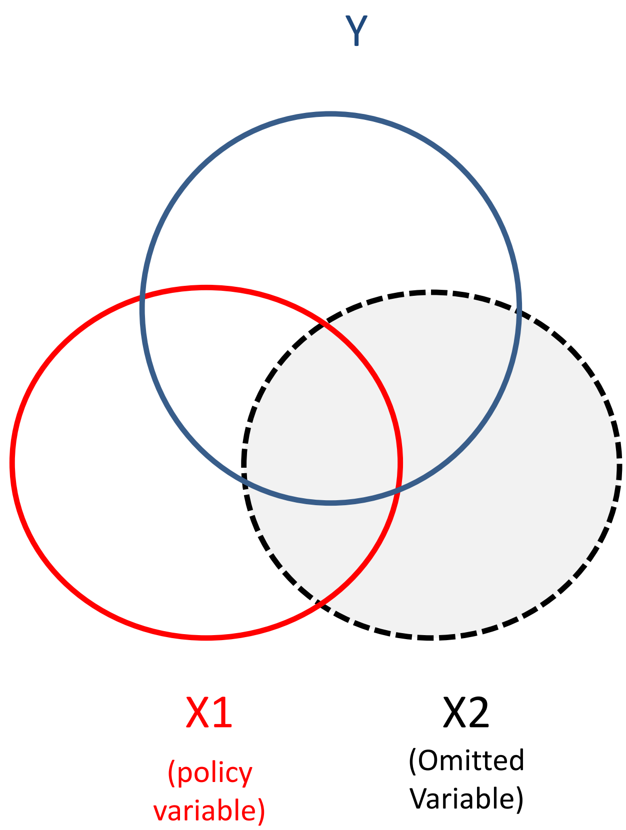
Fundamentals of

PROGRAM EVALUATION

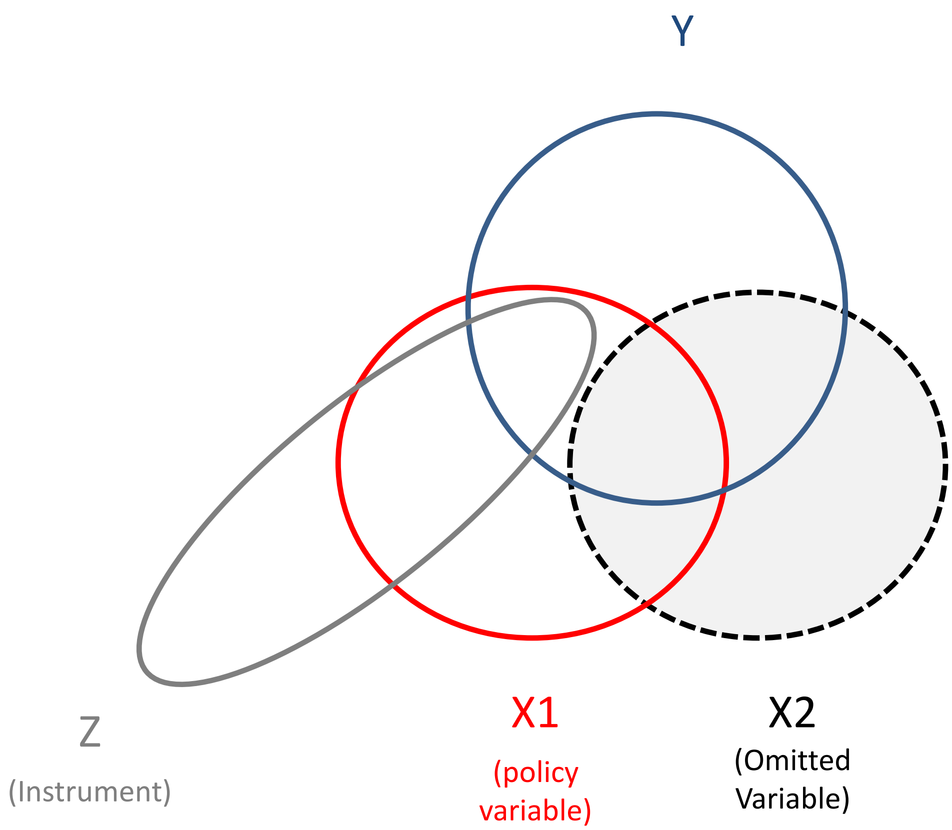
JESSE LECY

THREE MINUTE VERSION:

WE HAVE AN OVB PROBLEM:

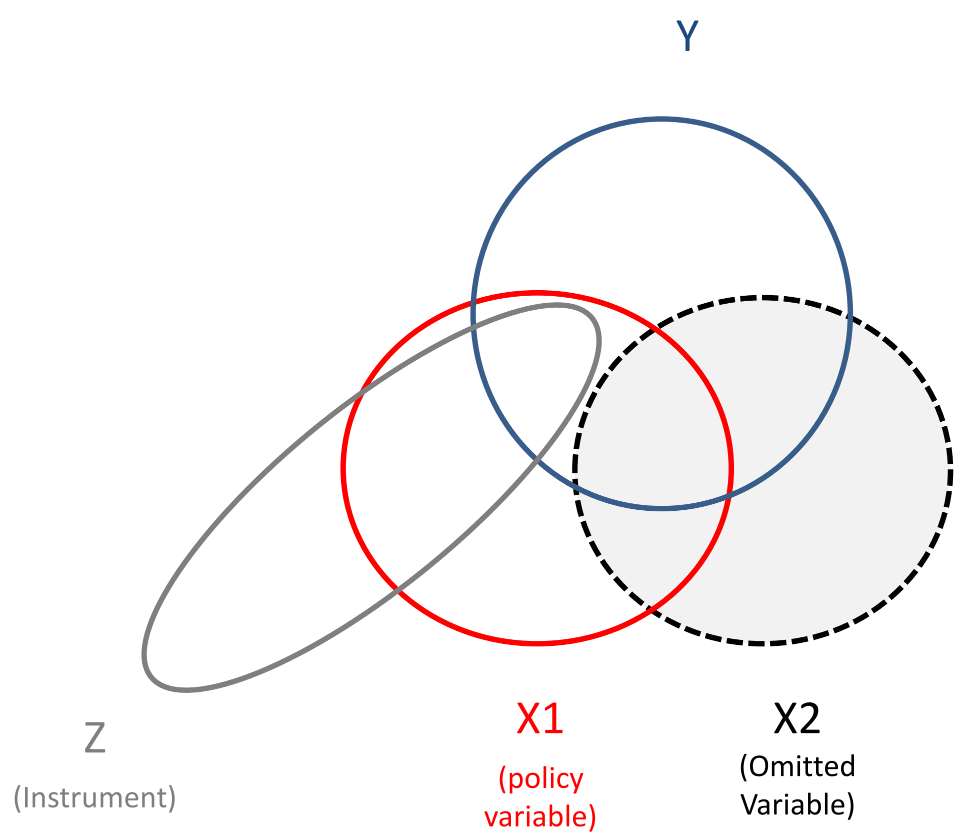


WE HAVE AN INSTRUMENTAL VARIABLE:



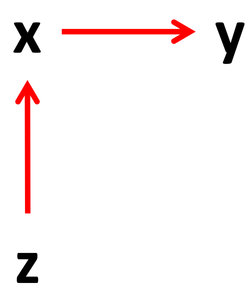
Z is correlated with X1 (the policy variable), uncorrelated with X2 (the omitted variable).

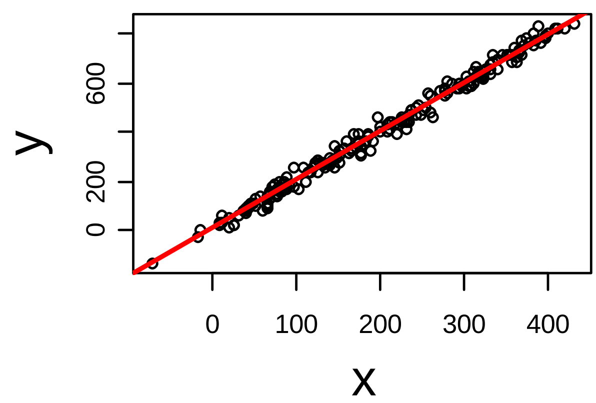
WE HAVE AN INSTRUMENTAL VARIABLE:



Z is correlated with Y only THROUGH X1.

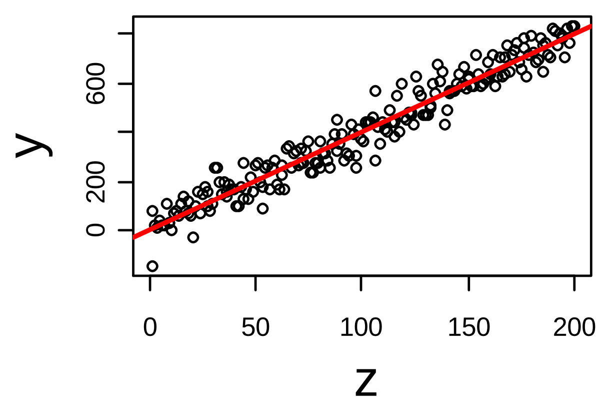
There is no causal relationship between Z and Y.

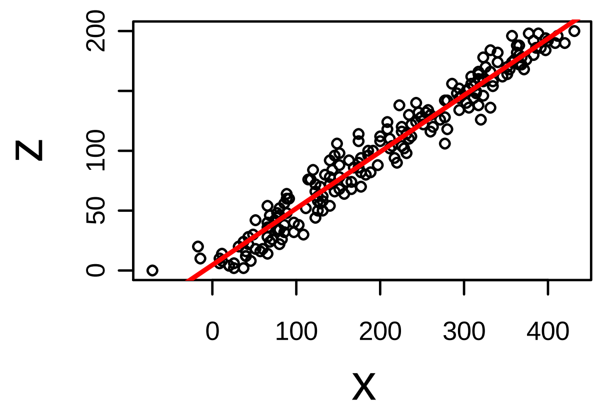




Correlation between Y and Z is natural extension of this causal structure.

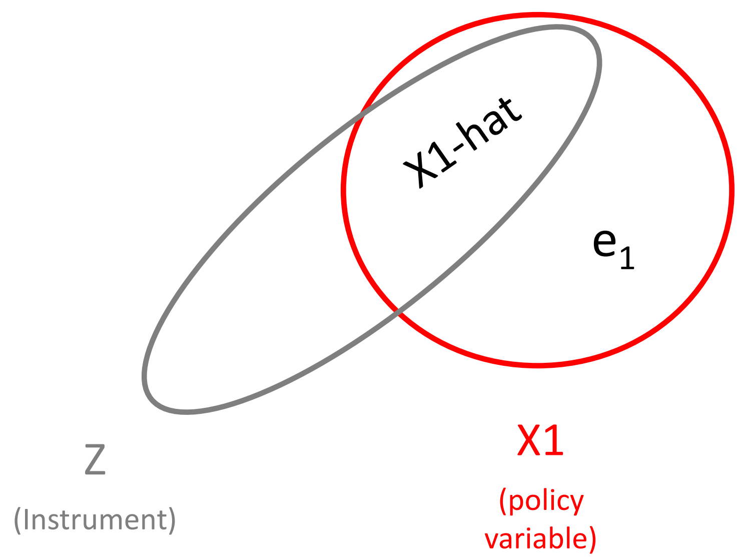
In this data, Y has no direct causal relationship with Z, but there will be a strong correlation in the data because of Y and Z are both correlated with X.





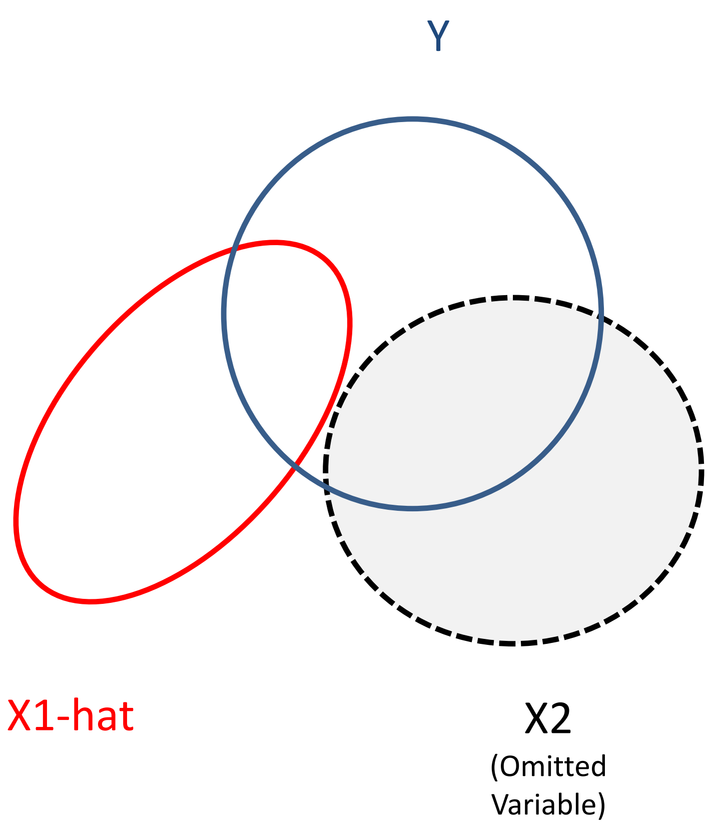
STEP 1:

Partition X1 into parts correlated and uncorrelated with Z



$$X1 = a_0 + a_1 * Z + e_1$$

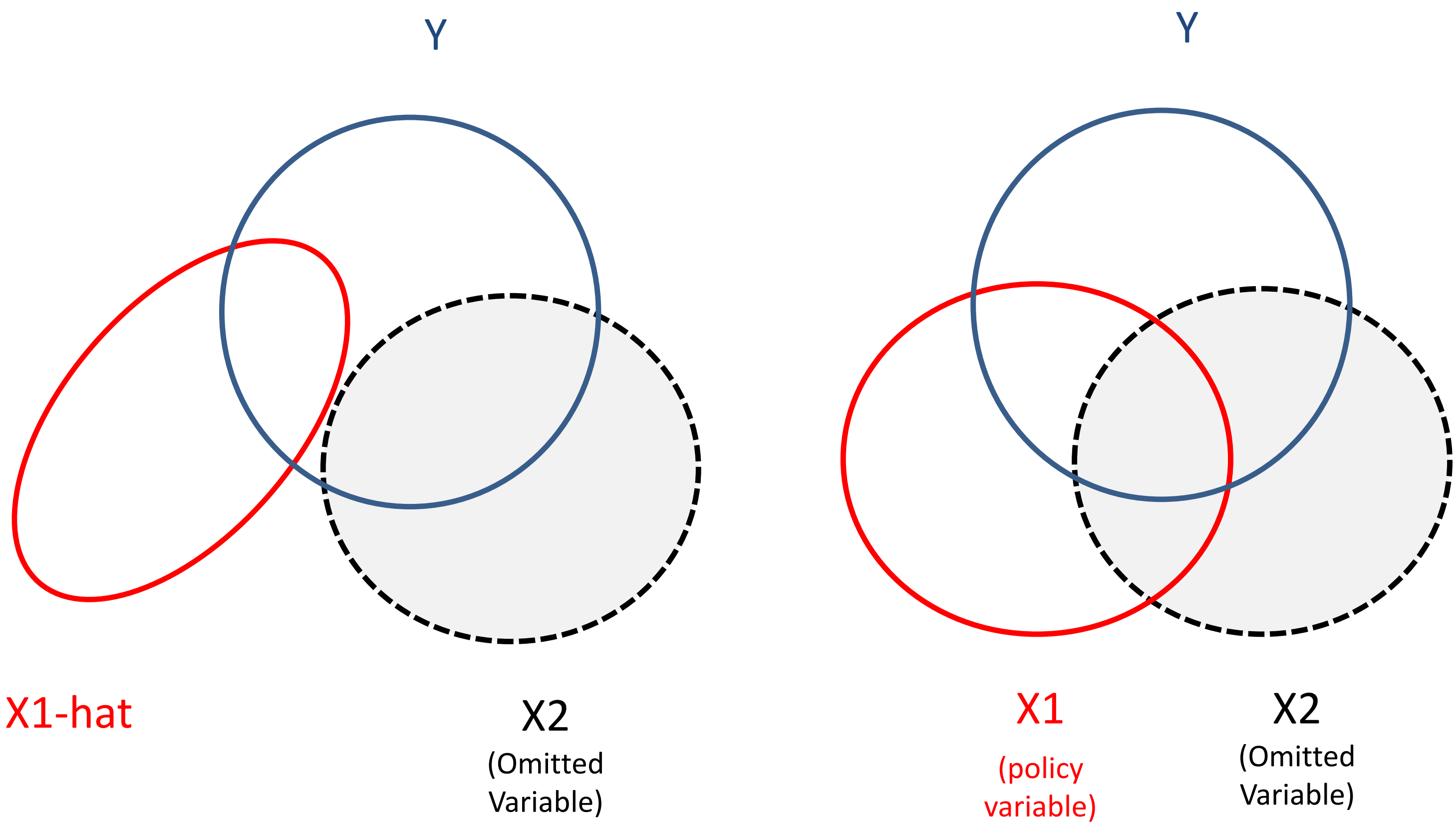
STEP 2:



$$Y = b_0 + b_1{*}X1\text{-hat} + e_2$$

Keep only the part of X1 that is correlated with Z.

OVB MITIGATED:



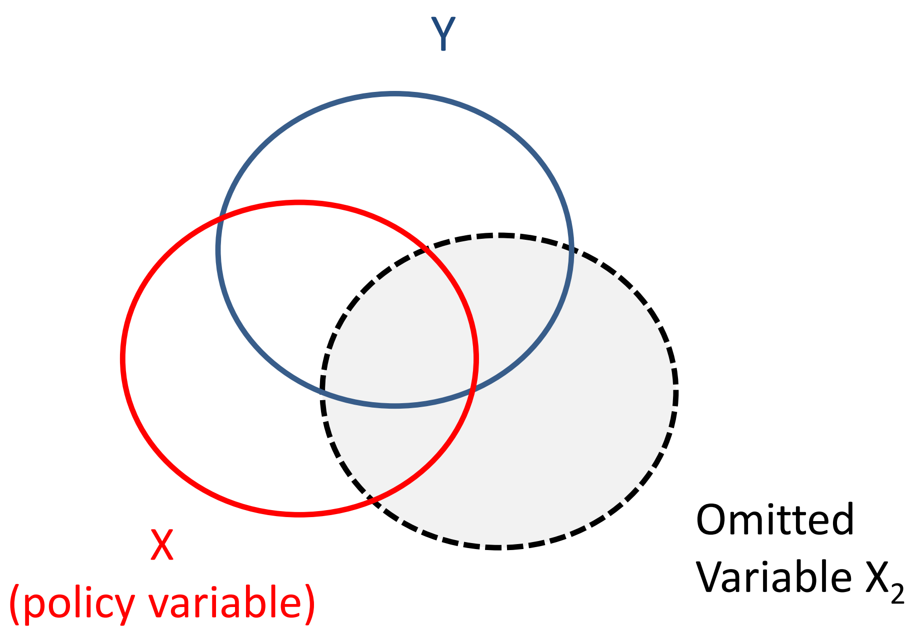
$Y = B_0 + B_1*X1 + B_2*X2 + e$ Full Model

$Y = b_0 + b_1*X1\text{-hat} + e_2$ IV Model

$b_1 \approx B_1$ Naïve slope close to the true slope

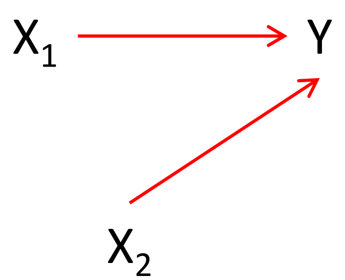
FULL VERSION:

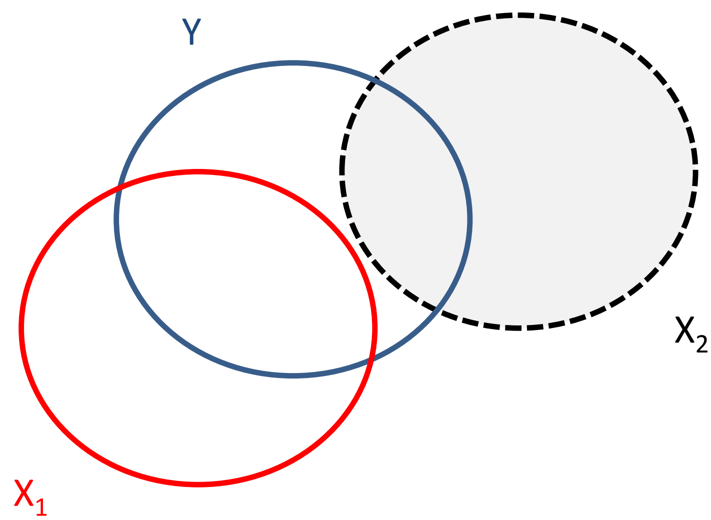
WE HAVE USED A CORRELATION MODEL TO EXAMINE OMITTED VARIABLE BIAS



But there are many possible causal models:

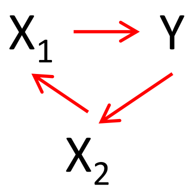
Case 1: Not Problematic



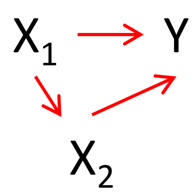


BUT THERE ARE MANY POSSIBLE CAUSAL MODELS:

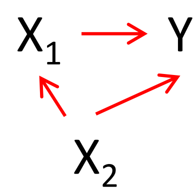
Case 2



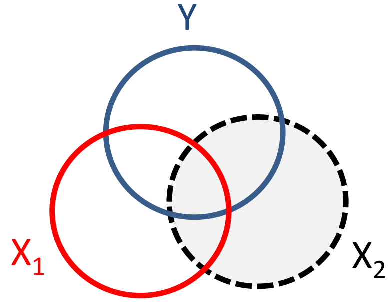
Case 3



Case 4

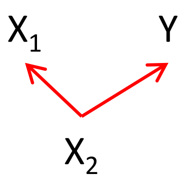


These are problematic when we omit X_2 because our slope on X_1 will not be biased

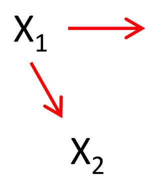


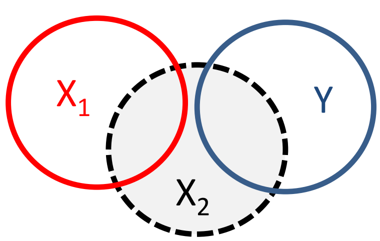
LURKING VARIABLES…SPURIOUS CORRELATIONS

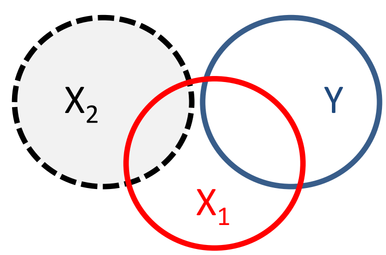
Case 5



Case 6



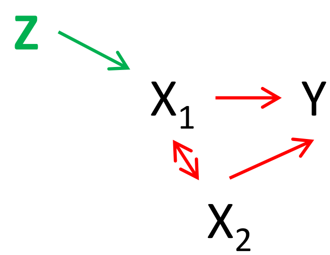




Don’t worry about these cases for now

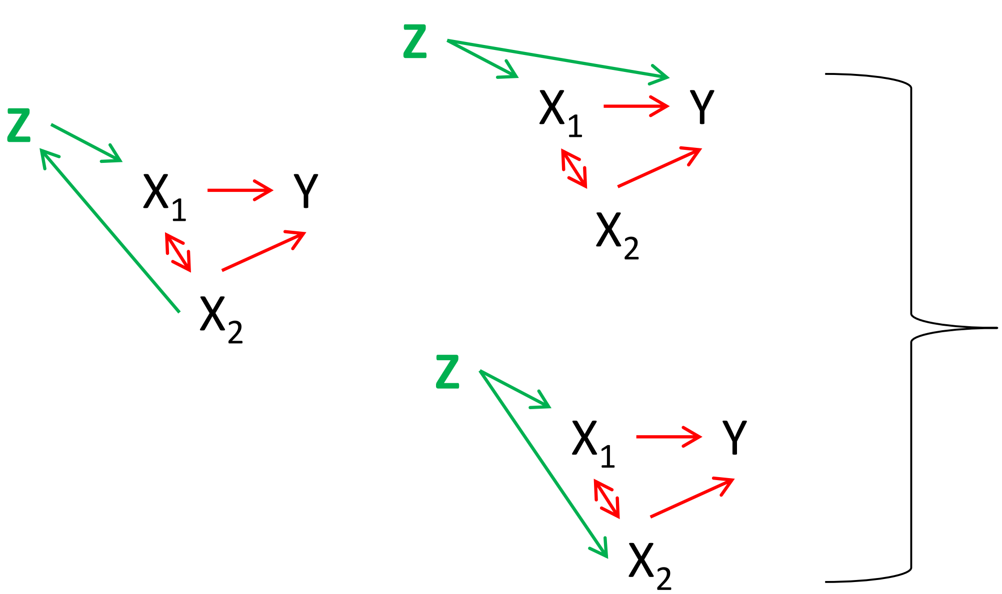
SO HOW DO WE FIX THIS?

INSTRUMENTAL VARIABLES EXPLOIT EXOGENEITY:



Exogeneous

Exogenous: Variable that is correlated with the policy variable but NOT with the omitted variable.



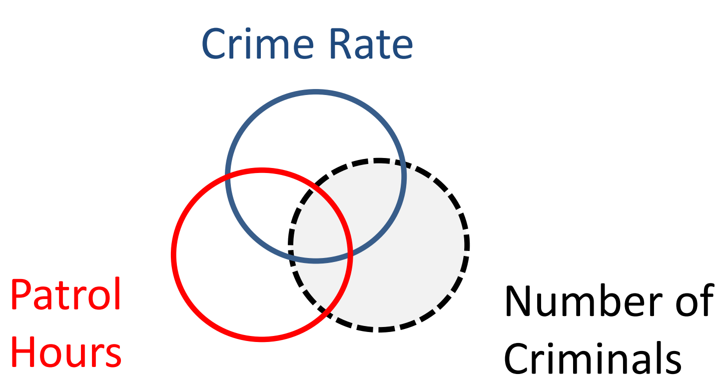
NOT Exogeneous

EXAMPLE:

Where have all the criminals gone?

Steven Levitte makes the claim that increasing policing only reduces crime slightly.

Why will patrol hours and the number of criminals be correlated? Will there be a positive or negative relationship?

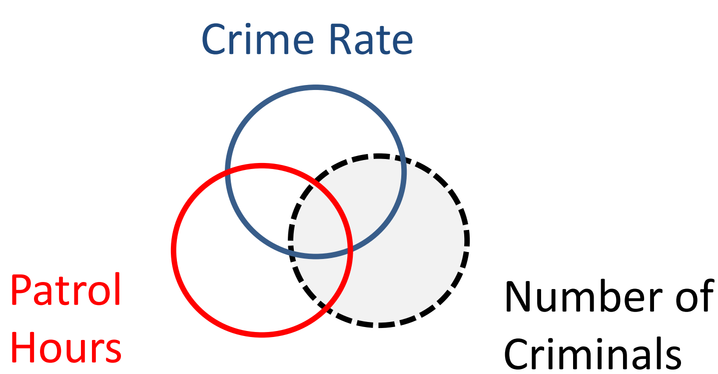


EXAMPLE:

Where have all the criminals gone?

Steven Levitte makes the claim that increasing policing only reduces crime slightly. The problem with the estimation:

"Cities with high crime rates, therefore, may tend to have large police forces, even if police reduce crime. Detroit has twice as many police officers per capita as Omaha, and a violent crime rate over four times as high, but it would be a mistake to attribute the differences in crime rates to the presence of the police. Similarly, within a particular city, if more police are hired when crime is increasing, a positive correlation between police and crime can emerge, even if police reduce crime."

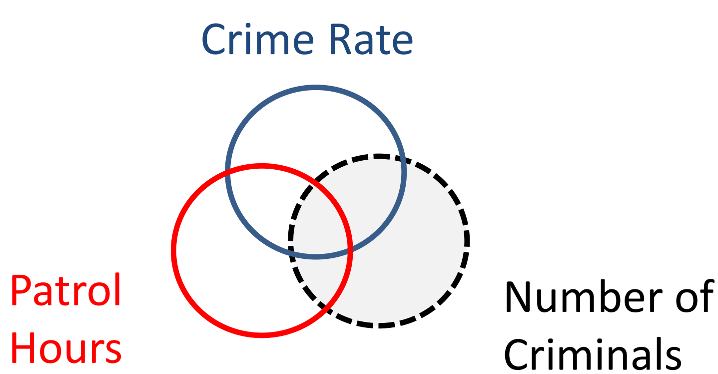


EXAMPLE:

Where have all the criminals gone?

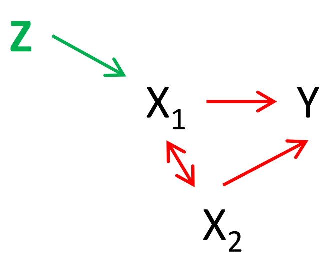
Steven Levitte makes the claim that increasing policing only reduces crime slightly.

Where would we find a measure of the number of criminals?

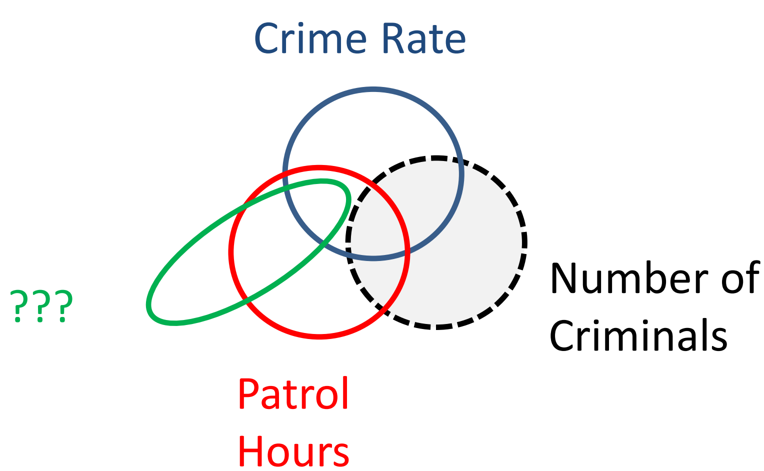


THE FIX:

Where have all the criminals gone?



We need to find an exogenous variable correlated to policing intensity but uncorrelated with crime rates

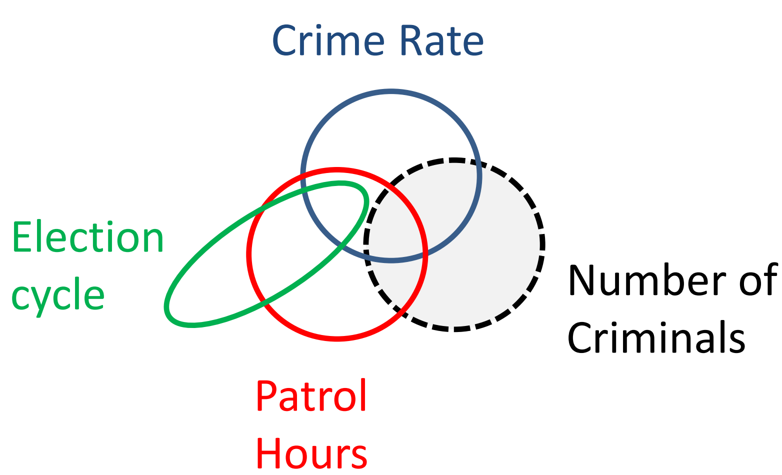


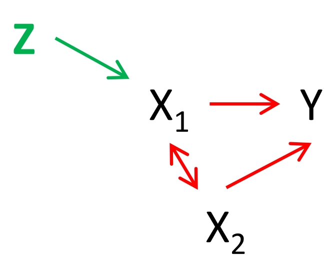
THE FIX:

Where have all the criminals gone?

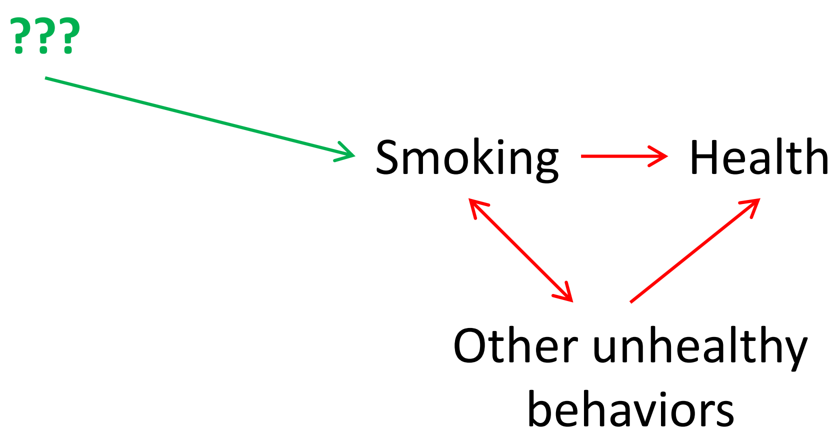
“The primary innovation of the paper is the approach used to break the simultaneity between police and crime. In order to identify the effect of police on crime, a variable is required that affects the size of the police force, but does not belong directly in the crime "production function." The instrument employed in this paper is the timing of mayoral and gubernatorial elections.”

Levitt, 1997

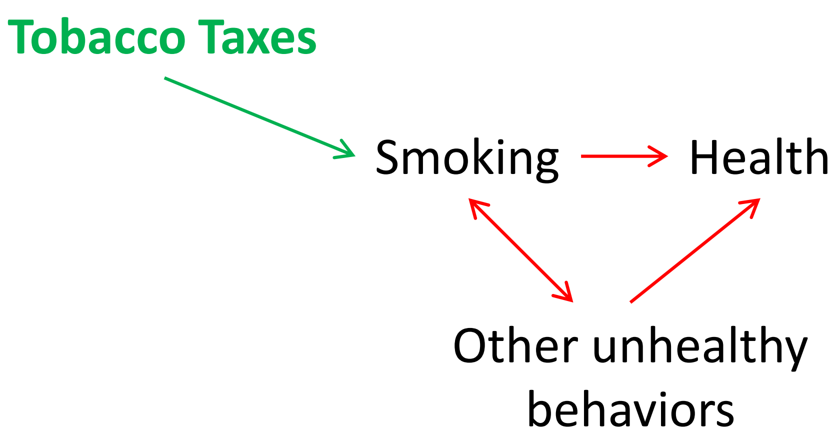




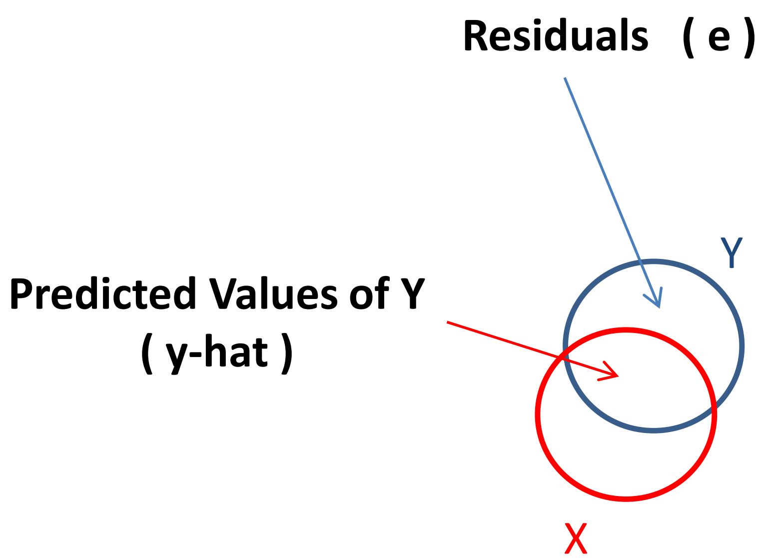
ANOTHER EXAMPLE:



ANOTHER EXAMPLE:

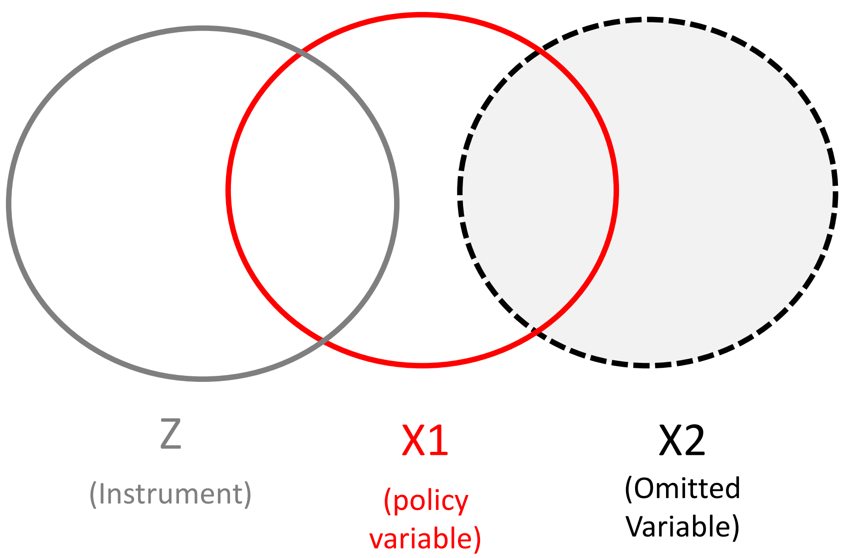


HOW DOES IT WORK?



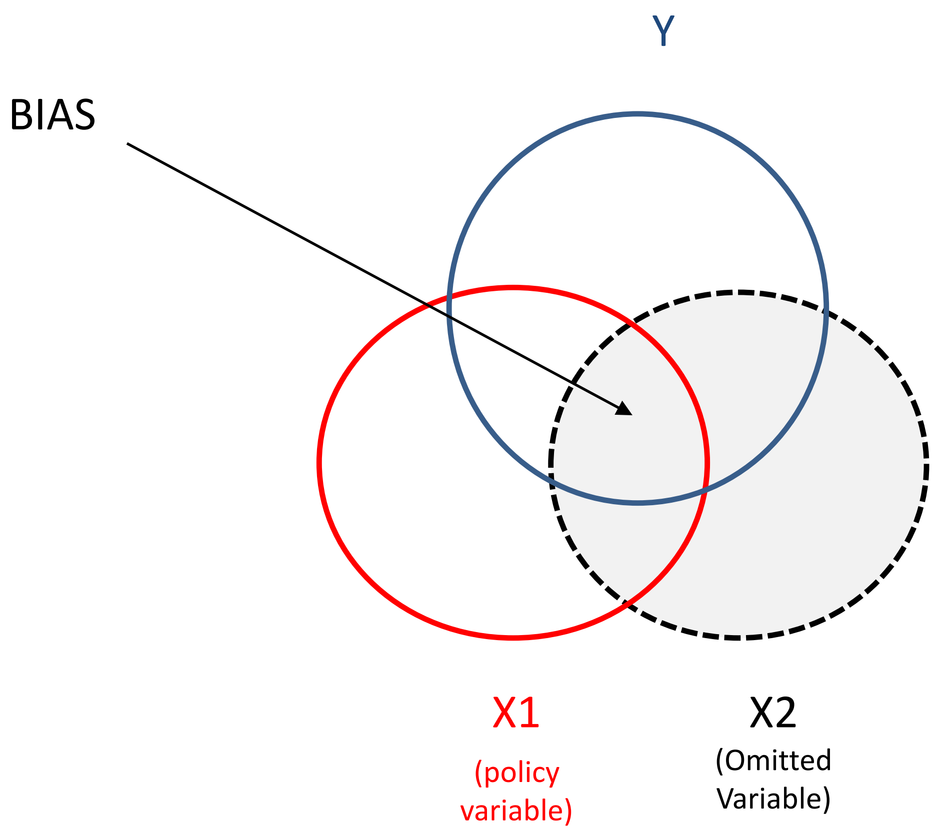
Recall, Y is partitioned into explained and unexplained portions

EXOGENOUS CRITERIA:



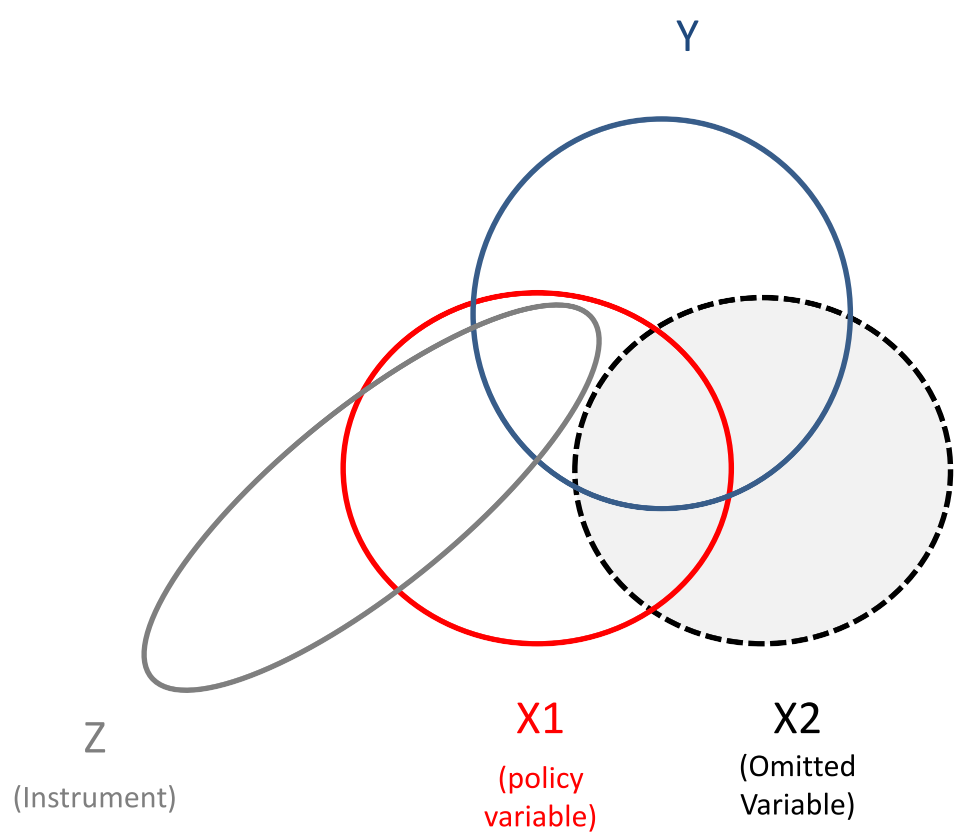
Z is uncorrelated with X2

EXOGENOUS CRITERIA:



X2 is a problematic omitted variable

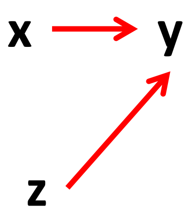
EXOGENOUS CRITERIA:

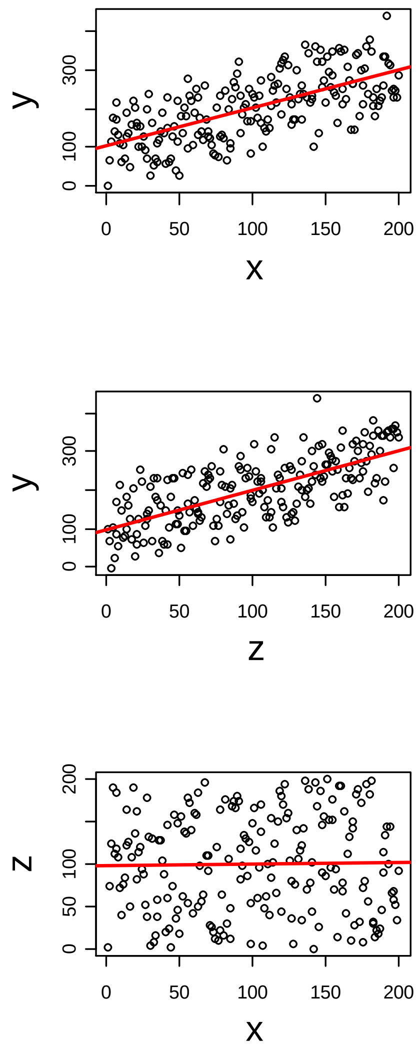


Z is correlated with Y, but only THROUGH X1

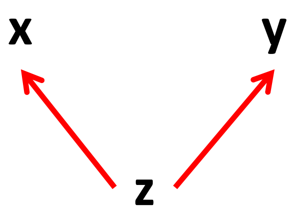
RECALL:

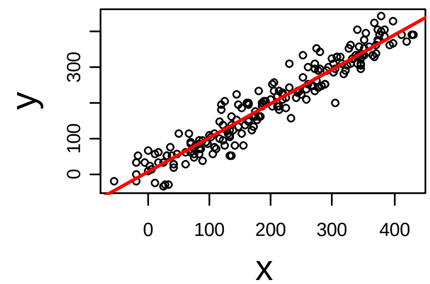
Example #1

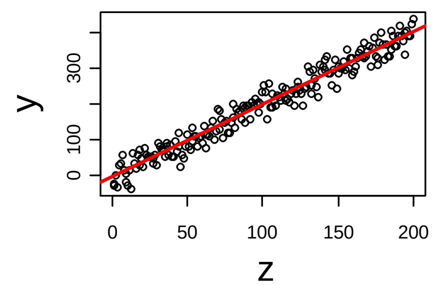

x
y
z

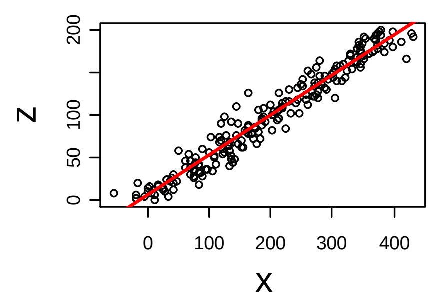

y
0
100
300
0
50
100
150
200
x
y
0
100
300
0
50
100
150
200
z
z
0
50
100
200
0
50
100
150
200
x

Example #2

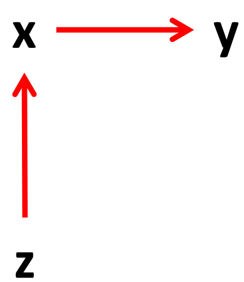

x
y
z

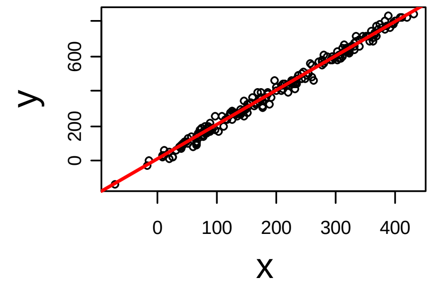

y
0
100
300
0
100
200
300
400
x

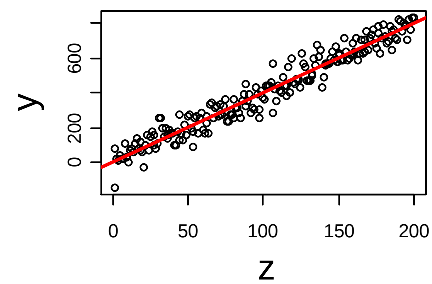

y
0
100
300
0
50
100
150
200
z

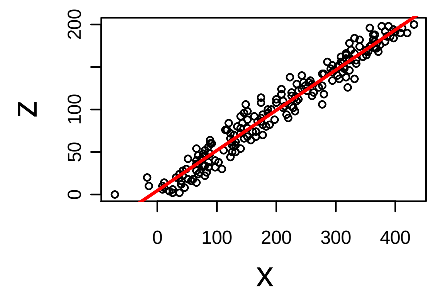

z
0
50
100
200
0
100
200
300
400
x

Example #3

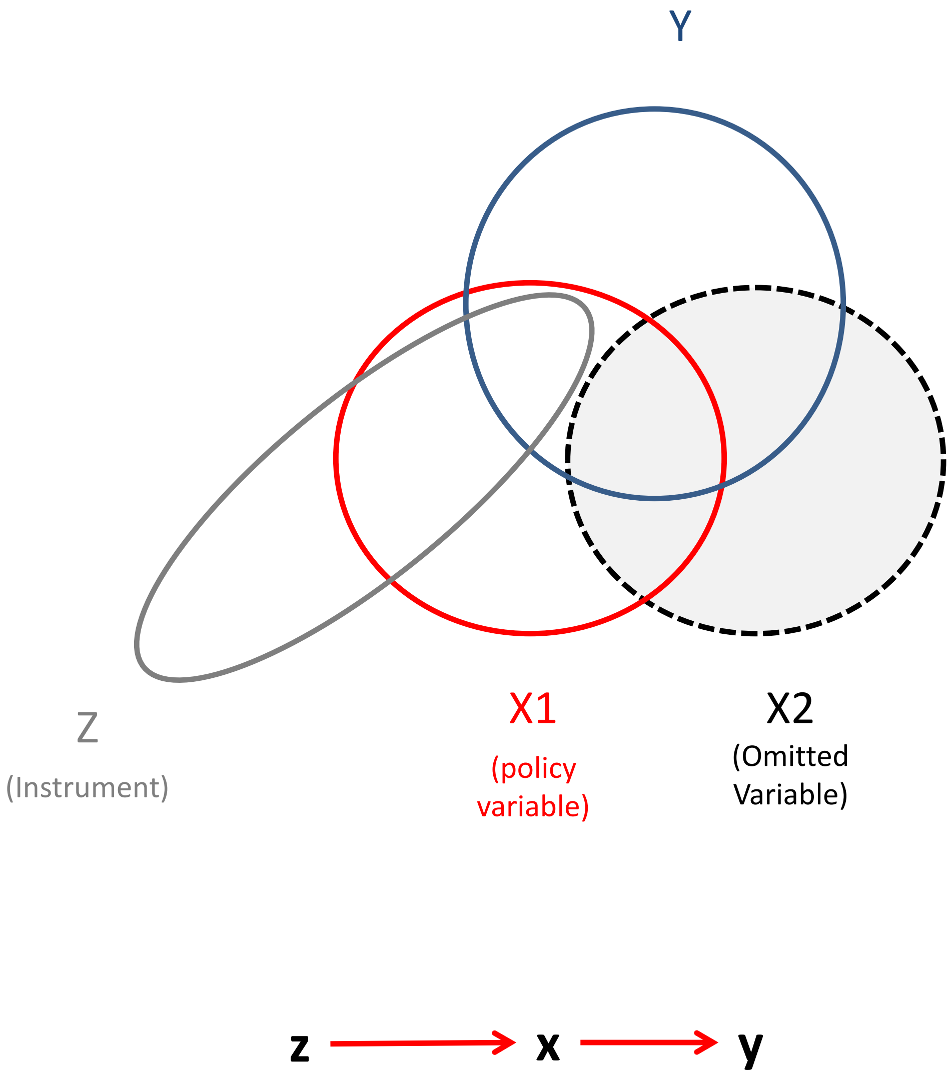

x
y
z


y
0
200
600
0
100
200
300
400
x


y
0
200
600
0
50
100
150
200
z


z
0
50
100
200
0
100
200
300
400
x

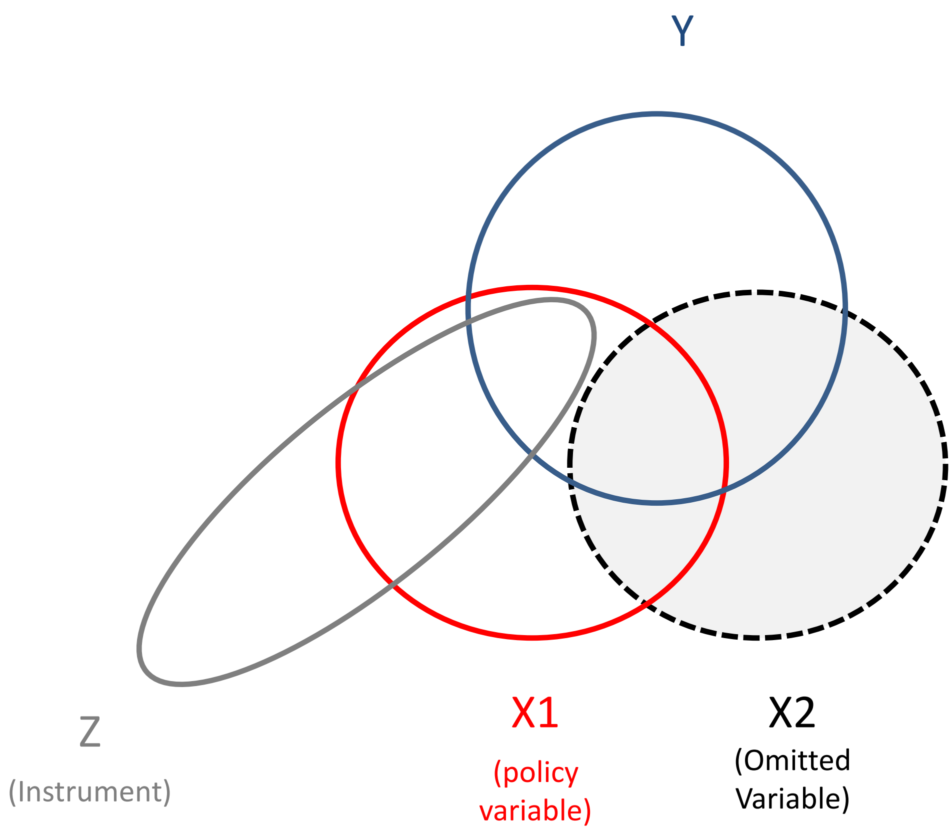
EXOGENOUS CRITERIA:



This is a natural property of correlation structures.

We exploit this property to isolate the uncontaminated portion of X1!

EXOGENOUS CRITERIA:



$$z \rightarrow x \rightarrow y$$

THIS IS HOW INSTRUMENTAL VARIABLES WORK:

$$True\ Model: Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon, \quad Corr(X_1, X_2) \neq 0$$

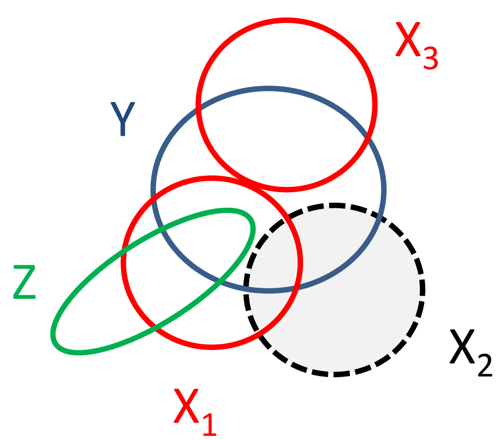
$$Biased\ Model: Y = b_0 + b_1 X_1 + b_3 X_3 + e$$

$$Two-Stage\ Least\ Squares$$

$$Stage 1: X_1 = \alpha_0 + \alpha_1 Z + \alpha_2 X_3 + e$$

$$Stage 2: Y = \gamma_0 + \gamma_1 \hat{X}_1 + \gamma_3 X_3 + e$$

$$Now: \gamma_1 \approx \beta_1$$



WE USE INSTRUMENTAL VARIABLES WHEN:

1. **We know that we have an omitted variable problem**

2. **We don’t have a way to include the omitted variable**

3. **We have an exogenous variable**

- **But how do we know that we have an exogenous variable???**

- Instrumental variables are justified by assumption, i.e. a good story. There is a test you can run if you have multiple instrumental variables, but often we rely on the story.

PROCEDURE FOR IV APPROACH:

1) X2 is the omitted variable – if it is omitted it means we don't have it so we can't include it in the model.

2) X1 is the policy variable.

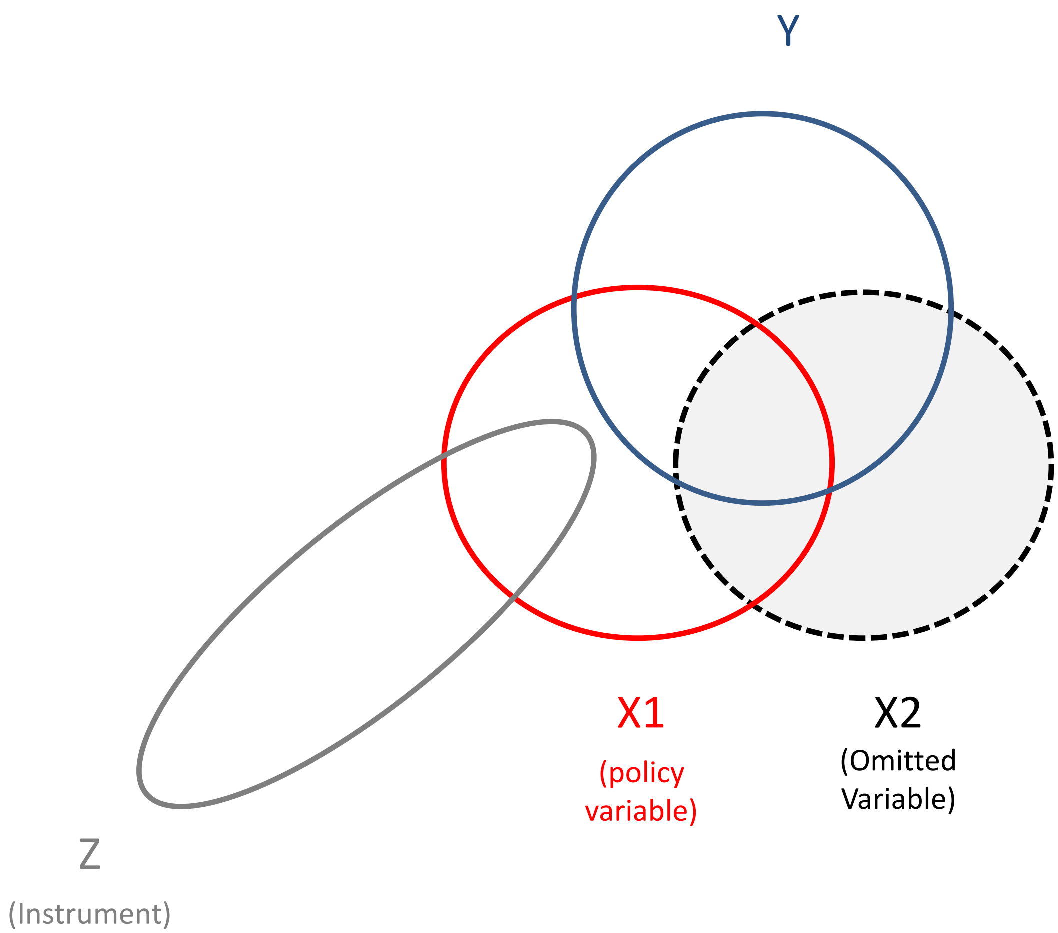
3) Z is the exogeneous variable.

4) X3, X4, X5, etc. are all included in the model, but are not affected by the omitted variable in the same way that the policy variable is.

5) In the first stage, we predict model X1 with the exogeneous variable and the other variables.

6) In the second stage, we ONLY include predicted values of X1, because these predicted values will be independent of the influences of the omitted variable since Z and X2 are uncorrelated.

EXOGENOUS CRITERIA:



This case would result from a weak correlation between Z and X1.

Z will not work as an instrument here.

THE THEORY:

- Using the instrument, we surgically remove the portion of X_1 that is "contaminated" by the omitted variable.

- Instrumental variables rely on the assumption of exogeneity, but this assumption can be tested statistically if there is more than one exogenous variable.

- The models needs at least on exogenous variable for every independent variable that is endogeneous.

- We need good predicted values for $X_{1,}$ so the R-square of the regression must be high (an F-stat of over 10 is the rule of thumb).