Web-Mining Agents

Dr. Özgür Özçep

Universität zu Lübeck
Institut für Informationssysteme

Structural Causal Models

Slides prepared by Özgür Özçep

Part III: Causality in Linear SCMs and Instrumental Variables

UNIVERSITÄT ZU LÜBECK
INSTITUT FÜR INFORMATIONSSYSTEME

Literature

- J.Pearl, M. Glymour, N. P. Jewell: Causal inference in statistics – A primer, Wiley, 2016.
  (Main Reference)
- J. Pearl: Causality, CUP, 2000.
- B. Chen & Pearl: Graphical Tools for Linear Structural Equation Modeling, Technical Report R-432, July 2015

Causal Inference in Linear SCMs

- All techniques and notions developed so far applicable for any SCM
- Of importance are linear SCMs
  - Equations of form $Y = a_0 + a_1X_1 + a_2X_2 + \ldots a_nX_n$
  - In focus of traditional causal analysis (in economics)

- Assumption for the following
  - All variables depending linearly on others (if at all)
  - Error variables (exogenous variables) have Gaussian/Normal distribution

Want to learn something about Gauss?

Why Gaussian?

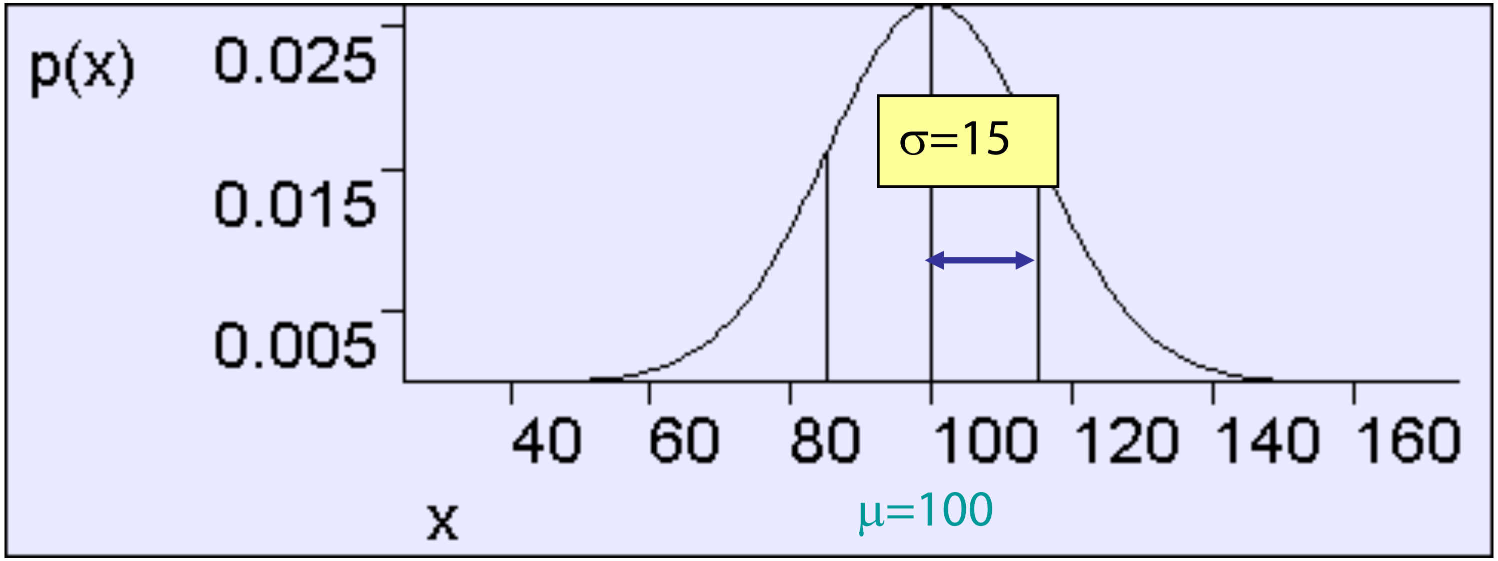
- Andrew Moore: “Gaussians are as natural as Orange Juice and Sunshine”

(http://www.cs.cmu.edu/~awm/tutorials)

(Used in the following slides on Gaussians)

- Proves useful to model RVs that are combinations of many (non)-measured influences
- Makes life easy because
    1. Efficient representation
    2. Substitute probabilities by expectations
    3. Linearity of expectations
    4. Invariance of regression coefficients

General Gaussian



$$E[X] = \mu$$

$$\mathrm{Var}[X] = \sigma^2$$

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Also known as the normal distribution or Bell-shaped curve

Shorthand: We say X ~ N(μ,σ^2) to mean "X is distributed as a Gaussian with parameters μ and σ^2".

In the above figure, X ~ N(100,15^2)

ÖÖ: So need only specify μ,σ^2

Bivariate Gaussians

Write r.v. $\mathbf{X} = \begin{pmatrix} X \\ Y \end{pmatrix}$ Then define $X \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to mean

$$p(\mathbf{x}) = \frac{1}{2\pi \, \|\boldsymbol{\Sigma}\|^{1/2}} \exp\left(-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \, \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

Where the Gaussian's parameters are…

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2{}_x & \sigma_{xy} \\ \sigma_{xy} & \sigma^2{}_y \end{pmatrix}$$

ÖÖ: Covariance matrix in 2 dimensions
$\sigma_{XY} = E[(X-E(X))(Y-E(Y))]$

Where we insist that Σ is symmetric non-negative definite

It turns out that E[X] = μ and Cov[X] = Σ. (Note that this is a resulting property of Gaussians, not a definition)*

*This note rates 7.4 on the pedanticness scale

ÖÖ: So need only specify 5= 2*2 + 2(2-1)/2 paramters

Multivariate Gaussians

ÖÖ: So, it is sufficient to consider pairwise correlation
Of Xi, Xj (next to their expectations and variances)
2*m + m(m-1)/2 => efficient representation of joint
distribution of $X_1 \ldots X_m$

Write r.v. $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_m \end{pmatrix}$

Then define $X \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to mean

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{m/2} \| \boldsymbol{\Sigma} \|^{1/2}} \exp\left(-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

Where the Gaussian's parameters have…

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2{}_1 & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{12} & \sigma^2{}_2 & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1m} & \sigma_{2m} & \cdots & \sigma^2{}_m \end{pmatrix}$$

Where we insist that Σ is symmetric non-negative definite

Again, E[X] = μ and Cov[X] = Σ. (Note that this is a resulting property of Gaussians, not a definition)

Why Gaussian?

- Andrew Moore: “Gaussians are as natural as Orange Juice and Sunshine”

(http://www.cs.cmu.edu/~awm/tutorials)

(Used in the following slides on Gaussians)

- Proves useful to model RVs that are combinations of many (non)-measured influences
- Makes life easy because
  1. Efficient representation
  2. Substitute probabilities by expectations

Substitute Probabilities by Expectations

- $P(X)$ becomes $E[X]$
- $P(Y|X)$ becomes $E[Y|X]$

(Conditional expectation defined as expected

$$E[Y|X=x] = \sum_y y\, P(Y=y|X=x) \qquad)$$

→ Can use regression to determine causal relations

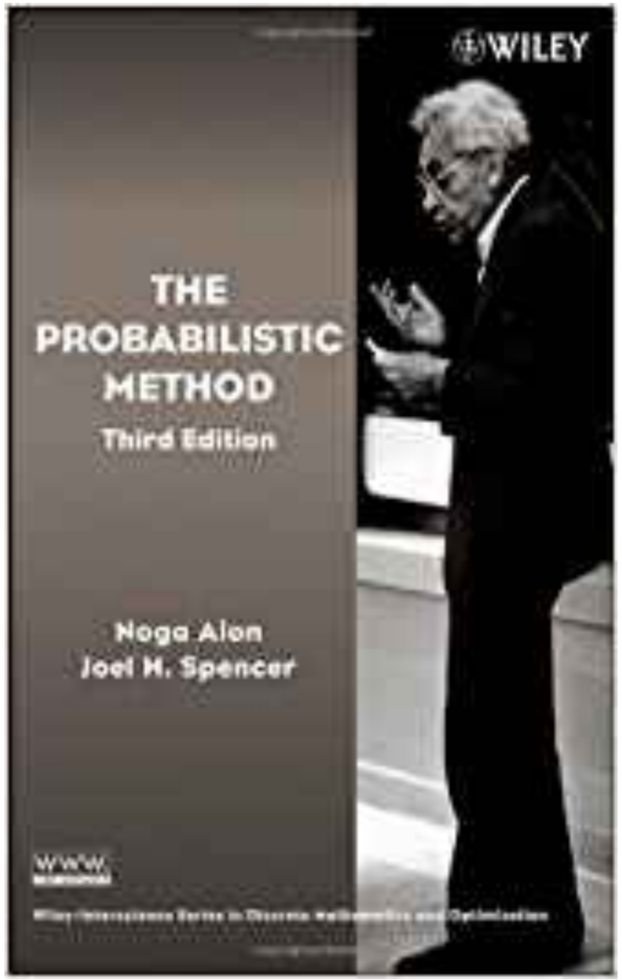
- $E[Y|X]$ defines a function $Y = f(X)$
- By regression we circumvent the problem of calculating the probabilities required for $E[Y|X]$

So, we will be guessing the deep/hidden structure (linear SCMs equations) as far as needed for our tasks – instead of working on level of probabilities

But remember also other direction

- Use probabibility to infer „crisp properties“

- Toy example:
  - If you know that the expected value for a RV is 0.5 (for RV in [0,1])
  - then you know (for sure) that there must be instances with value ≥ 0.5.



Why Gaussian?

- Andrew Moore: "Gaussians are as natural as Orange Juice and Sunshine"

(http://www.cs.cmu.edu/~awm/tutorials)

(Used in the following slides on Gaussians)

- Proves useful to model RVs that are combinations of many (non)-measured influences
- Makes life easy because
  1. Efficient representation
  2. Substitute probabilities by expectations
  3. Linearity of expectations
  4. Invariance of regression coefficients

Linearity of Expectations

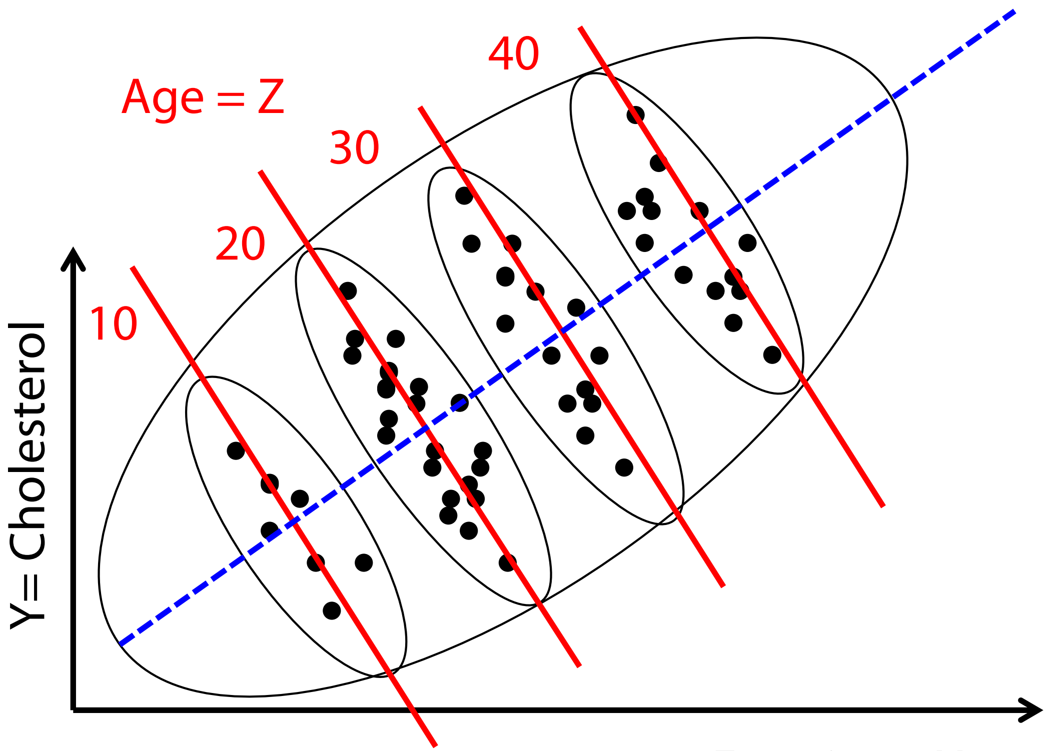
- Expectations can be written as linear combinations
  - $E[Y|X_1=x_1,X_2=x_2, \ldots, X_n=x_n] = r_0 + r_1x_1 + \ldots + r_nx_n$
  - Each of the slopes r_i are partial regression coefficients
  - Example and Notation

    $r_i = R_{Y X_i . X_1 \ldots X_{i-1}, X_{i+1}, \ldots X_n}$

    $=$ slope of Y on X_i when fixing all other X_j $(j \neq i)$
  - r_i does not depend on the values of the X_i but only on which set of X_is (the set of regressors) was chosen
  - This independency is also part of a continuous version of the Simpson's paradox (next slides)

Slope Constancy

- Measure weakly exercise and cholesterol in different age groups



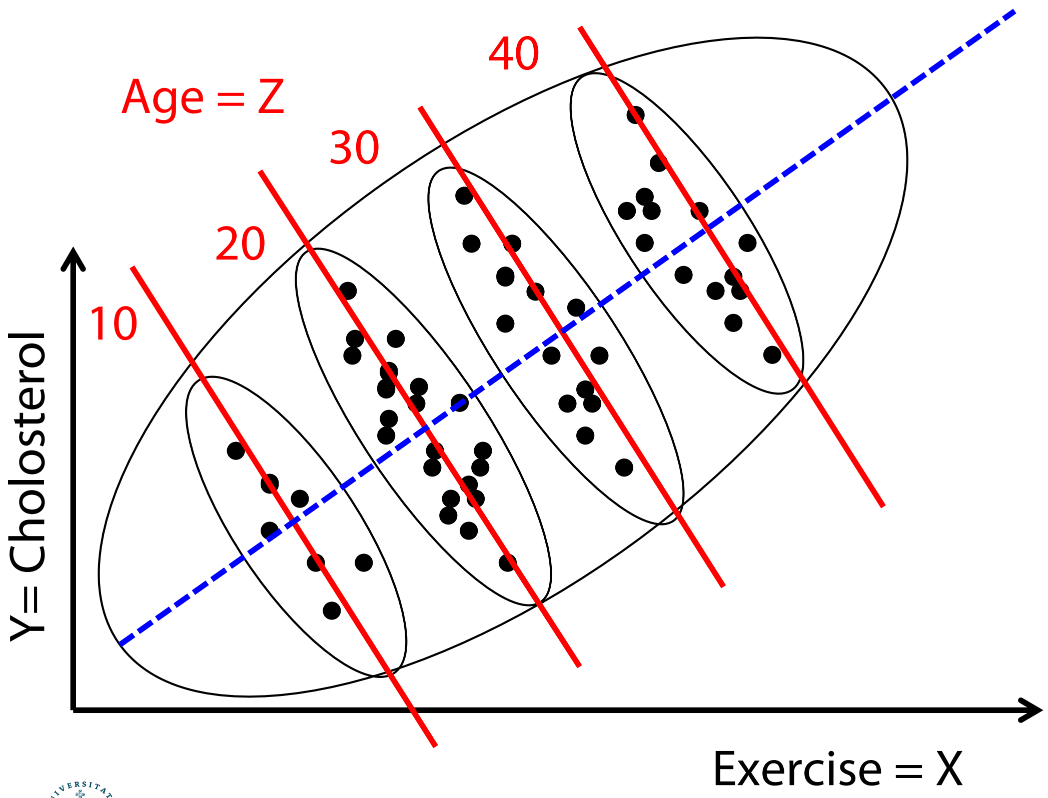
- $Y = r_0 + r_1 X + r_2 Z$
- $r_1 = R_{YX.Z} < 0$
- Z-fixed slope for Y,X independent of Z (and negative)

- Ignoring Z (regressing Y w.r.t X only) leads to combined positive slope R_{YX}

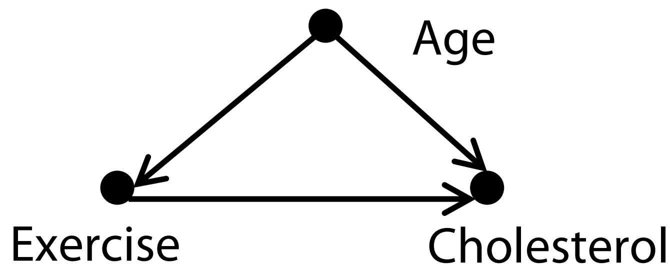
→ Simpson's paradox

Resolving the Paradox

- Measure weakly exercise and cholesterol in different age groups



- Age a cofounder of Exercise and Cholosterol
- Need to condition on Age=Z to find correct P(Y|do(X))



Regression Coefficients and Covariance

- Usually one finds (partial) regression coefficients by sampling
- But there exist formulae expressing connections to statistical measures such as covariance

- $\sigma_{XY} = E[(X-E[X])(Y-E[Y])]$ (covariance of X and Y)
- $\rho_{XY} = \sigma_{XY}/(\sigma_X \sigma_Y)$ (Correlation)
- Note: $\sigma_{XY} = 0 = \rho_{XY}$ iff X and Y are (linearly) independent

Theorem (Orthogonality principle)

If $Y = r_0 + r_1X_1 + \ldots + r_kX_k + \varepsilon$

then the best (least-square error minimizing) coefficients r_i (for any distributions X_i) result when $\sigma_{\varepsilon Xi} = 0$ for all $1 \leq i \leq k$

Regression Coefficients and Covariance

- Assume w.l.o.g. $E[\varepsilon] = 0$
- $Y = r_0 + r_1X + \varepsilon$ (*)
- $E[Y] = r_0 + r_1E[X]$ (by applying E)
- $XY = Xr_0 + r_1X^2 + X\varepsilon$ (by multiyplying (*) with X)
- $E[XY] = r_0E[X] + r_1E[X^2] + E[X\varepsilon]$ (by applying E)
- $E[X\varepsilon] = 0$ (by orthogonality)
- Solving for r_0 and r_1
  - $r_0 = E[Y] – E[X](\sigma_{XY}/\sigma_{XX})$
  - $r_1 = \sigma_{XY}/\sigma_{XX}$

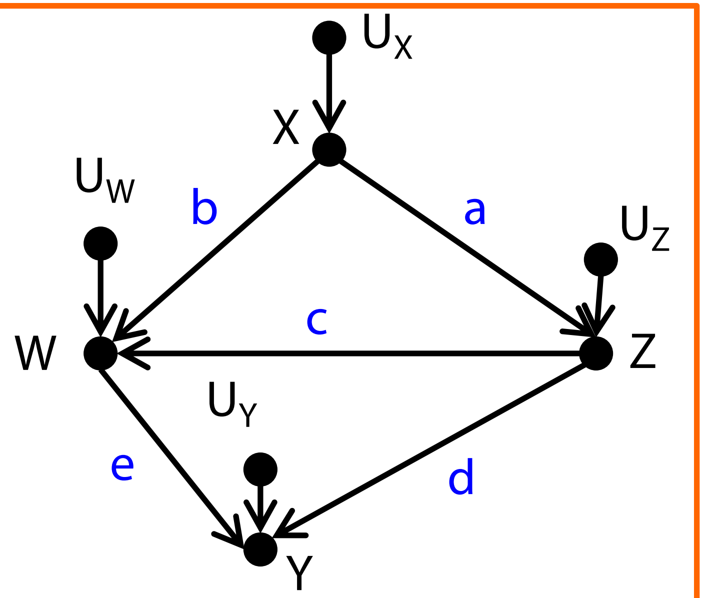
Similar derivations for multiple regression

Path Coefficients (Example)

Example

- Linear SCM
  - $X = U_X$
  - $Z = aX + U_Z$
  - $W = bX + cZ + U_W$
  - $Y = dZ + eW + U_Y$
- Graph of SCM as usual
- But now additional information by edge labels:

  Path Coefficients



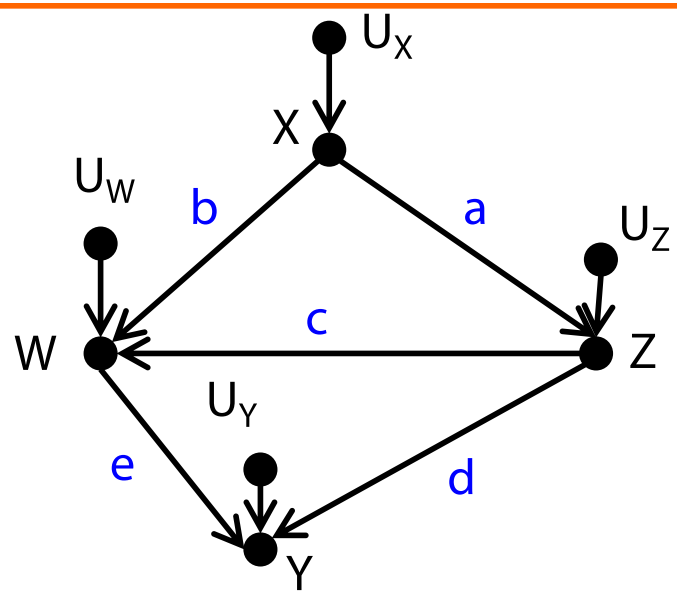
Linearity assumption makes association of coefficient to edge a well-formed operation

Path Coefficients (Example)

Example

- Linear SCM
  - $X = U_X$
  - $Z = aX + U_Z$
  - $W = bX + cZ + U_W$
  - $Y = dZ + eW + U_Y$
- Graph of SCM as usual
- But now additional information by edge labels:

  Path Coefficients

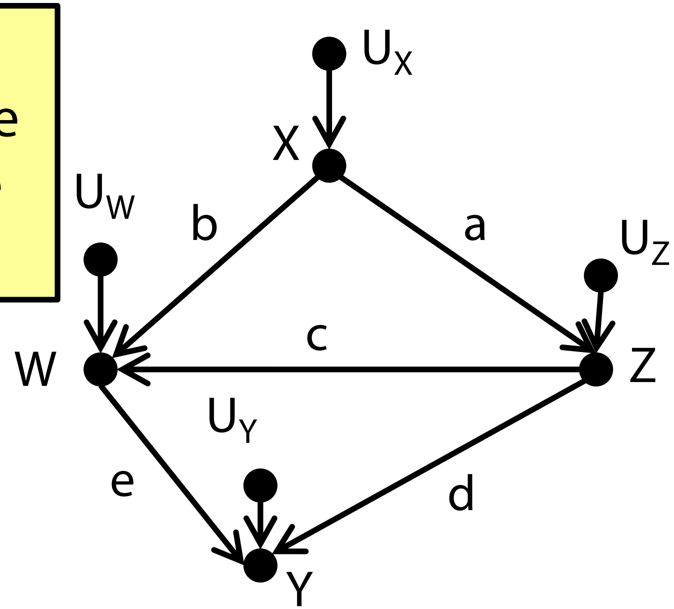


Warning from the beginning:
Path coefficients (causal) ≠ regression coefficients (descriptive)

Path Coefficients (Semantics)

- Linear SCM
  - $X = U_X$
  - $Z = aX + U_Z$
  - $W = bX + cZ + U_W$
  - $Y = dZ + eW + U_Y$

Note: CDE does not depend on the exact change of Z but only its change rate Z=+1



- Q: What is the semantics of the path coefficients on edge Z-Y?
- A: Direct effect CDE on Y of change Z=+1

$$CDE = E[Y|do(Z=z+1), do(W=w)] - E[Y|do(Z = z), do(W=w)]$$

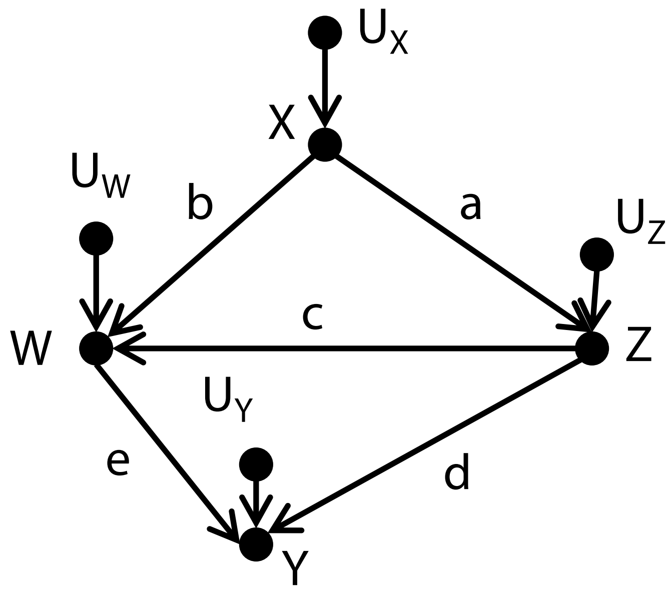
$$= d(z+1) + ew + E[U_Y] - (dz + ew + E[U_Y])$$

$$= d = \text{label on Z-Y edge}$$

We used the linearity of E
$E[aX + bY] = aE[X] + bE[Y]$

Total Effect in Linear Systems (Example)

- Linear SCM
  - $X = U_X$
  - $Z = aX + U_Z$
  - $W = bX + cZ + U_W$
  - $Y = dZ + eW + U_Y$



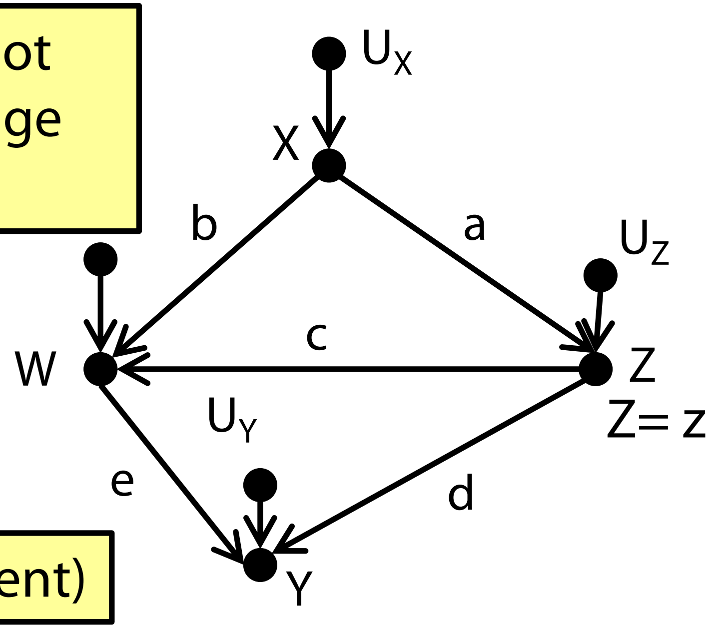
Total effect = ACE

- Q: What is the total effect of Z on Y?
- A: Sum of coefficient products over each directed Z-Y path
  - Directed path 1: Z-d->Y; product = d
  - Directed path 2: Z-c->W-e->Y; product =ec
  - Total effect = d + ec

Total Effect in Linear Systems (Intuition)

- Linear SCM
  - $X = U_X$
  - $Z = aX + U_Z$
  - $W = bX + cZ + U_W$
  - $Y = dZ + eW + U_Y$

Note 2: Total effect does not depend on the exact change of Z but only its rate Z=+1



Note 3: Holds for any linear SCM (U_is may be dependent)

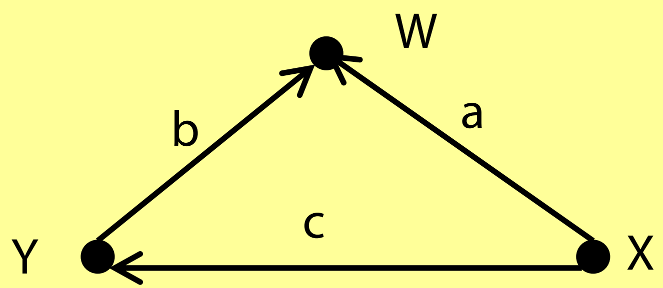
- Q: What is the total effect of Z on Y?
- A: Sum of coefficient products over each directed Z-Y path
  - Total effect τ: Intervene on Z and express Y by Z
  - $Y = dZ + eW + U_Y = dZ + e(bX + cZ + U_W) + U_Y$

$$= (d+ec)Z + ebX + U_Y + eU_W = \tau Z + U$$

Note 1: X, U_Y, U_W do not depend on Z

Note 4

- We followed (Bollen 1989) and summed over directed paths for calculating total effect in linear SCMs
- In book of Pearl, Glymour & Jewell (p.82-83) summation over non-backdoor paths
  - Not clear to me (maybe wrongly applied Wright‘s rule)
  - Consider SCM
    - W = bY + aX
    - Y = cX
    - ACE = c (and not c + b*a)



K. Bollen: Structural Equations with latent variables. New York, 1989.

Addendum and Historical Note to Note 4

- Earliest use of graphs in causal analysis in (Wright 1920)
- Wright path tracing for calculating covariances in linear SCMs

  $\sigma_{XY} = \sum_p \text{product}(p)$

  - where all p are X-Y paths not containing a collider and
  - product(p) = product of all structural coefficients and covariances of error terms

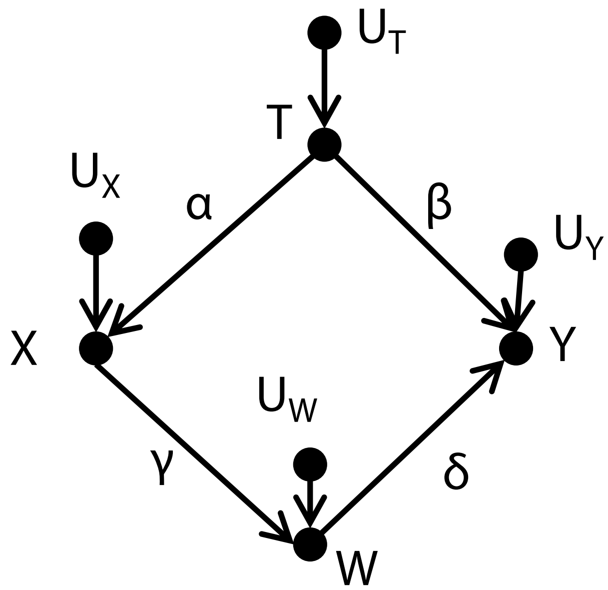
S. Wright. Correlation and Causation.
Journal of Agricultural Research 20, 557-585, 1921.

Identifying Structural Coefficients

- What if path coefficients are not known apriori or are not testable?
- One has to identify only those relevant for the specific task, e.g., total effect of X to Y or direct effect of Z on X
- For those required for the task one can use linear regression on the data
    1. Identify relevant variables for linear regression
    2. Identify within linear equation coefficients for the specific task

Direct Effect in Incomplete Linear Systems

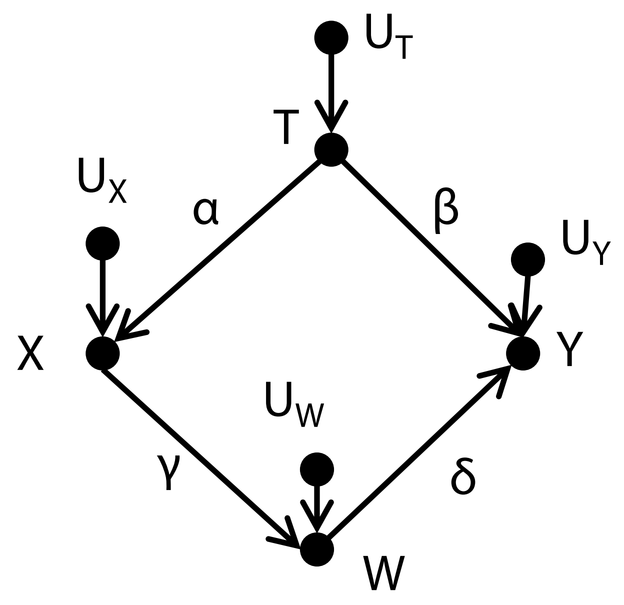
- Q: Direct effect of X on Y?
- A: Here, direct effect = 0
  - There is no edge from X to Y
  - Which amounts to path coefficient for X-Y edge = 0



Total effect in Incomplete Linear Systems

- Q: Total effect of X on Y?
- Now path coefficients not necessarily known (greek letters)
- Recall: With backdoor criterion identify variable set Z to adjust for

$GCE = P(y|do(x)) = \sum_z P(y \mid x,z)P(z)$



- Use backdoor to identify variables to regress for
- Here Z = {T}, so do linear regression on X,T:
  - $Y(X,T) = r_X X + r_T T + \varepsilon$
  - r_X = total effect of X on Y

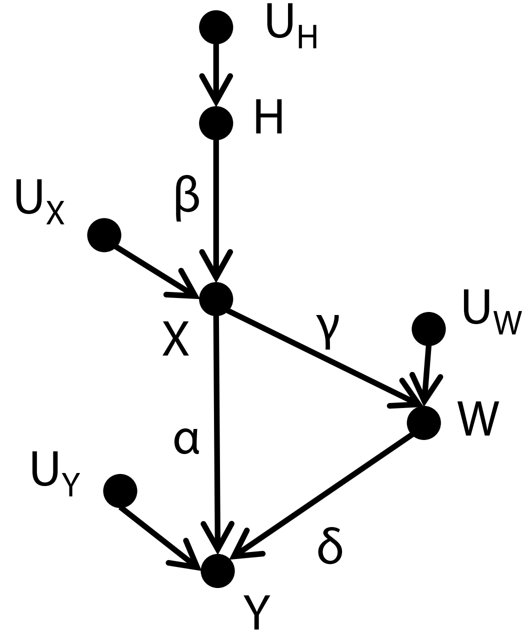
Wake-Up: Why not also regress for W?

- linear regression equation ≠ structural equation
- Regression coefficients handmade
- Path coefficients nature-made

Direct Effect in Incomplete Linear Systems

- Q: Direct effect of X on Y?

- A: In general find blocking variables Z for
  - X-Y backdoor paths or, more generally,
  - Indirect X-Y paths



- This can be achieved as follows
  - G_α = Graph G without edge X –α->Y
  - Z = variables d-separating X and Y

Here: Z = {W}

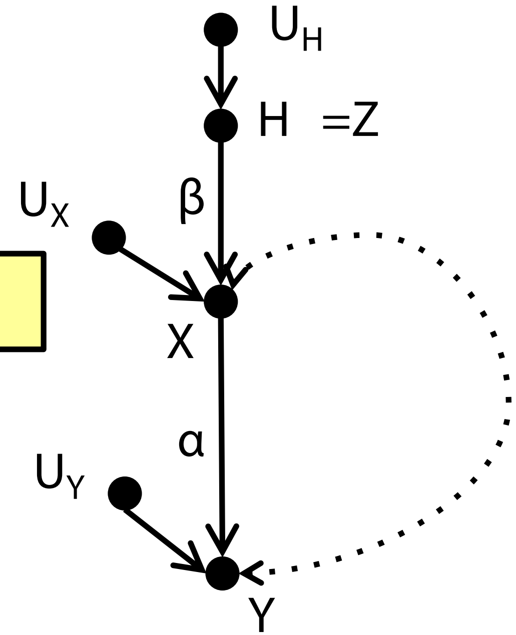
- $Y = r_X X + r_Z Z + \varepsilon$

Here: $Y = r_X X + r_W W + \varepsilon$

Direct effect of X on $Y = r_X =: \alpha$

Direct Effect in Incomplete Linear Systems

- Q: What if there are no d-separating Z?
- A:
  1. Find instrumental variables Z
     1. Z is d-connected to X in G_α and
     2. Z is d-separated from Y in G_α
  2. Regress $Y = r_1 Z + \varepsilon$
  3. Regress $X = r_2 Z + \varepsilon$
  4. $r_1/r_2 = \beta_{YZ}/\ \beta_{XZ} =: \alpha$ = direct effect of X on Y



Here: $Z = H$

Dashed arrow denotes an unobservable confounder

This is because: Since $Z = H$ emits no backdoors

- $r_2 = \beta$ and
- r_1 = total effect of Z on Y = $\beta\alpha$

Instrumental Variables (IVs)

- Usage of IVs to trace causal effects starts already in 1925 (econometrics)

  Wright. Corn and Hog correlations, Tech. Rep. 1300, US Department of Agriculture, 1925.

- Standard definitions in econometrics defined IVs w.r.t. single equation not w.r.t. parameter

Definition (classically according to economist's)

For an equation

$$Y = \alpha_1 X_1 + \ldots + \alpha_k X_k + U_Y \quad (*)$$

Z is instrumtenal variable for equation (*) iff

- Z is correlated with $X = \{X_1, \ldots X_k\}$ and
- Z is not correlated with U_Y

What‘s in a definition?

- The early economist‘s definition not (!) equivalent with our official definition
  - General question: What‘s a good definition*)?
  - Main problem with classical equation: too global
    - Full equation may not be identifiable, though some parameters are.

- The new definition is an example of a general interesting phenomenon
  - Many simplifications (clarifications/disambiguations) of (IV) research in econometrics by considering associated graph structure for SCM

*) An Addendum for Teatime on a Sunday ...

- More about „good definitions":

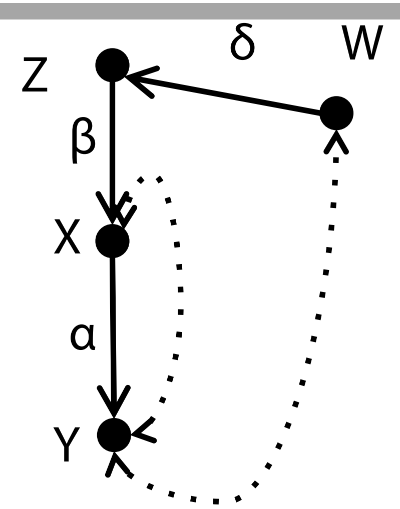
N. D. Belnap. On rigorous definitions. Philosophical Studies: An International Journal of Philosophy in the Analytic Tradition, 72(2,3):115–146, instrumental 1993.

- Rough summary: A definition is good (formally correct) if it fulfills
  1. eliminability (defined symbol can be replaced via old symbols)
  2. Non-creativity (no new sentences derivable in old language)

- The kind of goodness mentioned on slide before not (intended to be) captured by Belnap's explication.

Conditional IVs

- Z no IV anymore for α, because
  - Z not d-separated from Y
- But conditioning on W helps



C. Brito & J.Pearl: Generalized instrumental variables. In *Uncertainty in Artificial Intelligence, Proceedings of the Eighteenth Conference*, 85–93, 2002.

Definition (Brito & Pearl, 02) A variable Z is a conditional instrumental variable given set W for coefficient α (from X to Y) iff

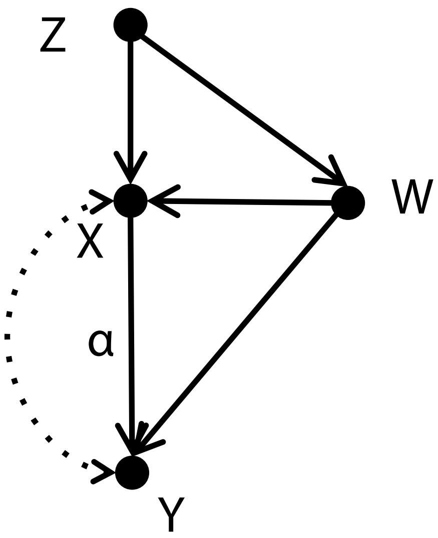
- Set of descendants of Y not intersecting with W
- W d-separates Z from Y in G_α
- W does not d-separate Z from X in G_α

If conditions fulfilled, then $\alpha = \beta_{YZ.W} / \beta_{XZ.W}$

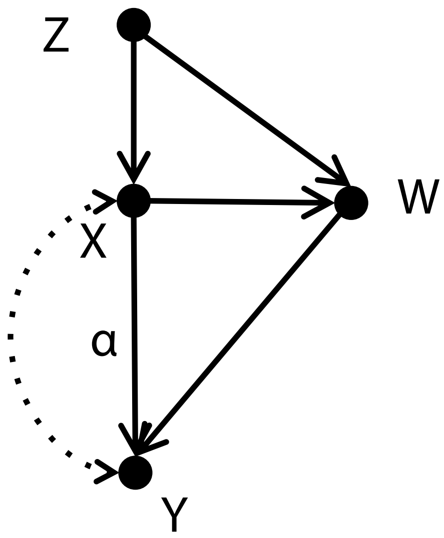
Conditional IVs (Examples)

Z instrument for α given W?

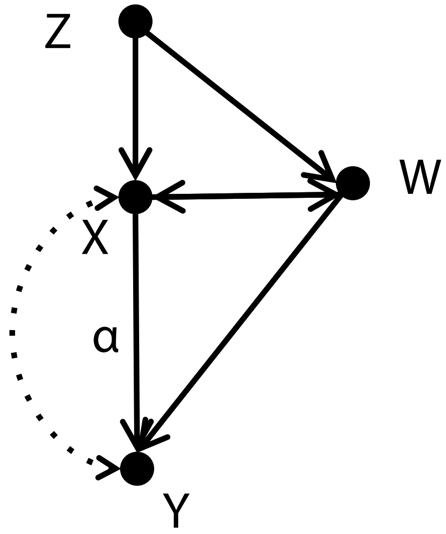
Definition Z is a conditional IV given set W for α iff
- Set of descendants of Y are not intersecting with W
- W d-separates Z from Y in G_α
- W does not d-separate Z from X in G_α



yes

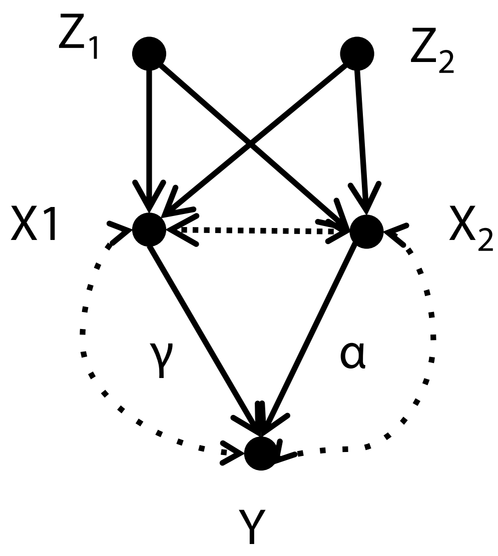


no



yes

Sets of IVs



- Sometimes need sets of instrumental variables
- Neither Z_1 nor Z_2 (on their own) are instrumental variables (for the identification of α or γ)
- Using both helps
  - Definition not trivial due to possible intersections of paths
    - Z_i -> .. -> X_i->Y and Z_j -> .. -> X_j->Y
- Using Wright's path tracing and solving for γ and α

$$\sigma_{Z1Y} = \sigma_{Z1X1}\gamma + \sigma_{Z1X2}\alpha$$
$$\sigma_{Z2Y} = \sigma_{Z2X1}\gamma + \sigma_{Z2X2}\alpha$$

Remember:
$\sigma_{XY} = \Sigma_p$ product(p) where all p are X-Y paths not containing a collider ...

Definition (Instrumental Set)

Set $\{Z_1, ..., Z_k\}$ is an instrumental set for path coefficients $\alpha_1,...,\alpha_k$, where X_i-α_k->Y, iff

1. For each i, Z_i is separated from Y in G' (where G' is G with edges $X_1 \rightarrow Y, ..., X_k \rightarrow Y$ deleted)
2. There are paths p_i: Z_i to Y containing $X_i \rightarrow Y$ $(1 \leq i \leq k)$, and for all paths p_i p_j ($i \neq j$ in $\{1,2,...k\}$) with any common RV V one of the following holds:
   - Both $p_i[Z_i...V]$ and $p_j[V...Y]$ point to V or
   - Both $p_j[Z_j...V]$ and $p_i[V...Y]$ point to V

$p_i[W...H]$ = subpath of p_i from W to H

Definition (Instrumental Set)

Condition 2. says:
Cannot merge two intersecting paths p_i and p_j to yield two unblocked paths: one must contain collider

2. There are paths p_i: Z_i to Y containing $X_i \rightarrow Y$ $(1 \leq i \leq k)$, and for all paths p_i p_j $(i \neq j$ in $\{1,2,\ldots k\})$ with any common RV V one of the following holds:
   - Both $p_i[Z_i \ldots V]$ and $p_j[V \ldots Y]$ point to V or
   - Both $p_j[Z_j \ldots V]$ and $p_i[V \ldots Y]$ point to V

$p_i[W \ldots H]$ = subpath of p_i from W to H

Theorem

Let $\{Z_1, ..., Z_k\}$ be an instrumental set for coefficients $\alpha_1 ... \alpha_k$ with X_i-α_k->Y.

Then: The equations below are linearly independent for almost all parameterizations of the model and can be solved to obtain expressions for $\alpha_1 ... \alpha_k$ in terms of the covariance matrix

$$\sigma_{Z1Y} = \sigma_{Z1X1}\alpha_1 + \sigma_{Z1X2}\alpha_2 + ... + \sigma_{Z1Xk}\alpha_k$$

$$\sigma_{Z2Y} = \sigma_{Z2X1}\alpha_1 + \sigma_{Z2X2}\alpha_2 + ... + \sigma_{Z2Xk}\alpha_k$$

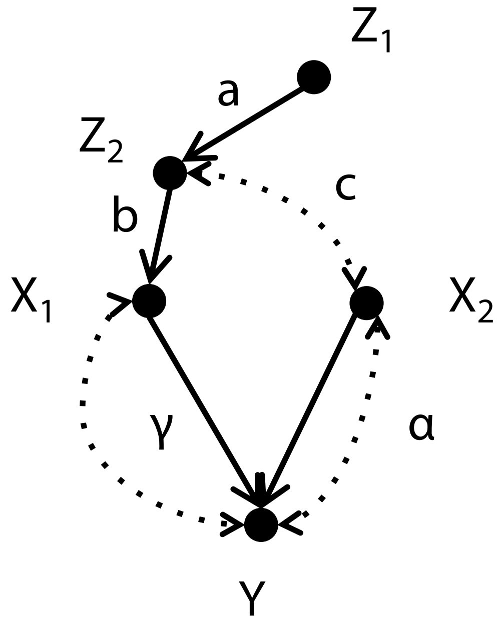
...

$$\sigma_{ZkY} = \sigma_{ZkX1}\alpha_1 + \sigma_{ZkX2}\alpha_2 + ... + \sigma_{ZkXk}\alpha_k$$

Ensuring linear independence:
- The rank of the covariance matrix has its maximum
- -> no information loss
- ensuring identifiability of parameters $\alpha_1 ... \alpha_k$.

Example: Instrumental sets (positive case)

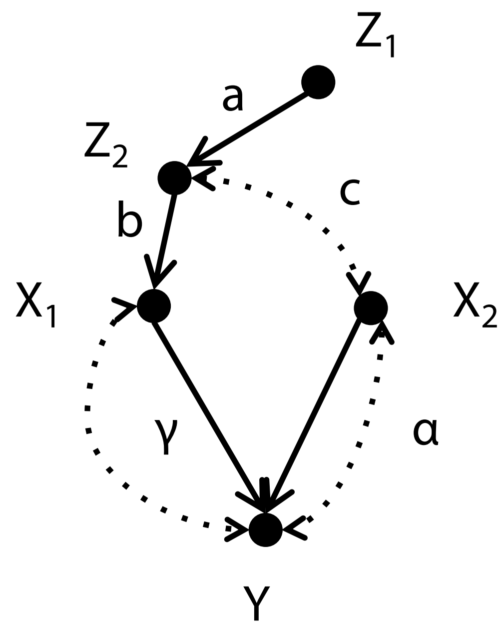
- $p_1 = Z_1 \rightarrow Z_2 \rightarrow X_1 \rightarrow Y$
- $p_2 = Z_2 \leftrightarrow X_2 \rightarrow Y$



- p_1 and p_2 satisfy condition 2 w.r.t. common variable $V = Z_2$
  - $p_1[Z_1 \ldots V] = Z_1 \rightarrow Z_2$ points to Z_2
  - $p_2[V \ldots Y] = p_2$ also points to Z_2
  - Z_2 as a collider blocks possible path merges of p_1 and p_2

Example: Instrumental sets (positive case)

- Algebraically
  - σ_{Z1Y} lacks influence of path $Z_2 \leftrightarrow X_2 \rightarrow Y$ and hence does not contain term $ac\alpha$
  - σ_{Z2Y} contains term $c\alpha$



- Applying Wright's rule

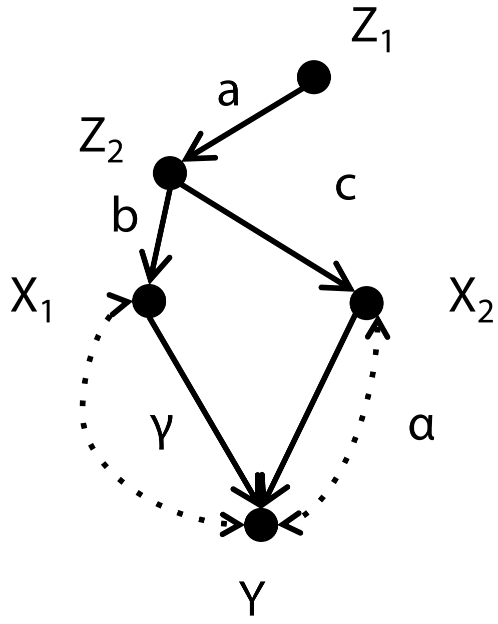
$$\sigma_{Z1Y} = \sigma_{Z1X1}\gamma + \sigma_{Z1X2}\alpha = \sigma_{Z1X1}\gamma + 0\alpha = ab\gamma$$

$$\sigma_{Z2Y} = \sigma_{Z2X1}\gamma + \sigma_{Z2X2}\alpha = b\gamma + c\alpha$$

- Solving linearly independent equations:
  - $\gamma = \sigma_{Z1Y}/\sigma_{Z1X1}$
  - $\alpha = \sigma_{Z2Y}/\sigma_{Z2X2} - \sigma_{Z2X1}\sigma_{Z1Y}/\sigma_{Z2X2}\sigma_{Z1X1}$

Example: Instrument sets (negative case)

- $p_1 = Z_1 \rightarrow Z_2 \rightarrow X_1 \rightarrow Y$
- $p_2 = Z_2 \rightarrow X_2 \rightarrow Y$



- Every path from Z_2 to Y is a "sub-path" of a path from Z_1 to Y

- Applying Wright's rule

$$\sigma_{Z2Y} = b\gamma + c\alpha$$

$$\sigma_{Z1Y} = ab\gamma + ac\alpha = a(b\gamma + c\alpha) = a\sigma_{Z2Y}$$

Conditional Instrumental Sets

- See C. Brito & J.Pearl: Generalized instrumental variables. In *Uncertainty in Artificial Intelligence, Proceedings of the Eighteenth Conference*, 85–93, 2002.

Mediation in Linear Systems

- Direct effect (DE) of X on Y mediated by Z
  - Remember in nonlinear case:

Definition The controlled direct effect (CDE) on Y of changing X from x to x′ is defined by

$P(Y= y| do(X=x), do(Z=z)) - P(Y= y| do(X=x'), do(Z=z))$

  - In linear case: Estimate path coefficient between X and Y as shown before (using, say, IVs)
- Total effect (τ) of X on Y
  - Estimate by regression as shown before
- Indirect effect of X on Y
  - $IE = \tau - DE$

(For non-linear systems need counterfactuals)