Non-Standard Databases and Data Mining

Instrumental Variables

Dr. Özgür Özçep
Universität zu Lübeck
Institut für Informationssysteme

Presented by Prof. Dr. Ralf Möller

Structural Causal Models

Slides prepared by Özgür Özçep

Part III: Causality in Linear SCMs and Instrumental Variables

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 are 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)

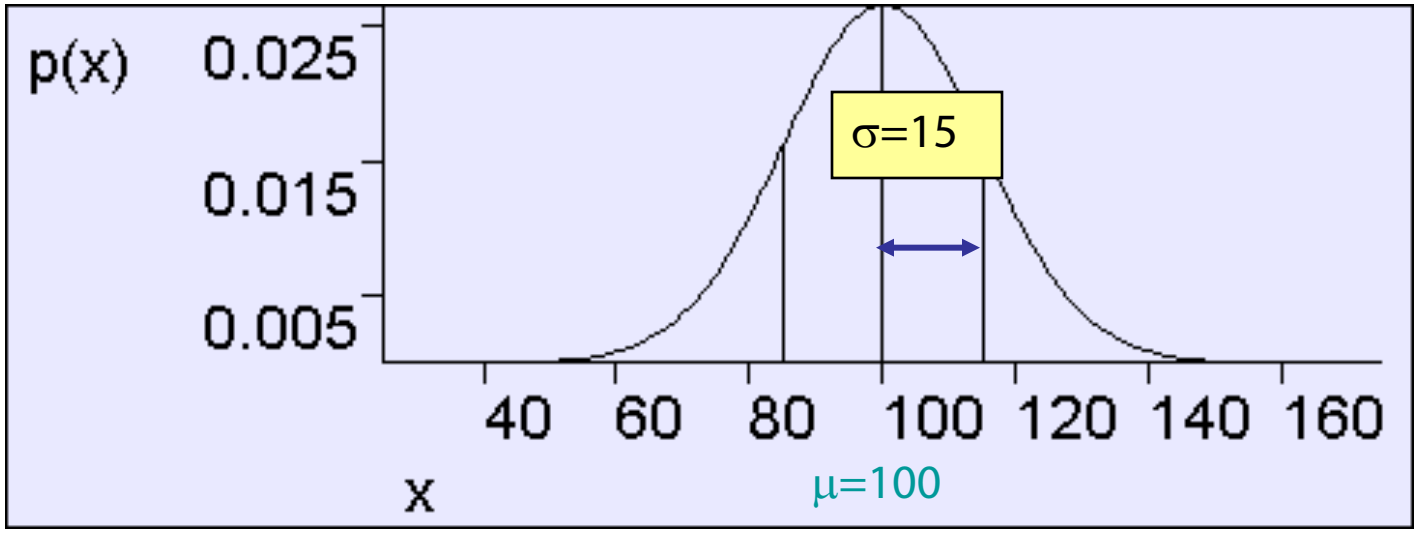
- Assumption for the following
  - All variables depending linearly on others (if at all)
  - Error variables (exogenous variables) have Gaussian/Normal distribution

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)

Need only specify μ,σ^2

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

Bivariate Gaussians

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

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

Where the Gaussian's parameters are…

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

Covariance matrix in 2 dimensions

$\sigma_{XY} = E[(X-E(X))(Y-E(Y))]$

Where we require that $\boldsymbol{\Sigma}$ is symmetric positive semidefinite

For positive semidefinite matrices M, the scalar $z^T M z$ is positive for every non-zero column vector z of real numbers

Non-zero determinant

It turns out that $E[X] = \mu$ and $Cov[X] = \boldsymbol{\Sigma}$. (Note that this is a resulting property of Gaussians, not a definition)*

*This note rates 7.4 on the pedanticness scale

So need only specify 2*N + N(N-1)/2 = 2*2 + 2(2-1)/2 = 5 parameters

Multivariate Gaussians

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

$$\text{Write r.v. } \mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_m \end{pmatrix} \qquad \text{Then define} \quad X \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad \text{to mean}$$

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

Gaussian's parameters …

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

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 follows

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

→ Can use regression to determine causal relations

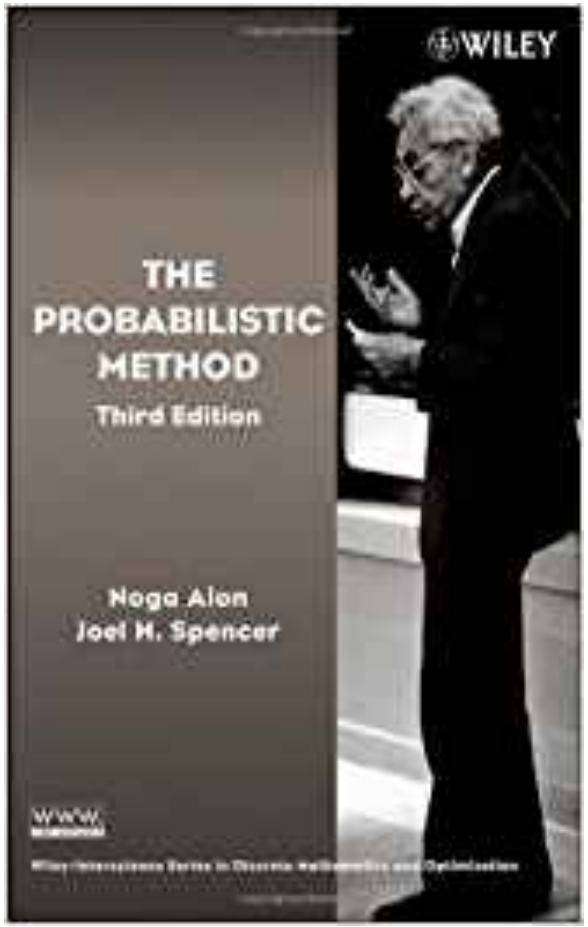
- $E[Y|X]$ defines a function $f(X,Y)$
- By regression we circumvent the problem of calculating the probabilities required for $E[Y|X]$

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 probabilities to infer „crisp properties“

- Toy Example:
  - If you know that the expected value of 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

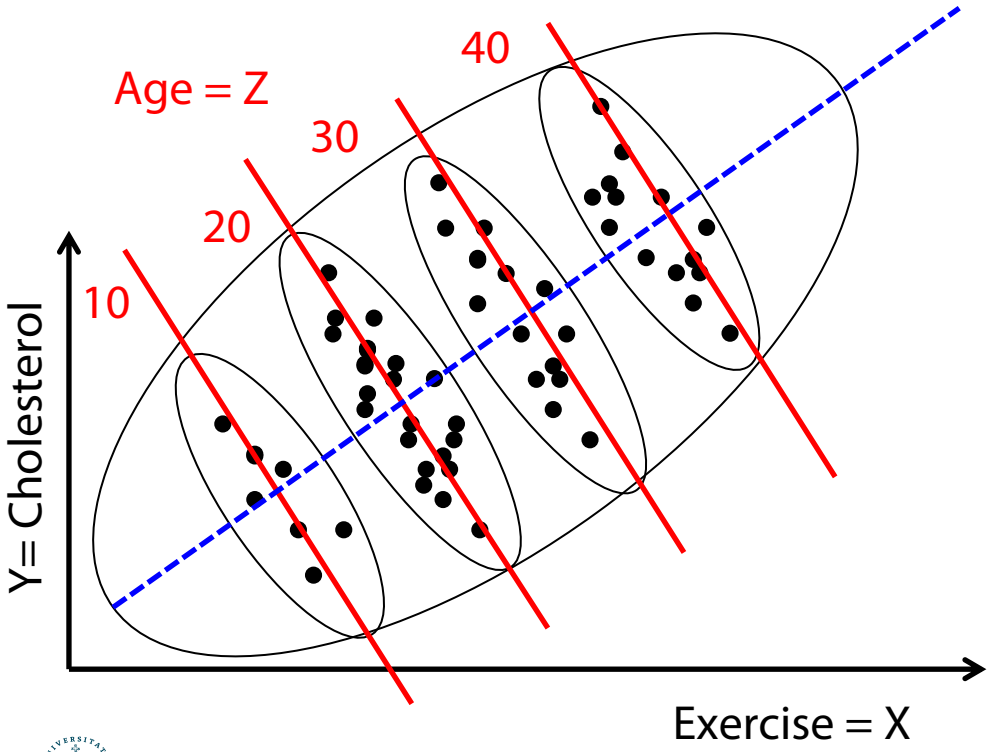
- Expectations are expressed as linear combinations
  - $E[Y|X_1=x_1, X_2=x_2, \ldots, X_n=x_n] = r_0 + r_1x_1 + \ldots + r_nx_n$
  - Each of the slopes r_i are partial regression coefficients
  - Example and Notation

    $r_i = \phi_{Y\,Xi\,.\,X1\ldots Xi-1,\,Xi+1,\ldots Xn}$

    = 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 weekly exercise and cholesterol in different age groups



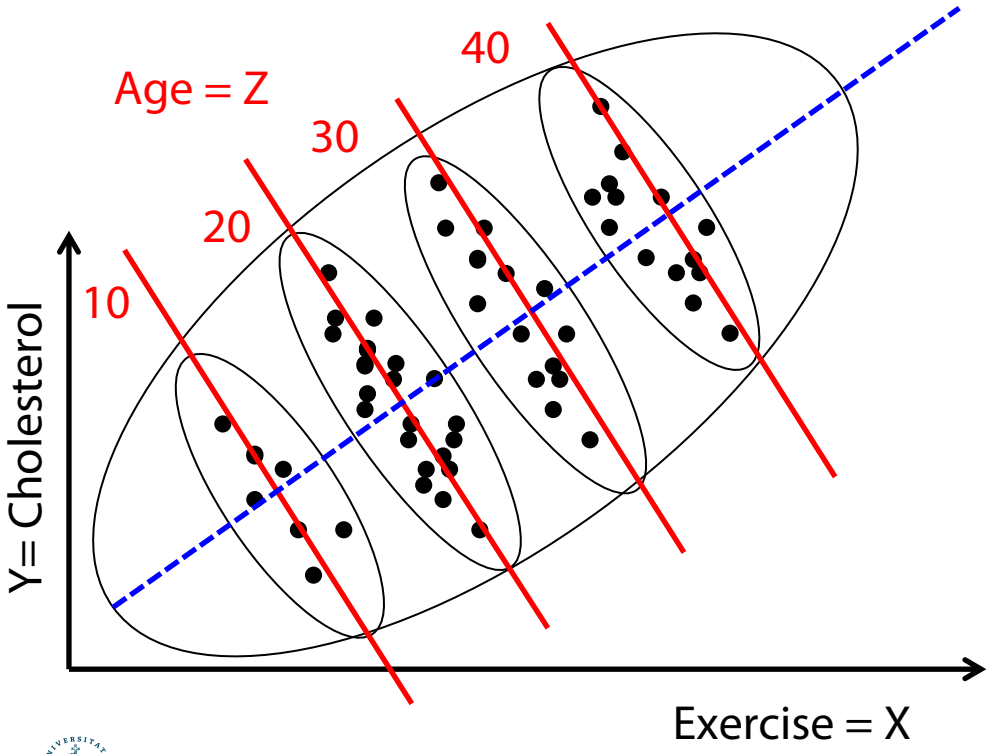
- $Y = r_0 + r_1 X + r_2 Z$
- $r_1 = R_{YX.Z} < 0$
- Z-fixed slope for Y,X independent of Z (and negative)

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

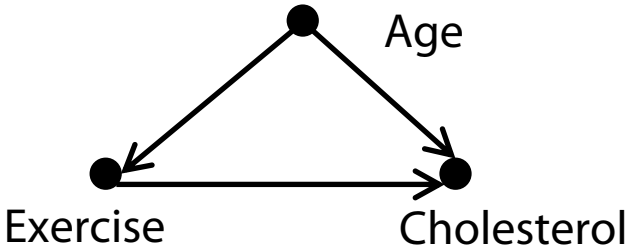
→ Simpson's paradox

Resolving the Paradox

- Measure weakly exercise and cholesterol in different age groups



- Age is a confounder of Exercise and Cholesterol
- 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

- $\sigma_{XY} = E[(X-E[Y])(Y-E[Y])]$ (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 independent

Orthogonality principle

Theorem

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 $\sigma_{\varepsilon X_i} = 0$ for all $1 \leq i \leq k$

Regression coefficients and covariance

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

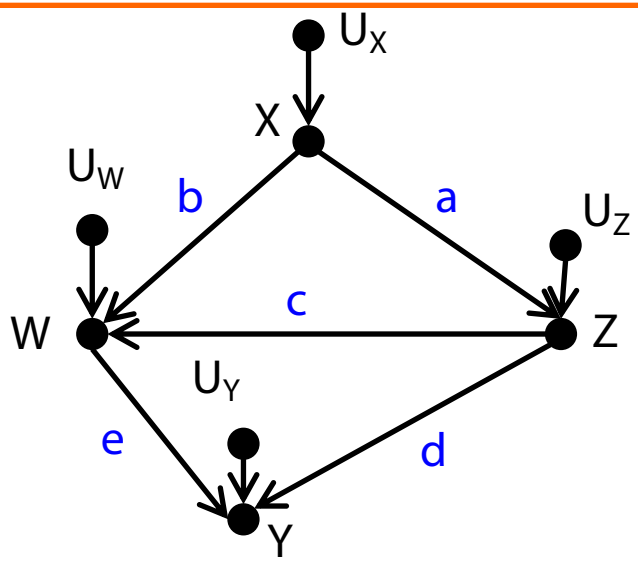
Similar derivations for multiple regression

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



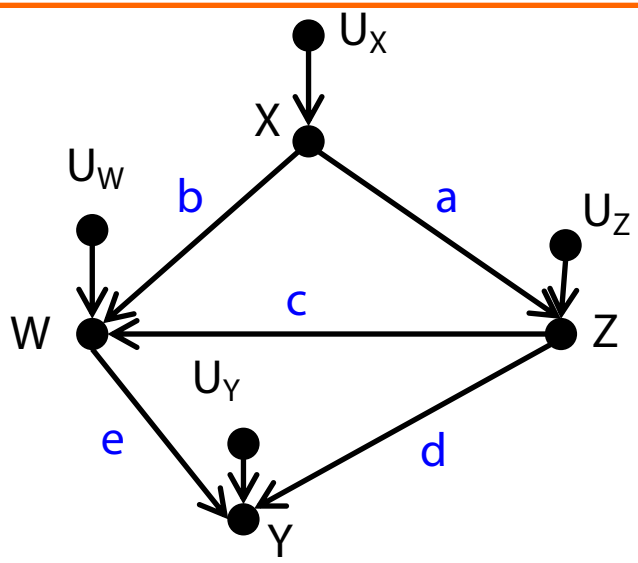
Linearity assumption makes association of coefficient to edge a well-formed operation

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

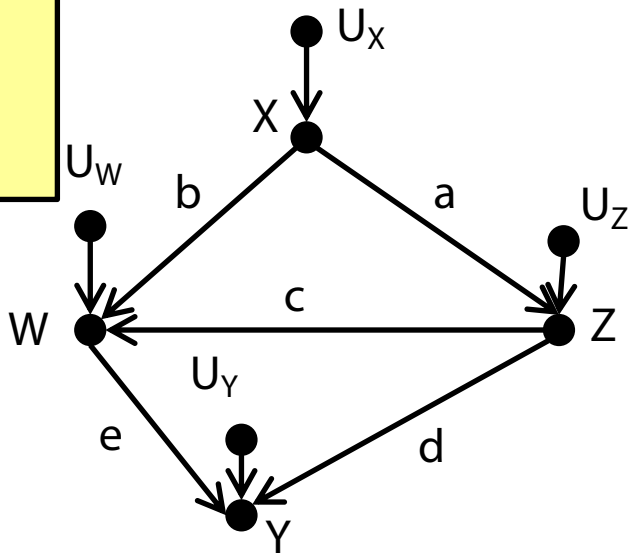


Warning from the beginning:
Path coefficients (causal) ≠ regression coefficients (descriptive)

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 rate Z=+1



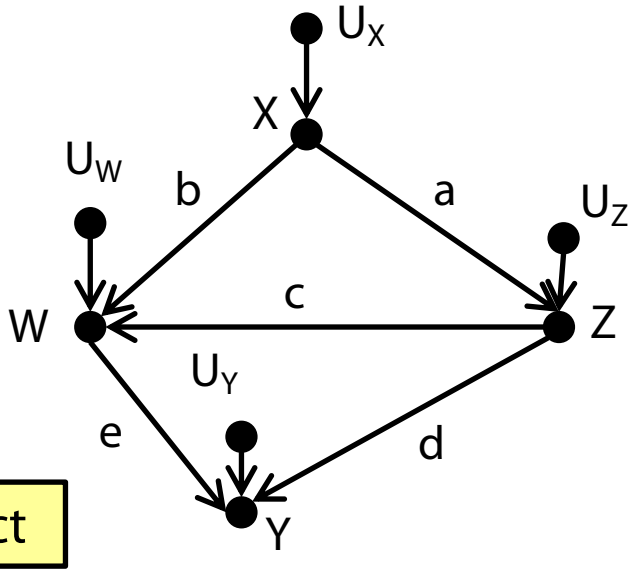
- Q: What is the semantics of the path coefficients on edge Z-Y?
- A: Causal Direct Effect (CDE) on Y of change Z=+1

$$CDE = E[Y|do(Z=z+1), do(W=w)] - E[Y|do(Z=z), do(W=w)]$$
$$= d(z+1) + ew + E[U_Y] - (dz + ew + E[U_Y])$$
$$= d = \text{label on Z-Y edge}$$

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

Total Effect in Linear Systems (Example)

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



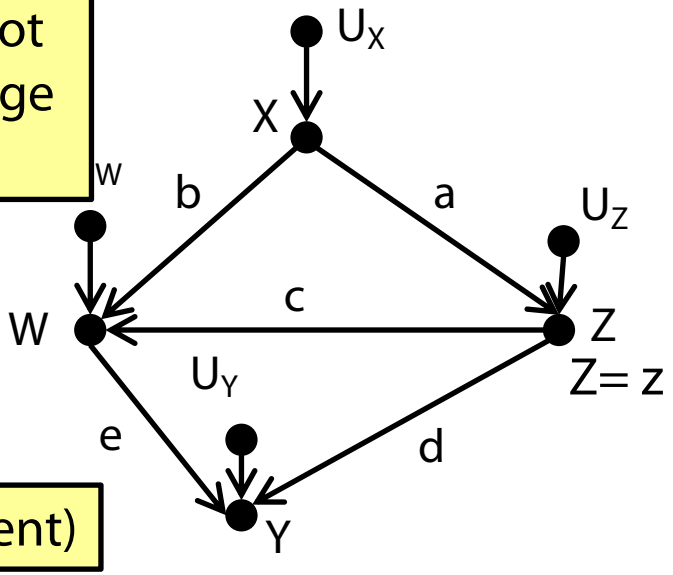
Total effect = general causal effect

- 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)

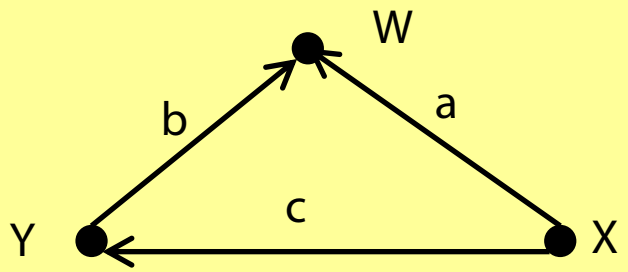
- Q: What is the total effect of Z on Y?
- A: Sum of coefficient products over each directed Z-Y path
  - Total effect τ: Intervene on Z and express Y by Z
  - $Y = dZ + eW + U_Y = dZ + e(bX + cZ + U_W) + U_Y$

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

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

Note 4

- We followed (Bollen 1989)) and summed over directed paths
- In book of Pearl, Glymour & Jewell (p.82-83) summation over non-backdoor paths
  - Seems to be an error (due to wrongly applied Wright's path 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

  $\sigma_{XY} = \sum_p$ 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

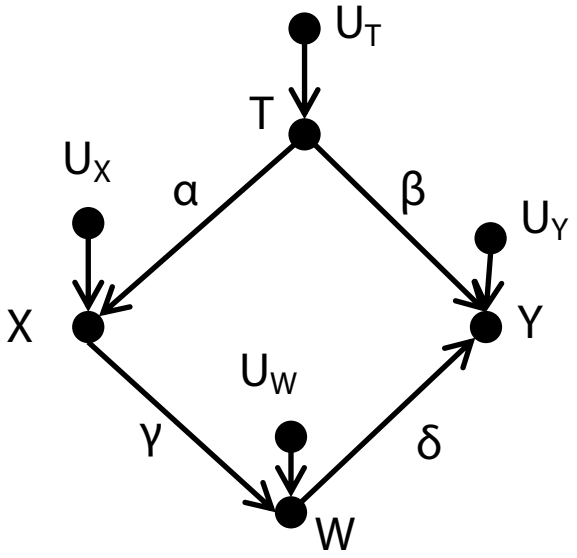
- What if path coefficients are not known apriori or are not testable?
- One has to identify only those relevant for the specific task, e.g., total effect of X to Y or direct effect of Z on X

- 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

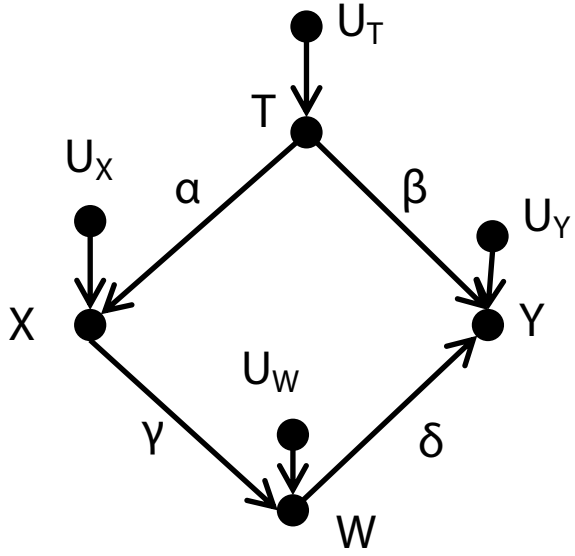
- Q: Direct effect of X on Y?

- 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 (GCE) of X on Y?
- Now path coefficients not necessarily known (Greek letters)
- Recall: With backdoor criterion identify Z to adjust for
  GCE = $P(y|do(x)) = \sum_z P(y \mid x,z)P(z)$



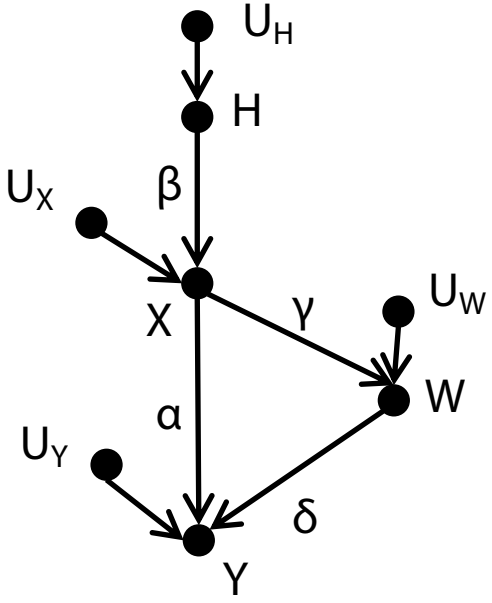
- Use backdoor to identify variables to regress for
- Here $Z = \{T\}$, so do linear regression on X,T:
  - $Y(X,T) = r_X X + r_T T + \varepsilon$
  - r_X = total effect of X on Y

- 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 and, 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}

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

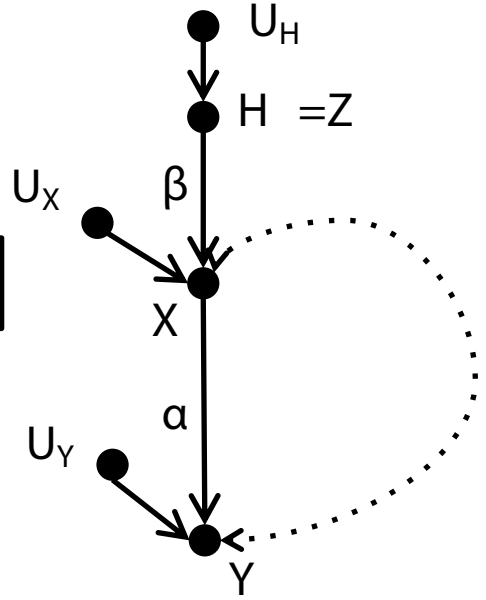
  Direct effect of X on Y= r_X =:α

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

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. $r_1/r_2 = \alpha$ = direct effect of X on Y



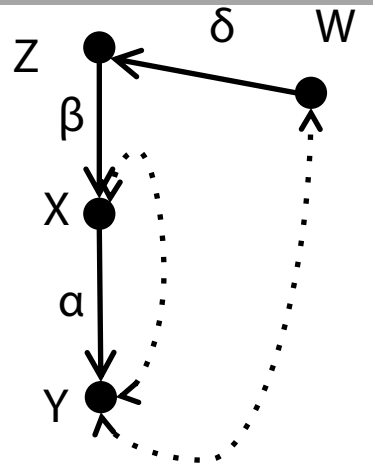
Here: Z = H

$$Y = \frac{r_1}{r_2}X - \frac{r_1}{r_2}\varepsilon + \varepsilon = \frac{r_1}{r_2}X + \varepsilon'$$

This is because
- Z = H emits no backdoors, so $r_2 = \beta$
- r_1 = total effect of Z on Y = $\beta\alpha$

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

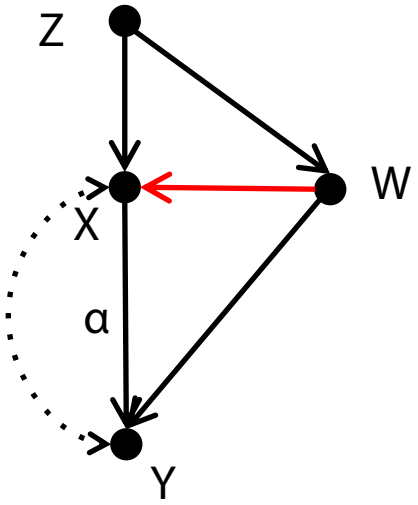
If conditions fulfilled, then $\alpha = \beta_{YZ.W} / \beta_{XZ.W}$

Conditional IVs (Examples)

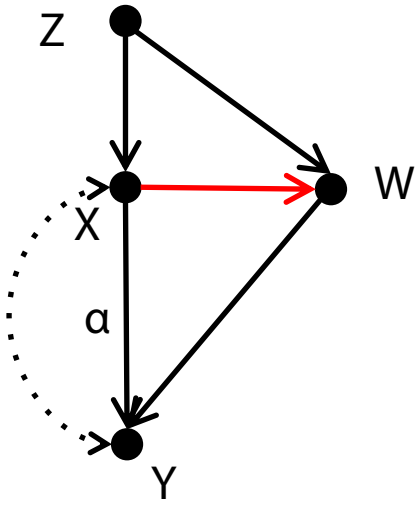
Z instrument for α given W?

Definition Z is a conditional IV given set W for α iff

- Set of descendants of Y 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

Summary

- Models can be incomplete
  - Unknown parameters
  - Unknown confounder structures
- Nevertheless, we can analyse certain direct and total causal effects
  - In come cases network structure and available parameters allow for conditioning on certain random variables
  - In case this is not possible, one can try to identify so-called
    - (Sets of) instrumental variables
    - (Sets of) Conditional instrumental variables