# Ch 6: Nodes, Edges, and Statistical (In)dependence

Prashant Shekhar, PhD
*Assistant Professor of Data Science*
*Department of Mathematics*
*Embry-Riddle Aeronautical University, FL, USA*
Email: shekharp@erau.edu

# Outline

# D-SEPARATION

# d-separation

- We say that two nodes in a directed acyclic graph (DAG) G are **d-separated** when all paths between them are **blocked**.
- A path between two nodes is regarded as blocked when there's a collider on a path between them or if there's a fork or a chain that contains another variable that we control for (or a descendant of such a variable).

# d-separation (More formally)

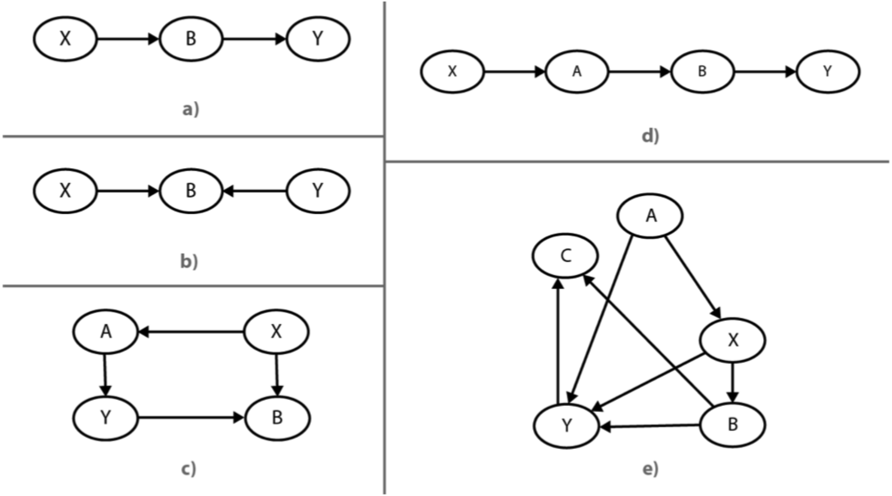
## d-separation

For any three disjoint sets of nodes $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$, a path between $\mathcal{X}$ and $\mathcal{Y}$ is blocked by $\mathcal{Z}$ in the following scenarios:

- If there's a fork, $i \leftarrow j \rightarrow k$, or a chain, $i \rightarrow j \rightarrow k$, in this path such that the middle node is $j \in \mathcal{Z}$.
- If there's a collider, $i \rightarrow j \leftarrow k$, on this path such that neither $j$ nor any of its descendants belong to $\mathcal{Z}$

In other words, if there's a chain or fork between $\mathcal{X}$ and $\mathcal{Y}$, we need to control for the middle node to close the path between $\mathcal{X}$ and $\mathcal{Y}$. If there's a collider between $\mathcal{X}$ and $\mathcal{Y}$, we should leave it uncontrolled altogether with all its descendants.

# What to control to make X and Y d-separated



# What to control to make X and Y d-separated

- **Part a**: To block the path between X and Y, *we need to control for B*.
- **Part b**: We have a collider situation so the path between X and Y is *already blocked*.
- **Part c**: Path through B is already blocked (collider). Hence we just need to *control for A*.
- **Part d** Since its a chain so we can *either control for A or B*
- **Part e**:
  - Path $X \leftarrow A \rightarrow Y$ can be blocked by controlling for A.
  - Path $X \rightarrow B \rightarrow C \leftarrow Y$ is already blocked (C is a collider).
  - Path $X \rightarrow B \rightarrow Y$ can be blocked by controlling for B.
  - The path $X \rightarrow Y$ cannot be blocked.

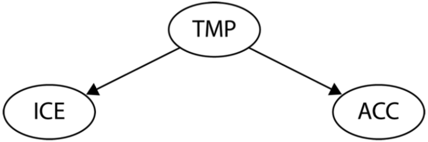
  Hence in part e, X and Y *can't be d-separated*

# Estimand

# Some definitions

- In statistical inference and machine learning, we often talk about estimates and estimators.
- **Estimates** are basically our best guesses regarding some quantities of interest given (finite) data.
- **Estimators** are computational devices or procedures that allow us to map between a given (finite) data sample and an estimate of interest. Estimators might be as simple as computing the arithmetic mean and as complex as a 550 billion- parameter language model. Linear regression is an estimator, and so is a neural network or a random forest model.
- **Estimand** is a quantity that we're interested in estimating. If an estimator is the how, an estimand is the what.

# An example




- It was observed in a city that the number of ice cream (ICE) sales and the number of accidental drownings (ACC) are correlated.
- However, later it was realized that ICE and ACC are not causally related. There is a confounding variable: temperature (TMP) that leads to more ice cream sales as well as more people to go to beaches. The variable relationships are represented in the above figure.
- In such a situation, the relationship between ICE and ACC is regarded as **spurious**.

# An example

- In order to predict the accidental deaths (ACC), we can fit our model on ice cream sales (ICE). i.e.,

$$ACC = f(ICE)$$

- What we want to estimate is the **causal effect of ICE on ACC**. In other words, **we want to understand what the change would be in ACC if we intervened on ICE**. Assuming discrete probability for simplicity, the estimand we are interested in is as follows:

$$P(ACC = acc|do(ICE = ice))$$

- If we used our naive model (without considering confounding), our estimand would be:

$$P(ACC = acc|do(ICE = ice)) = P(ACC = acc|ICE = ice)$$

## Example cont..

- However, we know that the relationship between ACC and ICE is spurious!
- To get the correct estimate of the causal effect of ICE on ACC, we need to control for temperature (TMP). i.e.,

$$ACC = g(ICE, TMP)$$

- This translates to $P(ACC = acc|do(ICE = ice)) =$

$$\sum_{tmp} P(ACC = acc|ICE = ice, TMP = tmp)P(TMP = tmp) \tag{1}$$

- In a more compact notation:

$$P(ACC|do(ICE)) = \sum_{temp} P(ACC|ICE, TMP)P(TMP) \tag{2}$$

# Example cont..

## Causal effect rule

The formula in (2) is an example of the so-called **causal effect rule**, which states that given a graph, $G$, and a set of variables, $Pa$, that are (causal) parents of $X$, the causal effect of $X$ on $Y$ is given by the following:

$$P(Y = y|do(X = x)) = \sum_z P(Y = y|X = x, Pa = z)P(Pa = z) \tag{3}$$

- In our example, TMP is the middle node in the fork between ICE and ACC, and so controlling for it blocks the non-causal path between ICE and ACC.
- All that we've done so far in this section has one essential goal – to find an estimand that allows us to compute unbiased causal effects from observational data.
- *Next, we'll focus on techniques that allow us to obtain causal estimands, given complete or partially complete graphs.*

## A refresher on the do-operator

- The do-operator informs us that we're working with interventional rather than observational distribution.
- In certain cases, interventional and observational distributions might be the same. For instance, if your true causal graph has a form of $X \rightarrow Y$, then $P(Y = y|do(X = x)) = P(Y = y|X = x)$.
- Whenever confounding appears, we need to adjust for the confounders' effects by controlling for additional variables in the right-hand side of the equation

# Back-door criterion

# Definition

The backdoor criterion aims at:

- Blocking spurious paths between our treatment and outcome nodes.
- At the same time, we want to leave all directed paths unaltered.
- Addtionally, dont want to create new spurious paths.

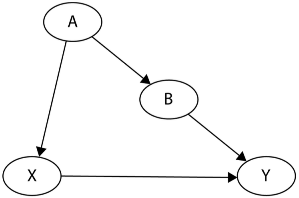
**back-door criterion (formally)**

A set of variables, $\mathcal{Z}$, satisfies the back-door criterion, given a graph $G$, and a pair of variables, if no node in $\mathcal{Z}$ is a descendant of $X$, and $\mathcal{Z}$ blocks all the paths between $X$ and $Y$ that contain an arrow into $X$.

We precisely did this in the ice cream example:

- We blocked all the paths between ICE and ACC that contained an arrow into ICE.
- TMP is not a descendant of ICE, so we also met the second condition.
- Finally, we didn't opened any new spurious paths

# Back-door and equivalent estimands




- Given the model presented in Figure above, which nodes should we control for in order to estimate the causal effect of X on Y?
- According to our definition of the back-door criterion, we need to block all the paths that have an arrow into X. We should not control for any descendants of X nor open any new paths. We can fulfill these conditions in three different ways:
  - Controlling for A
  - Controlling for B
  - Controlling for both A and B

# Back-door and equivalent estimands

For the problem on the previous slide we get 3 **equivalent estimands**:

$$
\begin{aligned}
P(Y=y|do(X=x)) &= \sum_a P(Y=y|X=x, A=a)P(A=a) \\
&= \sum_b P(Y=y|X=x, B=b)P(B=b) \\
&= \sum_b \sum_a P(Y=y|X=x, A=a, B=b)P(A=a)P(B=b)
\end{aligned}
$$

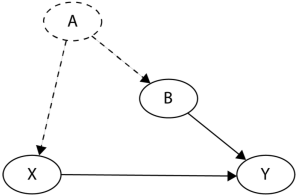
Please note, If it is sufficient to only control for one of the variables (A or B) to obtain a correct estimand for $X \rightarrow Y$, **we can essentially estimate the causal effect of X on Y even if one of the variables remains unobserved!**

# Back-door and equivalent estimands

## Equivalent estimands versus equal estimates

- Although for certain models we might find two or more equivalent estimands, estimates computed based on these (equivalent) estimands might differ slightly. This is natural in a finite sample size regime.
- Nonetheless, if your sample size is big enough, the differences should be negligible.
- Big differences might suggest an erroneous estimand, a lack of model convergence, or errors in the model code.

# Back-door adjustment with some unobserved variables




- Here the node A and two edges ($A \rightarrow B$ and $A \rightarrow X$) marked with dashed lines are all unobserved, yet we assume that the overall causal structure is known (including the two unobserved edges).
- In other words, we don't know anything about A or what the functional form of A's influence on X or B is. At the same time, we assume that A exists and has no other edges than the one presented above.
- Our estimand for the model presented in above figure would be identical to the second estimand for the fully observed model:

$$P(Y = y|do(X = x)) = \sum_b P(Y = y|X = x, B = b)P(B = b)$$

# Back-door adjustment with some unobserved variables

Things to note:

- Imagine that recording A is the most expensive part of your data collection process. Now, understanding the back-door criterion, you can essentially just skip recording this variable!
- One thing we need to remember is that to keep this estimand valid, we need to be sure that the overall causal structure holds. If we changed the structure a bit by adding a direct edge from A to Y, the preceding estimand would lose its validity.
- That said, if we completely removed A and all its edges from the model, our estimand would still hold. Can you explain why ?

## Back-door adjustment conclusion

- We learned about the back-door criterion and how it can help us build valid causal estimands.
- We saw that, in some cases, we might be able to build more than one valid estimand for a single model.
- We also demonstrated that the back-door criterion can be helpful in certain cases of unobserved confounding.

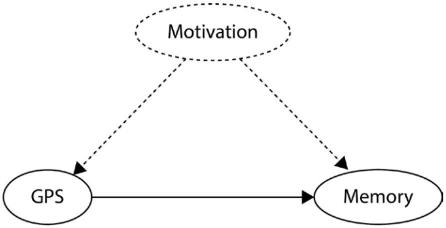
# Front-door criterion

# Problem at hand: Can GPS lead us astray

- In their 2020 study, Louisa Dahmani and Véronique Bohbot from McGill University showed that there's a link between GPS usage and spatial memory decline (Dahmani and Bohbot, 2020).
- Moreover, the effect is dose-dependent, which means that the more you use GPS, the more spatial memory decline you experience.
- The authors argue that their results suggest a causal link between GPS usage and spatial memory decline.
- However, we already know that something that looks connected does not necessarily have to be connected in reality.

# Problem at hand: Can GPS lead us astray: Alternate hypothesis

- It can be argued that GPS usage and spatial memory decline have a common cause - **low global motivation**.
- People with low global motivation are reluctant to learn new things (so they are not interested in remembering new information, including spatial information) and they try to avoid effort (hence, they prefer to use GPS more often).
- Hence a possible SCM for our problem could be:




- However, the model can't be deconfounded because the confounder (Motivation) is unobserved (dashed lines). **So the back-door criterion can't help us**.

# Front-door criterion to the rescue

- Since the backdoor criterion cant be used to deconfound our model, we need an alternative.
- In such cases we could possibly use the front-door criterion that relies on the concept of **mediation**.
- This would require us to find a variable that mediates the relationship between GPS usage and memory decline.

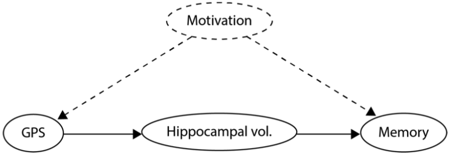
## Mediators and mediation

- We can say that the influence of one variable ($X$) on another ($Y$) is mediated by a third variable, $Z$ (or a set of variables, $\mathcal{Z}$), when at least one path from $X$ to $Y$ goes through $Z$.
- We can say that $Z$ **fully mediates** the relationship between $X$ and $Y$ when the only path from $X$ to $Y$ goes through $Z$.
- If there are paths from $X$ to $Y$ that do not pass through $Z$, the mediation is **partial**.

# Finding the mediator: case of London cabbies

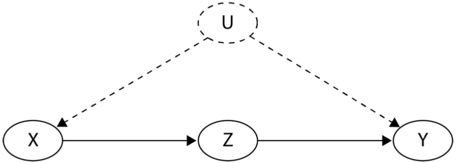
- London cab drivers need to pass a very restrictive exam checking their spatial knowledge and are not allowed to use any external aids in the process.
- It has been found that the experience as a taxi driver is related to hippocampus's volume (Maguire et al., 2000)
- The hippocampus is a pebble-sized (40-55 mm) brain structure, responsible for creating new memories – in particular, spatial memories (O'Keefe and Nadel, 1978).
- One study (Woollett and Maguire, 2011) showed that drivers who failed this exam did not show an increase in hippocampal volume. At the same time, in those who passed the exam, a systematic increase in hippocampal volume was observed during continual training over a 4-year period.
- Hence we can update our SCM as follows..

# Updated SCM




- In our new hypothetical model, we assume that hippocampal volume fully mediates the effects of GPS usage on a decline in spatial memory.
- The second important assumption we make is that motivation can only affect hippocampal volume indirectly through GPS usage. This assumption is critical in order to make the criterion that we're going to introduce next – the front-door criterion – useful to us.
- If motivation would be able to influence hippocampal volume directly, front-door would be of no help.

# Front-door criterion




Firstly looking at the relation between X and Z:

- There is one back-door path between them: $X \leftarrow U \rightarrow Y \leftarrow Z$. However it is already blocked due to the collider.
- Hence:

$$P(Z = z | do(X = x)) = P(Z = z | X = x)$$

Now analyzing effect of Z on Y ?

- Here the back-door path is open $Z \leftarrow X \leftarrow U \rightarrow Y$
- There's no collider on this path and U is unobserved, so we cannot control for it.

# Front-door criterion

- Fortunately, we can control for the other variable, X. A valid estimand of the causal effect of Z on Y is, therefore, the following:

$$P(Y = y|do(Z = z)) = \sum_{x'} P(Y = y|Z = z, X = x')P(X = x')$$

- We just blocked the back-door path from Z to Y by simply controlling for X. Now, we're ready to combine both estimands back together:

$$P(Y = y|do(X = x)) = \sum_{z} P(Y = y|do(Z = z))P(Z = z|do(X = x)) \tag{4}$$

$$= \sum_{z} P(Z = z|X = x) \sum_{x'} P(Y = y|X = x', Z = z)P(X = x') \tag{5}$$

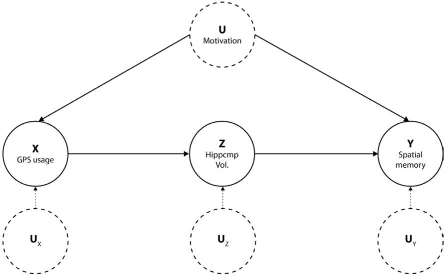
The above formula is called the **front-door formula/criterion/adjustment**.

## Front-door criterion

In general, we can say that a set of variables, $\mathcal{Z}$, satisfies the front-door criterion, given the graph, $G$, and a pair of variables, $X \rightarrow ... \rightarrow Y$, if the following applies (Pearl et al., 2016):

- $\mathcal{Z}$ intercepts all directed paths from X to Y.
- There are no open back-door paths from X to $\mathcal{Z}$
- All back-door paths from $\mathcal{Z}$ to Y are blocked by X.

# Implementing Front-door criterion (and others) in python




Here our SCM has following structural equations:

- $u \sim N(0, 4)$ with truncation such as $u \geq 0$
- $u_x \sim N(0, 5)$ with truncation such as $u \geq 0$
- $u_y, u_z \sim N(0, 2)$
- $x := 0.7u + u_x$
- $z := -0.6x + 0.25u_z$
- $y := 0.7z + 0.25u + 0.2u_y$

# Approaches

**Problem:** Estimate the causal effect of GPS Usage(X) on Spatial memory(Y)

## True Causal Effect

Based on the structural equation on previous slide, the **true causal effect** of GPS Usage on Spatial Memory is: -0.6x0.7 = -0.42

**Model 1: (Naive Approach)**: Here we directly model the relationship between variables without thinking about confounding and back-door paths. *Only applicable if we are able to observe*:

- GPS Usage (X)
- Spatial Memory (Y)

Here we fit the model directly on the **observational data**:

$$Y = a_0 + a_1 X \tag{6}$$

Hence the estimated causal effect is: $a_1$

# Approaches

**Model 2 (Front-door adjustment)**: Here we consider confounding while estimating the causal effect. *Only applicable if we are able to observe*:

- GPS Usage (X)
- Hippocampus Vol.(Z): mediator
- Spatial Memory (Y)

Here we fit the model directly on the **observational data**. Firstly, we model the causal effect of GPS usage (X) on Hippocampus Vol. (Z). Since there is no backdoor path for $X \rightarrow Z$, we directly model:

$$Z = b_0 + b_1 X$$

Next we measure the causal effect of Hippocampus Vol. (Z) on Spatial memory (Y). For this case, we do have a back-door path: $Z \leftarrow X \leftarrow U \rightarrow Y$. For blocking this back-door path we control for $X$. Hence we fit the model:

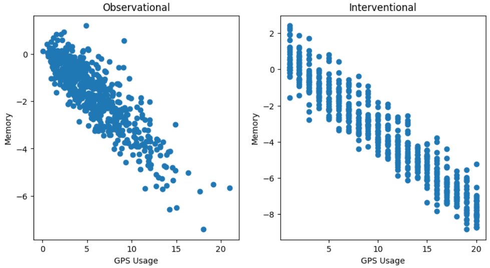
$$Y = c_0 + c_1 X + c_2 Z$$

Hence the estimated causal effect of X on Y is product of coefficient of X and Z: $b_1 c_2$

# Approaches

**Model 3 (Intervention based)**: Here we consider confounding while estimating the causal effect. Here we do intervention on X (GPS) usage to break the dependence of X on U and hence break the backdoor path $X \leftarrow U \rightarrow Y$. *Only applicable if we are able to create interventional samples for*:

- GPS Usage (X)
- Spatial Memory (Y)




Here we fit the model directly on the **interventional data**: $Y = d_0 + d_1 X$, producing the estimated causal effect: $d_1$.

# Approaches

**Model 4 (Back-door adjustment)**: Here we consider confounding while estimating the causal effect. *Only applicable if we are able to observe*:

- Motivation (U)
- GPS Usage (X)
- Spatial Memory (Y)

To estimate the causal impact of GPS Usage (X) on Spatial memory (Y), we will directly close the back-door path $X \leftarrow U \rightarrow Y$ by controlling for $U$:

$$Y = e_0 + e_1 U + e_2 X$$

Hence the causal effect of $X$ on $Y$ is $e_2$.

# Relative performance of approaches

Here we present the performance results of all 4 approaches for a random run (True causal effect = -0.42).

| Approach | Causal Effect (CE) | Percent Error $\lvert$(True CE-Estimated CE)/True CE$\rvert$ x 100% |
|---|---|---|
| Naive Approach | -0.342 | 18.42% |
| Front-door adjustment | -0.407 | 3.13% |
| Intervention based | -0.423 | 0.63% |
| Back-door adjustment | -0.420 | 0.03% |

If back-door adjustment and intervention based approaches have such low errors, why do we need front-door adjustment at all ?

## CODE

Refer to the [Python notebook] for implementation of:

- Data generation
- Model fitting and comparison for all 4 approaches

# DO-CALCULUS

# Introduction

- In the real world, not all causal graphs will have a structure that allows the use of the back-door or front-door criteria. Does this mean that we cannot do anything about them?
- Fortunately, no. Back-door and front-door criteria are special cases of a more general framework called do-calculus (Pearl, 2009).
- Moreover, do-calculus has been proven to be complete (Shpitser and Pearl, 2006), meaning that if there is an identifiable causal effect in a given DAG, G, it can be found using the rules of do-calculus.

**definition**

The $do-calculus$ is an axiomatic system for replacing probability formulas containing the $do-operator$ with ordinary conditional probabilities (so that it can be calculated from data). It consists of three axiom schemas that provide graphical criteria for when certain substitutions may be made.

# Some notations

Given a DAG G,

- $G_{\overline{X}}$: Modification of $G$, where we removed all the incoming edges to the nodes in $X$.
- $G_{\underline{X}}$: Modification of $G$, where we removed all outgoing edges from the nodes in $X$.
- $G_{\overline{X}\underline{Z}}$: Modification of $G$, where we removed all the incoming edges to the nodes in $X$ and all the outgoing edges from the nodes in $Z$.
- $G_{\overline{X,Z(W)}}$: Modification of $G$, where we removed all the incoming edges to the nodes in X. Additionally, blocked all incoming edges to nodes in $Z$ that aren't ancestors of nodes in $W$.

# 3 rules of do-calculus

**Rule1** When an observation can be ignored : if $Y \perp\!\!\!\perp_{G_{\overline{X}}} Z | X, W$:

$$P(Y = y | do(X = x), Z = z, W = w) = P(Y = y | do(X = x), W = w)$$

**Some intuition**:

- Considering the graph $G_{\overline{X}}$, hence considering intervention on $X$.
- In the original graph ($G_{\overline{X}}$) with no other condition we have $Y \perp\!\!\!\perp_{G_{\overline{X}}} Z | X, W$. i,e, $Y$ and $Z$ are $d-separated$ by $X \cup W$.
- Hence Z doesnt impact Y, causally or spuriously. Therefore, we can remove conditioning on $Z$ when analyzing probability of $Y$.
- In other words, we can ignore an observation of a quantity ($Z = z$ here) when it doesn't influence the outcome ($Y$) through any path.

# 3 rules of do-calculus

**Rule2** When intervention can be treated as an observation: if $Y \perp\!\!\!\perp_{G_{\overline{X}\underline{Z}}} Z|X, W$:

$$P(Y = y|do(X = x), do(Z = z), W = w) = P(Y = y|do(X = x), Z = z, W = w)$$

**Some intuition**

- Considering the graph $G_{\overline{X}}$, hence considering intervention on $X$.
- In the original graph ($G_{\overline{X}}$), with another condition of removing outgoing edges from $Z$, we have $Y \perp\!\!\!\perp_{G_{\overline{X}\underline{Z}}} Z|X, W$. i,e, $Y$ and $Z$ **become** $d-$*separated* by $X \cup W$.
- Hence in the original graph ($G_{\overline{X}}$), $Z$ only affected $Y$ causally (i.e., through directed paths)
- Hence in the original graph while writing $P(Y = y)$, do(Z=z) behind the conditioning bar can be replaced by $Z = z$ since there is no problem of spurious connection.
- Rule 2 can also be thought of as a generalization of the back-door criterion in which the set $\{do(X), W\}$ together form a back-door admissible set.

# 3 rules of do-calculus

- **Rule 3** When intervention can be ignored: $(Y \perp\!\!\!\perp Z | X, W)_{G_{\overline{X,Z(W)}}}$:

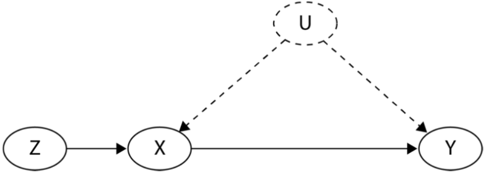
$$P(Y = y | do(X = x), do(Z = z), W = w) = P(Y = y | do(X = x), W = w)$$

**Some intuition**

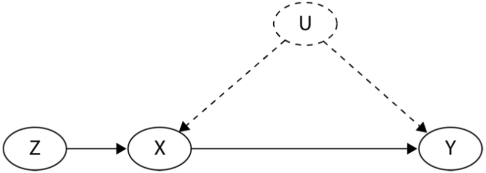
- Considering the graph $G_{\overline{X}}$, hence considering intervention on $X$.
- In the original graph $(G_{\overline{X}})$, with another condition of removing incoming edges to nodes in $Z$ that are not ancestors of $W$, gives us $(Y \perp\!\!\!\perp Z | X, W)_{G_{\overline{X,Z(W)}}}$
- Hence in the original graph $(G_{\overline{X}})$, $Z$ only affected $Y$ spuriously, through nodes that are part of $Z$ but aren't ancestors of $W$.
- Hence in the original graph $(G_{\overline{X}})$, for nodes in $Z$ that are ancestors of $W$, there is no spurious connection to $Y$ (for example due to it being a collider etc).
- Hence $do(Z = z)$ behind the conditioning bar can be ignored as there are no open causal paths between nodes in $Z$ and $Y$.

# Instrument Variables (IVs)

# Setup




- Instrumental variables (IVs) are a family of deconfounding techniques that are hugely popular in econometrics.
- In the above figure, we are interested in estimating the causal effect of $X$ on $Y$. You can see that:
  - We cannot use the back- door criterion here because $U$ is unobserved.
  - We cannot use the front-door criterion either because there's no mediator between $X$ and $Y$.




- Instrumental variable methods require a special variable called an **instrument** to be present in a graph. We will use Z to denote the instrument
- Our effect of interest is the causal effect of X on Y.

### Three conditions of IVs

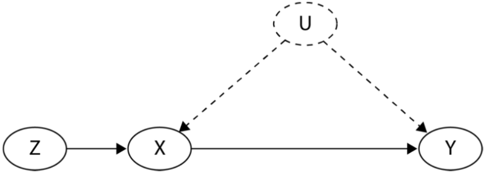
1. The instrument $Z$, is associated with $X$.
2. The instrument $Z$, doesn't affect $Y$ in any way except through $X$.
3. There are no common causes of $Z$ and $Y$.

- Think about the case where arrow between Z and X is reversed.

# More on IVs

- The first condition talks about association rather than causation.
- The nature of the relationship between $Z$ and $X$ determines how much information we'll be able to extract from our instrument.
- Theoretically speaking, we can even use instruments that are only weakly (non-directly) associated with X (in such a case, they are called proxy instruments).
- In certain cases, the only thing we'll be able to obtain will be the lower and upper bounds of the effect, and in some cases, these bounds might be very broad and, therefore, not very useful

# Example: SCM with structural equations




Our structural equations are as follows:

- $u \sim N(0,4)$ with truncation such as $u \geq 0$
- $z \sim N(0,2)$
- $u_x, u_y \sim N(0,2)$ $X := -0.2u + 0.7z + 0.1u_x$ $Y := 0.7u + 0.35x + 0.1u_y$

Hence, the true causal impact of $X$ on $Y$ is **0.35**.

# Using IV to estimate causal effect

In order to estimate the causal impact of $X$ on $Y$, we will compute the causal impact of:

- $Z$ **on** $Y$. $Z$ can effect Y through $Z \rightarrow X \leftarrow U \rightarrow Y$, however $X$ is a collider here, hence this path is blocked. Therefore only path remaining is $Z \rightarrow X \rightarrow Y$, so there is no con founding. For doing this we fit the model:

$$Y = a_0 + a_1 Z$$

- $Z$ **on** $X$: There is only one path so we fit the model:

$$X = b_0 + b_1 Z$$

Using these models, the overall causal effect of $X$ on $Y$ can be quantified as the ratio of coefficient of two models:

$$Causal(X \rightarrow Y) = a_1 / b_1$$

## CODE

Refer to the [Python notebook] for implementation of:

- Data generation for the given SCM
- Fitting models for causal impact of $Z$ on $Y$ and $Z$ on $X$, and accordingly computing the causal impact of $X$ on $Y$.

**Note**: The book asks to compute the causal impact of $Z$ on $Y$ and $X$ on $Y$ first, which is obviously incorrect.

Please note:

- Estimation using IVs can be extended to non-linear and non-parametric cases (for example, Li et al., 2022, and Carroll et al., 2004).
- All this makes IVs pretty flexible and broadly adopted, yet in practice, it might be difficult to find good instruments or verify that the necessary assumptions (for instance, a lack of influence of Z on Y other than through X) are met.