

Causalité

Vincent Arel-Bundock

Plan

- Théorie causale structurelle
- Graphes orientés acycliques
- Information causale vs. Information statistique
- Chemins
 - Fourchette
 - Chaîne
 - Collision
- Identification causale
- Exemples
- Simulations



Théorie causale structurelle

Théorie causale structurelle

Raisonnement contre-factuel:

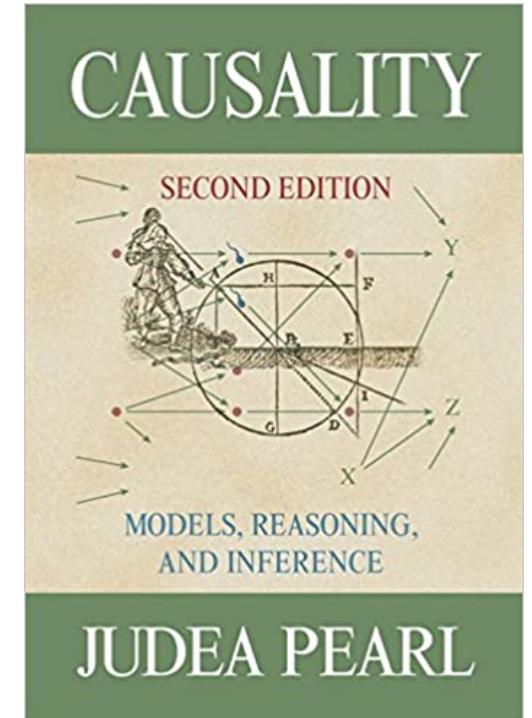
- Expérience de pensée des mondes alternatifs:
 - Avec aspirine
 - Avec placebo

Approche graphique:

1. Faire la liste des variables pertinentes pour la question de recherche.
2. Représenter graphiquement les relations causales entre ces variables.

Joie et allégresse!

- Est-ce que l'effet causal de X sur Y est identifiable?
- Le modèle statistique doit "ajuster" pour quelles variables?



Théorie causale structurelle: Exemple

Effet causal:

- Éducation → Idéologie politique

Variables pertinentes:

- Éducation
- Idéologie politique
- Revenu individuel
- Revenu des parents

Théorie causale structurelle: Exemple

Effet causal:

- Éducation → Idéologie politique

Variables pertinentes:

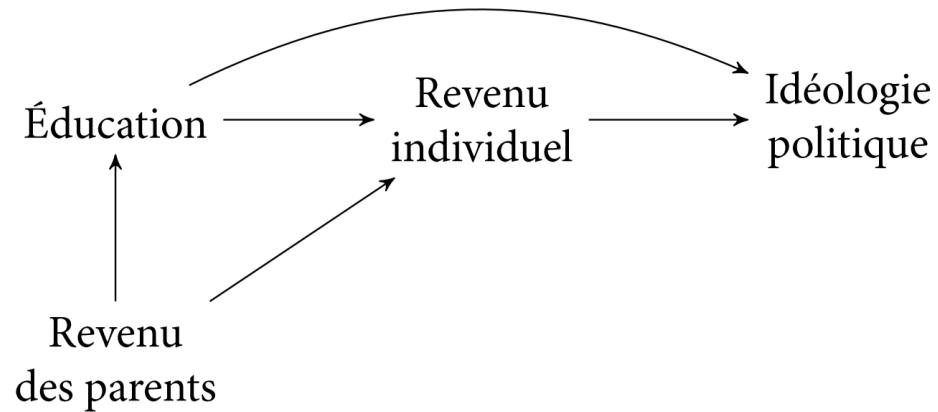
- Éducation
- Idéologie politique
- Revenu individuel
- Revenu des parents

Relations causales:

1. Éducation $\xrightarrow{+}$ Revenu individuel;
2. Revenu individuel $\xrightarrow{+}$ droite politique;
3. Revenu des parents $\xrightarrow{+}$ éducation;
4. Revenu des parents $\xrightarrow{+}$ Revenu individuel.

Source: Connaissances du chercheur, études antérieures, etc.

Théorie causale structurelle: Exemple

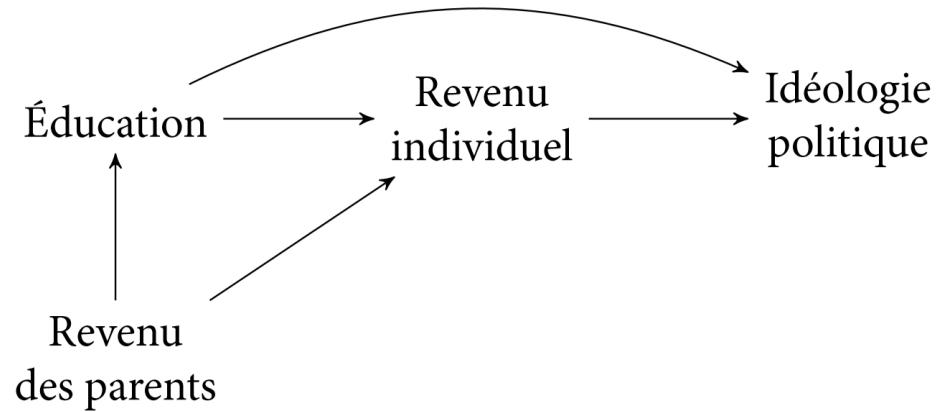


Relations causales:

1. Éducation $\xrightarrow{+}$ Revenu individuel;
2. Revenu individuel $\xrightarrow{+}$ droite politique;
3. Revenu des parents $\xrightarrow{+}$ éducation;
4. Revenu des parents $\xrightarrow{+}$ Revenu individuel.

Source: Connaissances du chercheur, études antérieures, etc.

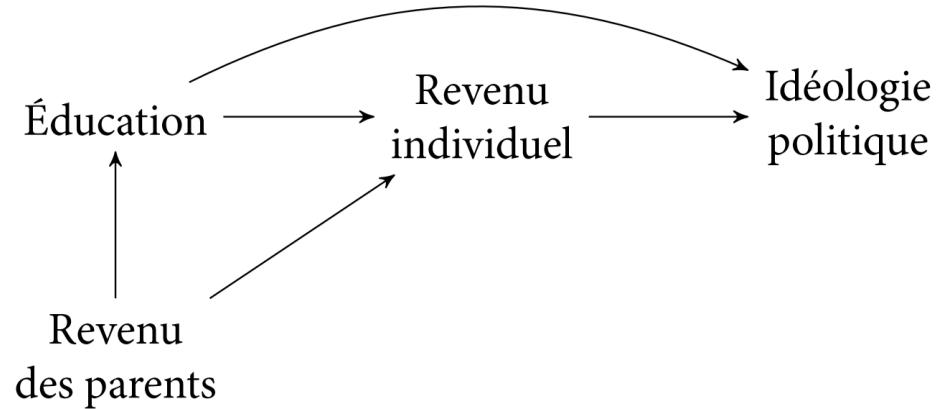
Théorie causale structurelle: Exemple



Quel modèle de régression multiple employer pour estimer l'effet causal de l'éducation sur l'idéologie?

- Aucun contrôle?
- Contrôle pour revenu individuel?
- Contrôle pour revenu des parents?
- Contrôle pour les deux?

Théorie causale structurelle: Exemple



Quel modèle de régression multiple employer pour estimer l'effet causal de l'éducation sur l'idéologie?

- ~~Aucun contrôle?~~
- ~~Contrôle pour revenu individuel?~~
- **Contrôle pour revenu des parents**
- ~~Contrôle pour les deux?~~

Pourquoi dessiner un graphique causal?

3 fonctions principales :

1. Transparency

- Révéler les postulats théoriques et les hypothèses de recherche de façon explicite et transparente.

2. Identification

- Est-il possible d'estimer l'effet causal de la variable indépendante sur la variable dépendante?

3. Modélisation

- Quelles variables de contrôle qui doivent être *incluses* dans un modèle de régression multiple, et quelles variables doivent en être *exclues*.



Graphes Orientés Acycliques

GRAPHES orientés acycliques

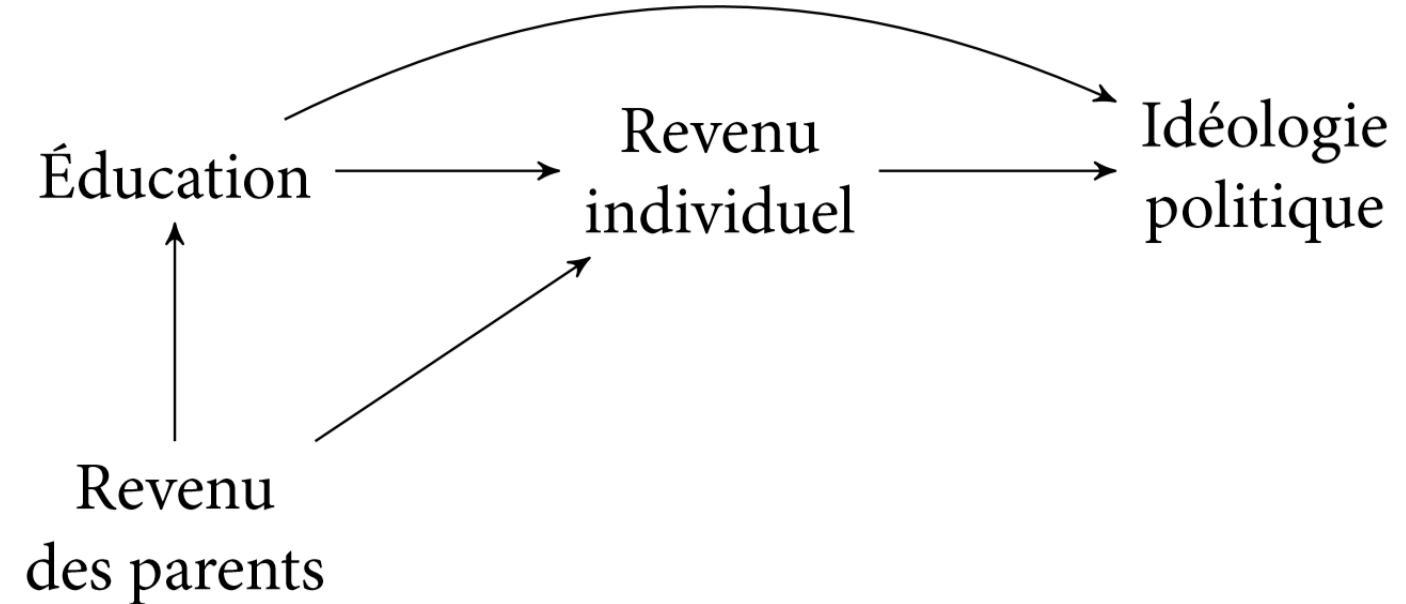
Représentation visuelle

Sommets (ou noeuds):

- Variables
- Phénomènes

Flèches:

- Relations causales



Graphes ORIENTÉS acycliques

Le GOA est *orienté* parce que:

- Les flèches indiquent la direction de la relation causale qui lie les variables.
- Les relations causales sont toujours unidirectionnelles.

A cause B , et non l'inverse:

$$A \rightarrow B$$

A cause B , B cause C , et A cause C :

$$A \rightarrow B \rightarrow C$$

Graphes ORIENTÉS acycliques

Le GOA est *orienté* parce que:

- Les flèches indiquent la direction de la relation causale qui lie les variables.
- Les relations causales sont toujours unidirectionnelles.

A cause B , et non l'inverse:

$$A \rightarrow B$$

A cause B , B cause C , et A cause C :

$$A \rightarrow B \rightarrow C$$

Définitions:

- Descendant
 - Une variable en aval dans le chemin
- Ancêtre
 - Une variable en amont dans le chemin

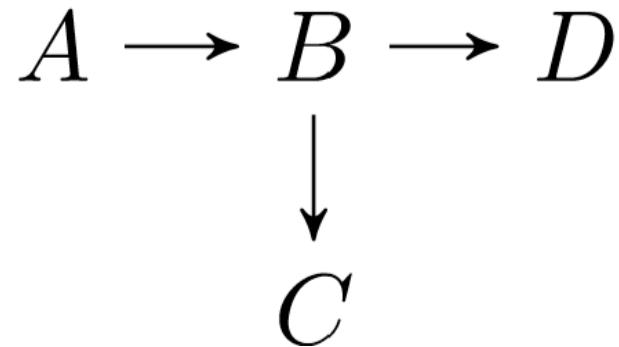
Un ancêtre cause un descendant, pas l'inverse.

Note: l'*absence* d'une flèche signifie qu'il n'y a *pas* de relation causale.

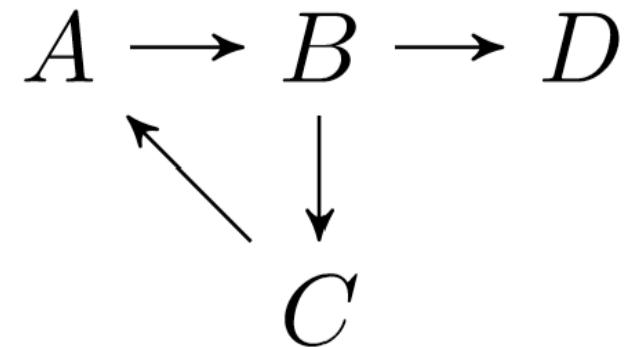
Graphes orientés ACYCLIQUES

Le GOA ne contient pas de chemin circulaire qui nous ramène au point de départ, et où toutes les flèches pointent dans la même direction.

Bon:



Mauvais:



Chemins



Effet causal vs. Information statistique

Pour analyser un GOA, il est utile de distinguer deux phénomènes: l'effet causal et l'information statistique

- L'effet causal circule de la cause à l'effet, mais jamais de l'effet à la cause.
- L'information statistique peut circuler dans les deux directions.

La cause peut nous donner de l'information sur l'effet, et l'effet peut nous donner de l'information sur la cause.

Effet causal vs. Information statistique

La pluie cause l'humidité du sol et non le contraire. La relation causale est unidirectionnelle.

Pluie → Humidité du sol

Par contre, si nous voyons que le sol est humide, nous pouvons déduire qu'il a plu récemment.

Le sol humide nous donne de l'information pertinente pour déduire (ou prédire) s'il y a eu de la pluie au cours des dernières heures. L'effet nous donne de l'information sur la cause.

Même si les relations causales sont toujours unidirectionnelles, l'information statistique peut parfois circuler dans les deux directions.

Effet causal vs. Information statistique

Les individus qui ont des comportements sexuels risqués sont plus nombreux à contracter une infection transmise sexuellement (ITS), et à sentir une brûlure à la miction:



La relation est unidirectionnelle. Par contre, l'information statistique circule dans les deux sens.

Effet causal vs. Information statistique

Les individus qui ont des comportements sexuels risqués sont plus nombreux à contracter une infection transmise sexuellement (ITS), et à sentir une brûlure à la miction:



La relation est unidirectionnelle. Par contre, l'information statistique circule dans les deux sens.

Chemin ouvert:

- Quand l'information statistique circule entre A et C .

Chemin fermé:

- Quand l'information statistique ne circule pas entre A et C .

Typologie des chemins



3 structures causales:

1. Fourchette: $A \leftarrow B \rightarrow C$
2. Chaîne: $A \rightarrow B \rightarrow C$
3. Collision: $A \rightarrow B \leftarrow C$

Est-ce que ces chemins sont ouverts ou fermés?

Fourchette: $A \leftarrow B \rightarrow C$

- 1 cause B
- 2 effets A et C

Mercure \longleftrightarrow Température \longrightarrow Crème glacée

Fourchette: $A \leftarrow B \rightarrow C$

- 1 cause B
- 2 effets A et C

Mercure \longleftrightarrow Température \longrightarrow Crème glacée

Fourchette = Ouverte

L'information statistique circule entre les extrémités:

- La colonne de mercure me permet de prédire le nombre de crèmes glacées vendues
- Le nombre de crèmes glacées vendues me permet de prédire la colonne de mercure

Fourchette: $A \leftarrow B \rightarrow C$

Mercure \longleftarrow Température \longrightarrow Crème glacée

Contrôler \approx "Connaître," "Observer" ou "Fixer"

Fourchette: $A \leftarrow B \rightarrow C$

Mercure \longleftarrow Température \longrightarrow Crème glacée

Contrôler \approx "Connaître," "Observer" ou "Fixer"

Si je connais la vraie température, la colonne de mercure est inutile pour prédire les crèmes glacées vendues.

Fourchette: $A \leftarrow B \rightarrow C$

Mercure \longleftarrow Température \longrightarrow Crème glacée

Contrôler \approx "Connaître," "Observer" ou "Fixer"

Si je connais la vraie température, la colonne de mercure est inutile pour prédire les crèmes glacées vendues.

Contrôler pour le maillon central d'une fourchette ferme le chemin.

L'information statistique ne circule plus.

Chaîne: $a \rightarrow b \rightarrow c$

Séquence de 2 relations causales:

1. A cause B
2. B cause C .



Chaîne: $a \rightarrow b \rightarrow c$

Séquence de 2 relations causales:

1. A cause B
2. B cause C .



Chaîne = Ouverte

L'information statistique circule entre les extrémités:

- Le comportement nous aide à prédire $\text{Pr}(\text{Brûlure})$
- La brûlure nous aide à prédire le comportement

Chaîne: $a \rightarrow b \rightarrow c$



Contrôler pour le maillon central d'une chaîne ferme le chemin.

Si un test sanguin démontre qu'une personne ne souffre pas d'une ITS, le comportement à risque ne nous aide plus à prédire si elle souffre de brûlure à la miction.

Lorsque nous contrôlons ou observons le maillon central d'une chaîne, ses deux extrémités ne "communiquent" plus: le chemin devient fermé.

Collision: $A \rightarrow B \leftarrow C$

- 2 causes A et C
- 1 effet B



Collision: $A \rightarrow B \leftarrow C$

- 2 causes A et C
- 1 effet B



Collision = Ouverte

L'information statistique ne circule pas entre les extrémités:

- Le biais d'un arbitre en faveur d'une équipe ne donne pas d'information sur la qualité du jeu de l'équipe.
- La performance d'une équipe nous en dit peu sur le biais potentiel de l'arbitre.

Collision: $A \rightarrow B \leftarrow C$



Contrôler pour le maillon central d'une collision ferme le chemin.

Exemple: on observe que l'équipe a gagné.

- Si une équipe a gagné même si elle a mal joué, les chances que l'arbitre soit biaisé sont plus élevées.
- Si une équipe a gagné même si l'arbitre n'était pas biaisé, les chances qu'elle ait bien joué sont plus hautes.

Connaître le maillon central d'une collision permet de faire le lien entre ses extrémités. Lorsque le maillon central est fixé, les deux extrémités "communiquent."

Fourchettes, chaînes et collisions

$A \leftarrow B \rightarrow C$ Ouvert

$A \rightarrow B \rightarrow C$ Ouvert

$A \rightarrow B \leftarrow C$ Fermé

Contrôle pour une variable dans un modèle de régression multiple, c'est tracer un cadre autour de la variable.

$A \leftarrow [B] \rightarrow C$ Fermé

$A \rightarrow [B] \rightarrow C$ Fermé

$A \rightarrow [B] \leftarrow C$ Ouvert

Un modèle statistique qui contrôle pour le maillon central du chemin renverse le flot d'information: la fourchette et la chaîne deviennent fermées, et la collision devient ouverte:

Chemins

Un chemin complexe est ouvert si et seulement si tous les maillons qui le composent sont ouverts.

Dès qu'un seul des maillons est fermé, le chemin dans son ensemble est fermé.

Exemple 1:

$$A \leftarrow B \leftarrow C \leftarrow D \rightarrow E \quad \text{Ouvert}$$

Exemple 2:

$$A \rightarrow B \leftarrow C \rightarrow D \rightarrow E \quad \text{Fermé}$$

Chemins

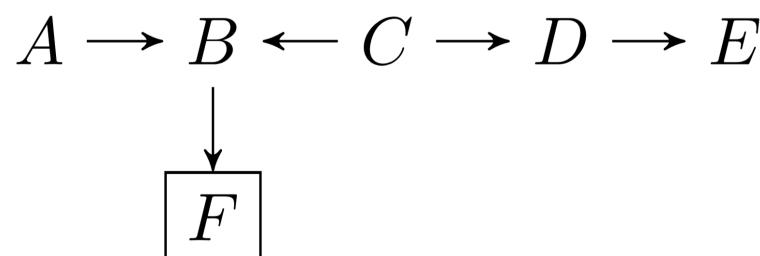
Si au moins un des maillons du chemin entre A et E est fermé, le chemin entier est fermé. Par exemple,

$A \leftarrow B \leftarrow C \leftarrow \boxed{D} \rightarrow E$ Fermé

$A \leftarrow \boxed{B} \leftarrow \boxed{C} \leftarrow \boxed{D} \rightarrow E$ Fermé

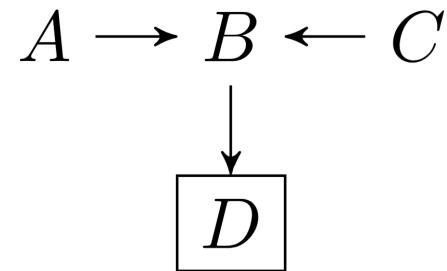
$A \rightarrow \boxed{B} \leftarrow C \rightarrow D \rightarrow E$ Ouvert

Puisque que contrôler pour le descendant d'une collision renverse le flot d'information, ce chemin est ouvert:



Cas spécial 1: Descendant d'une collision

Contrôler pour D ouvre le chemin entre A et C :



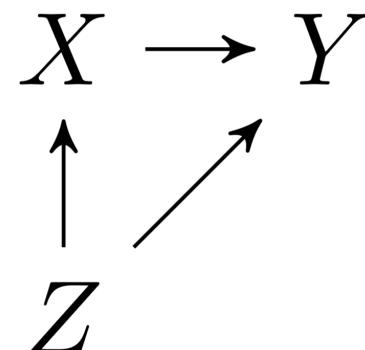
Le chemin entre A et C est fermé par la collision $A \rightarrow B \leftarrow C$.

Si on contrôle pour D , le flot d'information est renversé, et le chemin entre A et C devient (partiellement) ouvert.

Cas spécial 2: Chemin par la porte arrière

Définition:

1. Le chemin par la porte arrière lie la cause X à l'effet Y .
2. Une des extrémités du chemin pointe dans X .



Ce GOA comprend 2 chemins entre la cause X et l'effet Y :

1. $X \rightarrow Y$
2. $X \leftarrow Z \rightarrow Y$

Quel est le chemin par la porte arrière?

Est-ce que ce chemin est ouvert?

Et si je contrôle pour Z ?

Identification causale



Identification causale

Les deux conditions suivantes sont suffisantes pour que l'effet causal de X sur Y soit identifiable:

1. Le modèle statistique ne contrôle pas pour un descendant de X .

Quelles variables doit-on *exclure* du modèle?

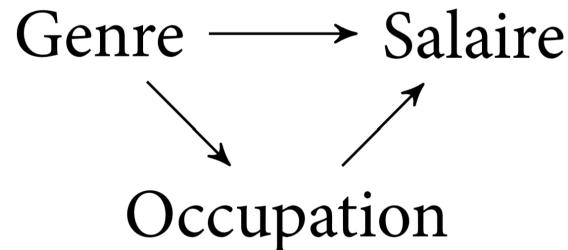
2. Tous les "chemins par la porte arrière" entre X et Y sont fermés.

Quelles variables doit-on *inclure* dans le modèle?

Condition #1: Ne pas contrôler pour les descendants de X

Un modèle statistique qui contrôle pour un descendant de X n'identifiera pas l'effet causal total de X sur Y .

Exemple:



Le genre peut avoir un effet sur le salaire à travers (au moins) deux chemins :

1. Une discrimination directe lors de la détermination des salaires.
2. Une "discrimination structurelle," qui passe indirectement à travers l'occupation.

Contrôler pour l'occupation nous donnerait un estimer "partiel" de l'effet "total" du genre sur les salaires.

Condition #2: Bloquer les chemins par la porte arrière

Un "chemin par la porte arrière" est un chemin qui remplit deux conditions:

1. Le chemin lie la cause X à l'effet Y .
2. Une des extrémités du chemin pointe dans X .

Quand les facteurs qui déterminent la valeur de X sont liés à Y (le chemin est ouvert), la condition #2 de l'identification causale est violée, et il est impossible d'estimer l'effet causal de X sur Y .

Condition #2: Bloquer les chemins par la porte arrière

Pour vérifier si tous les chemins par la porte arrière sont bloqués, il faut procéder en 2 étapes:

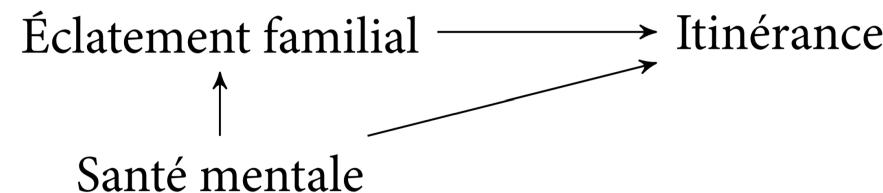
1. Trouver les chemins par la porte arrière
 - Faire la liste de tous les chemins qui lient la cause X à l'effet Y
 - Identifier les chemins dont une extrémité pointe dans X
2. Vérifier si ces chemins sont ouverts

Intuitivement, un chemin par la porte arrière représente "les causes de la cause" (le chemin pointe dans X).

Cette règle nous indique comment éliminer les relations fallacieuse ou les explications alternatives.

Condition #2: Bloquer les chemins par la porte arrière

Exemple: Effet causal de l'éclatement familial sur la probabilité qu'une personne devienne itinérante:



Ce GOA suggère qu'un trouble de santé mentale pourrait être une cause commune aux deux autres phénomènes; ce trouble pourrait augmenter la probabilité d'éclatement familial et d'itinérance :

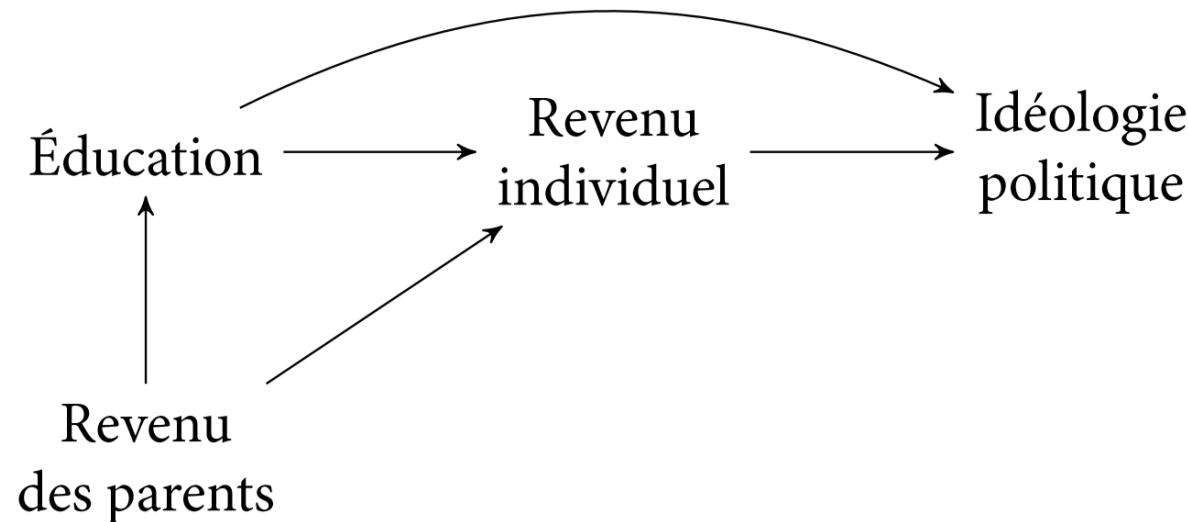
Éclatement familial ← **Santé mentale** → Itinérance

Pour estimer l'effet causal, il faut contrôler.

Exemples



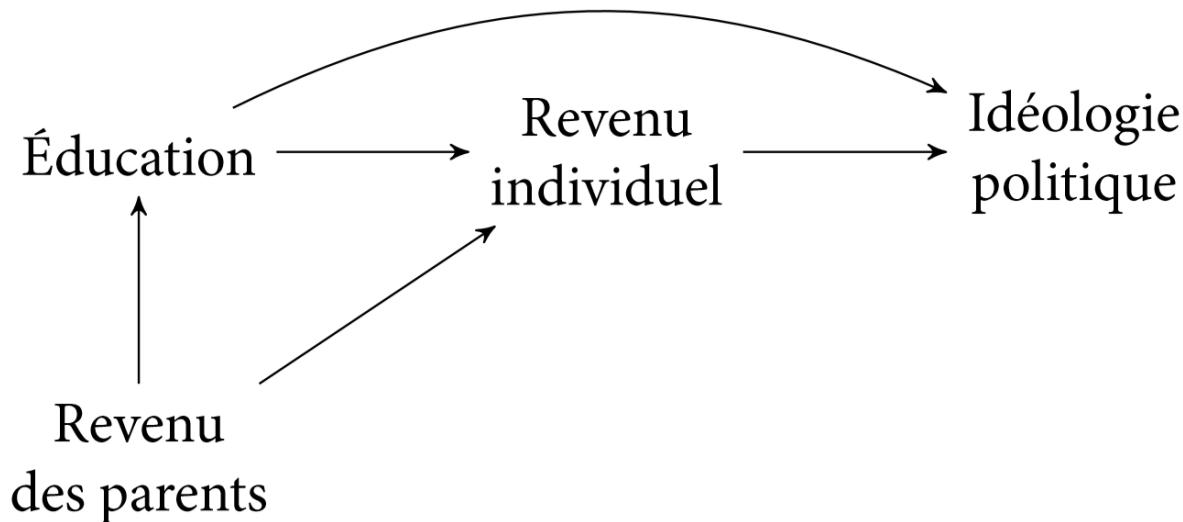
Éducation et idéologie politique



La règle #1 de l'identification causale dit qu'il ne faut pas contrôler pour les descendants de la cause qui nous intéresse ("Éducation").

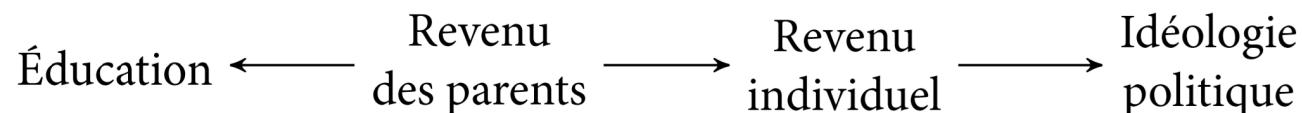
Notre modèle statistique ne devra *pas* contrôler pour la variable "Revenu individuel."

Éducation et idéologie politique

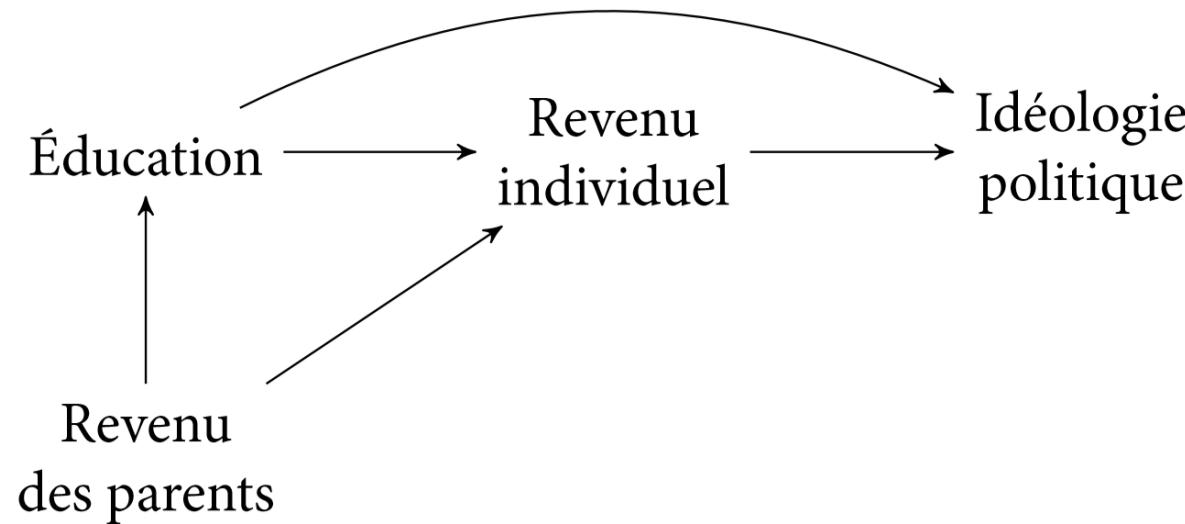


La règle #2 de l'identification causale dit qu'il faut fermer tous les chemins par la porte arrière.

Ici, il y a un chemin par la porte arrière entre la variable indépendante et la variable dépendante :



Éducation et idéologie politique

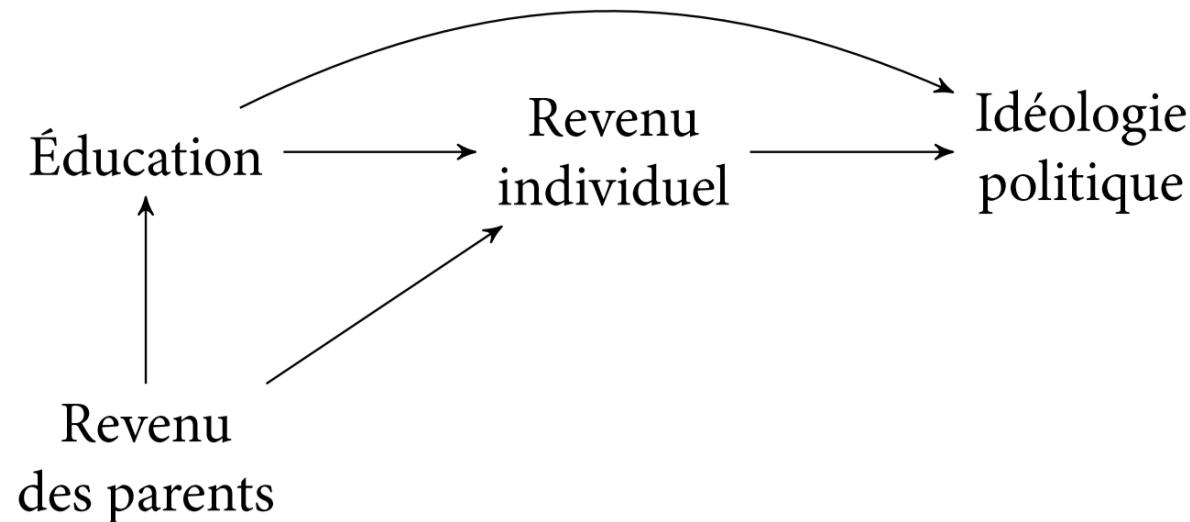


Pour fermer ce chemin, nous pourrions contrôler pour "Revenu des parents" ou pour "Revenu individuel."

Cependant, "Revenu individuel" est un descendant de "Éducation" : contrôler pour cette variable violerait la condition #1 de l'identification causale.

Identifier l'effet causal de "Éducation" sur "Idéologie politique" requiert que nous contrôlions pour la variable "Revenu des parents," et que nous ne contrôlions pas pour la variable "Revenu individuel."

Éducation et idéologie politique



Ces stratégies pourraient être exprimées par le modèle linéaire suivant:

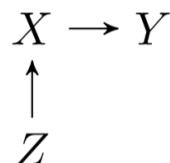
$$\text{Idéologie} = \beta_0 + \beta_1 \text{Éducation} + \beta_2 \text{Revenu des parents} + \varepsilon$$

Exemples

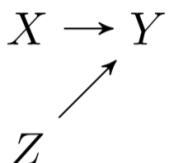
GRAPHIQUE 7.1. —

Six modèles qui permettent de produire un estimé non biaisé de l'effet causal de X sur Y . Les carrés identifient les variables qui doivent être contrôlées par le modèle statistique.

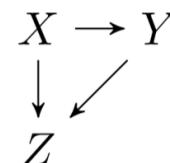
(A)



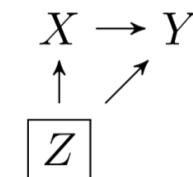
(B)



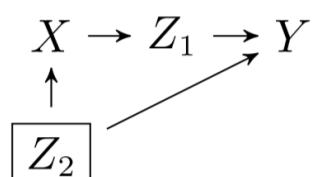
(C)



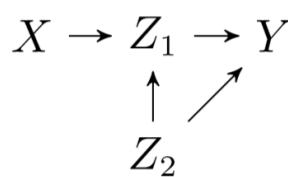
(D)



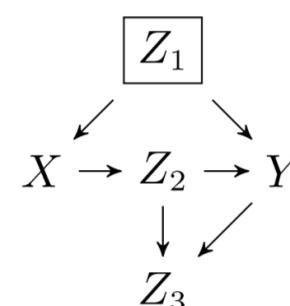
(E)



(F)



(G)



Exemples

$$X \rightarrow Y$$

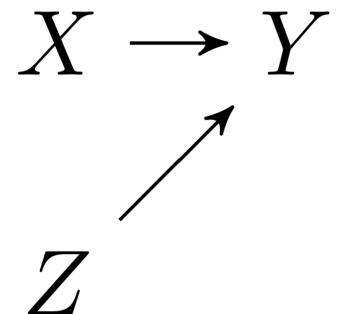
↑

$$Z$$

Il n'y a aucun chemin par la porte arrière : pas besoin de contrôler pour quoi que ce soit.

Afin d'estimer l'effet causal de X sur Y , il suffit de mesurer l'association bivariée entre ces variables.

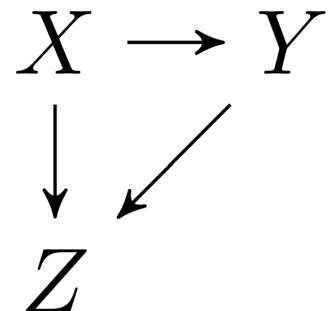
Exemples



Il n'y a aucun chemin par la porte arrière.

Il est inutile de contrôler ; la régression bivariée suffit.

Exemples



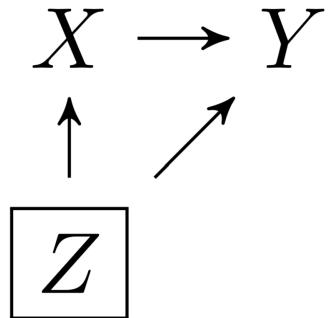
Deux chemins lient X à Y :

1. $X \rightarrow Y$. C'est le chemin causal qui nous intéresse.
2. $X \rightarrow Z \leftarrow Y$. C'est un chemin non causal.

Il n'y a aucun chemin par la porte arrière, car aucune extrémité ne pointe vers X .

Z est un descendant de X : contrôler pour Z violerait la première condition de l'identification causale, ouvrirait le flot d'information sur ce chemin non causal, et biaiserait nos résultats.

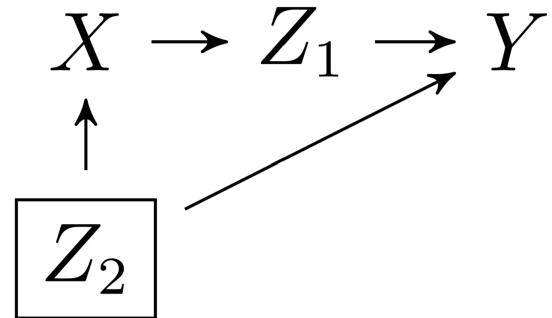
Exemples



Il y a un chemin ouvert par la porte arrière : $X \leftarrow Z \rightarrow Y$.

Il est essentiel de contrôler pour Z si on veut estimer l'effet causal de X sur Y .

Exemples

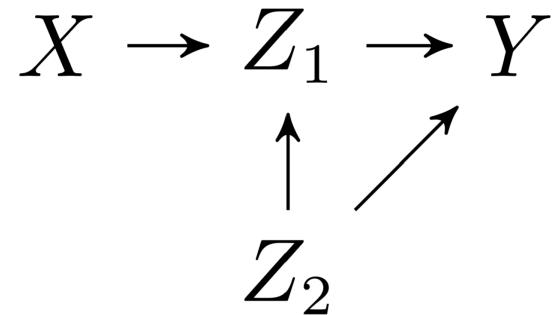


Il y a un chemin par la porte arrière: $X \leftarrow Z_2 \rightarrow Y$.

Il est essentiel de contrôler pour Z_2 si nous voulons estimer l'effet causal de X sur Y .

Il est aussi important de ne pas contrôler pour Z_1 , puisque cette variable est descendante de X .

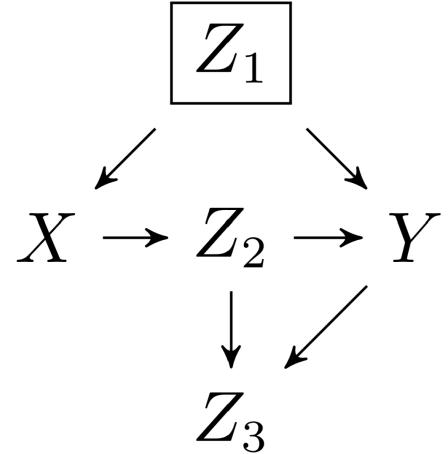
Exemples



Il n'y a pas de chemin ouvert par la porte arrière.

L'effet de X sur Y est identifiable, même si nous ne contrôlons pour aucune variable.

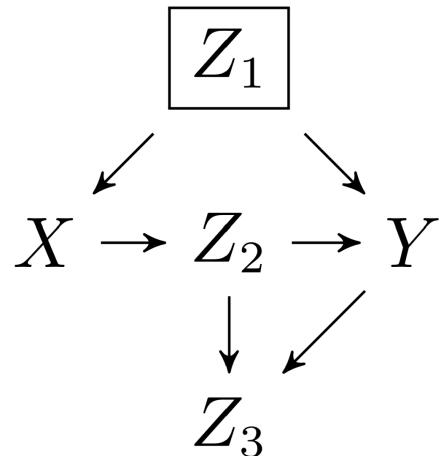
Exemples



Trois chemins lient les variables X et Y :

$$\begin{aligned} & X \rightarrow Z_2 \rightarrow Y \\ & X \rightarrow Z_2 \rightarrow Z_3 \leftarrow Y \\ & X \leftarrow Z_1 \rightarrow Y \end{aligned}$$

Exemples



Les variables Z_2 et Z_3 sont descendantes de X : il ne faut pas contrôler pour ces variables.

Le troisième chemin est ouvert, et se termine par une flèche qui pointe vers X . Il est essentiel de bloquer ce chemin par la porte arrière en contrôlant pour Z_1 .

Pour estimer l'effet causal de X sur Y , il faut contrôler pour Z_1 mais éviter de contrôler pour Z_2 ou Z_3 .

Simulations



Simulations dans R

La fonction `rnorm(n)` pige n nombres aléatoires dans une distribution normale.

```
rnorm(2)
```

```
## [1] -0.3573591 -0.1562602
```

```
rnorm(5)
```

```
## [1] 1.5050470 -0.6746082 -0.4212509 1.5089923 0.1607467
```

Simulations dans R

1 000 000 de nombres aléatoires

```
n <- 1000000
x <- rnorm(n)
```

```
x
```

```
## [1] -0.818235107 -1.053743458 -0.767747215 -1.292213028  0.195919291
## [6] -0.021567918  0.248218314 -0.310516393 -2.690361422 -1.567890774
## [11] -1.469900635 -0.448284418  0.703240794 -0.135033789  1.239794524
## [16]  0.114315128 -0.627329267  0.906613876  0.419354070 -0.704200338
## [21] -0.378695825 -0.814005125 -0.140967774 -0.155212044 -0.225369285
## [26]  0.826402849 -0.645247857  0.928260035 -0.069121420  2.153329130
## [31]  1.651351313  0.658844007 -0.612351690 -0.150719179  0.027866835
## [36] -1.319646993 -0.982339781 -1.188270369 -0.071317995 -2.050937313
## [41] -0.277320564  0.236186500  0.321099350  1.168991190  1.068908906
## [46]  1.620609924  0.363483203 -1.103337825  0.170465812 -1.995723878
## [51]  0.382998780  0.007205454  1.617768239 -0.514071427  0.436191887
## [56] -0.950773232  0.569958458  0.651799357  0.226756477 -0.207164635
## [61]  1.314671538 -0.002778056  0.196998001 -0.358561978  0.028559476
## [66] -0.861491603  0.143572022  0.213386508  1.984316168  0.586987171
```

Simulations dans R

Deux variables indépendantes...

```
X <- rnorm(n)  
Y <- rnorm(n)  
  
mod <- lm(Y ~ X)  
coef(mod)  
  
##   (Intercept)      X  
## -0.0018424669 -0.0005682175
```

...sont indépendantes.

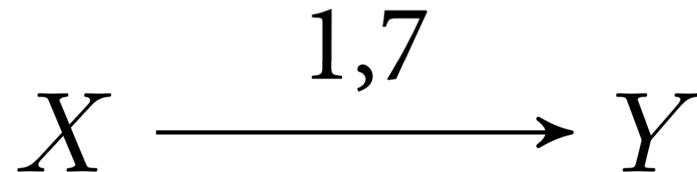
Simulation 1: Relation causale simple

Dans ce modèle, une augmentation d'une unité de X cause une augmentation de 1,7 unité de Y

$$X \xrightarrow{1,7} Y$$

Simulation 1: Relation causale simple

Dans ce modèle, une augmentation d'une unité de X cause une augmentation de 1,7 unité de Y .



```
n <- 100000
x <- rnorm(n)
y <- 1.7 * x + rnorm(n)
```

La fonction `lm` estime un modèle de régression linéaire:

```
mod <- lm(Y ~ X)
coef(mod)
```

```
## (Intercept)          X
## 0.005231824 1.704522418
```

Simulation 2: Chaîne

Dans une chaîne de relations causales linéaires,
l'effet causal est égal au produit des effets
individuels.



Le véritable effet de X sur Y dans ce GOA est
 $3 \cdot 0,5 = 1,5$

Simulation 2: Chaîne

Dans une chaîne de relations causales linéaires,
l'effet causal est égal au produit des effets
individuels.



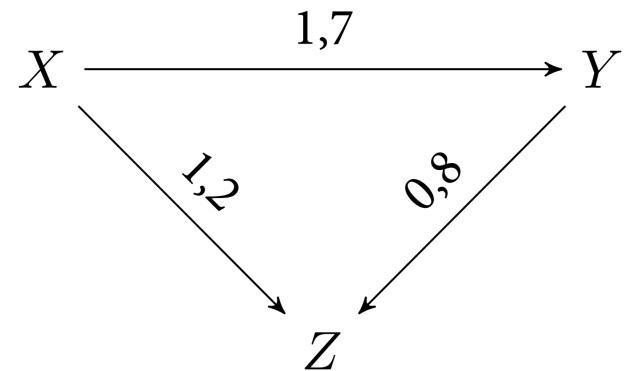
Le véritable effet de X sur Y dans ce GOA est
 $3 \cdot 0,5 = 1,5$

```
X <- rnorm(n)
Z <- 3 * X + rnorm(n)
Y <- 0.5 * Z + rnorm(n)
```

```
mod <- lm(Y ~ X)
coef(mod)
```

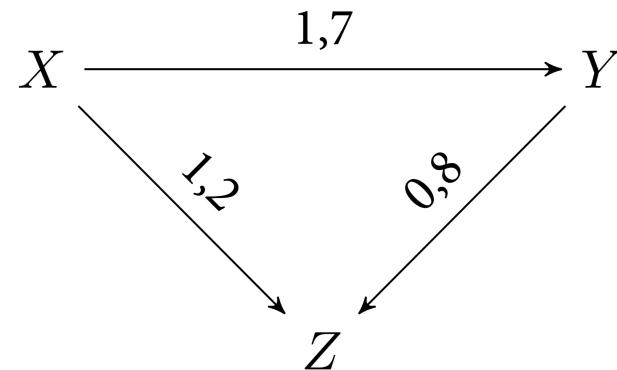
```
##   (Intercept)          X
## -0.001429681  1.500174936
```

Simulation 3: Collision



Dans ce modèle, il n'y a pas de chemin par la porte arrière : aucun contrôle n'est requis.

Simulation 3: Collision



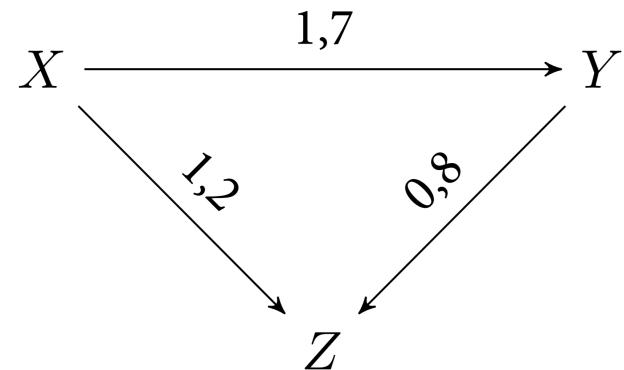
Dans ce modèle, il n'y a pas de chemin par la porte arrière : aucun contrôle n'est requis.

```
X <- rnorm(n)
Y <- 1.7 * X + rnorm(n)
Z <- 1.2 * X + 0.8 * Y + rnorm(n)

mod <- lm(Y ~ X)
coef(mod)
```

```
## (Intercept)          X
## 0.004385199 1.699741417
```

Simulation 3: Collision



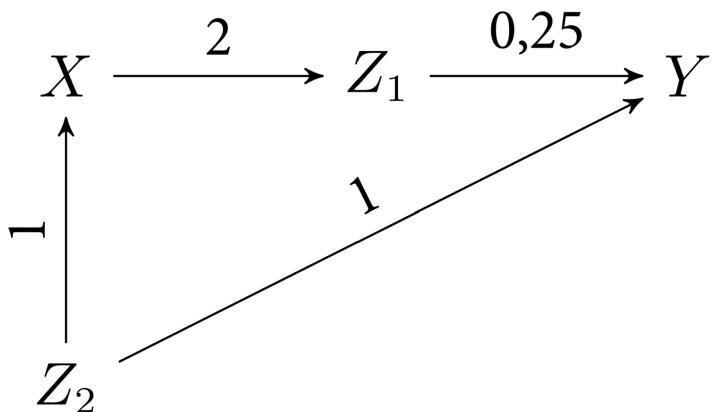
Z est un descendant de X : il est important de ne pas contrôler pour cette variable.

```
X <- rnorm(n)
Y <- 1.7 * X + rnorm(n)
Z <- 1.2 * X + 0.8 * Y + rnorm(n)

mod <- lm(Y ~ X + Z)
coef(mod)
```

	(Intercept)	X	Z
##	-0.004746742	0.456157856	0.485868711

Simulation 4: Fourchette



Le vrai effet causal de X sur Y est égal à:
 $2 \cdot 0,25 = 0,5$.

```
n <- 100000
Z2 <- rnorm(n)
X <- Z2 + rnorm(n)
Z1 <- 2 * X + rnorm(n)
Y <- 0.25 * Z1 + Z2 + rnorm(n)

mod <- list()
mod[[1]] <- lm(Y ~ X + Z2)
mod[[2]] <- lm(Y ~ X)
mod[[3]] <- lm(Y ~ X + Z2 + Z1)
```

Il faut:

- Contrôler pour Z_2 : Porte arrière
- Ne pas contrôler pour Z_1 : Descendant

Simulation 4: Fourchette

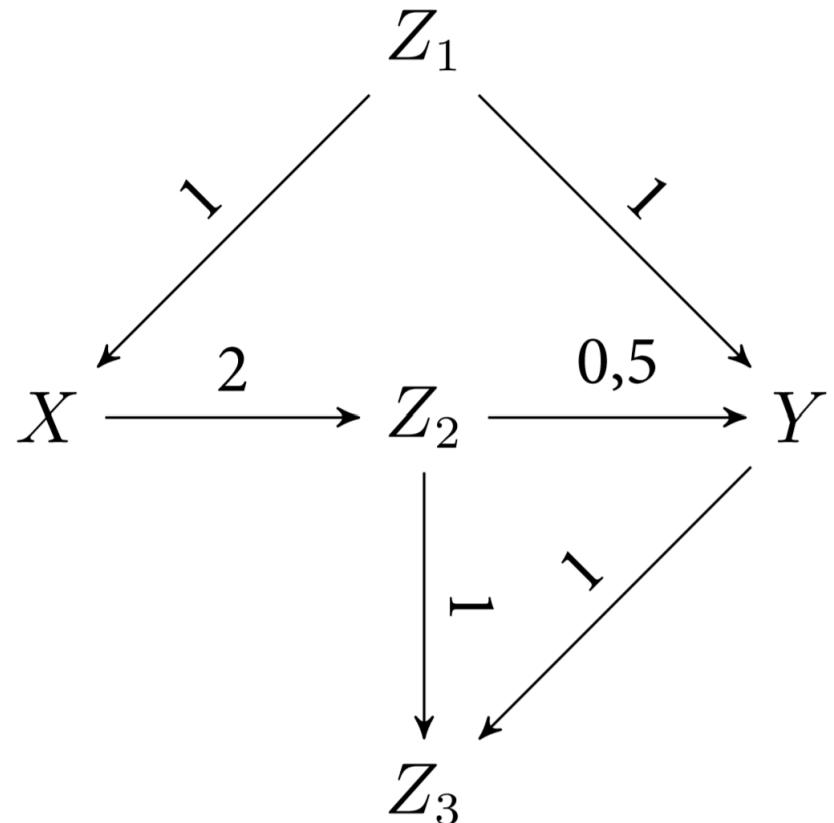
```
library(modelsummary)
```

```
## Loading required package: tables  
##  
## Attaching package: 'modelsummary'  
##  
## The following object is masked from 'package:tables':  
##  
##     All
```

```
modelsummary(mod)
```

	Model 1	Model 2	Model 3
(Intercept)	0.001	-0.001	0.001
	(0.003)	(0.004)	(0.003)
X	0.504	0.998	0.002
	(0.003)	(0.003)	(0.007)

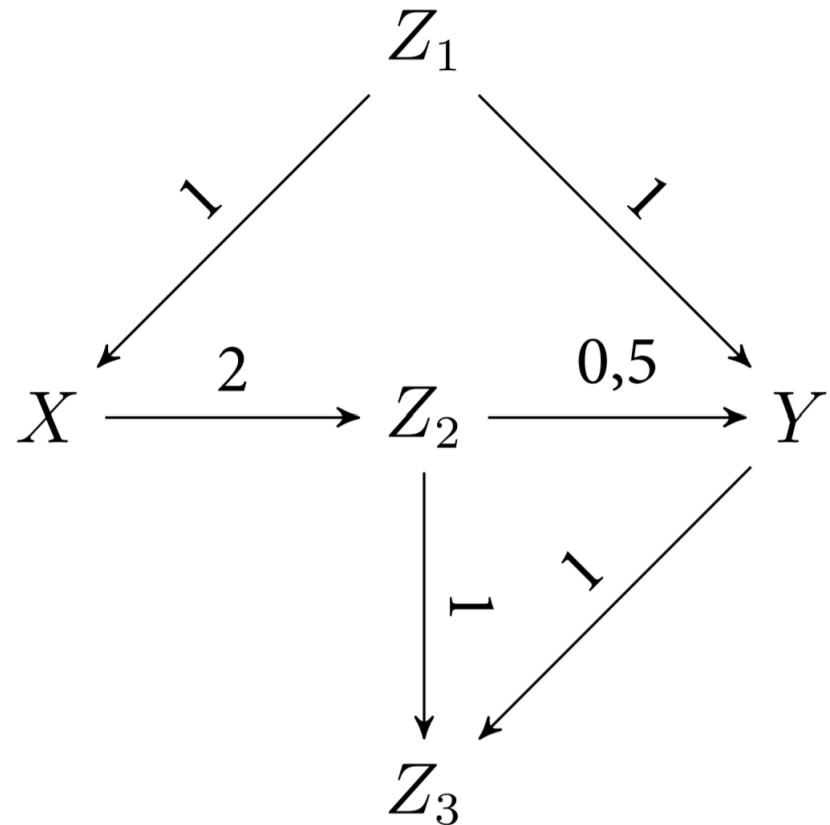
Simulation 5: Modèle complexe



Le vrai effet causal de X sur Y est égal à $2 \cdot 0.5 = 1$.

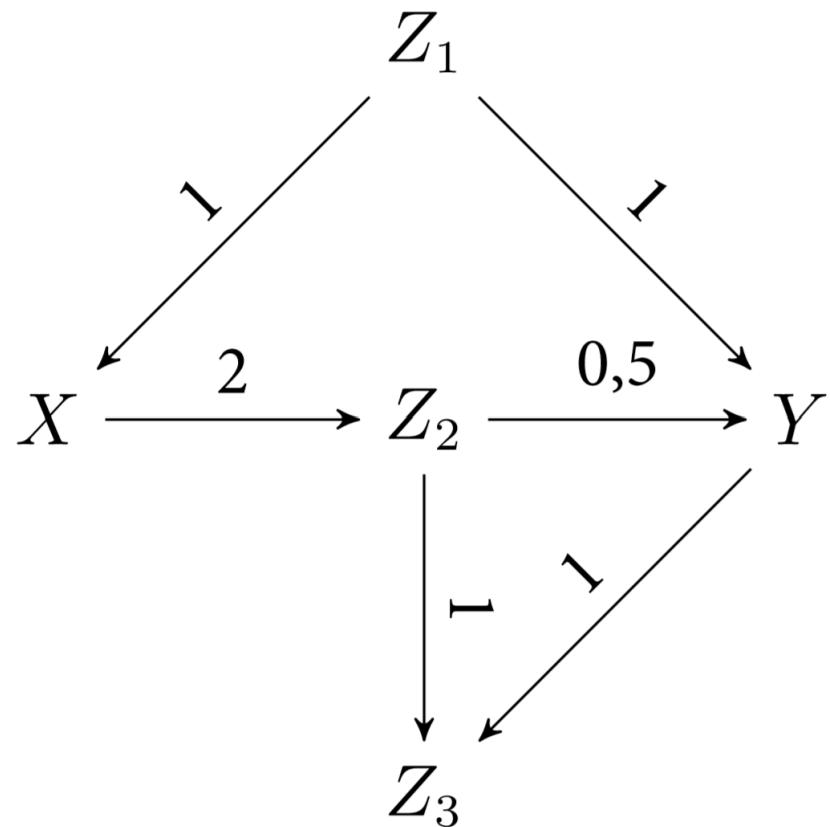
Un modèle statistique non biaisé devait *inclure* Z_1 , mais *exclure* Z_2 et Z_3 .

Simulation 5: Modèle complexe



```
Z1 <- rnorm(n)
X <- Z1 + rnorm(n)
Z2 <- 2 * X + rnorm(n)
Y <- 0.5 * Z2 + Z1 + rnorm(n)
Z3 <- Z2 + Y + rnorm(n)
```

Simulation 5: Modèle complexe



Nous estimons huit modèles avec différentes combinaisons des trois variables de contrôle Z_1 , Z_2 , et Z_3 :

```
mod <- list()
mod[[1]] <- lm(Y ~ X + Z1)
mod[[2]] <- lm(Y ~ X)
mod[[3]] <- lm(Y ~ X + Z2)
mod[[4]] <- lm(Y ~ X + Z3)
mod[[5]] <- lm(Y ~ X + Z1 + Z2)
mod[[6]] <- lm(Y ~ X + Z1 + Z3)
mod[[7]] <- lm(Y ~ X + Z2 + Z3)
mod[[8]] <- lm(Y ~ X + Z1 + Z2 + Z3)
```

Simulation 5: Modèle complexe

Nous estimons huit modèles avec différentes combinaisons des trois variables de contrôle Z_1 , Z_2 , et Z_3 :

```
modelsummary(mod)
```

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
(Intercept)	0.002	0.003	0.002	-0.001	0.000	-0.001	-0.000
	(0.004)	(0.004)	(0.004)	(0.003)	(0.003)	(0.002)	(0.002)
X	1.002	1.499	0.493	-0.154	-0.000	-0.229	0.204
	(0.004)	(0.003)	(0.008)	(0.005)	(0.007)	(0.004)	(0.005)
Z1	0.997				0.995	0.581	
	(0.005)				(0.004)	(0.003)	
Z2			0.503		0.501		-0.397
			(0.004)		(0.003)		(0.003)
Z3				0.472		0.411	0.597
						63 / 69	

Leçons



4 Leçons importantes

1. Quand nous connaissons le GOA, il est facile de voir quelles variables on doit inclure dans le modèle de régression.
2. Expériences aléatoires = WOW!
3. Biais post-traitement
4. Interprétation des variables de contrôle

Expériences aléatoires

Les expériences aléatoires sont souvent considérées comme le "Gold Standard" de l'inférence causale.

Si la valeur du traitement est déterminée de façon purement aléatoire, alors le traitement n'a aucun ancêtre.

$$X \rightarrow Y$$

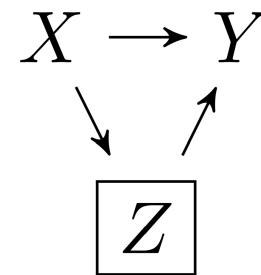
Aucune des flèches du GOA ne pointe vers la cause, et il n'existe aucun chemin par la porte arrière.

Dans une expérience où la valeur du traitement est aléatoire, la condition #2 de l'identification causale est *satisfait automatiquement!*

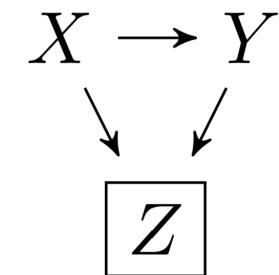
Biais post-traitement

Un "biais post-traitement" survient lorsqu'on contrôle pour un descendant de la cause, ce qui viole la première condition de l'identification causale.

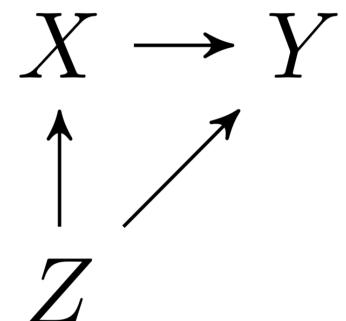
Bloquer un chemin causal.



Ouvrir un chemin *non* causal.



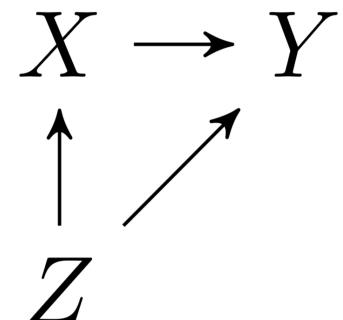
Ne pas interpréter les variables de contrôle



Pour estimer l'effet causal de X , il faut contrôler pour Z :

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon$$

Ne pas interpréter les variables de contrôle



Pour estimer l'effet causal de X , il faut contrôler pour Z :

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon$$

Pour estimer l'effet causal de Z , il ne faut pas contrôler pour X :

$$Y = \alpha_0 + \alpha_1 Z + \nu$$

Deux modèles différents!

UNITED LEADERSHIP

for a

STRONGER AMERICA

JEB2016.COM

