

Les graphes orientés acycliques (GOA)

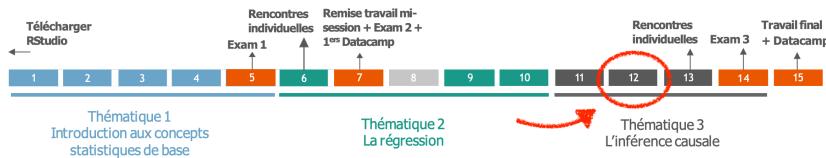
Méthodologie quantitative POL-2000

Laurence-Olivier M. Foisy & Adrien Cloutier

Aujourd'hui

Cours 12

- Présentation: Introduction à la causalité: Le problème fondamental de l'inférence causale
- Les graphes orientés acycliques (GOA)
 - Termes clés
 - Les types de chemins
 - Exemples



Les graphes orientés acycliques (GOA)

Pourquoi?

Les trois fonctions des GOA

1. Déterminer si c'est possible d'identifier l'effet causal de la VD sur la VI
2. Identifier les variables de contrôle qui doivent être incluses dans le modèle statistique
3. Dessiner des GOA force le chercheur à révéler ses postulats théoriques de façon explicite et transparente

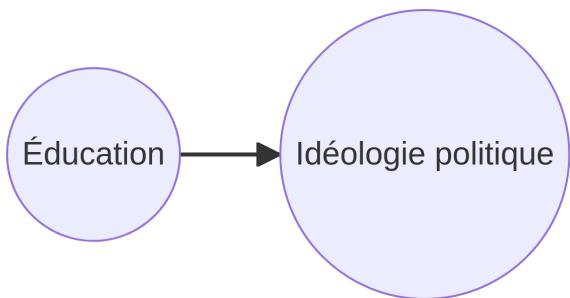
Comment?

- Par la connaissance, la littérature, l'intuition

Produire un GOA:

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

La relation entre l'éducation et l'idéologie politique



Étape 1: Faire la liste des variables pertinentes

1. Éducation augmente le revenu individuel
2. Revenu individuel augmente l'appui aux partis politiques qui promettent des baisses d'impôts
3. Revenu des parents augmente l'éducation de leurs enfants
4. Revenu des parents augmente le revenu de leurs enfants

Effet causal, ce qu'on veut étudier:



Variables pertinentes:

Éducation, Variable indépendante, cause

Idéologie politique, Variable dépendante, effet

Revenu des parents

Revenu individuel

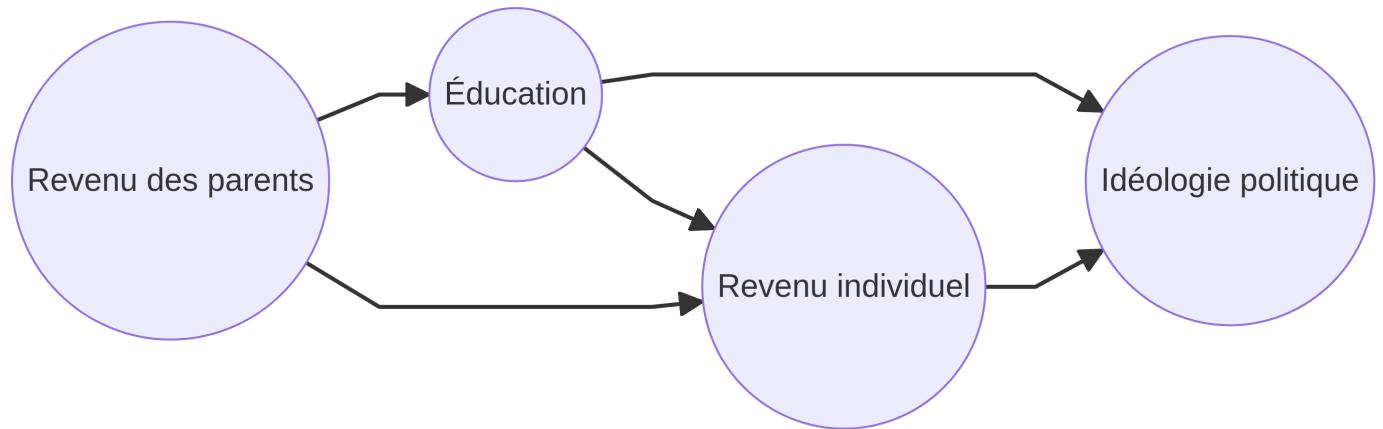
Étape 2: Représenter les relations causales

Sources: la littérature scientifique, les connaissances du chercheur, la théorie, la logique, la passion, etc.

Relations causales:



Étape 3: Dessiner le GOA



Les deux caractéristiques des GOA

1. Orientés

2. Acycliques

Ces deux caractéristiques sont essentielles pour la validité causale des GOA. La preuve mathématique fonctionne seulement si ces deux conditions sont respectées.

Orientés

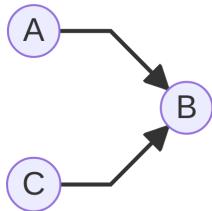
Les relations causales d'un GOA sont toujours unidirectionnelles

Quand deux relations causales se suivent, on dit qu'elles forment un chemin



- C est une variable descendante de A et B est une variable descendante de A
- A et B sont l'ancêtre de C

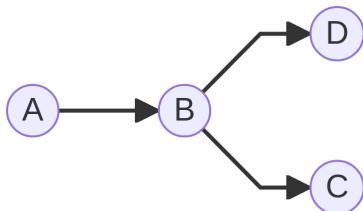
Si il n'y a pas de chemin entre A et C alors il n'y a pas de causalité



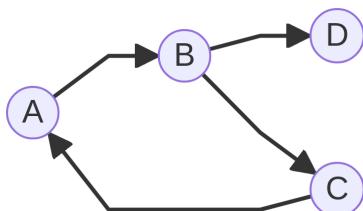
Acycliques

- Il ne peut pas y avoir de circuit
- Les flèches ne doivent pas nous faire revenir sur nos pas, pas de loops

Bon exemple:



Mauvais exemple:

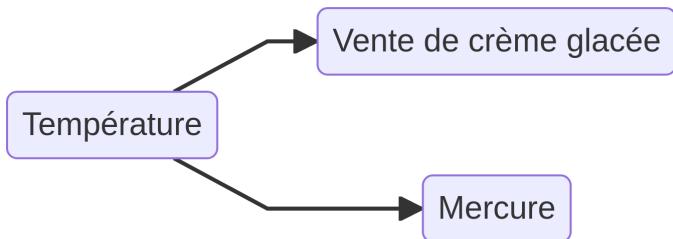


Effet causal vs information statistique



Effet causal vs information statistique

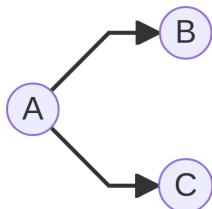
La causalité circule dans un sens seulement mais l'information statistique peut circuler dans les deux sens



Savoir que le mercure est élevé peut donner une bonne indice que la vente de crème glacée va augmenter mais ça ne veut pas dire que le mercure cause la vente de crème glacée

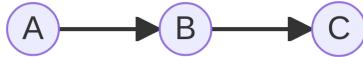
Trois types de chemins

1. Fourchette



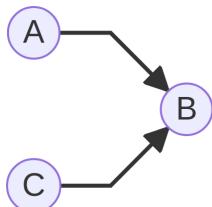
Ouvert!

2. Chaîne



Ouvert!

3. Collision



Fermé!

Conclusions

1. Les chaînes et fourchettes sont ouvertes mais les collisions sont fermées
2. Quand un modèle de régression contrôle le maillon central, il renverse le flot d'information.
Un chemin fermé devient ouvert et un chemin ouvert devient fermé

Chemin ouvert ou fermé

- Ouvert == l'information statistique circule entre A et C.
- Fermé == l'information statistique ne circule **PAS** entre A et C.

Contrôler pour le maillon central d'un chemin RENVERSE le flot d'information:

- Chemin ouvert devient fermé.
- Chemin fermé devient ouvert.



Chemin ouvert ou fermé

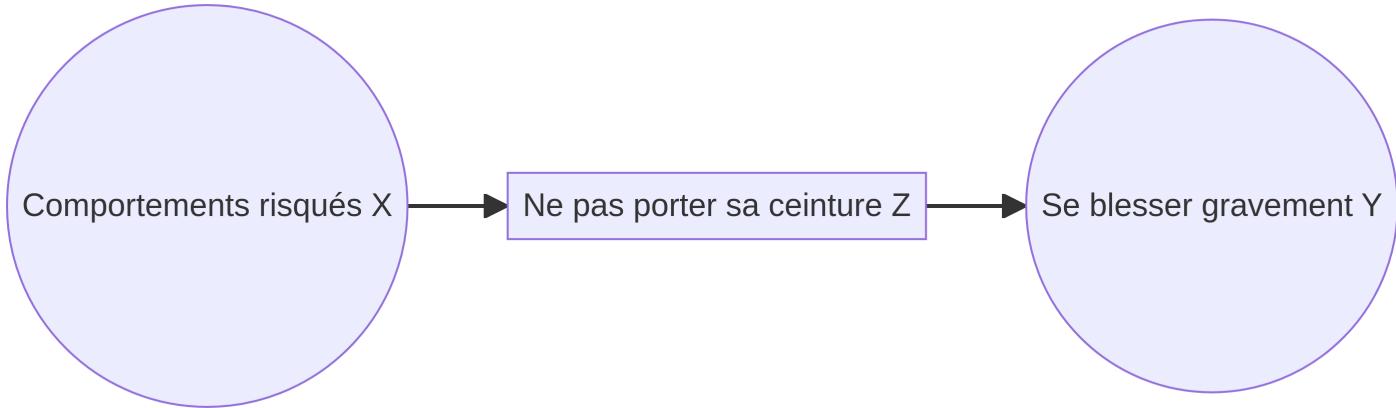
Pour signifier qu'on contrôle pour une variable, on l'encadre

Non contrôlé



L'information statistique circule

Contrôlé



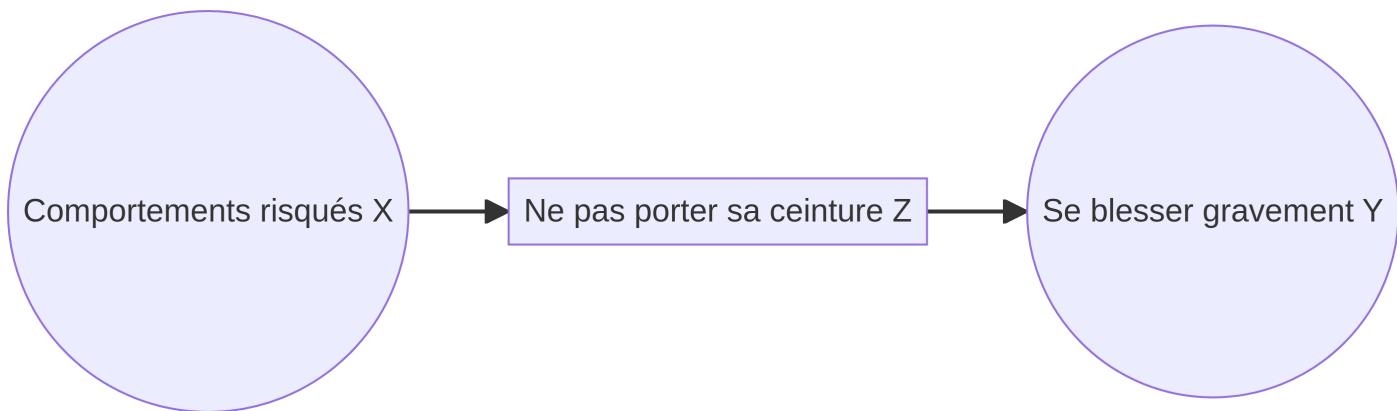
L'information statistique ne circule plus

Chemin ouvert ou fermé



Ici le chemin est ouvert donc l'information statistique circule entre les deux extrémité. Savoir que le comportement est risqué donne de l'information sur les chances de se blesser gravement et savoir que la personne est blessée gravement indique qu'il y a peut-être eu un comportement risqué.

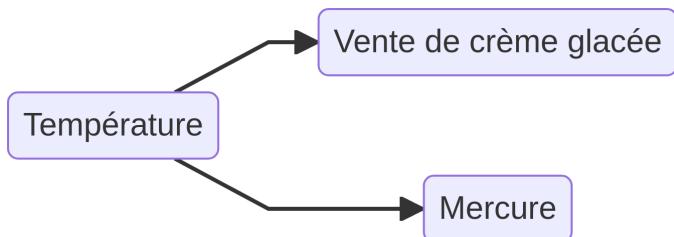
Chemin ouvert ou fermé



Ici le chemin est fermé donc l'information statistique ne circule pas entre les deux extrémité.
Si on sait que la personne portait sa ceinture et qu'elle s'est blessée gravement, ça ne nous donne pas d'information sur le comportement risqué.

Fourchette

Une fourchette c'est une cause avec deux effets



L'information statistique circule entre les deux extrémité donc le chemin est ouvert

Si on contrôle le maillon central, on ferme le chemin

Quand on fixe le maillon central, la température, les variation de mercure nous donne plus aucune information sur les ventes de crème glacée

Chaîne

Une séquence de deux relations causales



Une chaîne est ouverte, l'information y circule entre les extrémités

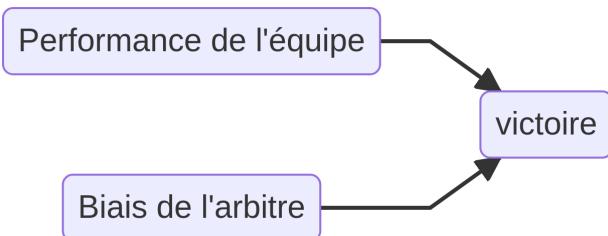
Connaître la cause donne de l'info sur l'effet

Connaitre l'effet donne de l'info sur la cause

Contrôler le maillon central ferme le chemin. Si je sais qu'un patient a une maladie cardio-vasculaire mais pas de diabète alors je sais que la cause n'est pas la surconsommation de sucre

Collision

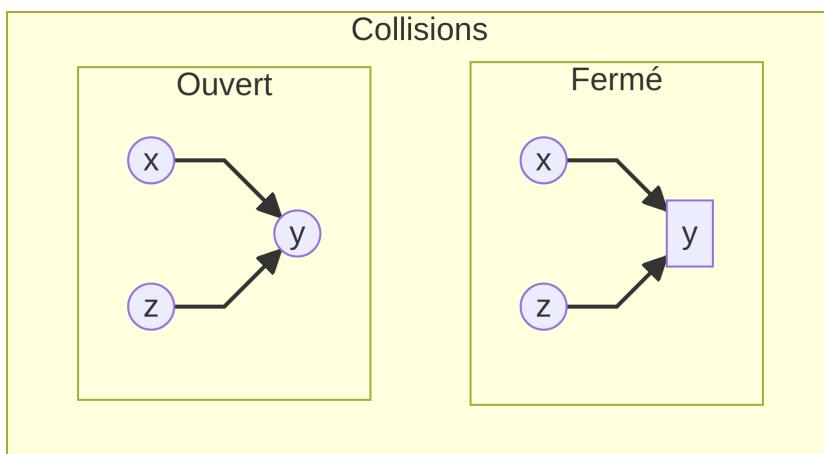
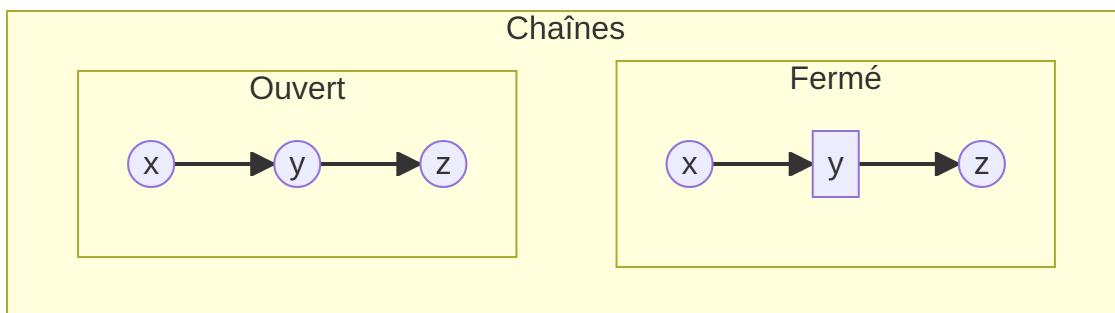
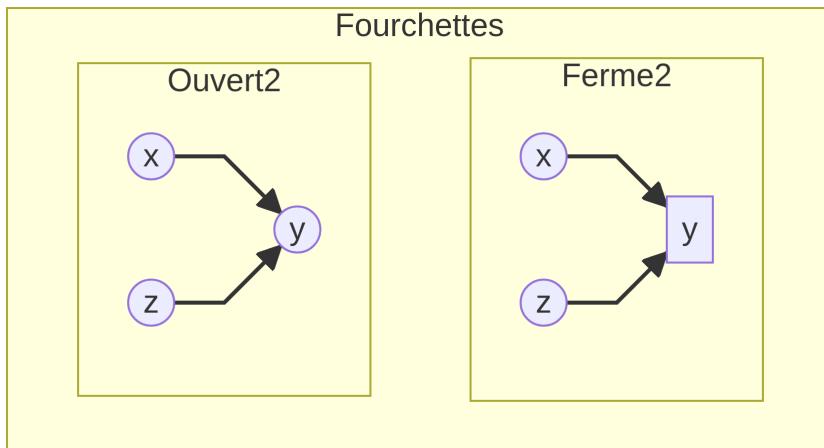
Deux causes et un effet



La collision est fermée parce que l'info statistique ne circule pas des deux bords.

Par contre si on contrôle le milieu, Admetton qu'on sait que la performance de l'équipe était nulle et qu'ils ont gagné, ça peut indiquer que l'arbitre était potentiellement biaisé

Fourchettes, Chaînes et collisions



Un chemin peut être composé d'une infinité de combinaison de fourchettes, chaînes, et collisions.
Le chemin est ouvert si tout est ouvert et fermé si un des lien est fermé.

Les contrôles

Les deux conditions

1. Ne pas contrôler pour un descendant de X
 - Permet de déterminer quels variables exclure du modèle statistique
2. Fermer tous les chemins par porte arrière (Contrôler les maillons centraux)
 - Permet de déterminer quelles variables inclure dans le modèle statistique

Ne pas contrôler les descendants de X

Disons que le genre influence le salaire et l'occupation et que l'occupation influence le salaire. Si on contrôle pour l'occupation, on vient juste observer l'effet du genre sur le salaire mais on obstrue l'effet du genre sur l'occupation sur le salaire. On coupe une part importante de l'effet du genre sur le salaire.



Fermer les chemins par la porte arrière

C'est quoi?

1. Un chemin qui lie X à Y
2. Une des extrémités pointe vers X

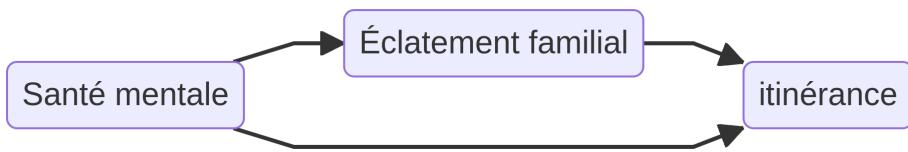
Les backdoors sont les causes de notre cause. On ne veut pas voir les effets des causes sur nos causes.

Fermer les backdoors

1. Faire une liste de tous les chemins qui lient la cause à l'effet
2. Est-ce que certains chemins pointent vers X?
3. Est-ce que ces chemins sont ouverts ?

Fermer les chemins par la porte arrière (Suite)

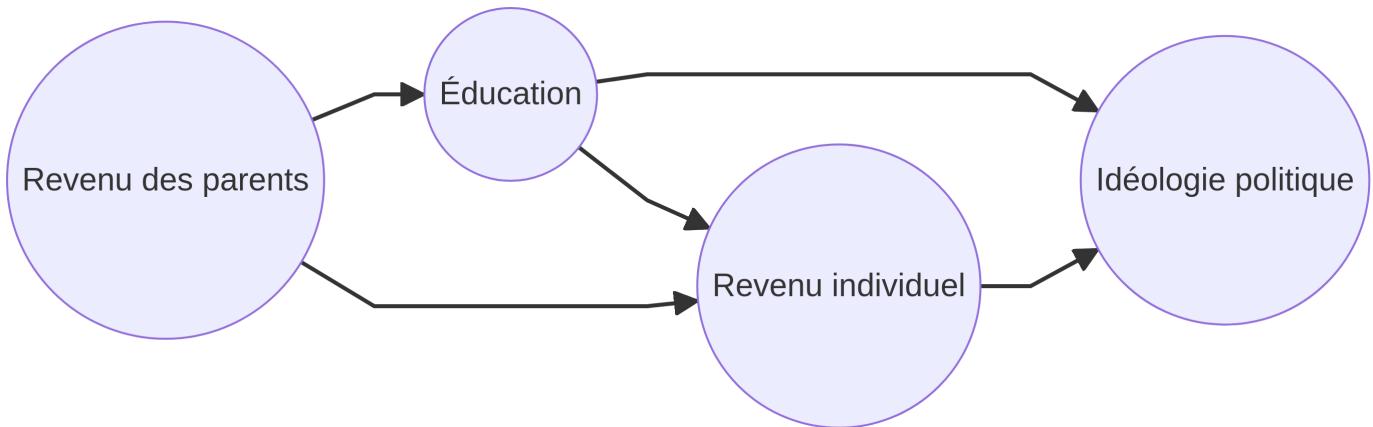
Relations de l'éclatement familial sur l'itinérance



Ici si on contrôle pour la santé mentale. On sait que c'est vrm l'éclatement familiale qui a causé l'itinérance et non la santé mentale qui a causé l'éclatement qui a causé l'itinérance

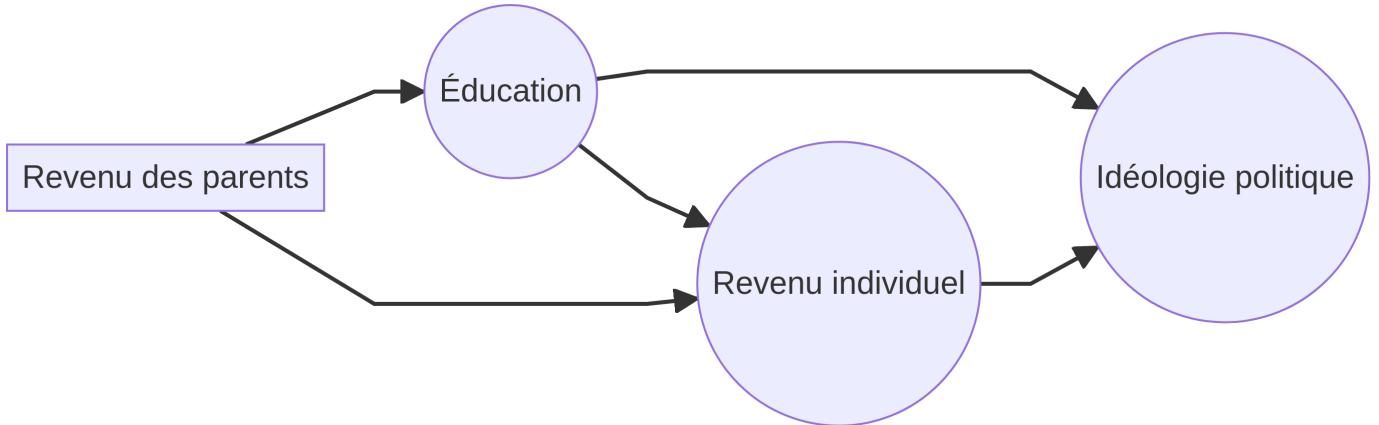
Exemple

1. Ne pas contrôler pour un descendant de X
2. Fermer tous les chemins par porte arrière



Exemple (Réponse)

1. Ne pas contrôler pour un descendant de X
2. Fermer tous les chemins par porte arrière



De la pratique

Comment déterminer si un chemin est ouvert ou fermé

- Les chaînes et les fourchettes sont ouvertes
- Les collisions sont fermées
- Un chemin composé uniquement de chaînes et de fourchettes est ouvert
- Dès qu'il y a au moins une collision, le chemin est fermé

Ouvert ou fermé?

$$X \rightarrow Z_1 \rightarrow Z_2 \rightarrow Z_3 \rightarrow Y$$

Ouvert!

$$X \rightarrow Z_1 \rightarrow Z_2 \rightarrow Z_3 \rightarrow Y$$

Ouvert ou fermé?

$$X \rightarrow Z_1 \leftarrow Z_2 \rightarrow Z_3 \rightarrow Y$$

Fermé!

$$X \rightarrow Z_1 \leftarrow Z_2 \rightarrow Z_3 \rightarrow Y$$

Ouvert ou fermé?

$$X \leftarrow Z_1 \leftarrow Z_2 \rightarrow Z_3 \rightarrow Y$$

Ouvert!

$$X \leftarrow Z_1 \leftarrow Z_2 \rightarrow Z_3 \rightarrow Y$$

Ouvert ou fermé?

$$X \rightarrow Z_1 \leftarrow Z_2 \rightarrow Z_3 \leftarrow Y$$

Fermé!

$$X \rightarrow Z_1 \leftarrow Z_2 \rightarrow Z_3 \leftarrow Y$$

Rappelez-vous

Contrôler pour le maillon central d'une collision ouvre un chemin

$$X \rightarrow \boxed{Z_1} \leftarrow Z_2 \rightarrow Z_3 \rightarrow Y$$

Rappelez-vous

Contrôler pour le maillon central d'une collision ouvre un chemin

$$X \leftarrow Z_1 \leftarrow \boxed{Z_2} \rightarrow Z_3 \rightarrow Y$$

Contrôler ou pas?

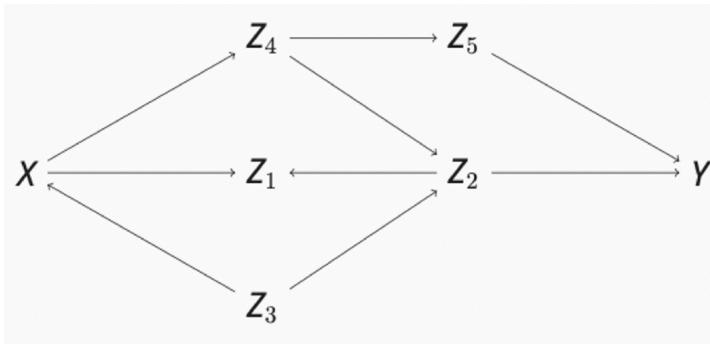
La marche à suivre pour savoir quoi contrôler

1. Faire la liste des variables pour lesquelles il faut **ÉVITER** de contrôler

- Tous les descendants de X
- Faire la liste de tous les chemins par porte arrière
 - Tous les chemins qui vont de X à Y et qui ont une flèche qui pointe vers X
 - Est-ce qu'il y a des chemins par porte arrière qui sont ouverts?
 - Dès qu'il y a une collision, le chemin est fermé
 - Est-il possible de bloquer ces chemins?
 - Donc... est-ce que l'effet causal de X sur Y est identifiable?

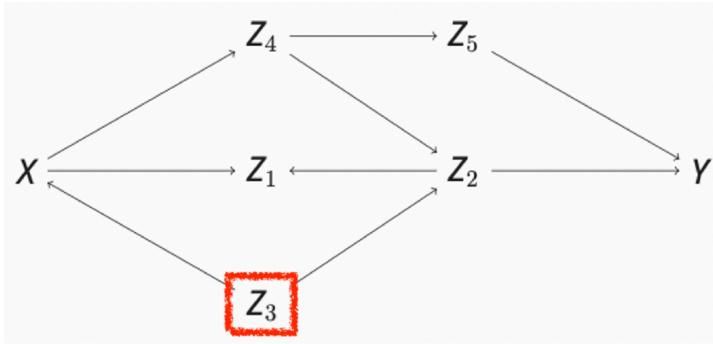
Contrôler ou pas?

- Y a-t-il des descendants de X?
- Y a-t-il des chemins par porte arrière?
- Y a-t-il des chemins par porte arrière ouverts?
- Est-il possible de bloquer ces chemins?
- L'effet causal de X sur Y est-il identifiable?



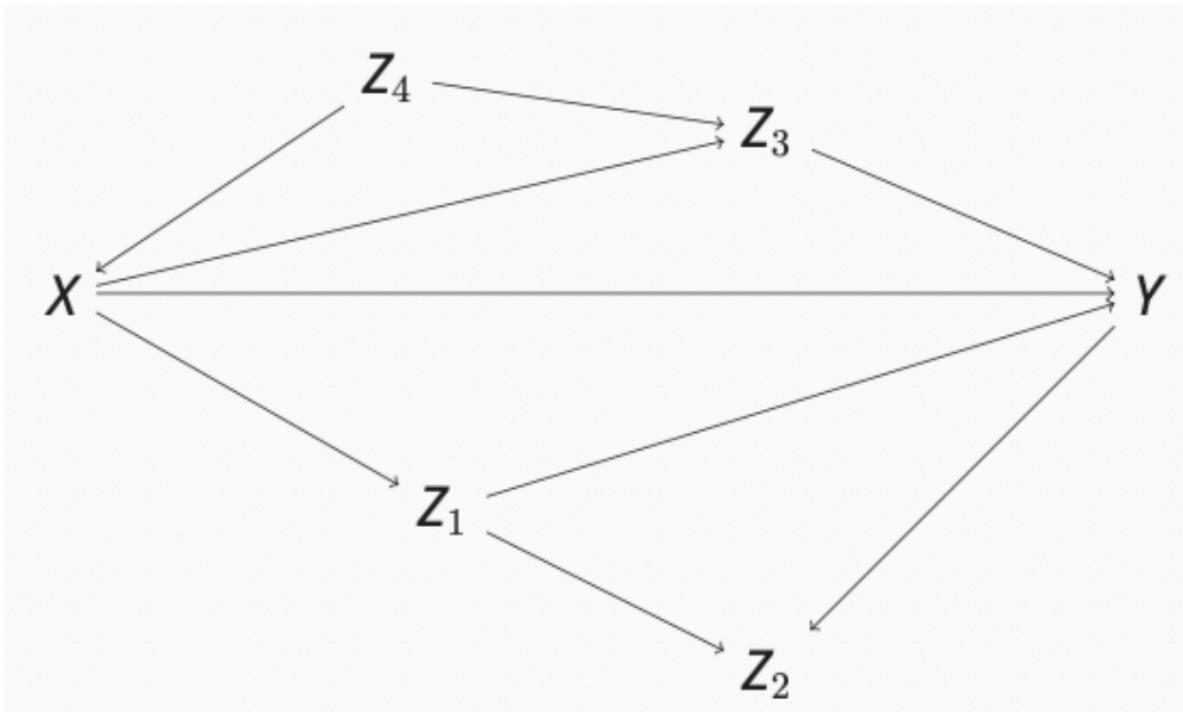
Contrôler ou pas?

- Y a-t-il des descendants de X? **Oui, Z1, Z2, Z4 et Z5**
- Y a-t-il des chemins par porte arrière? **Oui, Z3**
- Y a-t-il des chemins par porte arrière ouverts? **Oui, Z3**
- Est-il possible de bloquer ces chemins? **Oui, si on contrôle Z3**
- L'effet causal de X sur Y est-il identifiable? **Si on contrôle Z3!**



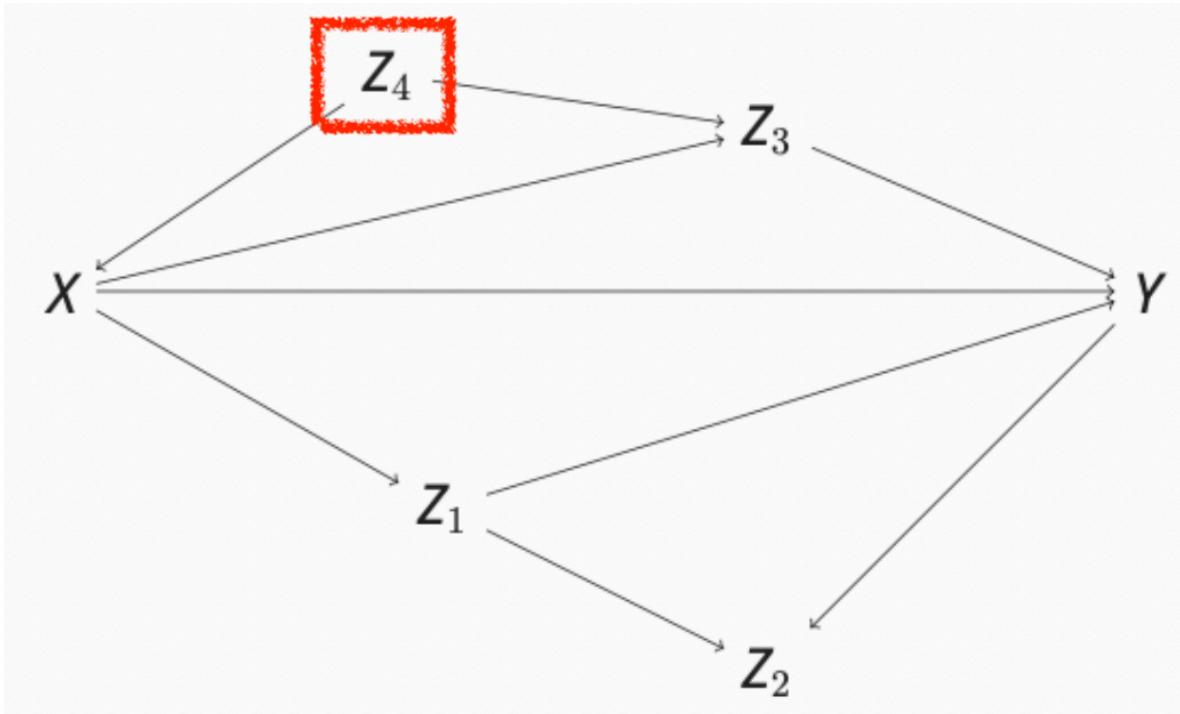
Contrôler ou pas?

1. Y a-t-il des descendants de X?
2. Y a-t-il des chemins par porte arrière?
3. Y a-t-il des chemins par porte arrière ouverts?
4. Est-il possible de bloquer ces chemins?
5. L'effet causal de X sur Y est-il identifiable?



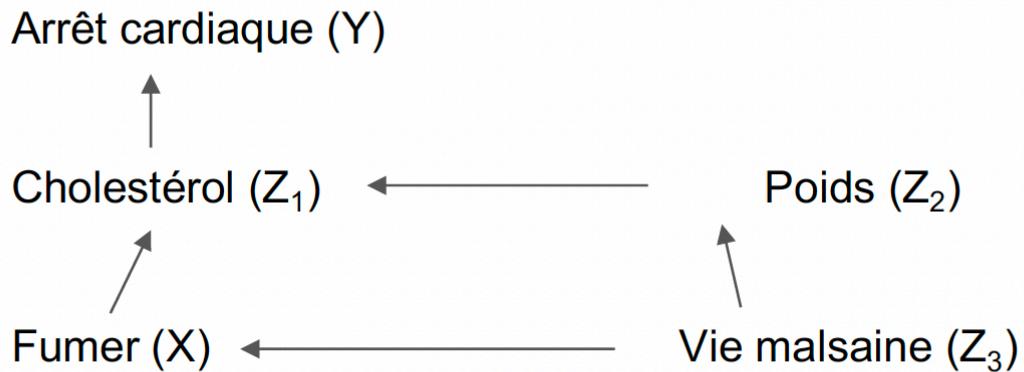
Contrôler ou pas?

1. Y a-t-il des descendants de X? **Oui, Z₁, Z₂, Z₃**
2. Y a-t-il des chemins par porte arrière? **Oui, Z₄**
3. Y a-t-il des chemins par porte arrière ouverts? **Oui, Z₄**
4. Est-il possible de bloquer ces chemins? **Oui, si on contrôle Z₄**
5. L'effet causal de X sur Y est-il identifiable? **Oui, si on contrôle Z₄**



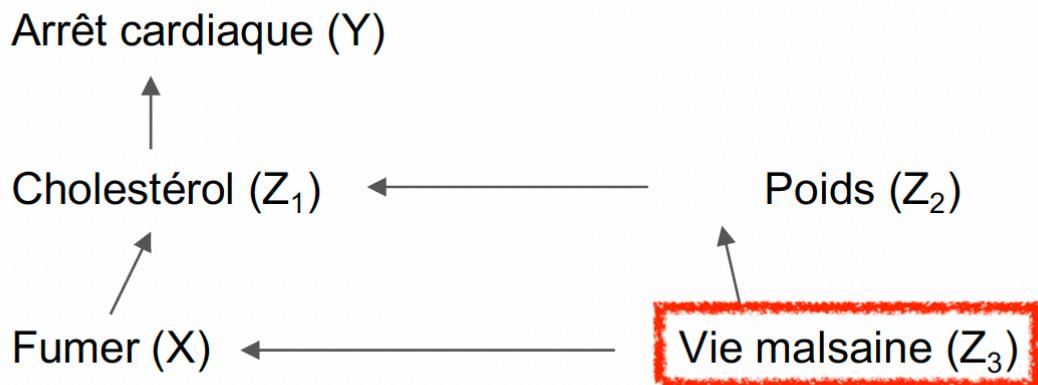
Contrôler ou pas?

1. Y a-t-il des descendants de X?
2. Y a-t-il des chemins par porte arrière?
3. Y a-t-il des chemins par porte arrière ouverts?
4. Est-il possible de bloquer ces chemins?
5. L'effet causal de X sur Y est-il identifiable?



Contrôler ou pas?

1. Y a-t-il des descendants de X? **Oui, Z_1**
2. Y a-t-il des chemins par porte arrière? **Oui, Z_3**
3. Y a-t-il des chemins par porte arrière ouverts? **Oui, Z_3**
4. Est-il possible de bloquer ces chemins? **Oui, si on contrôle Z_3**
5. L'effet causal de X sur Y est-il identifiable? **Oui, si on contrôle Z_3**



En résumé

- Dessiner un GOA:
 1. Faire la liste des variables pertinentes pour la question de recherche
 2. Représenter graphiquement les relations causales entre ces variables
- Contraintes:
 1. Orienté
 2. Acyclique
- Types de chemins
 1. Fourchette - ouvert
 2. Chaîne - ouvert
 3. Collision - fermé
- Contrôles:
 1. Ne pas contrôler pour un descendant de X
 2. Fermer tous les chemins par porte arrière

Travail de fin de session

Travail de fin de session

Vous devrez inclure un GOA!

- Dessiner les noeuds et les flèches selon la logique
 - Vous pourrez ensuite lire des articles scientifiques pour compléter
- En plus de votre X et Y, votre GOA doit inclure 3 autres variables qui pourraient être associées à la VD.
 - Il n'est pas nécessaire que ces variables existent dans la base de données obligatoire
 - Mais au moins une!
 - Votre base de données choisit devrait être votre premier réflexe
 - Si votre GOA indique qu'aucun contrôle est nécessaire... aucun contrôle est nécessaire!

Travail de fin de session

- Si votre résultat n'est pas significatif?
 - Cool! C'est AUSSI une contribution à la science
- Petit n (échantillon) et grand N (population)
- `_constant` == l'intercept

Table 1: Test d'hypothèse sur le faible niveau de confiance envers la police

	Faible niveau de confiance envers la police	
	(1)	(2)
Moins qu'un DES	0.04* (0.02)	0.01 (0.02)
Revenu bas		0.04 (0.02)
Femme		-0.06*** (0.01)
Immigrant		0.02 (0.02)
<code>_constant</code>	0.31*** (0.01)	0.35*** (0.01)
n	1,536	1,398

Source: Étude électorale canadienne de 2011

Méthode: Régressions linéaires multiples

Variable dépendante: Faible niveau de confiance envers la police (variable catégorielle)

Variables indépendantes: Toutes les variables indépendantes sont dichotomisées

* p<0.05; ** p<0.01; *** p<0.001

Simulations

```
n <- 100000
X <- rnorm(n)
Y <- 1.7 * X + rnorm(n)
```

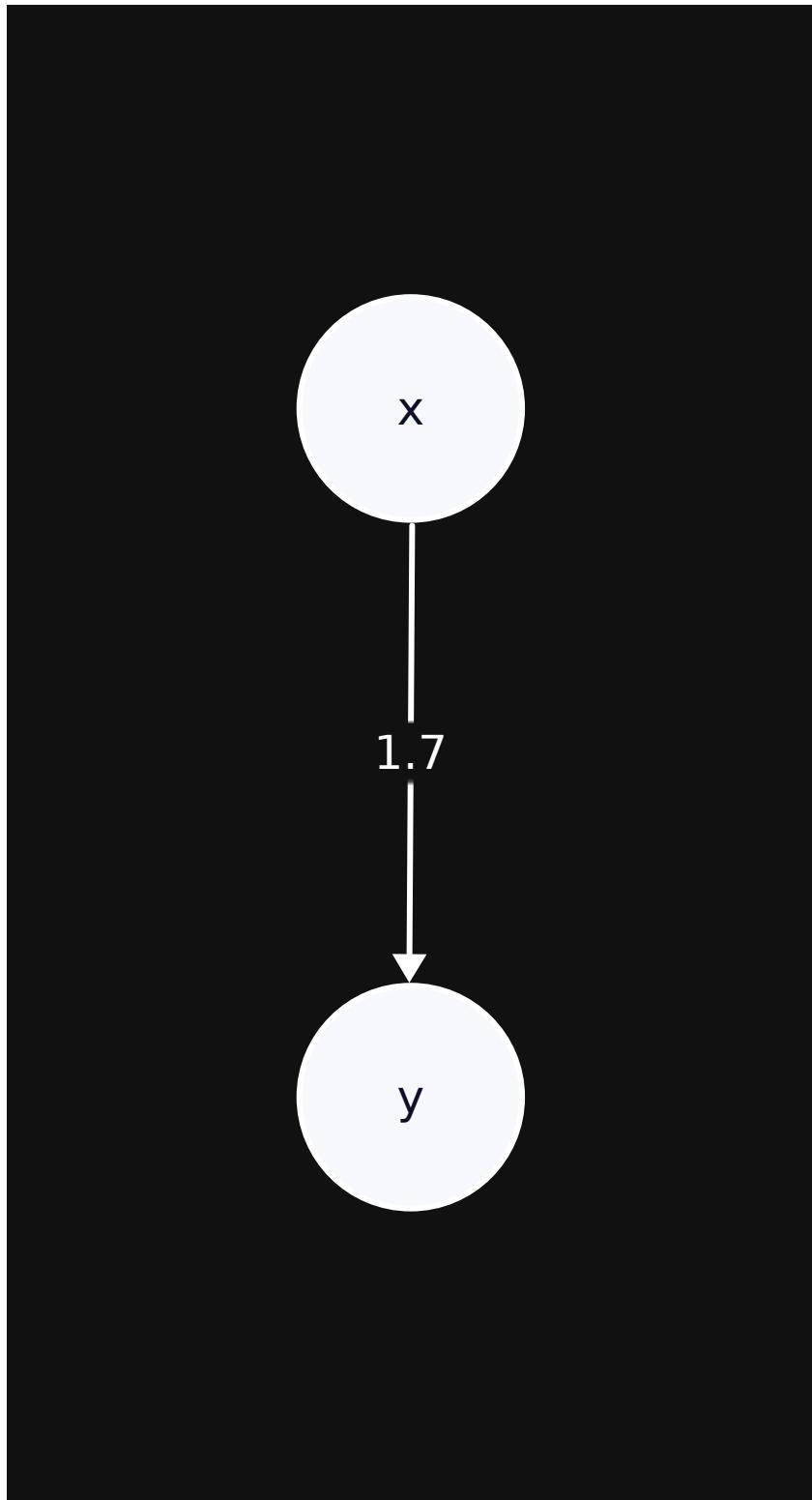
```
mod <- lm(Y ~ X)
coef(mod)
```

```
n <- 100000
X <- rnorm(n)
Y <- 1.7 * X + rnorm(n)
```

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

```
coefs <- coef(mod)
c(coefs["X"])
```

X
1.698916



Simulation Chaîne

```
X <- rnorm(n)
Z <- 3 * X + rnorm(n)
Y <- 0.5 * Z + rnorm(n)
mod <- lm(Y ~ X)
coef(mod)
```

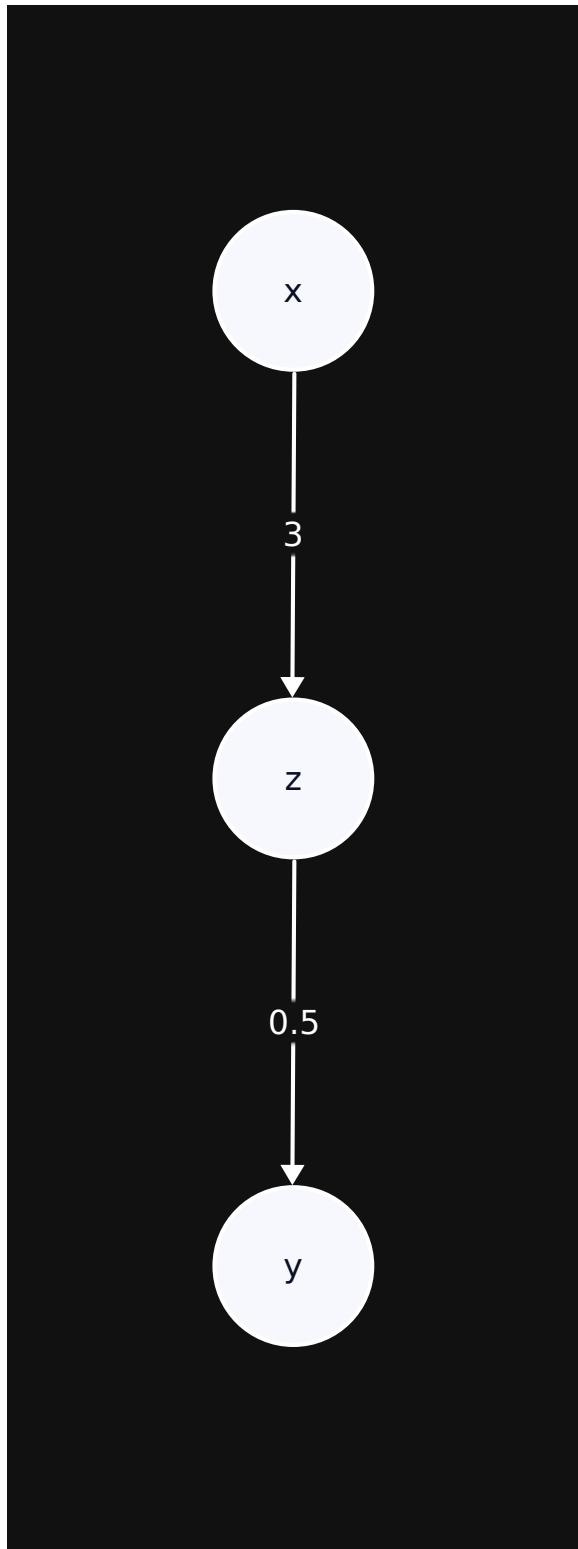
```
n <- 100000
X <- rnorm(n)
Z <- 3 * X + rnorm(n)
Y <- 0.5 * Z + rnorm(n)
mod <- lm(Y ~ X)
coefs <- coef(mod)
c(coefs["X"])
```

X
1.495695

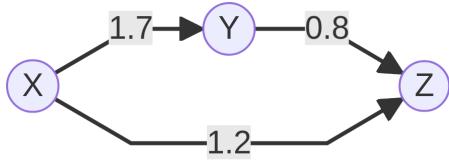
```
X <- rnorm(n)
Z <- 3 * X + rnorm(n)
Y <- 0.5 * Z + rnorm(n)
mod <- lm(Y ~ X + Z)
coef(mod)
```

```
n <- 100000
X <- rnorm(n)
Z <- 3 * X + rnorm(n)
Y <- 0.5 * Z + rnorm(n)
mod <- lm(Y ~ X + Z)
coefs <- coef(mod)
c(coefs["X"], coefs["Z"])
```

X Z
-0.00106409 0.49984081



Simulation Collision



```
X <- rnorm(n)
Y <- 1.7 * X + rnorm(n)
Z <- 1.2 * X + 0.8 * Y + rnorm(n)
mod <- lm(Y ~ X)
coef(mod)
```

```
n <- 100000
X <- rnorm(n)
Y <- 1.7 * X + rnorm(n)
Z <- 1.2 * X + 0.8 * Y + rnorm(n)
mod <- lm(Y ~ X)
coefs <- coef(mod)
c(coefs["X"])
```

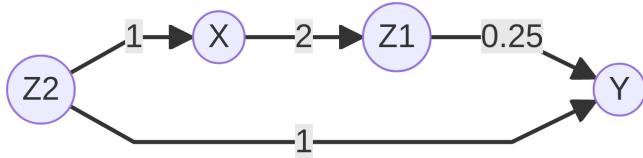
```
X
1.701785
```

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

```
n <- 100000
X <- rnorm(n)
Y <- 1.7 * X + rnorm(n)
Z <- 1.2 * X + 0.8 * Y + rnorm(n)
mod <- lm(Y ~ X + Z)
coefs <- coef(mod)
c(coefs["X"], coefs["Z"])
```

```
X          Z
0.443713 0.490478
```

Simulation Fourchette



```

Z2 <- rnorm(n)
X <- Z2 + rnorm(n)
Z1 <- 2 * X + rnorm(n)
Y <- 0.25 * Z1 + Z2 + rnorm(n)
mod <- lm(Y ~ X + Z2)
coef(mod)
  
```

```

n <- 100000
Z2 <- rnorm(n)
X <- Z2 + rnorm(n)
Z1 <- 2 * X + rnorm(n)
Y <- 0.25 * Z1 + Z2 + rnorm(n)
mod <- lm(Y ~ X + Z2)
coefs <- coef(mod)
c(coefs["X"], coefs["Z2"])
  
```

X Z2
0.5020560 0.9977387

```

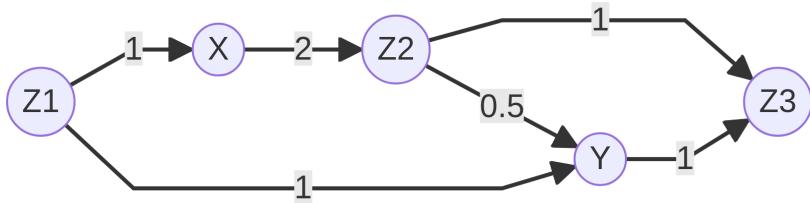
mod <- lm(Y ~ X)
coef(mod)
  
```

```

n <- 100000
X <- rnorm(n)
Y <- 1.7 * X + rnorm(n)
Z <- 1.2 * X + 0.8 * Y + rnorm(n)
mod <- lm(Y ~ X + Z)
coefs <- coef(mod)
c(coefs["X"])
  
```

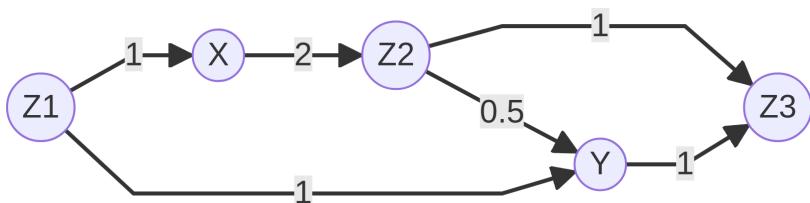
X
0.4484242

Simulation complexe



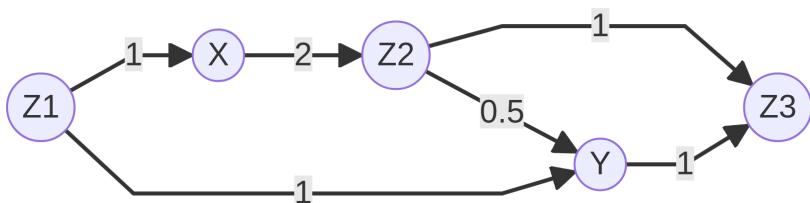
1. Y a-t-il des descendants de X?
2. Y a-t-il des chemins par porte arrière?
3. Y a-t-il des chemins par porte arrière ouverts?
4. Est-il possible de bloquer ces chemins?
5. L'effet causal de X sur Y est-il identifiable?

Simulation complexe



1. Y a-t-il des descendants de X? **Oui, Z2 et Z3**
2. Y a-t-il des chemins par porte arrière? **Oui, Z1**
3. Y a-t-il des chemins par porte arrière ouverts? **Oui, Z1**
4. Est-il possible de bloquer ces chemins? **Oui, si on contrôle Z1**
5. L'effet causal de X sur Y est-il identifiable? **Oui, si on contrôle Z1**

Simulation complexe - PREUVE



```

Z1 <- rnorm(n)
X <- Z1 + rnorm(n)
Z2 <- 2 * X + rnorm(n)
Y <- 0.5 * Z2 + Z1 + rnorm(n)
Z3 <- Z2 + Y + rnorm(n)
mod <- lm(Y ~ X + Z1)
coef(mod)

```

```

n <- 100000
Z1 <- rnorm(n)
X <- Z1 + rnorm(n)
Z2 <- 2 * X + rnorm(n)
Y <- 0.5 * Z2 + Z1 + rnorm(n)
Z3 <- Z2 + Y + rnorm(n)
mod <- lm(Y ~ X + Z1)
coefs <- coef(mod)
c(coefs["X"], coefs["Z1"])

```

X	Z1
0.9976535	1.0052492

```

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

```

```

n <- 100000
Z1 <- rnorm(n)
X <- Z1 + rnorm(n)
Z2 <- 2 * X + rnorm(n)
Y <- 0.5 * Z2 + Z1 + rnorm(n)
Z3 <- Z2 + Y + rnorm(n)
mod <- lm(Y ~ X)
coefs <- coef(mod)
c(coefs["X"])

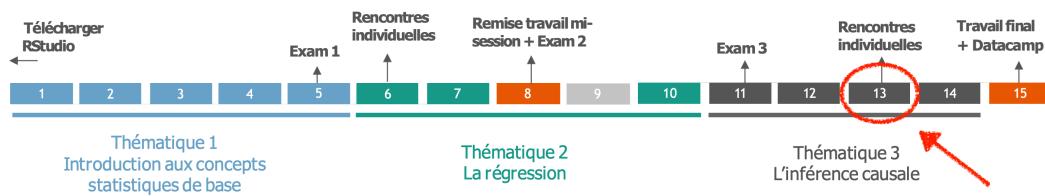
```

X	
1.494276	

Prochain cours

Messages

1. Rencontres individuelles la semaine prochaine!!
 - Pas de cours en classe.
 - Prenez rendez-vous déjà avec Camille Pelletier (Mac), Alexandre Bouillon (PC), Adrien Cloutier.
 - Arrivez avec votre GOA, du code, des graphiques, un début de travail!
2. Dans 2 semaines: Exam 3 (18 avril). Ce sera exactement le même format que l'examen 2 (papier crayon). Ils ont 1h encore. Ça peut couvrir la matière du cours 1 jusqu'à ton cours 12 sur les GOA (inclusivement). Pour étudier:
 - Faire les Datacamp
 - Faire les lectures
 - Relire les diapos
 - Être capable de rouler le code qu'on a fait en classe et en comprendre la logique.



Messages

3. Dans 3 semaines: Travail final + Datacamp complété pour le vendredi 26 avril 23h59.
4. Prennez de l'avance dans le travail final. Coder peut être long.