Analyse de données 2

Module d'introduction à l'analyse causale

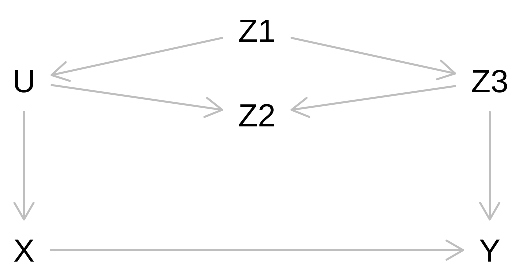
Avril 2022

Consignes

- Vous pouvez travailler en binôme, mais vous redirigerez individuellement vos réponses. Si vous travaillez avec quelqu'un, merci d'indiquer son nom dans vos solutions.
- Vous êtes encouragés à préparer vos solutions en utilisant R Markdown (fichier `.Rmd` et rendu `.pdf` ou `.html`). Dans ce cas vous serez amenés à utiliser la syntaxe LaTeX pour les réponses nécessitant d'écrire des formules mathématiques.
- Vous rendrez vos solutions en utilisant le dépôt moodle.

Exercice 1

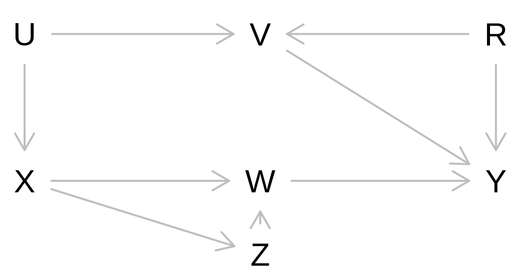
On considère le DAG



1. Donner la liste de tous les chemins ouverts entre X et Y.
2. D'un point de vue intuitif, quel(s) chemin(s) faut-il bloquer pour identifier l'effet causal de X sur Y?
3. Si on n'observe pas U, peut-on identifier l'effet causal de X sur Y?
4. Donner les ensembles d'ajustement pour le couple (X, Y).

Exercice 2

On considère le modèle causal donné par le DAG suivant



et par les lois de probabilité conditionnelles $P(\text{variable}|\text{pa}(\text{variable}))$ associées aux équations structurelles suivantes:

- $U = 1 + \epsilon_U$, avec $\epsilon_U \sim \mathcal{N}(0, 0.5)$
- $X = U + \epsilon_X$, avec $\epsilon_X \sim \mathcal{N}(0, 0.5)$
- $R = 2 + \epsilon_R \sim \mathcal{N}\epsilon(0, 0.5)$
- $V = U + R + \epsilon_V$, avec $\epsilon_V \sim \mathcal{N}(0, 0.5)$
- $Z = X + \epsilon_Z$, avec $\epsilon_Z \sim \mathcal{N}(0, 0.5)$
- $W = Z + 2X + \epsilon_W$, avec $\epsilon_W \sim \mathcal{N}(0, 0.5)$
- $Y = 2V + W + R + \epsilon_Y$, avec $\epsilon_Y \sim \mathcal{N}(0, 0.5)$.

1. Simuler un échantillon de $n = 1000$ observations $(u_i, x_i, r_i, v_i, z_i, w_i, y_i)$ selon le modèle ci-dessus.
2. Dans ce modèle causal, la variable X a-t-elle un effet causal sur Y?
3. Donner l'expression de la loi de probabilité jointe $\mathbb{F}(u, r, v, z, w, y|do(x)) = \mathbb{F}(U = u, R = r, V = v, Z = z, W = w, Y = y|do(X = x))$ en utilisant la définition d'opérateur $do(.)$ vue en cours.
4. A l'aide de l'équation de $\mathbb{F}(u, r, v, z, w, y|do(x))$ trouvée au point précédent, donner l'expression de la loi de probabilité $\mathbb{F}(y|do(x))$.
5. Simuler un $n = 1000$ observations (y_i^1) de la variable $Y|do(X = 1)$ et $n = 1000$ observations (y_i^0) de la variable $Y|do(X = 0)$.
6. Estimer l'effet causal moyen $\mathbb{E}(Y|do(X = 1) - Y|do(X = 0))$ à l'aides des observations (y_i^1), (y_i^0) simulées ci-dessus.
7. Donner un ensemble d'ajustement pour identifier $\mathbb{P}(y|do(x))$.
8. Donner une formule d'ajustement permettant d'identifier $\mathbb{P}(y|do(x))$ à l'aide des variables d'ajustement trouvées ci-dessus.
9. Estimer le modèle linéaire donnant Y en fonction de X et expliquer les résultats en les comparant avec la réponse à la question 4. Indication: ne pas oublier le résultat montré dans l'exercice 1 de la feuille de TD!

Exercice 3

La Figure 1 montre la preuve du critère frontdoor. Donner une justification pour chaque passage numeroté en rouge. Exemple: La relation d'indépendance (1) suit du fait que les chemins du type $Z \to A \to W$ et $Z \to Y \leftarrow W$ sont bloqués par A.

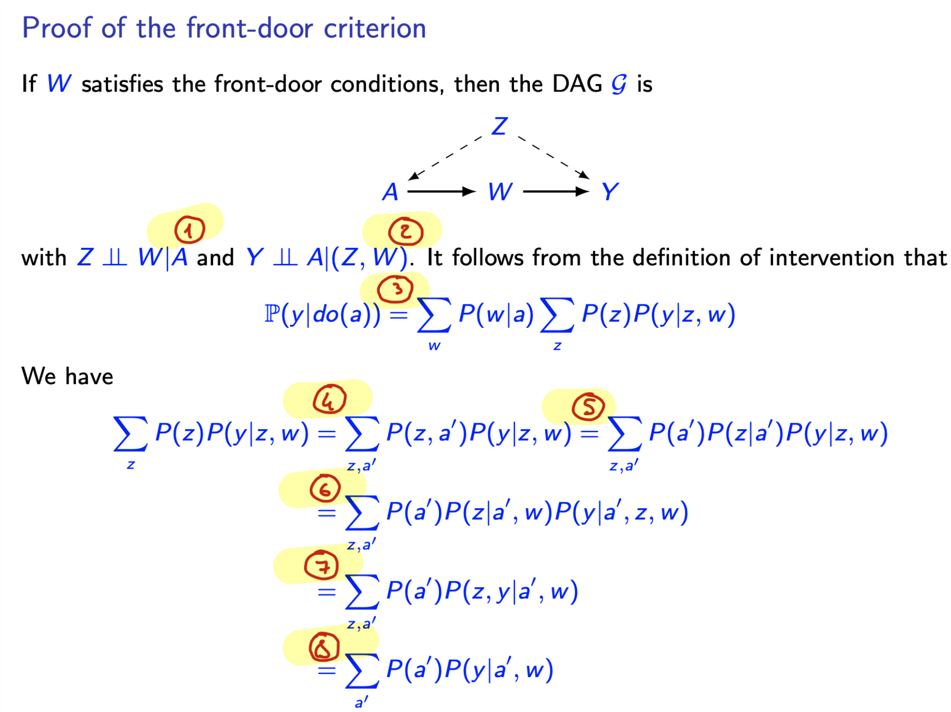


Figure 1: Preuve critère frontdoor, diapo 54 cours.