Rappel TD_8

Marion Brouard, Pauline Leveneur

December 5, 2025

- **Objectif du rappel :**
  1. Faire un point sur l'interprétation d'une variable binaire
  2. Rappeler le lien entre régressions linéaires et effets causaux
  3. Revoir quelques conditions pour l'estimation d'effets causaux dans une régression multiple

1 Interprétation d'une variable binaire

Pour D $\in \{0, 1\}$:

$$
\begin{aligned}
&Y(D) = \alpha_0 + D\beta_D + \epsilon \\
&\Rightarrow \hat{\alpha_0} = \frac{1}{n_0} \sum_{i, Di=0} Y_i = \bar{Y_0} \\
&\Rightarrow \hat{\beta_D} = \Big(\frac{1}{n_1} \sum_{i, Di=1} Y_i - \frac{1}{n_0} \sum_{i, Di=0} Y_i\Big) = \bar{Y_1} - \bar{Y_0}
\end{aligned}
$$

Avec n_0 le nombre d'individus pour lesquels $D = 0$ et n_1 le nombre d'individus pour lesquels $D = 1$.

Ainsi, la constante $\hat{\alpha}_0$ s'interprète comme la moyenne de l'outcome Y lorsque D=0 *(attention à l'interprétation quand il y a plusieurs variables binaires dans la régression : il s'agit de la moyenne pour la population dont toutes les variables binaires sont nulles).* $\hat{\beta_D}$ s'interprète comme la différence des moyennes entre les deux groupes. Ce coefficient peut s'interpréter comme un effet causal seulement sous certaines conditions. *(voir section 2)*

Exemple: Si l'outcome Y est le salaire, et D le sexe (1 pour les hommes, 0 pour les femmes), alors, $\hat{\alpha}_0$ est la moyenne des salaires des femmes, et $\hat{\beta}_D$ correspond à la différence moyenne des salaires entre les hommes et les femmes. Si on ajoute une variable D_2 correspondant au nombre d'années d'expérience, alors $\hat{\alpha}_0$ est la moyenne des salaires pour les femmes sans expérience ($D_2 = 0$).

2 Lien entre régressions linéaires et effets causaux

2.1 Rappel du cours

Attention : Le modèle causal ne peut pas être estimé tel quel, il faut passer par une régression linéaire classique en utilisant les MCO. Sous certaines conditions on trouve que les effets causaux peuvent être estimés par MCO.

- **Modèle causal (Théorique)** :
  $Y(D) = \zeta + D\delta^T + \eta$
  avec $E(\eta) = 0$ *(Toujours vrai tant qu'il y a une constante)*
  et $cov(D, Y(0)) \neq 0$ en général *(Si pas de conditions supplémentaires)*

En effet, on peut retrouver ce modèle théorique à partir de Y :

$$\begin{aligned} Y &= Y(0)(1-D) + Y(1)D \\ &= Y(0) + D(Y(1) - Y(0)) \\ &= Y(0) + D\Delta \\ &= Y(0) + D\Delta + E(Y(0)) - E(Y(0)) + D\delta^T - D\delta^T \\ &= \underbrace{E(Y(0))}_{\zeta} + D\delta^T + \underbrace{Y(0) - E(Y(0)) + D(\Delta - \delta^T)}_{\eta} \\ &= \zeta + D\delta^T + \eta \end{aligned}$$

On remarque ici assez facilement que $cov(D, \eta)$ est différent de 0 si on ne fait pas d'hypothèses supplémentaires. On trouve $cov(D, \eta) = 0$ si et seulement si $cov(Y(0), D) = 0$

Donc la régression linéaire classique permet d'estimer le modèle causal uniquement si $cov(Y(0), D) = 0$ **(Biais de sélection nul)**
On a alors $(\zeta, \delta^T, \eta) = (\alpha_0, \beta_D, \epsilon)$

- **Modèle linéaire de régression théorique de Y sur D (Théorique)** (simple prédiction linéaire) :
  $Y = \alpha_0 + D\beta_D + \epsilon$
  avec $E(\epsilon) = 0$
  et $cov(D, \epsilon) = 0$ (*Toujours vrai: CPO*)

2.2 Dans le cadre du TD

Dans le TD, il y a deux variables binaires. Comme dans le cours, on distingue deux modèles :

- **Modèle causal (Théorique)** :
  $Y(D) = \zeta_0 + D_1\delta_{01} + D_2\delta_{02} + \eta$
  avec $E(\eta) = 0$

- **Modèle linéaire de régression théorique de Y sur D (Théorique)** (simple prédiction linéaire) :
  $Y = \alpha_0 + D_1\beta_{01} + D_2\beta_{02} + \epsilon$
  avec $E(\epsilon) = 0$ et $cov(D_1, \epsilon) = cov(D_2, \epsilon) = 0$ (CPO)

  - En l'absence de biais de sélection (ie $Cov(D, Y(0)) = 0$) : les MCO permettent d'estimer des effets causaux (ie $\beta_{01} = \delta_{01}$ et $\beta_{02} = \delta_{02}$).
  - Si biais de sélection: $\alpha_0 + D_1\beta_{01} + D_2\beta_{02}$ = meilleur prédicteur linéaire de Y par D (pas d'effet causal)

3 Quelques conditions pour l'estimation d'effets causaux

On cherche à estimer l'effet causal de la variable D_2. Pour cela, il faut également que les autres variables du modèle remplissent certaines conditions (Attention : uniquement si celles-ci sont corrélées avec D_2). Dans le TD 8 nous avons vu deux exemples (endogénéité et erreur de mesure d'une autre variable) qui empêche l'estimation d'effets causaux.

- Si D_1 est endogène ($Cov(D_1, \eta) \neq 0$) alors le coefficient associé à D_2 n'estime pas un effet causal (Rappel : Si $cov(D1, D_2) \neq 0$) *(question 3, TD 8)*

- Si D_1 est exogène ($Cov(D_1, \eta) = 0$) mais mesuré avec erreur ($\tilde{D}_1 = D_1 + \nu$) alors le coefficient associé à D_2 n'estime pas un effet causal (Rappel : Si $cov(D1, D_2) \neq 0$) *(question 5, TD 8)*

4 Compléments: lien entre $\mathbf{Cov}(Y(0), D) = 0$ et $\mathbf{Cov}(\eta, D) = 0$

Dans le modèle causal vu en cours (effet du traitement linéaire et homogène) :

$$\begin{aligned} Y &= Y(0) + D\Delta \\ &= \zeta + D\delta^T + \eta \end{aligned}$$

avec $\eta = Y(0) - E(Y(0)) + D(\Delta - \delta^T)$ (cf calculs section 2.1)

Dans ce cadre, nous allons montrer que $\text{Cov}(Y(0), D) = 0 \iff \text{Cov}(\eta, D) = 0$.
Partons de $\text{Cov}(D, \eta)$ et reprenons la definition de η dans la réécriture du modèle causal:

$$\begin{aligned} \text{Cov}(D, \eta) = 0 &\iff \text{Cov}(D, \underbrace{Y(0) - E(Y(0)) + D(\Delta - \delta^T)}_{\eta}) = 0 \\ &\iff \text{Cov}(D, Y(0)) - \underbrace{\text{Cov}(D, E(Y(0))}_{=0} + \text{Cov}(D, D(\Delta - \delta^T)) = 0 \end{aligned}$$

Montrons que $\text{Cov}(D, D(\Delta - \delta^T)) = \underbrace{E(D^2(\Delta - \delta^T))}_{(1)} - \underbrace{E(D)E(D(\Delta - \delta^T))}_{(2)} = 0$

(1) $E(D^2(\Delta - \delta^T)) = E(D(\Delta - \delta^T)|D = 1)P(D = 1) + \underbrace{E(D(\Delta - \delta^T)|D = 0)}_{=0} P(D = 0) = \underbrace{E(\Delta - \delta^T|D = 1)}_{=0} P(D = 1) = 0$

La dernière simplification vient du fait que $E(\Delta|D = 1) = E(Y(1) - Y(0)|D = 1)$ et $E(\delta^T|D = 1) = E(E(Y(1) - Y(0)|D = 1)|D = 1) = E(Y(1) - Y(0)|D = 1)$ (lois des espérances itérées)

(2) $E(D(\Delta - \delta^T)) = E(D(Y(1) - Y(0) - E(Y(1) - Y(0)|D = 1))) = E(D(Y(1) - Y(0))) - \underbrace{E(D(E(Y(1) - Y(0)|D = 1)))}_{= E(D(Y(1) - Y(0)))} = 0$

Ainsi

$$\text{Cov}(D, \eta) = 0 \iff \text{Cov}(D, Y(0)) - \underbrace{\text{Cov}(D, E(Y(0))}_{=0} + \underbrace{\text{Cov}(D, D(\Delta - \delta^T))}_{=0} = 0$$

Et donc $\boxed{\text{Cov}(D, \eta) = 0 \iff \text{Cov}(D, Y(0)) = 0}$ dans le cadre linéaire