Variables de contrôle et matching

Pratiques de la Recherche en Économie

Florentine Oliveira
2025-02-18

Reminder

L'objectif est d'estimer l'effet d'un **traitement** D sur une variable d'**outcome** Y.

- par exemple l'effet d'avoir un master sur le salaire

L'objectif est d'estimer l'effet d'un **traitement** D sur une variable d'**outcome** Y.

- par exemple l'effet d'avoir un master sur le salaire

Peut-on en déduire l'effet du traitement en comparant l'outcome des individus traités à celui des individus non traités?

- par exemple peut-on quantifier l'effet d'avoir un master sur le salaire en comparant les individus qui ont un master à ceux qui n'en ont pas?

L'objectif est d'estimer l'effet d'un **traitement** D sur une variable d'**outcome** Y.

- par exemple l'effet d'avoir un master sur le salaire

Peut-on en déduire l'effet du traitement en comparant l'outcome des individus traités à celui des individus non traités?

- par exemple peut-on quantifier l'effet d'avoir un master sur le salaire en comparant les individus qui ont un master à ceux qui n'en ont pas?

Dans 99,99999% des cas, NON!: le groupe des traités et celui des contrôles ne sont en général pas comparables

- par exemple, les femmes et les enfants issus de milieux sociaux favorisés sont plus susceptibles de faire de hautes études, et ces caractéristiques peuvent aussi influencer le salaire

L'objectif est d'estimer l'effet d'un **traitement** D sur une variable d'**outcome** Y.

- par exemple l'effet d'avoir un master sur le salaire

Peut-on en déduire l'effet du traitement en comparant l'outcome des individus traités à celui des individus non traités?

- par exemple peut-on quantifier l'effet d'avoir un master sur le salaire en comparant les individus qui ont un master à ceux qui n'en ont pas?

Dans 99,99999% des cas, NON!: le groupe des traités et celui des contrôles ne sont en général pas comparables

- par exemple, les femmes et les enfants issus de milieux sociaux favorisés sont plus susceptibles de faire de hautes études, et ces caractéristiques peuvent aussi influencer le salaire

Dans 0,00001% des cas, il est possible de comparer l'outcome moyen des individus traités et des individus témoins si le traitement est distribué de façon aléatoire (RCT; mais cela est très onéreux, pose des questions éthiques, etc.).

Cette séance

1. Modèle de régression linéaire multivarié

    1.1. Biais de variable omise (OVB)
    1.2. Hypothèses
    1.3. Estimateur
    1.4. Bonnes et mauvaises variables de contrôle
    1.5. Coefficient de détermination (R2)

2. Matching

    2.1. Intuition
    2.2. Méthodes

3. Causalité

1. Modèle de régression linéaire multivarié

1. Modèle de régression linéaire multivarié

1.1. Biais de variable omise (OVB)

Notre modèle de base s'écrit

$$Y_i = \alpha + \beta D_i + \varepsilon_i \qquad (1)$$

Par exemple, Y_i désigne le salaire de l'individu i, D_i une dummy qui représente le fait d'avoir un master ou non, et ε_i le terme d'erreur.

1. Modèle de régression linéaire multivarié

1.1. Biais de variable omise (OVB)

Notre modèle de base s'écrit

$$Y_i = \alpha + \beta D_i + \varepsilon_i \qquad (1)$$

Par exemple, Y_i désigne le salaire de l'individu i, D_i une dummy qui représente le fait d'avoir un master ou non, et ε_i le terme d'erreur.

Supposons qu'il existe une variable W_i, par exemple une variable binaire égale à 1 si i est une Femme.

1. Modèle de régression linéaire multivarié

1.1. Biais de variable omise (OVB)

Notre modèle de base s'écrit

$$Y_i = \alpha + \beta D_i + \varepsilon_i \qquad (1)$$

Par exemple, Y_i désigne le salaire de l'individu i, D_i une dummy qui représente le fait d'avoir un master ou non, et ε_i le terme d'erreur.

Supposons qu'il existe une variable W_i, par exemple une variable binaire égale à 1 si i est une Femme.

W_i est implicitement contenue dans le terme d'erreur dans le modèle (1).

1. Modèle de régression linéaire multivarié

1.1. Biais de variable omise (OVB)

Notre modèle de base s'écrit

$$Y_i = \alpha + \beta D_i + \varepsilon_i \qquad (1)$$

Par exemple, Y_i désigne le salaire de l'individu i, D_i une dummy qui représente le fait d'avoir un master ou non, et ε_i le terme d'erreur.

Supposons qu'il existe une variable W_i, par exemple une variable binaire égale à 1 si i est une Femme.

W_i est implicitement contenue dans le terme d'erreur dans le modèle (1).

Or, les femmes sont en moyenne davantage éduquées que les hommes.

1. Modèle de régression linéaire multivarié

1.1. Biais de variable omise (OVB)

Notre modèle de base s'écrit

$$Y_i = \alpha + \beta D_i + \varepsilon_i \qquad (1)$$

Par exemple, Y_i désigne le salaire de l'individu i, D_i une dummy qui représente le fait d'avoir un master ou non, et ε_i le terme d'erreur.

Supposons qu'il existe une variable W_i, par exemple une variable binaire égale à 1 si i est une Femme.

W_i est implicitement contenue dans le terme d'erreur dans le modèle (1).

Or, les femmes sont en moyenne davantage éduquées que les hommes.

Donc D_i est corrêlé à ε_i, ou dit autrement $\mathbb{E}(\varepsilon_i | D_i) \neq 0$

1. Modèle de régression linéaire multivarié

1.1. Biais de variable omise (OVB)

Notre modèle de base s'écrit

$$Y_i = \alpha + \beta D_i + \varepsilon_i \qquad (1)$$

Par exemple, Y_i désigne le salaire de l'individu i, D_i une dummy qui représente le fait d'avoir un master ou non, et ε_i le terme d'erreur.

Supposons qu'il existe une variable W_i, par exemple une variable binaire égale à 1 si i est une Femme.

W_i est implicitement contenue dans le terme d'erreur dans le modèle (1).

Or, les femmes sont en moyenne davantage éduquées que les hommes.

Donc D_i est corrêlé à ε_i, ou dit autrement $\mathbb{E}(\varepsilon_i | D_i) \neq 0$

🚨 **= Biais de variable omise** 🚨

1. Modèle de régression linéaire multivarié

1.1. Biais de variable omise (OVB)

On a le modèle

$$Y_i = \alpha + \beta D_i + \varepsilon_i$$

où la (ou les) variable W_i est omise (donc appartient à ε_i).

La modèle *multivarié* s'écrit

$$Y_i = \gamma + \delta D_i + \phi W_i + \nu_i$$

On a donc la relation:

$$\beta = \underbrace{\delta}_{\text{"vrai" estimateur}} + \underbrace{\phi\pi}_{OVB}$$

où π est le coefficient de la régression de W_i sur D_i ($W_i = \lambda + \pi D_i$)

1. Modèle de régression linéaire multivarié

1.1. Biais de variable omise (OVB)

Exemple tiré du DM

	Log(Wage)	Women	Log(Wage)
(Intercept)	7.282***	0.442***	7.433***
	(0.004)	(0.004)	(0.004)
At least Bac	0.481***	0.097***	0.514***
	(0.006)	(0.006)	(0.006)
Women			−0.341***
			(0.006)
Num.Obs.	31835	31835	31835
R2 Adj.	0.173	0.009	0.260
* p < 0.1, ** p < 0.05, *** p < 0.01			

1. Modèle de régression linéaire multivarié

1.2. Hypothèses du modèle multivarié

H_1 Linéarité: le modèle est linéaire dans les paramètres: $\frac{\partial y_i}{\partial x_{ik}} = \beta_k, \forall k = 1, \ldots, K$

H_2 Échantillon Aléatoire: l'échantillon est aléatoire et représentatif de la population.

H_3 **Exogeneité *conditionnelle***: Conditionnellement aux contrôles W, D est exogène

Formellement, $\mathbb{E}(\varepsilon_i | D, W) = 0$

H_4 Variation: il y a suffisamment de variation dans X où $X = (1 \;\; D \;\; W)$

- Dit autrement, chaque variable explicative apporte une information qui lui est propre
- Formellement, les explicatives ne sont pas colinéaires (cas multivarié: $(X'X)$ est inversible)

H_5 Les erreurs ε_i sont sphériques: H_{5a} homoscédasticité & H_{5b} Absence d'autocorrélation

1. Modèle de régression linéaire multivarié

1.2. Hypothèses du modèle multivarié

L'hypothèse d'indépendance conditionnelle, ou ***Conditional Independance Assumption (CIA)***, aussi appelée sélection sur les observables, indique que:

- conditionellement à des variables explicatives W_i, les outcomes potentiels $\{Y_{0i}, Y_{1i}\}$ sont indépendants du traitement D_i
- dit autrement, en contrôlant par les variables W_i, le traitement D_i est *as-good-as random*

Formellement, dans le framework des outcomes potentiels, l'hypothèse d'identification devient:

$$\{Y_{0i}, Y_{1i}\} \perp D_i | W_i$$

On a donc:

$$\begin{aligned} \textbf{Biais de Sélection} &= \mathbb{E}(Y_{0i}|W_i, D_i = 1) - \mathbb{E}(Y_{0i}|W_i, D_i = 0) \\ &= \mathbb{E}(Y_{0i}|W_i) - \mathbb{E}(Y_{0i}|W_i) \\ &= 0 \end{aligned}$$

1.3. Estimateur dans le cas multivarié

L'estimateur MCO dans le cas multivarié s'écrit:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

Maths

C'est toujours l'estimateur *BLUE*: **B**est **L**inear **U**nbiased **E**stimator

1.4 Bonnes et mauvaises variables de contrôle

Une *bonne* variable de contrôle doit:

- contribuer à **expliquer la variable dépendante**
- par une **information qui lui est propre**
- **ne pas être impactée par le traitement** d'intérêt

Si elles satisfont ces conditions, les variables de contrôle permettent:

- d'atténuer le risque de biais de variable omise
- gagner en précision

1. Modèle de régression linéaire multivarié

1.4 Bonnes et mauvaises variables de contrôle

Une *mauvaise* variable de contrôle est:

Une **variable non pertinente** est une variable qui ne contribue pas à expliquer l'outcome.

- le paramètre estimé sera donc nul 😐
- l'estimateur reste sans biais tant que la variable non pertinente n'est pas corrélée à ε_i 😐
- l'estimateur est moins précis 😢

Une **variable redondante** si l'information qu'elle contient est déjà contenue dans une autre variable

- lorsque deux variables sont très corrélées, difficile de distinguer l'effet "propre" de chacune, donc les estimateurs de ces deux variables seront très imprécis
- dans le cas extrême de colinéarité, le modèle n'est pas identifiable 😢

Un **mauvais contrôle** (*bad control*) est une variable qui est également affectée par le traitement

- l'estimateur peut être biaisé 😢

1.5 Coefficient de détermination (R2)

Le coefficient de détermination, ou R^2, informe sur la **qualité** de la régression linéaire, i.e. la **part de la variance de l'outcome expliquée par les** X.

Formellement,

$$
\begin{aligned}
R^2 &= \frac{SCE}{SCT} = \frac{\sum_i (\hat{y}_i - \bar{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2} \\
&= 1 - \frac{SCR}{SCT} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2}
\end{aligned}
$$

NB: le R^2

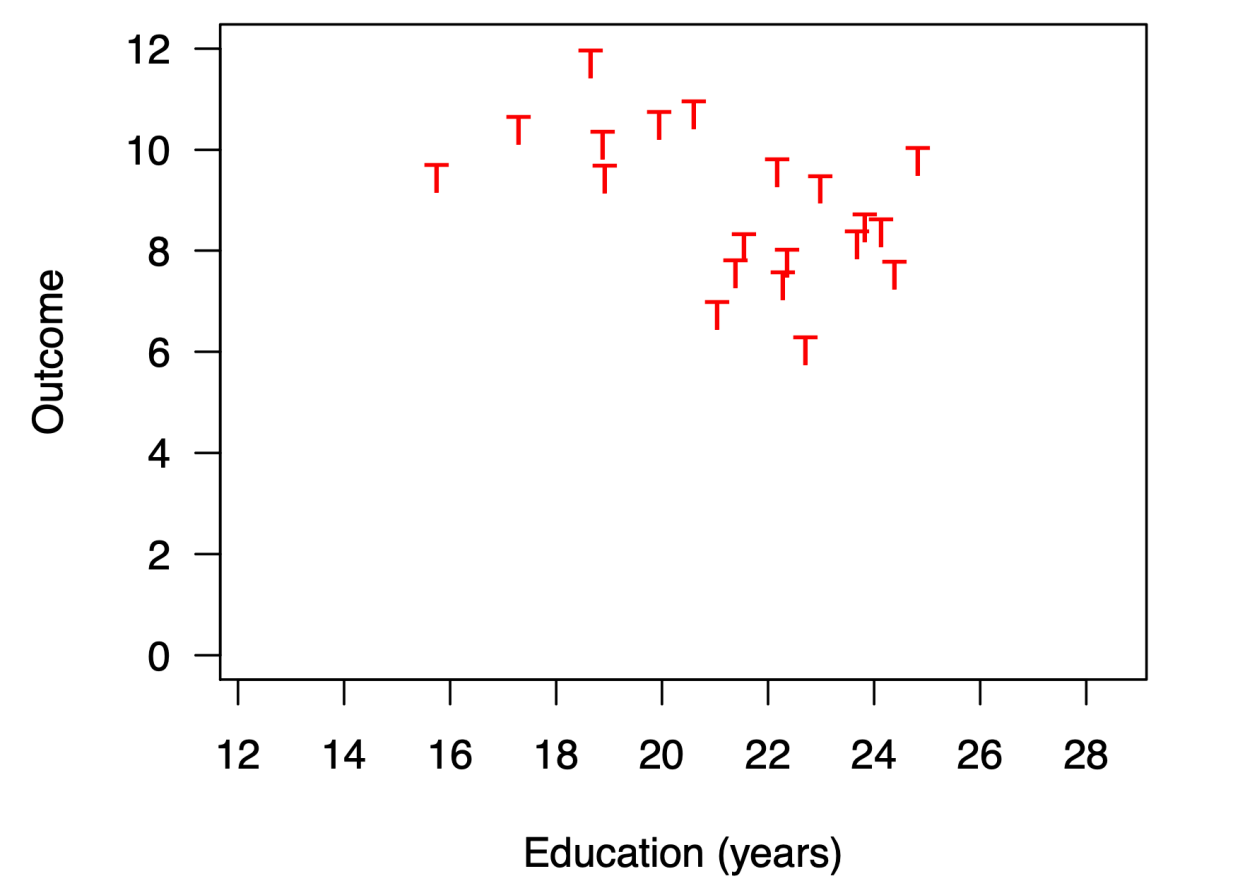
- augmente **mécaniquement** avec le nombre de varibales explicatives
  - comparer les R^2 ajustés lorsqu'on compare différents modèles
  - $R^2_{adj} = 1 - (1 - R^2)\frac{N-1}{N-K-1}$
- est spécifique à un sample
- est très informatif pour faire de la prédiction mais n'informe en rien sur la causalité

2. Matching

2. Matching

2.1. Intuition

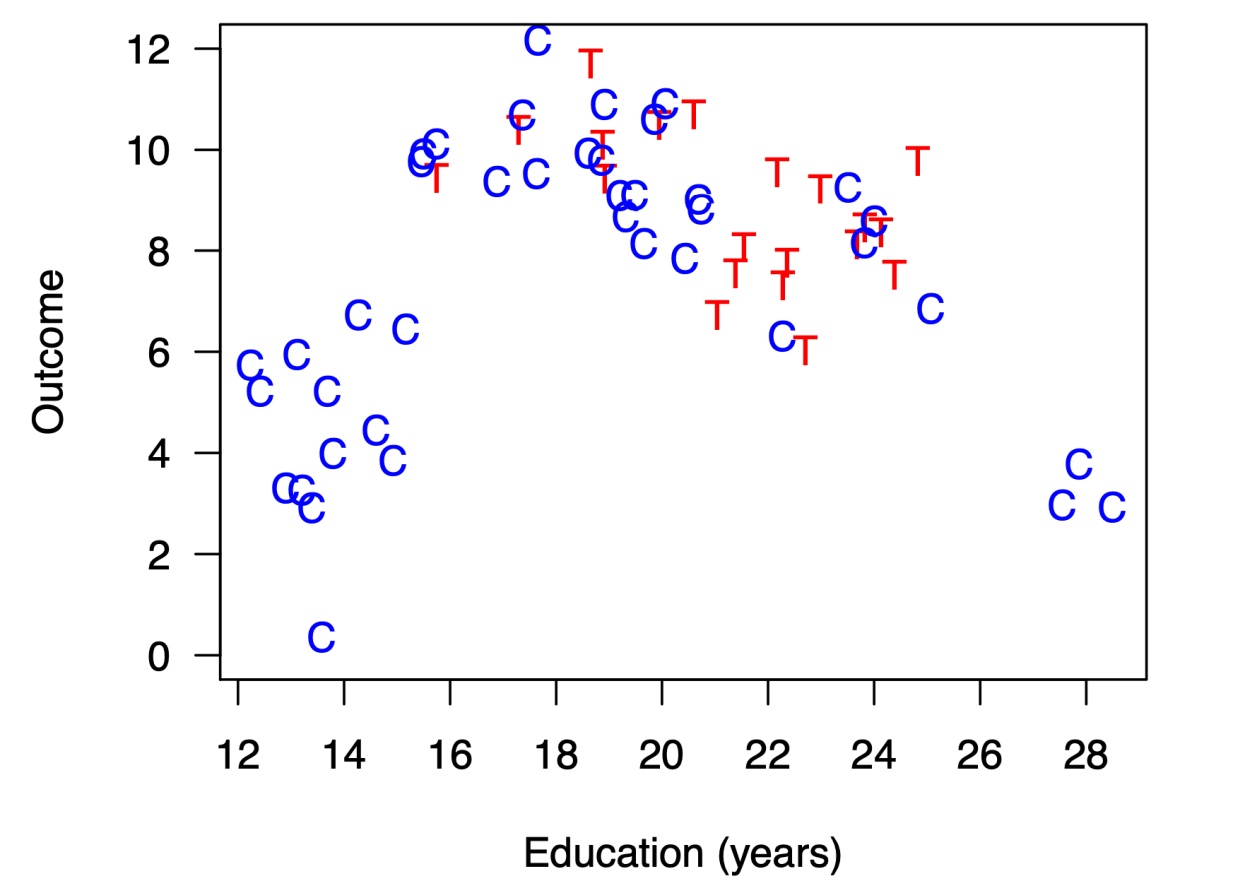
Lorsque l'on **contrôle** par des observables W, on peut avoir de l'*imbalance* (Ho, Imai, King,Stuart, 2007, Political Analysis)



2. Matching

2.1. Intuition

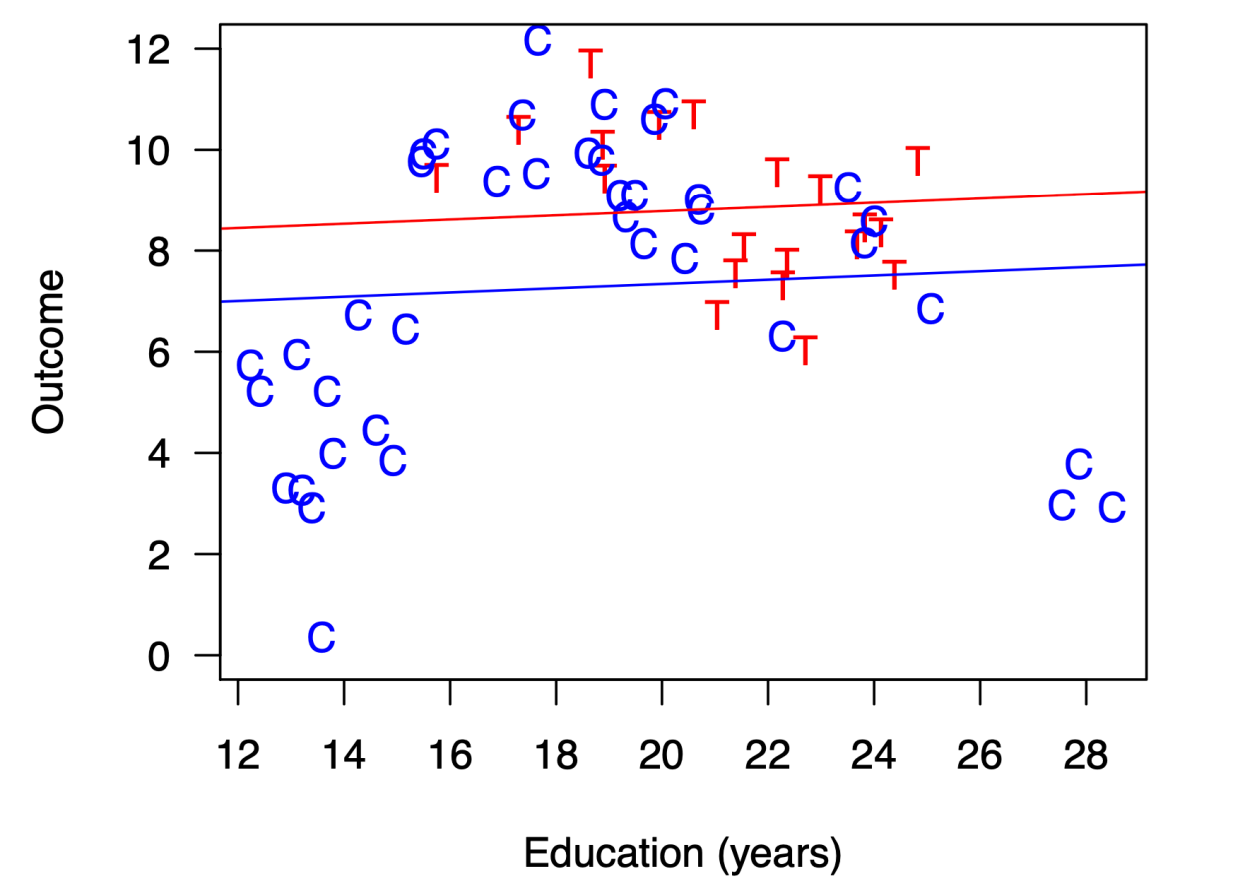
Lorsque l'on **contrôle** par des observables W, on peut avoir de l'*imbalance* (Ho, Imai, King,Stuart, 2007, Political Analysis)



2. Matching

2.1. Intuition

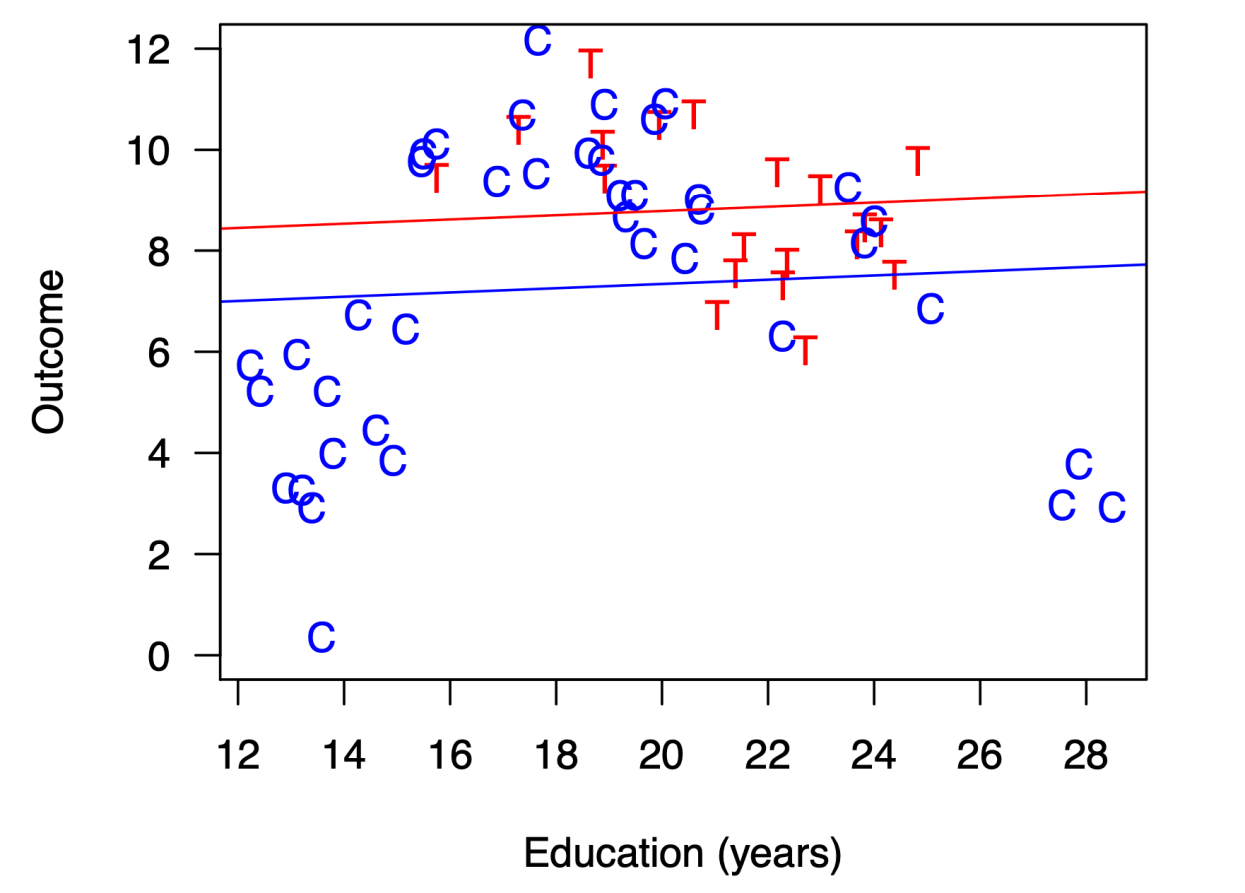
Lorsque l'on **contrôle** par des observables W, on peut avoir de l'*imbalance* (Ho, Imai, King,Stuart, 2007, Political Analysis)



2. Matching

2.1. Intuition

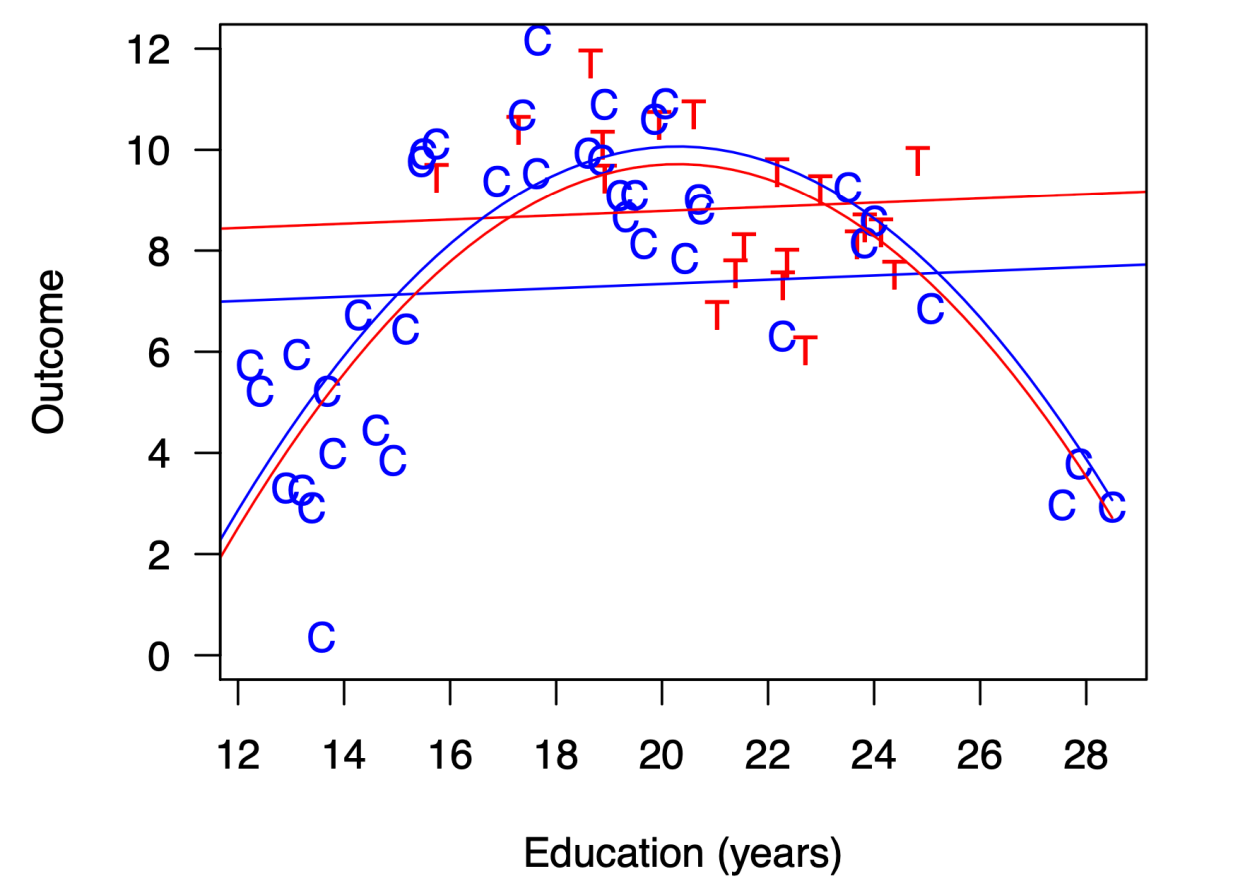
Lorsque l'on **contrôle** par des observables W, on peut avoir de l'*imbalance* (Ho, Imai, King,Stuart, 2007, Political Analysis)



2. Matching

2.1. Intuition

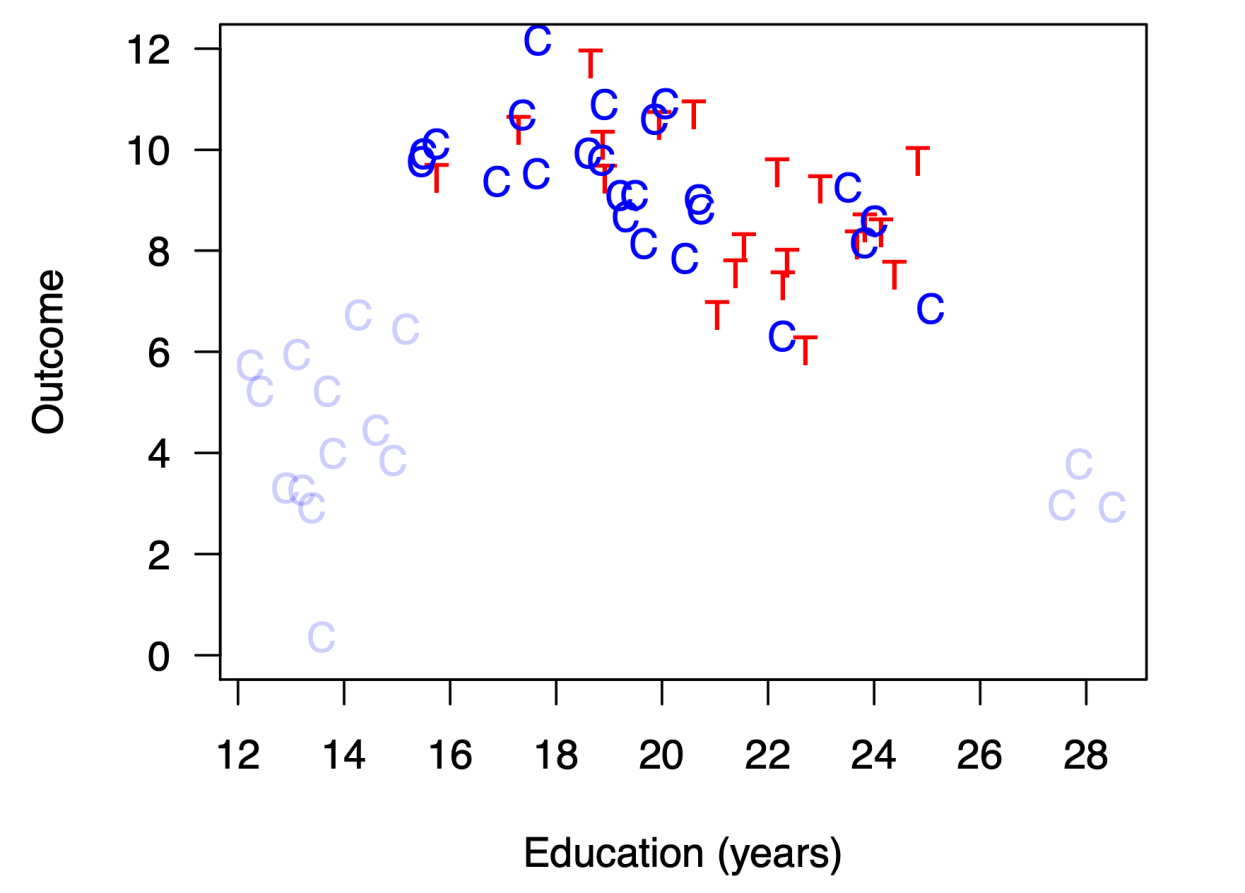
Lorsque l'on **contrôle** par des observables W, on peut avoir de l'*imbalance* (Ho, Imai, King,Stuart, 2007, Political Analysis)



2. Matching

2.1. Intuition

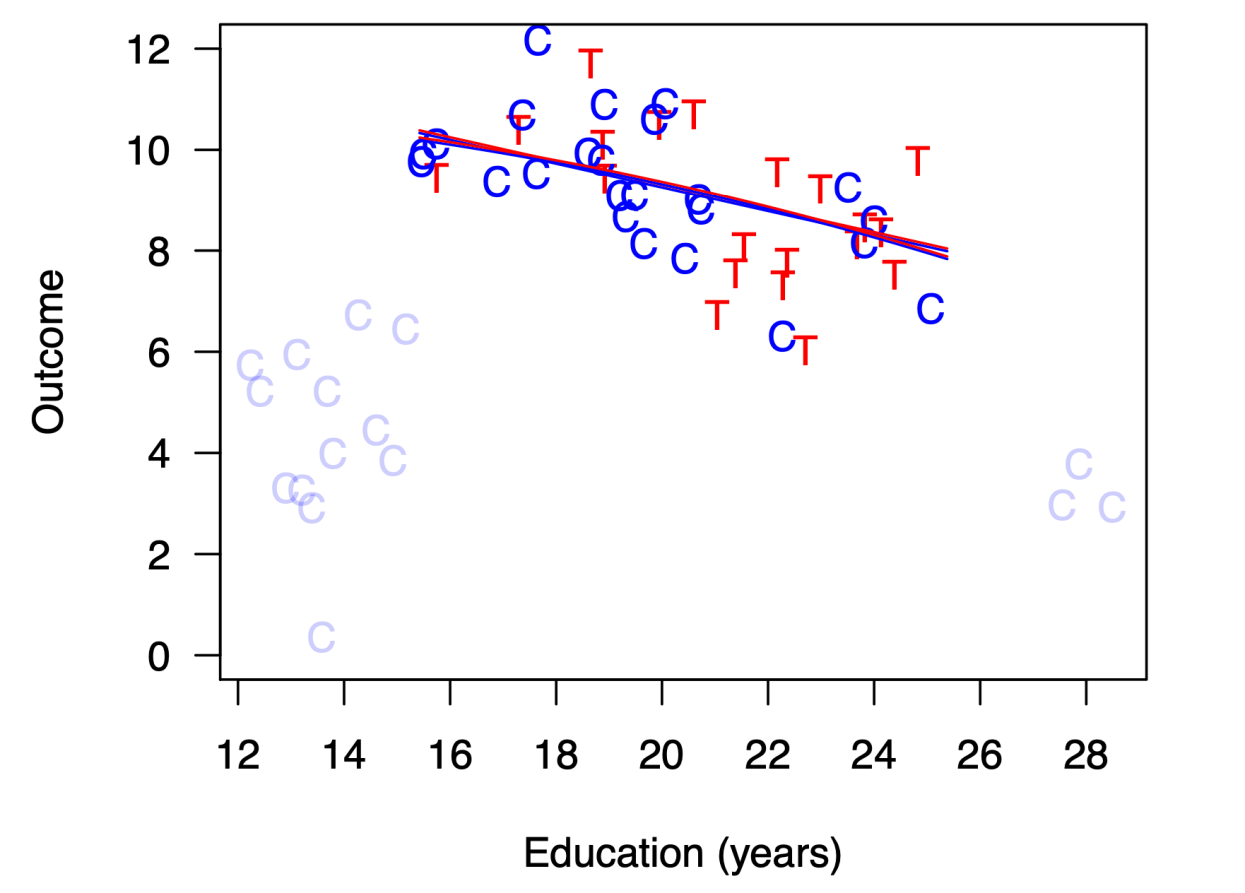
Lorsque l'on **contrôle** par des observables W, on peut avoir de l'*imbalance* (Ho, Imai, King,Stuart, 2007, Political Analysis)



2. Matching

2.1. Intuition

Lorsque l'on **contrôle** par des observables W, on peut avoir de l'*imbalance* (Ho, Imai, King,Stuart, 2007, Political Analysis)



2. Matching

2.2. Méthodes de matching

Les estimateurs de *matching* construisent artificiellement un **groupe** d'individus non traîtés qui a les mêmes caractéristiques que le groupe d'individus traités

- Assure un support commun des variables explicatives

Différents méthodes/algorithmes de matching:

- Matching exact: on cherche les individus *strictement identiques* (très restrictif!)
- *Propensity score matching*: on assigne à chaque individu une probabilité d'être traîté et on matche ceux avec des scores proches
- Plus proches voisins (*Nearest neighboors*): on cherche l'individu non-traîté le plus proche d'un individu traité selon certaines métriques
- ... NB: des combinaisons de méthodes sont possibles!

Inconvénients:

- certaines observations n'ont pas de match: on estime l'effet de traitement lorsque c'est "faisable"

3. Causalité

3. Causalité

Les méthodes de **matching** sont des méthodes d'identification qui reposent sur la CIA, c'est à dire que la sélection dans le traitement est uniquement liée à des variables **observables** X.

L'hypothèse d'indépendance conditionnelle revient à dire qu'en **contrôlant** (i.e. en "tenant compte de") **par un ensemble de variables** X_i, D est ***as good as random***.

Sous l'hypothèse d'indépendance conditionnelle, l'effet estimé est l'**effet moyen du traitement conditionnel** (***Conditional* ATE** - CATE)

En réalité, il s'agit d'une **hypothèse d'identification très forte**:

- elle suppose d'inclure toutes les variables explicatives X qui expliquent la corrélation entre ε_i et D_i
  - Problème: nombre de ces variables sont **inobservables**

Recap: OLS

Hypothèse d'identification: CIA

- Intuition: conditionnellement aux catactéristiques W par lesquelles on *"contrôle"*, le traitement est aléatoire
- Formellement:, $\mathbb{E}(\varepsilon_i | D_i, W_i) = 0$

Comparaison: parmi les individus qui ont les mêmes caractéristiques par lesquelles on contrôle, on compare les individus traîtés à ceux qui ne le sont pas.

Modèle: $Y = X\beta + \varepsilon$, où $X = (1 D W)$

Estimateur: $\hat{\beta} = (X'X)^{-1}X'Y$

Implémentation sur `R`:

- `lm` pour estimer les paramètres du modèle
- `summary` pour afficher le résultat de l'estimation
- `coeftest`, argument `vcov = vcovHC(fit, type = 'HC0')` pour obtenir des se robustes à l'hétéroscédasticité
- `stargazer` ou `modelsummary` pour exporter les résultats en une table $\LaTeX$

Sources

Causal inference: The Mixtape, Scott Cunningham

Annexe

Estimateur MCO dans le cas multivarié

Le modèle multivarié s'écrit: $y_i = \beta_1 + \beta_2 D_i + \sum_{k=1}^{K} w_{ik}\gamma_k + \varepsilon_i$

On peut l'écrire sous forme matricielle: $Y = \beta_1 + \beta_2 D + W\gamma + \varepsilon$

où W est une matrice contenant les K variables de contrôle.

On peut finalement réécrire le modèle:

$$Y = X\beta + \varepsilon$$

où

$$X = \begin{bmatrix} 1 & D_1 & W_{1,1} & \cdots & W_{1,K} \\ 1 & D_2 & W_{2,1} & \cdots & W_{2,K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & D_n & W_{n,1} & \cdots & W_{n,K} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \gamma_1 \\ \vdots \\ \gamma_K \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Estimateur MCO dans le cas multivarié

L'estimateur des MCO est celui qui minimise la somme des carrés des résidus:

$$\min_{\beta} \varepsilon'\varepsilon = (Y - X\beta)'(Y - X\beta) = Y'Y - 2\beta'X'Y + \beta'X'X\beta$$

La condition du premier ordre est:

$$\frac{\partial(Y'Y - 2\beta'X'Y + \beta'X'X\beta)}{\partial\beta} = -2X'Y + 2X'X\beta = 0$$

Si $(X'X)$ est inversible (cf H4) , alors

$$(X'X)\beta = X'Y \implies \beta = (X'X)^{-1}X'Y$$

Back