unisanté
Centre universitaire
de médecine générale
et santé publique · Lausanne

Forum de Statistique Unisanté

Comment sélectionner les variables explicatives dans un modèle statistique: Les études descriptives, prédictives et causales

Patrick Taffé, PhD, PD

DFRI / Division de biostatistique
Unisanté
patrick.taffe@unisante.ch

Lausanne, le 24 mars 2022

Plan de l’exposé

1. Avant-propos : L’analyse de régression linéaire
2. Facteur confondant versus facteur pronostique
3. Les différents types d’études : descriptives, prédictives et causales
4. La sélection des variables explicatives dans les différents types d’études
5. Illustration au moyen des données de LaLonde
6. Conclusions

1. Avant-propos : L’analyse de régression linéaire

La question à laquelle nous allons tenter d’apporter une réponse est :

« Comment sélectionner les variables explicatives dans un modèle statistique »

La réponse dépend du type d’étude envisagé :

« Descriptive, prédictive ou causale »

Nous allons nous placer dans le contexte de l'analyse de régression dont le but est d'établir une relation entre un outcome/variable dépendante et des régresseurs/variables explicatives :

$$Y_i = \underbrace{E(Y_i \mid X_{1i} = x_{1i}, X_{2i} = x_{2i}, ..., X_{ki} = x_{ki}; \beta_0, \beta_1, ..., \beta_k)}_{\text{modèle de régression}} + \varepsilon_i$$

où $E(Y_i \mid X_{1i} = x_{1i}, X_{2i} = x_{2i}, ..., X_{ki} = x_{ki}; \beta_0, \beta_1, ..., \beta_k)$ représente l'espérance conditionnelle, i.e. la moyenne de l'outcome pour une valeur fixée des régresseurs $X_{1i} = x_{1i}, X_{2i} = x_{2i}, ..., X_{ki} = x_{ki}$, les coefficients $\beta_0, \beta_1, ..., \beta_k$ sont des paramètres d'intérêt inconnus à estimer et ε_i est le terme d'erreur habituel, dont on suppose souvent la Normalité conditionnelle :

$$\varepsilon_i \mid X_{1i} = x_{1i}, X_{2i} = x_{2i}, ..., X_{ki} = x_{ki} \sim N(0, \sigma_\varepsilon^2).$$

Dans le cas où l'on a postulé un modèle simple de régression linéaire multiple (forme fonctionnelle linéaire, sans interactions), le modèle s'écrit :

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \varepsilon_i$$

L'erreur ε_i existe pour plusieurs raisons :

1. L'erreur permet de prendre en compte l'effet conjoint sur la variable dépendante des facteurs explicatifs non-inclus dans le modèle ; remarquons qu'un modèle de régression inclut toujours un nombre fini de variables explicatives et constitue, par la force des choses, une représentation simplifiée de la réalité.

2. Il y a toujours des erreurs de mesures. Par exemple, dans une expérience où l'on contrôle les variables d'input $X_{1i} = x_{1i}, X_{2i} = x_{2i}, ..., X_{ki} = x_{ki}$, la variable dépendante Y est presque toujours mesurée avec une certaine erreur de mesure qui sera prise en compte par l'erreur. On suppose, en revanche, que les variables explicatives sont mesurées sans erreurs (ou avec de très petites erreurs de mesure).

3. L’erreur permet de prendre en compte les erreurs de spécification du modèle ; en effet la relation de régression postulée entre les variables explicatives $X_{1i} = x_{1i}, X_{2i} = x_{2i}, ..., X_{ki} = x_{ki}$ et la variable dépendante Y, e.g. linéaire, n’est qu’approximative, la « vraie » relation fonctionnelle n’étant pas connue et certainement pas exactement linéaire.

Le tableau suivant résume les hypothèses du modèle de régression linéaire multiple (Greene, 2008 ; Taffé, 2014) :

Hypothèses du modèle de régression linéaire multiple :

Hypothèse 1 : l'« *Existence* »

Pour toute valeur $x_1, x_2, ..., x_k$ des variables explicatives $X_1, X_2, ..., X_k$, Y est une variable aléatoire avec une certaine distribution de probabilité ayant une espérance $\mu_{Y|X_1,..., X_k}$ et une variance $\sigma^2_{Y|X_1,..., X_k}$ finies.

Hypothèse 2 : l'« *Indépendance* »

Les variables Y_i, $i = 1,..., n$, sont indépendantes les unes des autres conditionnellement à $X_{i1} = x_{i1}, ..., X_{ik} = x_{ik}$.

Hypothèse 3 : l'« *Exogénéité* »

La condition suivante : $E(\varepsilon_i \mid X_{1i} = x_{1i}, ..., X_{ki} = x_{ki}) = 0$, $i = 1,..., n$, est vérifiée lorsque les variables explicatives X_j, $j = 1, \ldots, k$, sont exogènes, ce qui implique l'absence de corrélation entre l'erreur ε_i et $X_{1i}, ..., X_{ki}$.

Le tableau suivant résume les hypothèses du modèle de régression linéaire multiple :

Hypothèse 4 : la « *Linéarité* »
L'espérance de Y_i, conditionnelle à $X_{i1} = x_{i1}, ..., X_{ik} = x_{ik}$,
$E(Y_i \mid X_{1i} = x_{1i}, ..., X_{ki} = x_{ki}) = \mu_{Y|X_1,..., X_k}$, est linéaire dans les paramètres.

Hypothèse 6 : la « *Normalité* »
La distribution de Y_i, conditionnelle à $X_{i1} = x_{i1}, ..., X_{ik} = x_{ik}$, est Normale :
$f(Y_i \mid X_{1i} = x_{1i}, ..., X_{ki} = x_{ki}) = N(\mu_{Y|X_1,... X_k}, \sigma_\varepsilon^2)$

Hypothèse 7 : l'« *absence d'erreurs de mesure* »
Les régresseurs $X_{i1} = x_{i1}, ..., X_{ik} = x_{ik}$ sont mesurés sans erreur.

Hypothèse 8 : l'« *absence de multicolinéarité parfaite* »
Les régresseurs $X_{i1} = x_{i1}, ..., X_{ik} = x_{ik}$ sont linéairement indépendants.

L'hypothèse la plus fondamentale est celle d'exogénéité :

$$E(\varepsilon_i \mid X_{1i} = x_{1i}, \ldots, X_{ki} = x_{ki}) = 0\,,\ i = 1, \ldots, n$$

car c'est sur cette hypothèse que repose la plupart des résultats concernant les « bonnes » propriétés des estimateurs (absence de biais, convergence, Normalité asymptotique).

En remaniant l'équation de régression, l'on peut exprimer l'erreur ε_i comme suit :

$$\varepsilon_i = Y_i - \underbrace{E(Y_i \mid X_{1i} = x_{1i}, X_{2i} = x_{2i}, ..., X_{ki} = x_{ki}; \beta)}_{\text{modèle de régression}}$$

En calculant l'espérance conditionnelle, l'on obtient la *condition d'exogénéité* :

$$E(\varepsilon_i \mid X_{1i} = x_{1i}, X_{2i} = x_{2i}, ..., X_{ki}) = 0\,,\ i = 1,...,n$$

à condition que le modèle ait été correctement spécifié, i.e. :

- Qu'il intègre tous les facteurs confondants les plus importants
- Que la relation fonctionnelle de chacune des variables soit adéquate (y compris toutes les interactions existantes)
- Qu'il n'y ait pas de problème de simultanéité entre l'une des variables explicatives X_{ji} et l'outcome Y_i (i.e. une variation de Y_i implique une variation X_{ji} et réciproquement)
- Que les régresseurs $X_{1i} = x_{1i}, X_{2i} = x_{2i}, ..., X_{ki} = x_{ki}$ ont été mesurés sans erreurs

Remarque

Une conséquence importante de l'hypothèse d'exogénéité est que l'investigateur peut inférer les caractéristiques de la distribution de Y_i conditionnelle à X_i (i.e. $f(y \mid x)$), telles que l'espérance conditionnelle et la variance conditionnelle, sans avoir à modéliser le processus ayant généré les régresseurs X_i, ce qui simplifie considérablement la tâche du modélisateur.

Pour illustrer ce point, considérons la définition alternative suivante de l'exogénéité.

La variable X_i est exogène au sens faible lorsque la distribution jointe $f(y, x; \theta)$ peut se factoriser comme suit :

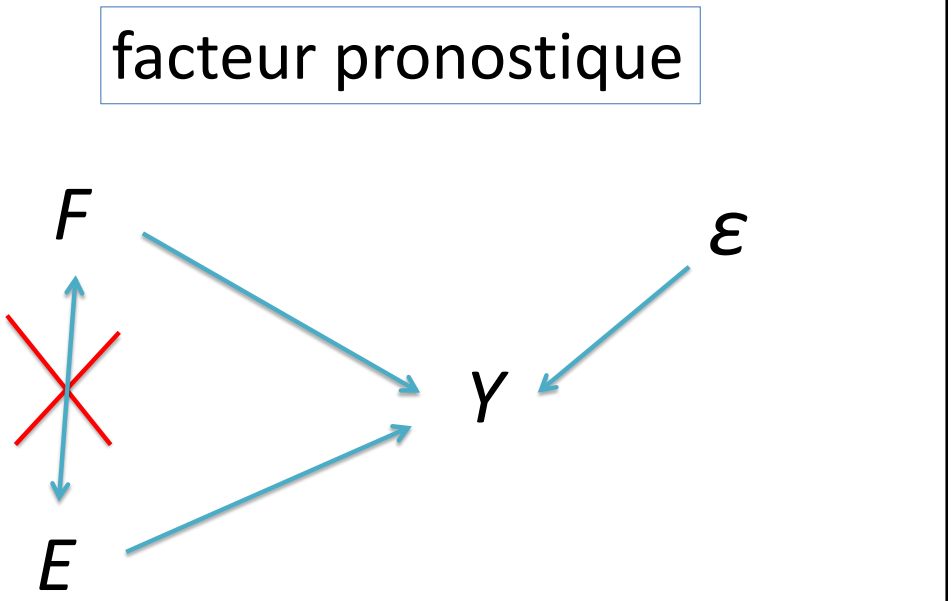
$$f(y, x; \theta) = f_{Y|X}(y \mid x; \theta_1) f_X(x; \theta_2)$$

où les paramètres $\theta_1 \in \Theta_1$ et $\theta_2 \in \Theta_2$, $\theta \in \Theta = \Theta_1 \times \Theta_2$, sont libres de varier séparément.

2. Facteurs confondants versus facteurs pronostiques

En fonction de l'objectif de l'étude (descriptive, prédictive ou causale), il est parfois important de distinguer les facteurs confondants des facteurs pronostiques.

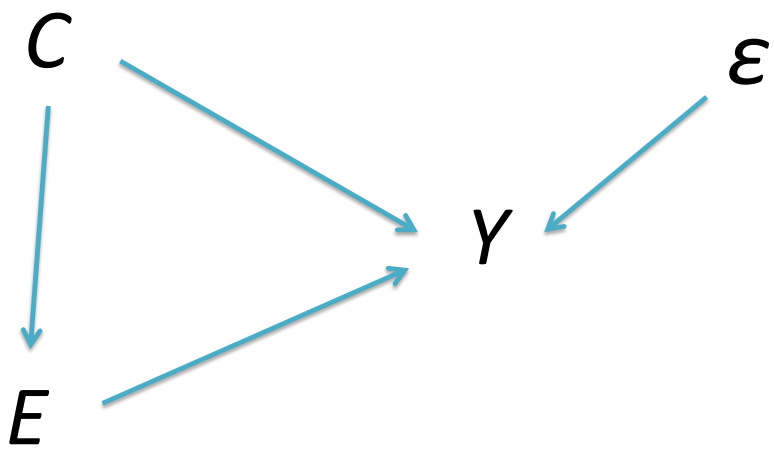
Pour cela, il est utile d'utiliser un DAG (Directed Acyclic Graph) :



Un facteur pronostique *F* est une variable qui est cause de l'outcome *Y*, mais n'est pas associée avec l'exposition.

Bien évidemment, pour que *F* soit cause de *Y* il faut que *F* soit définit avant *Y*.

facteur confondant



Un facteur confondant *C* est une variable qui est à la fois cause de l'outcome *Y* et cause de l'exposition *E*.

Bien évidemment, pour que *C* soit à la fois cause de *Y* et cause de *E*, il faut que *C* soit définit avant *Y* et *E*.

Remarquons que dans le DAG le sens des flèches est fondamental (on reviendra sur ce point plus loin).

3 La sélection des variables explicatives

On peut distinguer au moins trois différents types d'études dont les objectifs d'inférence diffèrent :

- études « descriptives » ou « d'associations »
- études « prédictives »
- études « étiologiques »

3.1 Les études descriptives

Dans une étude descriptive ou d'associations, l'objectif est avant tout de décrire aussi simplement que possible les relations entre les variables considérées (i.e. les corrélations partielles ou associations conditionnelles) au moyen d'un modèle descriptif.

On peut s'intéresser, par exemple, à la relation de régression entre une variable dépendante (e.g. la pression systolique) et des variables indépendantes (âge, genre, BMI, etc.) :

$$SBP = \beta_0 + \beta_1 age + \beta_2 sexe + \beta_3 BMI + \varepsilon$$

Cette description permet, par exemple, d'identifier des sous-groupes d'individus, e.g. à haut, moyen, et bas risque d'hypertension.

3.2 Les études prédictives

Dans une étude prédictive ou pronostique, l'investigateur cherche avant tout à construire un modèle permettant de prédire le mieux possible un outcome future.

Un modèle prédictif est utile typiquement pour faire de la classification, comme classer des patients arrivant à l'hôpital en fonction de leur niveau de risque,

ou diagnostiquer certaines affections, par exemple sur la base des pixels d'une image,

détecter des lésions cancéreuses.

3.3 Les études étiologiques

Dans une étude étiologique ou causale, l'objectif principal est d'estimer l'effet causal d'une intervention (un traitement ou une exposition particulière) sur un outcome.

Le gold standard est sans conteste l'essai randomisé (RCT).

Dans le contexte d'une étude observationnelle, l'objectif d'une analyse causale est d'obtenir un « effect-size » de même amplitude que si l'on avait pu randomiser.

Il est donc crucial dans d'une étude observationnelle que l'effet soit estimé avec un biais aussi faible que possible et les corrélations partielles ou la performance prédictive du modèle passent au second plan.

4 La sélection des variables explicatives dans les différents types d'études

La sélection des variables explicatives (ou covariables) à intégrer dans un modèle de régression dépend de l'objectif de l'étude (Hernan et al., 2006; Sauerbrei et al., 2007 ; Shmueli, 2010) :

- étude « descriptive » ou « d'associations »
- étude « prédictive »
- étude « étiologique »

4.1 La sélection des régresseurs dans les études descriptives

Dans une étude descriptive/d'association, l'objectif est avant tout de décrire aussi simplement que possible les relations entre les variables considérées (e.g. calculer des corrélations partielles ou associations conditionnelles) au moyen d'un modèle descriptif.

Dans une étude descriptive, toutes les variables explicatives d'intérêt et potentiellement prédictives sont considérées

la sélection des variables à retenir dans le modèle (descriptif) final repose largement sur des considérations purement statistiques :

- test de significativité des paramètres (e.g. p-value ≤ 0.05)
- goodness of fit du modèle (e.g. coefficient de détermination)

$$SBP = \beta_0 + \beta_1 age + \beta_2 sexe + \beta_3 BMI + \varepsilon$$

Dans un modèle descriptif, le coefficient associé à une variable explicative mesure le degré d'association entre cette variable et l'outcome ;

par exemple, sous un modèle Normal multivarié, le coefficient associé à une variable explicative permet de tester la corrélation partielle.

$$SBP = \beta_0 + \beta_1 age + \beta_2 sexe + \beta_3 BMI + \varepsilon$$

Lorsque le modèle est linéaire, le coefficient associé à une variable explicative mesure le contraste entre deux sous-populations hétérogènes définies par une variation unitaire de la variable en question.

Pour le montrer, considérons l'exemple de la relation entre pression systolique (SBP) et l'âge, le genre et le BMI et calculons l'impact d'un accroissement unitaire du BMI sur SBP :

$$SBP_1 = \beta_0 + \beta_1 BMI + \beta_2 age + \beta_3 sexe + \varepsilon_1$$

$$SBP_2 = \beta_0 + \beta_1 (BMI + 1) + \beta_2 age + \beta_3 sexe + \varepsilon_2$$

$$\Rightarrow \quad SBP_2 - SBP_1 = \beta_1 + (\varepsilon_2 - \varepsilon_1)$$

$$\Leftrightarrow \quad \beta_1 = (SBP_2 - SBP_1) - (\varepsilon_2 - \varepsilon_1)$$

De cette dernière équation, il apparait clairement que β_1 mesure l'impact d'un accroissement unitaire du BMI au niveau individuel uniquement lorsque la condition suivante est vérifiée :

$$\varepsilon_2 = \varepsilon_1$$

C'est-à-dire, lorsque que toutes les variables non-mesurées (représentées par l'erreur) et qui affectent le niveau de l'outcome ont été fixées.

$$\Leftrightarrow \quad \beta_1 = (SBP_2 - SBP_1) - (\varepsilon_2 - \varepsilon_1)$$

Lorsque la condition $\varepsilon_2 = \varepsilon_1$ n'est pas vérifiée, β_1 mesure le contraste entre deux sous-populations hétérogènes qui diffèrent non seulement de par une variation unitaire de la variable d'intérêt (le BMI dans notre exemple)…

mais aussi de par les variables non mesurées qui sont aussi associées avec l'outcome (facteurs confondants et pronostiques non mesurés).

C’est seulement lorsque tous les facteurs confondants ont été intégrés dans le modèle et que l’hypothèse d’exogénéité est vérifiée que β_1 admet une interprétation claire.

En effet, lorsque l’hypothèse d’exogénéité est valide, l’on a :

$$E(SBP_2 - SBP_1 \mid age, sexe) = \beta_1 + \underbrace{E(\varepsilon_2 - \varepsilon_1 \mid age, sexe)}_{=0 \text{ sous H3}} = \beta_1$$

ce qui peut aussi s’écrire :

$$\begin{aligned} E(SBP_2 - SBP_1 \mid age, sexe) &= E(SBP_2 \mid age, sexe) - E(SBP_1 \mid age, sexe) \\ &= \beta_1 \end{aligned}$$

de sorte que sous l’hypothèse d’exogénéité β_1 mesure le contraste des moyennes des SBP calculées dans les deux sous-populations définies par une variation unitaire du BMI.

Autrement dit, lorsque tous les facteurs confondants ont été mesurés et inclus dans le modèle et que le modèle ait été correctement spécifié, de sorte que l'hypothèse d'exogénéité soit vérifiée,

β_1 mesure l'impact *ceteris paribus* d'un accroissement unitaire du BMI, i.e. comme si l'on avait pu contrôler tous les facteurs affectant l'outcome :

=> Il s'agit d'un effet causal

En effet, même si les facteurs pronostiques n'ont pas tous été inclus dans le modèle, leurs effets n'affectent pas le contraste calculé.

En pratique, dans une étude descriptive la question est :

« l'hypothèse d'exogénéité est-elle valide ? »

Malheureusement, compte tenu de la sélection des variables basée sur des tests de significativité, il est très probable que pas tous les facteurs confondants n'aient été identifiés et intégrés dans le modèle, de sorte que l'hypothèse d'exogénéité est violée dans ce type d'étude.

Admettons que l’investigateur ait pu identifier et intégrer tous les facteurs confondants les plus importants dans son modèle.

Peut-il prétendre a l’exogénéité ?

Malheureusement, encore une fois ce n'est probablement pas le cas…

car pour pouvoir prétendre qu'il y a exogénéité il ne suffit pas d'avoir intégré dans le modèle tous les facteurs confondants les plus importants,

il faut encore que le modèle de régression ait été correctement spécifié, c'est-à-dire que la relation fonctionnelle postulée entre l'outcome et les régresseurs soit correcte, y compris toutes les interactions existantes (de tous les ordres : 2, 3, etc.).

C'est une hypothèse très/trop forte, qui n'est que rarement justifiable en pratique et c'est en se tournant vers des méthodes d'analyse non-paramétriques…

notamment la méthode du score de propensité (Rosenbaum & Rubin, 1983; Dehejia & Wahba, 1999, 2002; Austin, 2011), qu'une solution peut être apportée.

4.2 La sélection des régresseurs dans les études prédictives

Dans une étude prédictive ou pronostique, l'investigateur cherche avant tout à construire un modèle permettant de prédire le mieux possible un outcome future.

Dans une étude prédictive l'on considère tous les facteurs de risque potentiels et la sélection du modèle (i.e. des variables explicatives et de la forme fonctionnelle) repose essentiellement sur un critère de performance prédictive, comme l'erreur quadratique moyenne de prédiction :

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(SBP_i - \hat{SBP}_i)^2$$

où SBP_i correspond à une valeur future de pression systolique mesurée sur l'individu i et $\hat{SBP}_i = f(\text{age}, \text{sexe}, \text{BMI}, \ldots; \hat{\beta})$ est la prédiction fournie par le modèle.

L’objectif, dans une étude prédictive, est d’optimiser le critère de performance prédictive adopté,

par exemple, minimiser l’erreur quadratique moyenne de prédiction.

Il existe d’autres critères de performance prédictive, comme l’erreur absolue moyenne de prédiction, les critères AIC et BIC, etc.

Afin d'identifier les composantes de l'erreur quadratique moyenne de prédiction, il est utile de considérer la décomposition suivante (Bias-Variance decomposition) :

$$\text{Modèle : } Y_i = f(x_i \mid \beta) + \varepsilon_i, \quad \varepsilon_i \mid X_i \sim N(0, \sigma_\varepsilon^2)$$

$$MSE = E\left[\left(Y_i - \hat{f}(x_i \mid \hat{\beta})\right)^2 \mid X\right]$$

$$= \underbrace{\left[f(x_i \mid \beta) - \mathrm{E}\left(\hat{f}(x_i \mid \hat{\beta}) \mid X\right)\right]^2}_{Bias^2} + \underbrace{Var\left[\hat{f}(x_i \mid \hat{\beta}) \mid X\right] + \sigma_\varepsilon^2}_{\text{Variance components}}$$

où le premier terme de l'équation représente le carré du biais dans l'estimation de la relation fonctionnelle inconnue $f(x_i \mid \beta)$, le deuxième la variance de l'estimateur et σ_ε^2 la variance de l'erreur.

Modèle : $Y_i = f(x_i \mid \beta) + \varepsilon_i, \quad \varepsilon_i \mid X_i \sim N(0, \sigma_\varepsilon^2)$

$$MSE = E\left[\left(Y_i - \hat{f}(x_i \mid \hat{\beta})\right)^2 \mid X\right]$$

$$= \underbrace{\left[f(x_i \mid \beta) - \mathrm{E}\left(\hat{f}(x_i \mid \hat{\beta}) \mid X\right)\right]^2}_{Bias^2} + \underbrace{Var\left[\hat{f}(x_i \mid \hat{\beta}) \mid X\right] + \sigma_\varepsilon^2}_{\text{Variance components}}$$

Sur la base de cette décomposition, l'on en déduit qu'un modèle prédictif pas assez complexe risque d'introduire un biais,

tandis qu'un modèle prédictif trop complexe va générer une variance trop grande.

Comment procéder pour construire un bon modèle prédictif ?

Il s'agit d'un processus itératif.

L’investigateur commence par formuler un premier modèle :

$$Y_i = \hat{f}(x_i \mid \beta) + \varepsilon_i, \quad \varepsilon_i \mid X \sim N(0, \sigma_\varepsilon^2)$$

Afin d’évaluer de façon objective la performance prédictive de ce modèle, il faut qu’il utilise deux échantillons différents,

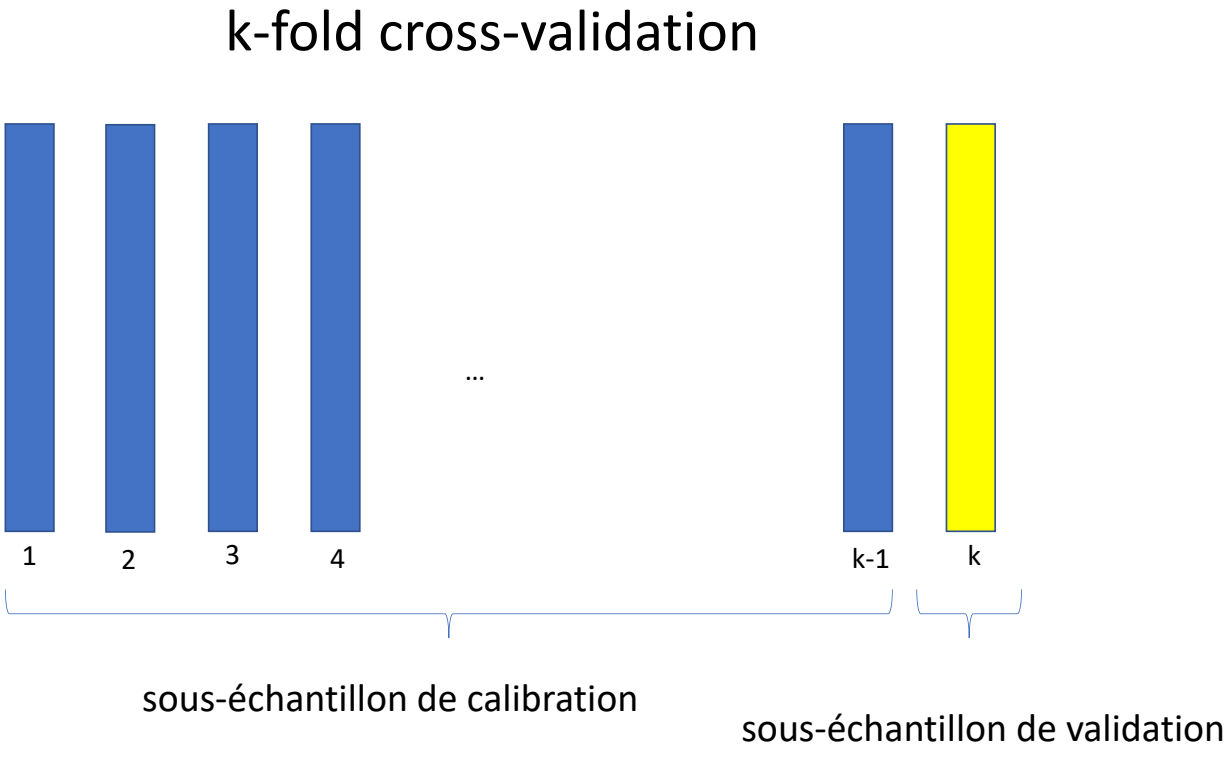
le premier, l’échantillon de calibration, avec lequel il va estimer les paramètres du modèle,

le deuxième, l’échantillon de validation, avec lequel il va calculer le critère de performance prédictive.

Afin que les résultats ne dépendent pas d'un partage particulier de l'échantillon, l'investigateur va utiliser la méthode de validation croisée (Shao J,1993), par exemple la méthode « *k*-fold cross-validation ».

Alternativement, il pourra utiliser la méthode du bootstrap (George, 2000 ; Altman and Royston, 2000).

Dans la méthode « *k*-fold cross-validation », l'échantillon original est divisé de manière aléatoire en *k* sous-échantillons de tailles égales. Dans une première étape, les *k* – 1 premiers sous-échantillons servent d'échantillon de calibration et le dernier d'échantillon de validation :



Puis, ce processus est répété *k* - 1 fois, en utilisant à chaque fois un autre sous-échantillon de validation, de sorte à avoir utilisé une seule fois tous les sous-échantillons pour la validation.

Finalement, la moyenne des *k* erreurs quadratiques moyennes (MSE) est calculée

et utilisée comme valeur du critère de performance prédictive de ce premier modèle.

Comme ce premier modèle n'est pas forcément celui qui a la performance prédictive la meilleures, l'investigateur va considérer un deuxième modèle, différents du premier.

Ce deuxième modèle pourra comporter d'autres variables explicatives et/ou une autre relation fonctionnelle et/ou d'autres interactions.

Puis il considérera un troisième modèle et ainsi de suite…

L’investigateur peut considérer autant de modèles prédictifs qu’il le veut…

et parmi tous les modèles investigués, il retiendra celui performant le mieux en termes d’optimisation du critère de performance prédictive (i.e. celui minimisant l’erreur quadratique moyenne de prédiction).

Remarquons que le modèle est écrit, ici, sous une forme très générale :

$$SBP_i = f(\text{age}, \text{sexe}, \text{BMI}, \ldots; \beta) + \varepsilon$$

où f est une fonction aussi complexe que nécessaire, mais aussi simple que possible ☺, pour obtenir une « bonne » performance prédictive.

$$SBP_i = f(\text{age}, \text{sexe}, \text{BMI}, \ldots; \beta) + \varepsilon$$

Typiquement, le modèle peut intégrer plus de paramètres à estimer que d'observations…

l'estimation se faisant par des méthodes de « régularisation » ou « shrinkage » (LASSO, Ridge, Elastic net, Réseaux de neurones, etc.),

et le choix de la forme fonctionnelle *f* est complètement libre, pour autant que le modèle retenu optimise le critère de performance prédictive.

$$SBP_i = f(\text{age}, \text{sexe}, \text{BMI}, \ldots; \beta) + \varepsilon$$

Dans un modèle purement prédictif, les variables explicatives retenues dans le modèle final ne sont pas forcément statistiquement significatives,

puisque le critère de sélection du modèle cible la performance prédictive (et non pas les associations statistiquement significatives).

De la même façon, les facteurs confondants d'une association ne sont pas pris en compte.

Par conséquent le modèle ne peut pas répondre à une question causale…

$$SBP_i = f(\text{age}, \text{sexe}, \text{BMI}, \ldots; \beta) + \varepsilon$$

En ce qui concerne le choix de la forme fonctionnelle *f*, il existe une énorme littérature sur le développement de formes fonctionnelles flexibles,

comme les méthodes de Regression Tree, Random Forest, Bagging, Boosting, Réseaux de neurones, etc.

Ces méthodes statistiques ont, pour certaines, été développées déjà au siècle passé et ont tout récemment trouvé un énorme regain d'intérêt, compte tenu des puissances de calculs disponibles aujourd'hui.

Elles sont connues/ont été popularisées sous le vocable de « Machine Learning » (Hastie et al. (2009) The Elements of Statistical Learning).

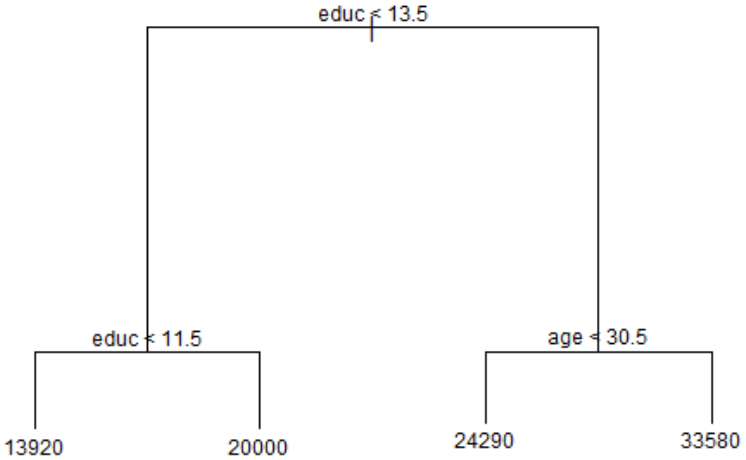
Par exemple, la méthode Random Forest

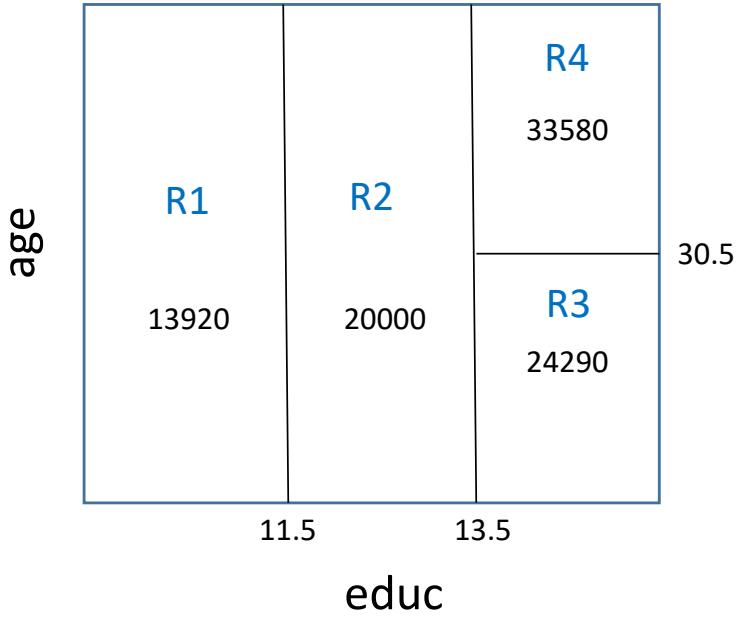
qui consiste à générer un grand nombre d’arbres de régression, au moyen d’échantillons bootstrap,

et à calculer les prédictions en considérant la moyenne des valeurs prédites pour chaque observation dans chaque arbre.

Pour se fixer les idées, considérons un exemple simple où la variable dépendante est le revenu annuel et les variables explicatives sont l'âge et le niveau d'éducation (i.e. nombre d'années de formation) :

Regression Tree





La forme fonctionnelle correspondante est simplement donnée par :

$$f(x_i;\beta) = \sum_{j=1}^{N} \beta_j I(x_i) \in R_j$$

où $I(x_i)$ est une fonction indicatrice d'appartenance de l'observation *i* à la région R_j .

Il s'agit donc d'une fonction en escaliers dont les marches seront d'autant plus étroites que le nombre de régions sera grand.

Dans le cas de la méthode Random Forest les marches sont d'autant plus étroites que l'on aura généré un grand nombre d'arbres.

On peut bien sûr calculer des contrastes entre des sous-groupes d’individus,

mais il ne s’agira ni de contrastes statistiquement significatifs,

ni de contrastes causaux,

simplement de contrastes de valeurs prédites.

4.3 La sélection des régresseurs dans les études étiologiques

Dans une étude observationnelle étiologique ou causale, l'objectif principal est d'estimer l'effet causal d'une intervention (un traitement ou une exposition particulière) comme si contrefactuellement l'on avait pu randomiser.

Par conséquent, le choix des variables à introduire dans le modèle causal repose avant tout sur l'identification des facteurs confondants les plus importants…

mais aussi les variables pour lesquelles il ne faut pas ajuster, comme les facteurs intermédiaires, les médiateurs et les colliders.

L'identification de ces différents facteurs peut se faire avantageusement sur la base d'un graphe acyclique ou DAG (Directed Acyclic Graph) (Hernan et al., 2002; Greenland et al., 1999).

Dans une étude causale, une variable statistiquement non-significative peut très bien être conservée dans le modèle afin de contrôler un biais potentiel de confusion.

L'estimation de l'effet causal doit se faire autant que possible de façon non-paramétrique, car un modèle paramétrique extrapole les effets...

et risque de fournir des estimations biaisées (Dehejia & Wahba, 1999 ; Ho et al., 2007).

A cette fin, la méthode du score de propensité est très utile.

Remarquons que la définition classique en épidémiologie d'un facteur confondant :

une variable associée à la fois à l'exposition et à l'outcome

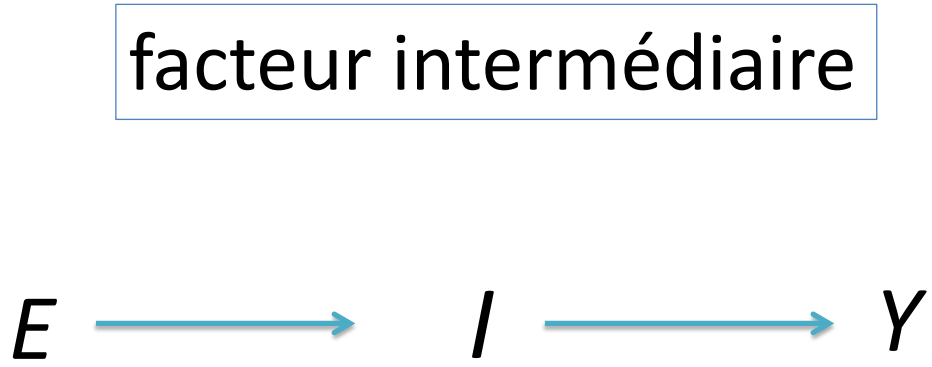
n'est pas suffisante pour identifier un facteur confondant…,

car elle ne permet pas de distinguer les facteurs confondants des facteurs intermédiaires, des médiateurs et des colliders.

L'utilisation de DAGs permet simplement de mettre en lumière ces concepts.

Le DAG

L'exemple suivant illustre au moyen d'un DAG une situation où une variable est *associée à la fois à l'exposition et à l'outcome* et pourtant l'ajustement pour cette variable dans un modèle de régression induit un biais, alors que sans ajustement il n'y a pas de biais :

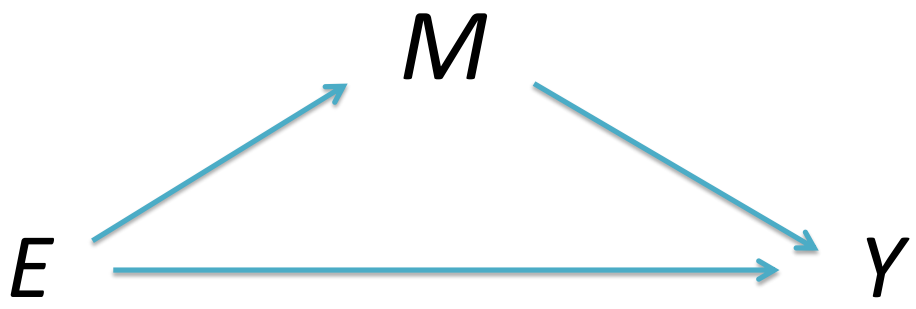


L'ajustement par le facteur intermédiaire (*I*) dans un modèle de régression masque l'effet de l'exposition (*E*) sur l'outcome (*Y*) et introduit un biais.

Sans ajustement, le coefficient associé à l'exposition mesure l'effet causal de *E* sur *Y*.

La figure ci-dessous illustre une autre situation où une variable est *associée à la fois à l'exposition et à l'outcome* et pourtant l'ajustement pour cette variable dans un modèle de régression induit un biais :

la médiation



L'ajustement par le médiateur (*M*) masque une partie de l'effet de l'exposition (*E*) sur l'outcome (*Y*).

Dans ce cas, le coefficient associé à l'exposition ne mesure pas l'effet causal total de *E* sur *Y*, mais uniquement l'effet direct.

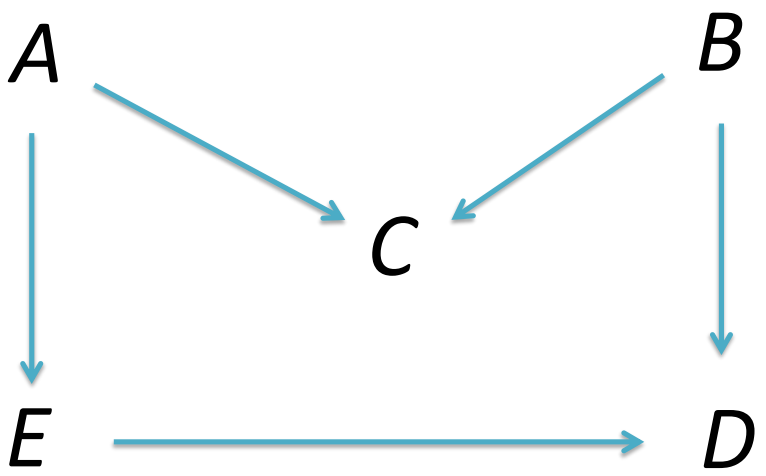
Pour estimer l'effet indirect de *E* sur *Y* il faut aussi modéliser la variable endogène *M*.

L'effet total s'obtient comme la somme de l'effet direct et de l'effet indirect.

La médiation correspond à une situation où l'exposition a un effet direct ainsi qu'un effet indirect sur l'outcome, par le biais d'un médiateur (Ten Have & Joffe, 2010).

L'exemple suivant illustre encore une autre situation où une variable est *associée à la fois à l'exposition et à l'outcome* et pourtant l'ajustement pour cette variable dans un modèle de régression induit un biais, alors que sans ajustement il n'y a pas de biais :

exemple d'un collider bloquant le « backdoor path » EACBD



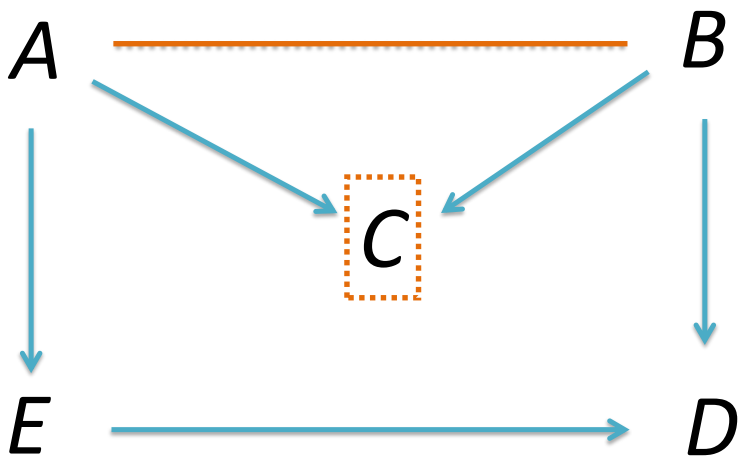
Dans cet exemple, *A* affecte directement *E* et *C*, *B* affecte directement *C* et *D*, mais *A* et *B* sont indépendants (i.e. ils ne sont pas associés marginalement).

Le facteur *C* est appelé « collider » et bloque le « backdoor path » (i.e. passage) *EACBD*.

L'estimation directe de l'impact de *E* sur *D* (sans ajustement) est sans biais,

puisque le facteur *C* bloque le passage *EACBD* ; il n'y a pas de biais de confusion.

exemple d'ajustement inutile même délétère



En revanche, l'ajustement pour le facteur *C* génère une association artificielle entre *A* et *B* (marginalement *A* et *B* ne sont pas associés), et du coup entre *E* et *D* (puisque le « backdoor path » *EABD* n'est pas bloqué), ce qui engendre un biais de confusion.

Pour obtenir une estimation sans biais de l'impact de *E* sur *D*, lorsque l'on ajuste pour *C*, il faut aussi ajuster pour *A* ou pour *B* afin de bloquer le « backdoor path » artificiel entre *A* et *B*.

Cet exemple illustre que parfois l'ajustement pour une variable peut engendrer la nécessité d'ajuster pour une autre afin de ne pas générer un biais de confusion.

L'exemple du « collider » illustre clairement que dans le cadre d'une analyse causale la sélection des variables d'ajustement ne peut pas se faire sur la base de critères purement statistiques (tests de significativité, p-values…),

mais requière une connaissance contextuelle afin d'identifier correctement les variables pour lesquelles il faut ajuster (les facteurs confondants)

et celles pour lesquelles il ne faut pas ajuster (les « colliders », les médiateurs, les facteurs intermédiaires).

Enfin, l’identification des facteurs confondants est essentielle mais pas encore suffisante pour obtenir un effet causal.

En effet, les erreurs de spécification du modèle entraînent immanquablement une distorsion dans les effets estimés

et il faut utiliser des méthodes d’analyse non-paramétriques pour obtenir cet effet, comme la méthode du score de propensité (Dehejia & Wahba, 1999; Austin, 2011).

5 Illustration au moyen des données de LaLonde

Objectif : Evaluer l'impact d'un programme d'occupation temporaire (i.e. l'exposition ou traitement d'intérêt) sur le revenu future (i.e. l'outcome d'intérêt) d'individus étant dans une situation socio-économique difficile.

Méthode : Pour répondre à sa question de recherche, LaLonde (1986) avait réalisé deux études, une étude expérimentale, dans laquelle il avait comparé les revenus moyens en 1978 des 185 personnes ayant participé au programme de réinsertion avec celui des 260 contrôles n'ayant pas bénéficié du programme,

et quelque années plus tard une étude observationnelle dans laquelle il avait repris le revenu moyen de 1978 des 185 personnes ayant participé au programme et comparé ce revenu avec celui (de 1978) d'un groupe de contrôle formé par les 2490 personnes ayant participé à l'enquête PSID (Panel Study of Income Dynamics) réalisée en 1986.

Dans l'étude observationnelle, pour prendre en compte les différences entre le groupe traité et le groupe non-traité, LaLonde avait considéré les variables d'ajustement suivantes (sélectionnées d'après la théorie économique) :

- l'âge (`age`)
- le niveau d'éducation (nb. d'années de formation) (`educ`)
- l'origine ethnique (black/other) (`black`)
- hispanique (oui/non) (`hispanic`)
- marié (oui/non) (`married`)
- variable indicatrice d'absence de diplôme (1/0) (`nodegree`)
- le revenu de 1974 (`re74`)
- le revenu de 1975 (`re75`)
- variable indicatrice de non emploi en 1974 (`u74`)
- variable indicatrice de non emploi en 1975 (`u75`)

Résultats de l’étude expérimentale

Dans l’étude expérimental, la comparaison des revenus de 1978 chez les 185 traités versus 260 non-traités a produit une différence des moyennes de 1794 $ en faveur de l’intervention.

Résultats de l'étude observationnelle

Les résultats dépendent de la méthode d'analyse adoptée :

1) La comparaison brute des revenus moyens de 1978 des 185 traités versus les 2490 non-traités a produit une différence de -15'2005 $:

```
------------------------------------
T |    N(re78)  mean(re78)
----------+-------------------------
0 |      2,490    21553.92
1 |        185    6349.144
------------------------------------
```

```
      Source |       SS       df       MS              Number of obs =    2675
-------------+------------------------------           F(  1,  2673) =  173.41
       Model |  3.9811e+10     1  3.9811e+10           Prob > F      =  0.0000
    Residual |  6.1365e+11  2673   229573194           R-squared     =  0.0609
-------------+------------------------------           Adj R-squared =  0.0606
       Total |  6.5346e+11  2674   244375671           Root MSE      =   15152

------------------------------------------------------------------------------
        re78 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           T |  -15204.78   1154.614   -13.17   0.000    -17468.81   -12940.75
       _cons |   21553.92   303.6414    70.98   0.000     20958.53    22149.32
------------------------------------------------------------------------------
```

Soit une baisse de revenu de -15204.78 $ (à contraster avec 1794 $).

2) L'estimation d'un modèle de régression linéaire simple produit les résultats suivants :

```
      Source |       SS       df       MS              Number of obs =    2675
-------------+------------------------------           F( 11,  2663) =  344.41
       Model |  3.8373e+11    11  3.4885e+10           Prob > F      =  0.0000
    Residual |  2.6973e+11  2663   101287377           R-squared     =  0.5872
-------------+------------------------------           Adj R-squared =  0.5855
       Total |  6.5346e+11  2674   244375671           Root MSE      =   10064

------------------------------------------------------------------------------
        re78 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           T |    4.157964   1013.976     0.00   0.997    -1984.102    1992.418
         age |   -91.83894     22.053    -4.16   0.000    -135.0817   -48.59621
        educ |    579.9227   104.1238     5.57   0.000     375.7509    784.0944
     married |    1225.366   585.9115     2.09   0.037     76.47799    2374.253
       black |    -482.818   497.8437    -0.97   0.332    -1459.017    493.3815
    hispanic |    2195.133   1091.664     2.01   0.044     54.53778    4335.727
    nodegree |    601.6139   646.3833     0.93   0.352    -665.8501    1869.078
        re74 |    .3129598    .031634     9.89   0.000        .25093    .3749896
        re75 |    .5435253   .0309048    17.59   0.000     .4829254    .6041252
         u74 |    2396.491   1024.492     2.34   0.019     387.6109    4405.372
         u75 |   -1468.213   947.2429    -1.55   0.121     -3325.62    389.1927
       _cons |    30.26114   1691.965     0.02   0.986    -3287.437    3347.959
------------------------------------------------------------------------------
```

Soit un accroissement du revenu de 4 $ (à contraster avec 1794 $).

3) Dans le cas où l'on a sélectionné les régresseurs sur la base de tests de significativité :

```
      Source |       SS           df       MS      Number of obs   =     2,675
-------------+----------------------------------   F(10, 2664)     =    379.57
       Model |  3.8397e+11         10  3.8397e+10   Prob > F        =    0.0000
    Residual |  2.6949e+11      2,664   101159316   R-squared       =    0.5876
-------------+----------------------------------   Adj R-squared   =    0.5860
       Total |  6.5346e+11      2,674   244375671   Root MSE        =     10058

------------------------------------------------------------------------------
        re78 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           T |  -860.3545   1052.138    -0.82   0.414    -2923.445    1202.736
         age |  -85.08692   22.00993    -3.87   0.000    -128.2452   -41.92863
        educ |   556.7908   74.54669     7.47   0.000     410.6155    702.966
     married |   1327.449   583.1378     2.28   0.023     184.0008    2470.898
    hispanic |   2627.694     1080.7     2.43   0.015     508.5983    4746.79
        re74 |   .3168921   .0316121    10.02   0.000     .2549054    .3788788
        re75 |   .5364448   .0311342    17.23   0.000     .4753952    .5974943
         u74 |   2556.751   1021.759     2.50   0.012       553.23    4560.271
         u75 |  -2442.298   1065.592    -2.29   0.022    -4531.769   -352.8278
   black_u75 |    2475.07   1226.064     2.02   0.044     70.93588    4879.205
       _cons |   122.8411   1283.533     0.10   0.924    -2393.981    2639.664
------------------------------------------------------------------------------
```

Soit une diminution du revenu de -860 $ (à contraster avec 1794 $).

4) Si l'on considère, un modèle de régression plus complexe, où la forme fonctionnelle des régresseurs continus (`age, educ, re_74_75`) a été déterminée au moyen de la méthode des polynômes fractionnels et où l'on a introduit des interactions entre certaines covariables (`blacku74, blacku75`) et considéré le revenu moyen des années 1974 et 1975 (`re_74_75`), l'on obtient :

```
      Source |       SS           df       MS      Number of obs   =     2,675
-------------+----------------------------------   F(12, 2662)     =    315.78
       Model |  3.8383e+11         12  3.1986e+10   Prob > F        =    0.0000
    Residual |  2.6963e+11     2,662   101290194   R-squared       =    0.5874
-------------+----------------------------------   Adj R-squared   =    0.5855
       Total |  6.5346e+11     2,674   244375671   Root MSE        =     10064

------------------------------------------------------------------------------
        re78 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           T |  -627.9992   1104.809    -0.57   0.570    -2794.371    1538.372
     Iage__1 |  -103.1298   21.80504    -4.73   0.000    -145.8864   -60.37328
    Ieduc__1 |   1129.315   180.0579     6.27   0.000      776.248    1482.383
     married |   1276.899   586.6718     2.18   0.030       126.52    2427.277
       black |   -841.736   521.7449    -1.61   0.107    -1864.802    181.3303
    hispanic |   2049.599   1091.367     1.88   0.060    -90.41346    4189.612
    nodegree |   255.4688   585.7932     0.44   0.663    -893.1871    1404.125
    Ire_7__1 |   .8480196     .02075    40.87   0.000     .8073319    .8887074
         u74 |   4666.319   1067.761     4.37   0.000     2572.594    6760.045
         u75 |   -4630.55   1048.497    -4.42   0.000    -6686.502   -2574.599
    blacku74 |   -2416.97   1956.525    -1.24   0.217    -6253.433    1419.493
    blacku75 |   4963.411   1832.781     2.71   0.007     1369.593    8557.229
       _cons |   19064.92   631.7692    30.18   0.000     17826.11    20303.73
------------------------------------------------------------------------------
Deviance:56887.902.
```

Soit une réduction du revenu de -629 $ (à contraster avec 1794 $).

Clairement, dans cet exemple (données Lalonde) l'analyse classique de régression linéaire multiple ne permet pas d'estimer l'effet causal de l'intervention…

5) Estimation par la méthode du score de propensité :

```
Treatment-effects estimation                    Number of obs      =      2,639
Estimator      : propensity-score matching      Matches: requested =          4
Outcome model  : matching                                      min =          4
Treatment model: logit                                         max =          5
------------------------------------------------------------------------------
             |               AI Robust
        re78 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
ATET         |
           T |
   (1 vs 0)  |   1779.104   802.6989     2.22   0.027     205.8428    3352.365
------------------------------------------------------------------------------
```

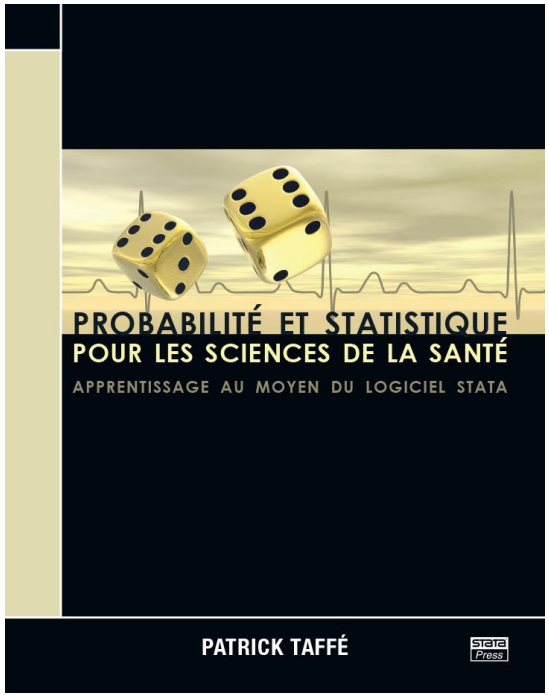
On obtient quasiment le même effet 1779 $ que dans l'étude randomisée 1794 $.

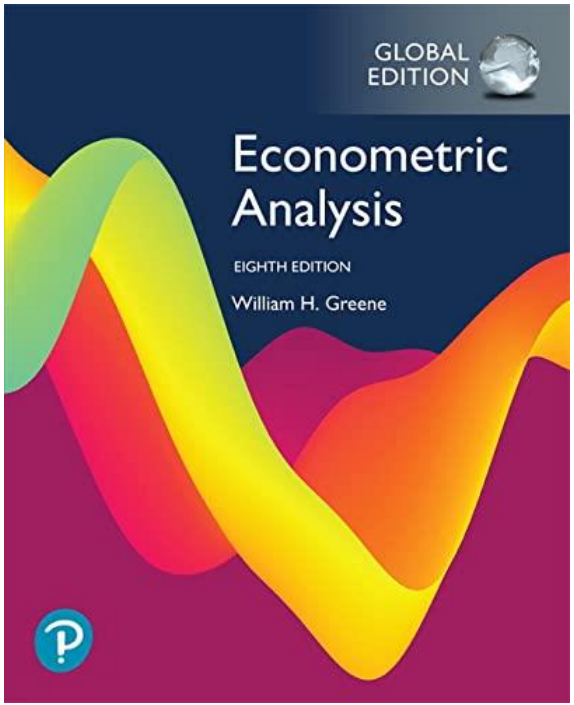
Manifestement, dans cet exemple, la méthode du score de propensité supplante les méthodes classiques d'analyse de régression (de l'outcome).

6 Conclusion

Nous avons illustré par ces propos qu'il était fondamental de distinguer le type d'étude, descriptive, prédictive ou causale, en fonction de ses objectifs d'inférence, car la sélection des variables explicatives et la modélisation en dépendent.

Pour plus de détails…





Références

- Altman DG and Royston P. **What do we mean by validating a prognostic model?** *Statistics in Medicine*, 19: 453-473, 2000.
- Austin PC. **An introduction to propensity score methods for reducing the effects of confounding in observational studies**. *Multivariate Behavioral Research* 2011; 46: 399-424.
- Dehejia RH and Wahba S. **Causal effects in nonexperimental studies: reevaluating the evaluation of training programs**. *Journal of the American Statistical Association* 1999; 94: 1053-1062.
- Dehejia RH and Wahba S. **Propensity score-matching methods for nonexperimental causal studies**. *Review of Economics and Statistics* 2002; 84: 151-161.
- George EI. **The variable selection problem**. *Journal of the American Statistical Association* 2000; 95: 1304-1308.
- Greene WH. (2008) ***Econometric Analysis***. Sixth Edition, Prentice Hall.
- Greenland S, Pearl J, and Robins JM. **Causal diagrams for epidemiologic research**. *Epidemiology* 1999; 10: 37-48.
- Hastie T, Tibshirani R, Friedman J. (2009) ***The Elements of statistical Learning***. Second Edition, Springer.
- Hernan MA, Hernadez-Diaz S, Werler MM et al. **Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology**. *American Journal of Epidemiology* 2002; 155:176-184.
- Hernan MA, Robins JM. **Estimating causal effects from epidemiological data**. *Journal of Epidemiology & Community Health* 2006; 60:578-586.
- Hernan MA. **A definition of causal effect for epidemiological research**. *Journal of Community Health* 2003; 58: 265-271.
- Ho DE, Imai K, King G, and Stuart EA. **Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference**. *Political Analysis* 2007; 15: 199-236.
- Holland PW. **Statistics and causal inference**. *Journal of the American Statistical Association* 1986, 81: 945-960.
- Imbens GW. **Nonparametric estimation of average treatment effects under exogeneity: a review**. *Review of Economics and Statistics* 2004; 86: 4-29.
- LaLonde R. **Evaluating the Econometric Evaluations of Training Programs**. *American Economic Review* 1986; 76: 604-620.
- Rosenbaum PR and Rubin D. **The central role of the propensity score in observational studies for causal effects**. *Biometrika* 1983; 70: 41-55.
- Rubin DB. **Which ifs have causal answers**. *Journal of the American Statistical Association* 1986a; 81: 961-962.
- Sauerbrei W, Royston P, Binder H. **Selection of important variables and determination of functional form for continuous predictors in multivariable model building**. *Statistics in Medicine* 2007; 26: 5512-5528.
- Shao J. **Linear model selection by cross-validation**. *Journal of the American Statistical Association* 1993; 88: 486-494.
- Shao J. **Linear model selection by cross-validation**. *Journal of the American Statistical Association* 1993; 88: 486-494.
- Shmueli G. **To Explain or to Predict?** *Statistical Science* 2010, 25:289-310.
- Stuart EA. **Matching methods for causal inference: a review and look forward**. *Statistical Science* 2010; 25: 1-21.
- Taffé P. (2014) ***Probabilité et Statistique pour les Sciences de la Santé***. Stata Press.
- Ten Have TR and Joffe MM. **A review of causal estimation of effects in mediation analyses**. *Statistical methods in medical Research* 2010, 21: 77-107.

Merci pour votre attention ☺