Une introduction à l'utilisation du score de propension
pour estimer l'effet causal d'une exposition à l'aide de données observationnelles

Les méthodes du score de propension sont des méthodes populaires pour estimer causal l'effet d'une exposition sur une issue à l'aide de données observationnelles (des données non randomisées). Les méthodes du score de propension sont particulièrement utiles dans le cas où on s'intéresse à une issue binaire relativement rare. Dans ce cas, il pourrait être difficile, voire impossible, d'ajuster de façon traditionnelle pour plusieurs variables d'ajustement à l'intérieur d'un modèle de l'issue en fonction de l'exposition. Dans cette chronique, j'effectue une brève introduction à aux méthodes associées au score de propension, accompagnée d'exemples de lignes de code en R.

Note : Avant d'exécuter les commandes « `require` », assurez-vous d'avoir préalablement installé le *package* nécessaire sur votre ordinateur à l'aide du menu « Packages », « Installer le(s) package(s)... » dans R.

Motivation

Au milieu des années 1970, un programme visant à donner une expérience de travail à des individus faisant face à des difficultés économiques et sociales a été mis en place aux États-Unis. L'efficacité de ce programme a été évaluée à l'aide d'une étude randomisée. Le jeu de données *lalonde*, disponible avec le *package MatchIt* dans R, contient l'information de 185 sujets ayant participé au groupe « traitement » de cette étude randomisée et de 429 sujets « contrôles » provenant de la population générale (n'ayant pas participé à l'étude randomisée). Le jeu de données ainsi généré correspond à ce qu'on aurait pu observer si une étude observationnelle avait plutôt été réalisée pour évaluer l'effet du programme. L'objectif de créer un tel jeu de données est d'évaluer la capacité de répliquer les résultats d'une étude randomisée avec des données observationnelles (par exemple, Dehejia & Wahba, 1999).

À l'aide de ces données, il est possible d'estimer l'effet de l'exposition (participer ou non au programme) sur le revenu des individus en 1978, en ajustant pour différentes variables pré-exposition, dont l'âge (en années), le nombre d'années d'éducation, le revenu en 1974 (en $ US), la race/ethnicité (africain-américain, hispanique ou autre), le statut matrimonial (marié ou non), ainsi que le fait d'avoir un diplôme du secondaire (oui ou non).

Les lignes de code suivantes permettent d'obtenir l'information nécessaire à la construction d'un tableau mettant en évidence les différences préexposition en fonction du groupe d'exposition (Tableau 1).

```
require(MatchIt);
head(lalonde);

tapply(lalonde$age, lalonde$treat, mean);
tapply(lalonde$educ, lalonde$treat, mean);
```

```
tapply(lalonde$re74, lalonde$treat, mean);
tapply(lalonde$age, lalonde$treat, sd);
tapply(lalonde$educ, lalonde$treat, sd);
tapply(lalonde$re74, lalonde$treat, sd);

table(lalonde$black, lalonde$treat);
prop.table(table(lalonde$black, lalonde$treat), 2);
table(lalonde$hispan, lalonde$treat);
prop.table(table(lalonde$hispan, lalonde$treat), 2);
table(lalonde$married, lalonde$treat);
prop.table(table(lalonde$married, lalonde$treat), 2);
table(lalonde$nodegree, lalonde$treat);
prop.table(table(lalonde$nodegree, lalonde$treat), 2);
```

Tableau 1 : Caractéristiques préexposition en fonction de l'exposition*.

	Non exposés	Exposés
Âge	28.0 (10.8)	25.8 (7.2)
Années d'éducation	10.2 (2.9)	10.3 (2.0)
Revenu en 1974	5619 (6788)	2095 (4886)
Africain-américain	87 (20%)	156 (84%)
Hispanique	61 (14%)	11 (6%)
Marié	220 (51%)	35 (19%)
Sans diplôme	256 (60%)	131 (71%)

* Pour les variables continues, la moyenne et l'écart type (entre parenthèses) sont rapportés, alors que pour les variables catégorielles, le nombre et le pourcentage (entre parenthèses) sont rapportés.

Le score de propension

Le score de propension a été introduit par Rosembaum et Rubin (1983). Les méthodes utilisant le score de propension ont initialement été conçues pour estimer l'effet d'une exposition binaire (0/1) sur une issue de type quelconque, mais des généralisations ont été proposées depuis pour les expositions catégoriques ou continues.

Afin de simplifier la présentation, nous considérons le cas d'une variable d'exposition *X* binaire (0/1) et d'un ensemble de covariables préexposition (*baseline*) potentiellement confondantes ***U***. Ces variables sont souvent conceptualisées comme étant des déterminants à la fois de l'exposition et de la réponse, *Y*. On supposera par ailleurs que de contrôler pour les variables ***U*** permettrait d'estimer l'effet (causal) de l'exposition sur la réponse.

L'idée du score de propension est de résumer l'information contenue dans les covariables ***U*** à l'intérieur d'une seule variable *e(x)* de sorte que le fait de contrôler pour *e(x)* permettrait également d'estimer l'effet d'intérêt. En fait, le score de propension est tel que la distribution des covariables préexposition est la même chez les sujets exposés que chez les sujets non exposés pour une valeur de *e(x)* donnée. Mathématiquement, $X \perp\!\!\!\perp \boldsymbol{U} \mid e(x)$, où $\perp\!\!\!\perp$ désigne l'indépendance statistique. Autrement dit, pour une valeur *e(x)* donnée, les sujets exposés et non exposés auront en moyenne les mêmes

caractéristiques initiales. Ainsi, le score de propension permet, d'une certaine manière, de simuler un contexte d'étude randomisée, puisque, en moyenne, la randomisation assure que la distribution des variables préexposition est la même chez les sujets exposés que chez ceux non exposés.

Le score de propension correspond simplement à la probabilité qu'un sujet soit exposé conditionnellement à ses variables préexposition, $e(x) = P(X = 1| \boldsymbol{U})$. En pratique, il est commun d'estimer le score de propension à l'aide d'un modèle de régression logistique, bien que des modèles plus sophistiqués puissent être utilisés. Le code R suivant permet d'estimer le score de propension à l'aide des données observées (on utilise la racine carrée du revenu en 1974; cette transformation permet de réduire l'influence des revenus extrêmes).

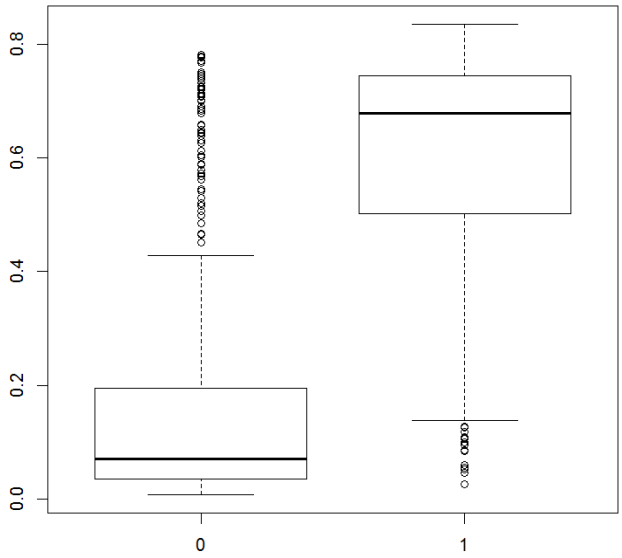
```
lalonde$sqrtre74 = sqrt(lalonde$re74);
ps = glm(treat ~ age + educ + sqrtre74 + black + hispan
      + married + nodegree,  data = lalonde,
       family = binomial(link = "logit"));
summary(ps);
```

À cette étape, il sera possible d'effectuer différents diagnostics usuels du modèle, notamment de déceler la présence de données influentes ou aberrantes. Il est par ailleurs pertinent de vérifier qu'il existe un chevauchement dans la distribution du score de propension des deux groupes d'exposition. S'il y absence de chevauchement sur une partie des données, l'inférence sera faite à partir d'extrapolations et donc les résultats pourraient ne pas être valides. En effet, le score de propension vise à assurer un équilibre dans les caractéristiques de préexposition *pour un score donné*. Si certaines plages de score ne sont présentes que pour l'un des deux groupes d'exposition, l'approche perd son sens sur ces plages.

Le code suivant permet de tracer un diagramme en boîte du score de propension estimé en fonction du statut d'exposition. On peut constater à la Figure 1 que les diagrammes se recoupent pratiquement sur l'ensemble des valeurs du score de propension. Cependant, là où les valeurs sont concentrées pour un groupe, elles sont très peu présentes pour l'autre. Il s'agit donc d'une situation où il est particulièrement difficile d'estimer correctement l'effet de l'exposition. Il serait possible d'éliminer les zones de non-recoupement pour éviter d'effectuer de l'extrapolation. Cependant, la population cible sur laquelle porteraient les inférences serait modifiée.

```
boxplot(ps$fitted~lalonde$treat);
```

Figure 1 : Score de propension estimé selon le niveau d'exposition.



Estimation de l'effet du traitement

Il existe trois méthodes principales pour estimer l'effet du traitement à l'aide du score de propension, soit l'appariement, la pondération et l'ajustement. Ces trois méthodes sont maintenant brièvement présentées.

Appariement

L'approche d'appariement est particulièrement populaire, puisque c'est celle qui se rapproche le plus d'une analyse de données provenant d'une étude randomisée. Selon cette approche, chaque sujet exposé est apparié à un ou plusieurs sujets non exposés en fonction du score de propension. Le package *MatchIt* en R est justement conçu pour effectuer l'appariement. Le code suivant permet de générer un jeu de données apparié en fonction d'un appariement 1-1 sans remise.

```
matchs = matchit(treat ~ ps$fitted, method = "nearest", data =
lalonde, distance = "mahalanobis");
summary(matchs);
m.data = match.data(matchs);
```

Dans le jeu de données résultant, il ne devrait plus y avoir d'association entre l'exposition et les variables de préexposition, tel qu'attendu d'une étude randomisée.

En construisant un tableau similaire au Tableau 1 sur le jeu de données appariées (Tableau 2), on peut constater que les caractéristiques préexposition sont beaucoup

mieux équilibrées, bien qu'il demeure un certain déséquilibre, particulièrement par rapport à la race/ethnicité. Une solution à ce problème pourrait être d'effectuer un appariement avec remise, où un même témoin (non exposé) peut être apparié à plus d'un exposé.

Tableau 2 : Caractéristiques préexposition en fonction de l'exposition sur l'échantillon apparié.*

	Non exposés	Exposés
Âge	25.0 (10.4)	25.8 (7.2)
Années d'éducation	10.6 (2.6)	10.3 (2.0)
Revenu en 1974	2298 (4300)	2096 (4887)
Africain-américain	87 (47%)	156 (84%)
Hispanique	39 (21%)	11 (6%)
Marié	37 (20%)	35 (19%)
Sans diplôme	117 (63%)	131 (71%)

* Pour les variables continues, la moyenne et l'écart type (entre parenthèses) sont rapportés, alors que pour les variables catégorielles, le nombre et le pourcentage (entre parenthèses) sont rapportés.

Si l'équilibre atteint est jugé satisfaisant, on peut estimer l'effet du traitement assez simplement, par exemple à l'aide d'un test de Student **non-pairé** (voir par exemple Stuart, 2008) sur le revenu en 1978. On obtient une différence de revenu de 673 $ (intervalle de confiance (IC) à 95% : -767 $ à 2114 $) en faveur des sujets exposés.

```
t.test(re78 ~ treat, data = m.data);
```

Un désavantage associé à l'appariement est la perte potentielle de certains sujets. Dans cet exemple, le groupe des sujets non exposés étant plus grand que le groupe des sujets exposés, il y a plusieurs sujets non exposés qui n'ont pas été appariés à un sujet exposés et qui ne sont donc pas considérés dans l'analyse.

Pondération

La deuxième approche se rapproche également d'une analyse de données provenant d'une étude randomisée. Elle consiste à utiliser une pondération telle que le statut d'exposition n'est plus associé aux variables préexposition dans l'échantillon pondéré. Pour ce faire, il suffit de pondérer chaque observation par l'inverse de la probabilité qu'elle reçoive le niveau d'exposition qu'elle a vraiment reçu. Autrement dit, les sujets exposés reçoivent un poids de $1/e(x)$, alors que les sujets non exposés reçoivent un poids de $1/(1 – e(x))$. Les poids peuvent donc se calculer comme étant $X/e(x) + (1 – X)/(1 – e(x))$.

```
w = lalonde$treat/ps$fitted + (1 - lalonde$treat)/(1 -
ps$fitted);
```

Toutefois, puisque cette approche peut créer des poids ayant des valeurs très élevées, il est souvent recommandé d'effectuer une troncation. Ceci évite qu'une ou plusieurs

observations pour lesquelles le poids est très important aient une influence indue sur les résultats. Par exemple, dans notre analyse, on peut constater qu'un des poids a une valeur supérieure à 30. Le code suivant effectue une troncation au 99^{e} percentile.

```
wt = pmin(w, quantile(w, 0.99));
```

À l'aide du code plus bas, on peut construire le Tableau 3 permettant de vérifier si la pondération a réussi à équilibrer les caractéristiques préexposition. On peut y constater que l'équilibre est meilleur que sur l'échantillon non pondéré, mais qu'il reste des déséquilibres importants, surtout par rapport au revenu de 1974. Dans une telle situation, il serait approprié de réviser le modèle utilisé pour le score de propension, par exemple en considérant d'y inclure des termes quadratiques ou des interactions.

```
require(SDMTools);
lalonde = cbind(lalonde, wt);
by(data = lalonde[, c("age", "wt")], lalonde$treat, FUN =
function(x){wt.mean(x$age, x$wt)});
by(data = lalonde[, c("age", "wt")], lalonde$treat, FUN =
function(x){wt.sd(x$age, x$wt)});
by(data = lalonde[, c("educ", "wt")], lalonde$treat, FUN =
function(x){wt.sd(x$educ, x$wt)});
by(data = lalonde[, c("educ", "wt")], lalonde$treat, FUN =
function(x){wt.sd(x$educ, x$wt)});
by(data = lalonde[, c("re74", "wt")], lalonde$treat, FUN =
function(x){wt.sd(x$re74, x$wt)});
by(data = lalonde[, c("re74", "wt")], lalonde$treat, FUN =
function(x){wt.sd(x$re74, x$wt)});

by(data = lalonde[, c("black", "wt")], lalonde$treat, FUN =
function(x){sum(x$black*x$wt)});
by(data = lalonde[, c("black", "wt")], lalonde$treat, FUN =
function(x){sum(x$black*x$wt)/sum(x$wt)});
by(data = lalonde[, c("hispan", "wt")], lalonde$treat, FUN =
function(x){sum(x$hispan*x$wt)});
by(data = lalonde[, c("hispan", "wt")], lalonde$treat, FUN =
function(x){sum(x$hispan*x$wt)/sum(x$wt)});
by(data = lalonde[, c("married", "wt")], lalonde$treat, FUN =
function(x){sum(x$married*x$wt)});
by(data = lalonde[, c("married", "wt")], lalonde$treat, FUN =
function(x){sum(x$married*x$wt)/sum(x$wt)});
by(data = lalonde[, c("nodegree", "wt")], lalonde$treat, FUN =
function(x){sum(x$nodegree*x$wt)});
by(data = lalonde[, c("nodegree", "wt")], lalonde$treat, FUN =
function(x){sum(x$nodegree*x$wt)/sum(x$wt)});
```

Tableau 3 : Caractéristiques préexposition en fonction de l'exposition sur l'échantillon pondéré.*

	Non exposés	Exposés
Âge	27.1 (10.9)	25.4 (6.7)
Années d'éducation	10.3 (2.7)	10.5 (2)
Revenu en 1974	4446 (6346)	2895 (5629)
Africain-américain	246 (40%)	248 (51%)
Hispanique	72 (12%)	68 (14%)
Marié	252 (41%)	121 (25%)
Sans diplôme	387 (63%)	300 (62%)

* Pour les variables continues, la moyenne et l'écart type (entre parenthèses) sont rapportés, alors que pour les variables catégorielles, le nombre et le pourcentage (entre parenthèses) sont rapportés.

Une approche simple permettant d'obtenir des tests statistiques tenant compte de la pondération (en fait, il s'agit d'une approche donnant des ICs conservateurs) est d'utiliser un estimateur robuste de l'erreur type, par exemple celui habituellement produit par les GEEs. À l'aide du code suivant, on obtient une estimation de l'effet de l'exposition de 384 $ (IC à 95% : -1073 $ à 1839 $).

```
require(geepacks);
mod.pond = geeglm(re78~treat, data = lalonde, weight = wt, id =
1:nrow(lalonde), family = gaussian(link = "identity"));
summary(mod.pond);
```

Ajustement

La dernière approche est la plus simple d'utilisation, il suffit d'ajuster un modèle pour la variable réponse en fonction du traitement en ajustant pour le score de propension. Cependant, cette approche ne permet pas un diagnostic naturel du niveau d'équilibre des caractéristiques préexposition. Plusieurs approches sont possibles pour l'ajustement, par exemple entrer le score de propension de façon linéaire dans le modèle, le diviser en catégories (par exemple, en quintiles) ou utiliser une modélisation flexible (par exemple, un *spline* cubique). Le code suivant effectue un ajustement à l'aide d'un *spline* cubique.

```
require(rms);
mod.ajust = lm(re78~treat + rcs(ps$fitted), data = lalonde);
summary(mod.ajust);
```

Avec cette approche, l'effet de l'exposition est estimé à 1356 $ (IC à 95% : -237 $ à 2950 $).

Conclusion

J'espère que cette brève introduction au score de propension et aux méthodes associées vous aura permis d'en apprendre un peu et vous incitera peut-être à vouloir en apprendre davantage sur le sujet. Tel qu'illustré, l'implantation de ces méthodes dans les logiciels n'est pas extrêmement difficile, puisqu'elle ne requiert généralement que des outils assez classiques (à l'exception de la méthode d'appariement). Par ailleurs, la méthode d'appariement et la méthode de pondération offrent des outils permettant de vérifier si l'équilibre dans les caractéristiques de préexposition est atteint. Toutefois, rien ne peut assurer que les groupes exposés et non exposés sont équilibrés par rapport à des caractéristiques qui n'ont pas été mesurées dans l'étude.

À titre d'information, l'effet de l'exposition estimé avec les données de l'étude randomisée est de 1794 $ (IC à 95% : 528 $ à 3060 $), l'approche par ajustement ayant donc donné les résultats les plus près de ceux de l'étude randomisée.

Remerciements

Je souhaite adresser des remerciements spéciaux à Myrto Mondor et à Steve Méthot pour leur révision minutieuse de mon texte. Leurs commentaires et corrections ont permis de bonifier substantiellement mon article. Je prends toutefois l'entière responsabilité de toutes fautes, imprécisions ou erreurs qui pourraient subsister dans le texte révisé.

Denis Talbot

Références :

Dehejia, R. H. & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448): 1053-1062.

Rosembaum, P. & Rubin, D. (1983). The central role of propensity score in observational studies for causal effects. *Biometrika*, 70:41-55

Stuart, E. A. (2008). Developing practical recommendations for the use of propensity scores: Discussion of 'A critical appraisal of propensity score matching in the medical literature between 1996 and 2003' by Peter Austin, Statistics in Medicine. *Statistics in medicine*, 27(12), 2062-2065.