Machine Learning pour l'économétrie : éléments de correction des exercices

Christophe Gaillac
Université d'Oxford & Nuffield College
CREST

Jérémy L'Hour
Capital Fund Management
CREST

15.1 La régression en tant qu'estimateur par pondération

1. Un premier argument est que $DY = DY_1$ presque sûrement. Un second argument est : $\mathbb{E}[DY_0] = [DX'\beta_0 + D\varepsilon] = \mathbb{E}[DX]'\beta_0$.
2. La valeur théorique est $\beta_0 = \mathbb{E}[(1-D)XX']^{-1}\mathbb{E}[(1-D)XY]$. Sa contrepartie empirique est donnée par :

$$\widehat{\beta} = \left[\frac{1}{n}\sum_{i=1}^{n}(1-D_i)X_iX_i'\right]^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}(1-D_i)X_iY_i\right]$$

   $\widehat{\beta}$ est également obtenu à partir d'une régression de Y sur X pour l'échantillon des non-traités.
3. A partir de la première question on peut montrer que :

$$\begin{aligned}\pi\mathbb{E}[Y_1 - Y_0|D=1] &= \mathbb{E}[D(Y_1-Y_0)]\\ &= \mathbb{E}[DY] - \mathbb{E}[DX]'\beta_0\\ &= \mathbb{E}[DY] - \mathbb{E}[DX]'\mathbb{E}[(1-D)XX']^{-1}\mathbb{E}[(1-D)XY]\\ &= \mathbb{E}[DY] - \mathbb{E}\left[(1-D)\mathbb{E}[DX]'\mathbb{E}[(1-D)XX']^{-1}XY\right]\\ &= \mathbb{E}[DY - (1-D)W_0Y]\end{aligned}$$

   si l'on note $W_0 := \mathbb{E}[DX]'\mathbb{E}[(1-D)XX']^{-1}X$.
4. Puisque W_0 est de dimension un, on peut écrire :

$$\begin{aligned}\mathbb{E}[(1-D)XW_0] &= \mathbb{E}[(1-D)XW_0']\\ &= \mathbb{E}\left[(1-D)XX'\mathbb{E}[(1-D)XX']^{-1}\mathbb{E}[DX]\right]\\ &= \mathbb{E}[(1-D)XX']\mathbb{E}[(1-D)XX']^{-1}\mathbb{E}[DX]\\ &= \mathbb{E}[DX]\end{aligned}$$

   Les caractéristiques du groupe témoin repondéré ont le même premier moment que celles du groupe traité.

5. Les poids sont : $\omega_j = n_1^{-1} \left[\sum_{i=1}^n D_i X_i\right]' \left[\sum_{i=1}^n (1-D_i) X_i X_i'\right]^{-1} X_j$. Nous utilisons la même astuce que précédemment :

$$\begin{aligned}
\sum_{j:D_j=0} X_j \omega_j &= \sum_{j=1}^n (1-D_j) X_j \omega_j' \\
&= n_1^{-1} \sum_{j=1}^n (1-D_j) X_j X_j' \left[\sum_{i=1}^n (1-D_i) X_i X_i'\right]^{-1} \left[\sum_{i=1}^n D_i X_i\right] \\
&= n_1^{-1} \sum_{i=1}^n D_i X_i \\
&= n_1^{-1} \sum_{i:D_i=1} X_i
\end{aligned}$$

   Disons que le premier élément de X est la constante, alors $\sum_{j:D_j=0} \omega_j = n_1^{-1} \sum_{i:D_i=1} 1 = 1$. Par conséquent, la somme des poids est égale à un.
6. Dans l'estimateur du contrôle synthétique, les poids somment à un et sont positifs, ce qui n'est pas forcément le cas ici : si les poids somment bien à un, ils peuvent être négatifs. En outre, par construction cet estimateur reproduit les caractéristiques du groupe traité, or cela n'est pas forcément le cas avec le contrôle synthétique. Enfin, mais cela n'est pas directement visible ici, le contrôle synthétique donne une solution sparse, ce qui n'est pas le cas ici.

15.2 Score orthogonal pour l'effet du traitement sur les traités

1. (a) D'abord, $\mathbb{E}[DY^{obs}] = \mathbb{E}[DY_1]$. Ensuite :

$$\begin{aligned}
\mathbb{E}[\exp(X'\beta_0)(1-D)Y^{obs}] &= \mathbb{E}\left[\frac{p(X)}{1-p(X)}(1-D)Y_0\right] \\
&= \mathbb{E}\left[\frac{p(X)}{1-p(X)}\mathbb{E}[(1-D)Y_0|X]\right] \\
&= \mathbb{E}\left[\frac{p(X)}{1-p(X)}\mathbb{E}[(1-D)|X]\mathbb{E}[Y_0|X]\right] \\
&= \mathbb{E}\left[p(X)\mathbb{E}[Y_0|X]\right] \\
&= \mathbb{E}\left[DY_0\right].
\end{aligned}$$

   Ainsi, $\mathbb{E}[DY^{obs}] - \mathbb{E}[\exp(X'\beta_0)(1-D)Y^{obs}] = \mathbb{E}[D(Y_1 - Y_0)]$. Finalement on note que $\mathbb{E}[D(Y_1 - Y_0)] = \tau_0 P[D=1] = \mathbb{E}[D\tau_0]$.

   (b) $p(X)$ peut être estimé par un logit, ce qui donne un estimateur de β_0. Ensuite $\hat{\tau}$ peut être estimé par :

$$\hat{\tau} = \frac{\frac{1}{n}\sum_{i=1}^n (D_i - (1-D_i)\exp(X'\hat{\beta}))Y_i^{obs}}{\frac{1}{n}\sum_{i=1}^n D_i}$$

2. (a) On peut montrer que :

$$\mathbb{E}[\partial_\beta m(W, \tau_0, \beta_0)] = -\mathbb{E}[X D Y_0] \neq 0.$$

   (b) La méthode la plus efficace est le maximum de vraisemblance, donnée par l'équation (LOGIT) à la section (2.3.2) du livre, où l'on remplace Y_i par D_i. L'estimateur $\hat{\tau}$ de τ_0 trouvé à la question 1.a où l'on utilise l'estimateur du maximum de vraisemblance $\widehat{\beta}$ sera bien asymptotiquement gaussien. Cela se montre simplement.

   (c) En grande dimension, il sera nécessaire d'utiliser des méthodes de machine learning telles que l'esimateur de β_0 ne sera pas asymptotiquement gaussien (par exemple, un Logit Lasso comme indiqué dans l'énoncé). Cela va créer un biais pour $\hat{\tau}$ puisqu'en général

$$\mathbb{E}[\partial_\beta m(W, \tau_0, \beta_0)] = -\mathbb{E}[X D Y_0] \neq 0.$$

   Cela est montré au chapitre 4.

3. (a) La CIA implique $Y_0 - X'\gamma_0 \perp D|X$, donc

$$\mathbb{E}[DX(Y_0 - X'\gamma_0)] = \mathbb{E}[X\mathbb{E}[\varepsilon|X]\mathbb{E}[D|X]] = 0.$$

   (b) Inspirés par le chapitre 5, on propose :

$$\psi(W, \tau, \beta, \gamma) = (D_i - \exp(X'\beta)(1 - D_i))\,(Y_i^{obs} - X_i'\gamma) - D_i\tau.$$

$$\mathbb{E}[\partial_\beta \psi(W, \tau_0, \beta_0, \gamma_0)] = -\mathbb{E}[X \exp(X'\beta_0)(1 - D)(Y^{obs} - X_i'\gamma)] = 0.$$

$$\mathbb{E}[\partial_\gamma \psi(W, \tau_0, \beta_0, \gamma_0)] = -\mathbb{E}[(D - (1 - D)\exp(X'\beta_0))X_i] = 0.$$

4. On peut se baser sur les estimateurs proposés à la section 5.3 et utiliser le théorème 5.1.

15.3 Modèle de vote

1. Une régression au moyen de forêts aléatoires honnêtes (car D est continu), avec random-split, est particulièrement adaptée aux formes des fonctions de base de $\mathcal{F}_{p,q}$ (hypercubes). Cet estimateur ne peut être utilisé que dans le cas de faible dimension, car le taux de convergence $n^{-1/(1+p\alpha_2\delta)}$ ne permet pas d'obtenir $p + q >> \log(n)$.

2. Dans ce cas, on peut utiliser un Lasso avec les régresseurs transformés :

$$\widetilde{X}_{t,i} = 1\{X_t \in C_{a_i,\epsilon}\}$$
$$\widetilde{Z}_{t,i} = 1\{Z_t \in C_{b_i,\epsilon}\}$$

   Ainsi le modèle linéaire

$$D_t = \gamma_0^\top \widetilde{X}_t + \delta_0^\top \widetilde{Z}_t + u_t,$$

$\widetilde{X}_t \in R^p$, $\widetilde{Z}_t \in R^q$, est particulièrement bien adapté. Il peut être estimé via :

$$(\widehat{\gamma}_0, \widehat{\delta}_0) \in \underset{(\gamma_0,\delta_0)\in R^{p+q}}{\operatorname{argmin}} \quad \frac{1}{n}\sum_{t=1}^{n}\left(D_t - \widetilde{X}_t'\gamma_0 - \widetilde{Z}_t'\delta_0\right)^2 + \frac{\lambda}{n}\left\|\widehat{Y}(\gamma_0,\delta_0)'\right\|_1,$$

où

$$\left\|\widehat{Y}(\gamma_0,\delta_0)'\right\|_1 = \sum_{j=1}^{p}\left|\widehat{Y}_j\gamma_{0,j}\right| + \sum_{j=1}^{q}\left|\widehat{Y}_{j+p}\delta_{0,j}\right|.$$

Les $\widehat{Y} \in \mathcal{M}_{p+q,p+q}(R)$ sont des paramètres de pénalisation.

3. De façon standard, en utilisant une régression logistique :

$$\tilde{S}_t := \ln\left(\frac{S_t}{1-S_t}\right) = g(X_t'\beta_0) + \tau_0 D_t + \xi_{L,t}.$$

4. On a joute l'équation linéaire suivante :

$$\widetilde{Z}_t = \Pi\widetilde{X}_t + \zeta_t,\ \zeta_t \perp X_t, \Pi \in \mathcal{M}_{p+q,p+q}(R).$$

Ainsi, on se retrouve avec :

$$\begin{aligned} D_t =& \widetilde{X}_t'\gamma_0 + \widetilde{X}_t'\Pi'\delta_0 + u_t + \zeta_t'\delta_0 \\ =& \widetilde{X}_t'(\gamma_0 + \Pi'\delta_0) + \rho_t^d, \quad \text{et donc} \\ D_t =& \widetilde{X}_t'\nu_0 + \rho_t^d, \quad \rho_t^d \perp X_t \end{aligned} \tag{15.1}$$

On note le paramètre de nuisance par $\eta = (\beta_0, \nu_0, \delta_0, \gamma_0)$. Soit

$$\begin{aligned} &m(W_t,\eta,\tau_0) = \\ &\left(\tilde{S}_t - \widetilde{X}_t'(\tau_0\nu_0) - g(\widetilde{X}_t'\beta_0) - \tau_0\left(D_t - \widetilde{X}_t'\nu_0\right)\right)\left(\widetilde{X}_t'\gamma_0 + \widetilde{Z}_t'\delta_0 - \widetilde{X}_t'\nu_0\right), \end{aligned}$$

alors les deux équations sont satisfaites, et en particulier :

$$\begin{aligned} &\mathbb{E}\left[\partial_{\beta_0} m(W_t,\eta,\tau_0)\right] = -\mathbb{E}\left[\zeta_t'\delta_0\widetilde{X}_t g(\widetilde{X}_t'\beta_0)\right] = 0 \ \ (\text{car } \zeta_t \perp X_t) \\ &\mathbb{E}\left[\partial_{\nu_0} m(W_t,\eta,\tau_0)\right] = -\mathbb{E}\left[\widetilde{X}_t\xi_t\right] = 0 \end{aligned}$$

5. Dans le chapitre correspondant, la normalité asymptotique de $\widehat{\tau}$ est prouvée pour le modèle affine-quadratique, qui impose que :
   (a) soit g est une fonction affine. Dans ce cas, on retombe dans le cas vu dans le chapitre correspondant.
   (b) soit g est une fonction quadratique, ce qui est permis par le théorème, mais requiert que l'on emploi un estimateur de β_0 pour un modèle à index non-liénaire, dans la première étape.

15.4 Hétérogénéité de l'écart salarial homme-femme

1. $\mathbb{E}[\ln W_i|X_i, F_i = 1] - \mathbb{E}[\ln W_i|X_i, F_i = 0]$ est l'écart de salaire moyen entre les hommes et les femmes pour la population ayant les caractéristiques X_i.
2. (a) X_i peut contenir beaucoup de variables : heures travaillées, expérience, expérience au carré, âge, type d'éducation, nombre d'années d'éducation, localisation géographique, nationalité, statut marital, nombre de jeunes enfants, nombre d'enfants en général, industrie, traits psychologiques tels que la conscience et l'ouverture, etc.
   (b) Dans ce cas, un simple estimateur MCO fonctionne grâce à l'hypothèse d'exogénéité.
   (c) Non, ce n'est pas le cas. Compte tenu de la structure sparse, on doit utiliser la procédure de double sélection vue en cours, en utilisant un Lasso dans les deux premières étapes – une brève description de la procédure est nécessaire ici.
   (d) $\mathbb{E}[\ln W_i|X_i, F_i = f] = \theta f + X_i'\beta$. Cela signifie que l'écart salarial est constant sur tout le support de X_i, ce qui est probablement déraisonnable.
3. (a) C'est vrai. En effet, dans ce cas $\mathbb{E}[\ln W_i|X_i, F_i = 1] - \mathbb{E}[\ln W_i|X_i, F_i = 0] = \theta(Z_i) = \sum_{k=1}^{K} \theta_k Z_{i,k}$. L'écart salarial varie donc de θ_k points de pourcentage lorsque $Z_{i,k}$ varie d'une unité. Étant donné que l'écart salarial global est négatif, une valeur positive de θ_k signifie que (par exemple) l'écart salarial est inférieur à la ligne de base dans la population pour laquelle $Z_k = 1$.
   (b) En utilisant la notation $\theta = (\theta_k)_{k=1,\dots,K}$, on a :

$$\ln W_i = \alpha + F_i Z_i'\theta + X_i'\beta + \varepsilon_i,$$

   On a donc un modèle linéaire avec $p + K$ covariables $(F_i Z_i', X_i')'$ et les $p + K$ équations normales sont les suivantes

$$\mathbb{E}\left[(\ln W_i - \alpha - F_i Z_i'\theta - X_i'\beta + \varepsilon_i)(F_i Z_i', X_i')'\right] = 0.$$

   Si l'on ne se pré-occupe que de θ, alors on ne conserve que

$$\mathbb{E}\left[(\ln W_i - \alpha - F_i Z_i'\theta - X_i'\beta + \varepsilon_i)F_i Z_i\right] = 0,$$

   en considérant β comme un paramètre de nuisance.
   (c) Une condition immunisée ψ pour $(\theta_1, \dots, \theta_K)$ prendra la forme

$$\mathbb{E}\left[(\ln W_i - \alpha - F_i Z_i'\theta - X_i'\beta + \varepsilon_i)(F_i Z_i - X_i'\gamma\right] = 0,$$

   avec les dérivées par rapport à β égales à zéro. Cela revient à utiliser la procédure de "double-sélection" mais avec K paramètres d'intérêts. Cela nécessitera $K + 2$ étapes :

   i. Les K premières étapes consistent à régresser chaque élément de Z_i sur X_i pour le sous-échantillon de femmes en utilisant un Lasso,
   ii. La $K+1$ième étape est une régression Lasso de $\ln W_i$ sur X_i,
   iii. La dernière étape est une régression de $\ln W_i$ sur $F_i Z_i$ et l'union de tous les éléments de X_i sélectionnés précédemment.

4. Exemple : Par rapport à la base de référence, avoir un enfant de 18 ans ou plus augmente l'écart salarial de 5 pp (l'écart salarial est plus négatif pour elles). Ainsi, les femmes qui ont un enfant de 18 ans ou moins gagnent 5 pp de moins que les hommes par rapport aux autres femmes.
5. Il y a un problème de test multiple. (Ces tableaux corrigent déjà les tests multiples, mais vous ne pouviez pas le savoir).
6. En utilisant que $\mathbb{E}\left[\hat{\mu}(x; X_1, \ldots, X_n)\right] = \mathbb{E}\left[T(x; X_1, \ldots, X_n)\right]$. Donc, nous avons :

$$
\begin{aligned}
&\left|\mathbb{E}\left[\hat{\mu}(x; X_1, \ldots, X_n)\right] - \mu(x)\right| \\
&= \left|\mathbb{E}\left[T(x; X_1, \ldots, X_n)\right] - \mu(x)\right| \\
&= \left| \sum_{i \in \{i_1, \ldots, i_s\}} \mathbb{E}\left[\frac{1\{X_i \in L(x)\}}{s|L(x)|} \ln W_i\right] - \mu(x) \right| \\
&= \left| \mathbb{E}\left[\ln W_i | X_i \in L(x)\right] \frac{1}{s} \sum_{i \in \{i_1, \ldots, i_s\}} \mathbb{E}\left[\frac{1\{X_i \in L(x)\}}{|L(x)|} | X_i \in L(x)\right] - \mu(x) \right| \\
&= \left|\mathbb{E}\left[\ln W_i | X_i \in L(x)\right] - \mathbb{E}\left[\ln W_i | X_i = x\right]\right| \leq C\mathrm{Diam}(L(x)).
\end{aligned}
$$

7. On choisit ces deux méthodes en fonction des indicateurs de performance Λ et $\overline{\Lambda}$, qui mesurent le degré d'hétérogénéité capturé par la procédure. Le tableau indique clairement que Random Forest et Elastic Net sont les meilleures. Le tableau 2 montre que la moyenne de $\mathbb{E}[\ln W_i | X_i, F_i = 1] - \mathbb{E}[\ln W_i | X_i, F_i = 0]$ (*i.e.* β_1) est négative, ce qui signifie que les femmes gagnent en moyenne moins que les hommes, mais que la pente de la BLP est significativement positive et proche de 1, donc il y a de l'hétérogénéité et son profil est assez bien décrit par les proxies du réseau Elatic et des forêts aléatoires.
8. D'après la figure 15.1 et le tableau 3, on voit qu'il existe un groupe de femmes pour lequel il n'y a pas d'écart salarial. Il s'agit des femmes ayant moins d'enfants et d'expérience que la moyenne. Ici, le paramètre d'intérêt dépend de la précision du proxy ML, ainsi que des splits. Nous ne pouvons donc apprendre que sur les caractéristiques de $\mathbb{E}[\ln W_i | X_i = \cdot, F_i = 1] - \mathbb{E}[\ln W_i | X_i = \cdot, F_i = 0]$ (hétérogénéité, sous-groupes qui bénéficient le moins et le plus et leurs caractéristiques) et non sur cette quantité

elle-même et précédemment faite. Comme cette procédure Generic ML dépend du fractionnement de l'échantillon, la p-value doit être adaptée pour prendre en compte ce caractère aléatoire supplémentaire.

9. (Bonus)

15.5 Sécheresse et incitations à économiser l'eau

1. (a) $\beta_1 = \mathbb{E}[\tau(X)] = \mathbb{E}[Y_1 - Y_0]$, l'effet moyen du traitement.
   (b) β_2 est le meilleur coefficient prédicteur linéaire (voir le cours pour la formule), le test H_0 : « $\beta_2 = 0$ » offre un test d'hétérogénéité. Lorsque cette hypothèse est rejetée, nous savons qu'il y a à la fois del'hétérogénéité et que le proxy ML permet de la capter en partie. Lorsque cette hypothèse n'est PAS rejetée, la conclusion n'est pas claire : cela peut être soit parce qu'il n'y a pas d'hétérogénéité, soit parce que le prédicteur proxy est faible (non corrélé avec le CATE).
2. (a) Elle peut être exprimée comme une fonction du carré de la corrélation entre le proxy ML et le véritable CATE multiplié par la variance du CATE. En maximisant cette quantité, on s'assure de sélectionner l'algorithme le plus corrélé au véritable CATE.
   (b) Gradient Boosting Machine, car son Λ correspondant est le plus grand.
3. (a) Oui, le test de H_0 : « $\beta_1 = 0$ » est rejeté à n'importe quel niveau de confiance. En moyenne, les ménages qui ont reçu une incitation à économiser l'eau, consomment environ 952 gallons (3604 litres) d'eau de moins que les ménages non traités, toute chose égale par ailleurs.
   (b) L'hypothèse que $\beta_2 = 0$ est rejetée pour l'algo 1. Pour l'algo 2, elle n'est pas rejetée (pour un test de 5% de niveau de confiance par exemple), ce qui signifie soit qu'il n'y a pas d'hétérogénéité, soit que l'algo 2 est trop faible. Puisque l'algo 2 est dominé par l'Algo 1 en terme de Λ, il faut donner plus de poids à l'algo 1. Notez que, comme expliqué dans le cours, pour un test de niveau α, la p-value affichée ici doit être inférieure à $\alpha/2$ pour être rejetée, afin de prendre en compte le fractionnement aléatoire des données.
4. (a) Voir la régression pour le GATES :

$$w(X)(D - p(X))Y = \sum_{k=1}^{5} \gamma_k G_k + \varepsilon,$$

   où l'effet de traitement le plus affecté est γ_1 et l'effet de traitement pour le moins affecté est estimé par γ_5. En effet, un calcul rapide montre que :

$$\gamma_k = \mathbb{E}[Y_1 - Y_0 | G_k].$$

   Cette régression peut être estimée par les MCO. Il s'agit simplement d'une moyenne de $w(X)(D - p(X))Y$ sur chaque groupe.

(b) Cette question ne peut pas recevoir de réponse rigoureuse car nous testons par groupe, donc même si l'effet moyen du traitement est significativement différent de zéro dans le groupe le moins affecté, l'effet du traitement (tel que mesuré par le CATE) peut être nul pour certains individus de ce groupe. De plus, les groupes sont basés sur les valeurs des prédicteurs proxy qui ne sont pas parfaitement corrélés avec le véritable CATE. Ainsi, même dans le groupe le plus affecté (tel que défini par les valeurs du prédicteur proxy), le véritable CATE peut être nul pour certains individus. Cependant, nous avons accepté la conclusion selon laquelle, avec l'algorithme 1, même les personnes les moins affectées ont un effet de traitement non nul, ce qui implique que l'effet de traitement est significatif pour la plupart des personnes.

(c) Pas vraiment, le test de la différence n'est rejeté à aucun niveau communément accepté (5%, 10%).

5. (a) Voir la régression pour le CLAN dans Chernozhukov et al. (2017) :

$$X = \sum_{k=1}^{5} \theta_k G_k + \varepsilon,$$

où la caractéristique moyenne des plus affectés est θ_1 et la caractéristique moyenne des moinsaffectés est estimée par θ_5. Elle peut être estimée par MCO.

(b) Tout d'abord,on remarque que ces coefficients estiment $P[VOTE = 1|G_1 = 1]$ et $P[VOTE = 1|G_5 = 1]$ mais puisque $P[G_1] = P[G_5] = .2$, le théorème de Bayes implique que :

$$\frac{P[VOTE = 1|G_1 = 1]}{P[VOTE = 1|G_5 = 1]} = \frac{P[G_1 = 1|VOTE = 1]}{P[G_5 = 1|VOTE = 1]},$$

donc nous pouvons avoir l'interprétation que les ménages où le vote est plus fréquent sont plus susceptibles d'être parmi les plus touchés par la campagne pro-sociale par rapport à être parmi les moins touchés. La différence est significative.

(c) La même remarque s'applique et il s'ensuit que les démocrates sont plus susceptibles d'être parmi les plus touchés (la différence est significative) par rapport aux moins touchés. Cette différence n'est pas significative pour les ménages républicains.

Notez que dans les trois cas, le test compare le plus et le moins affecté, et non chaque groupe à la population générale. Il se pourrait donc que les démocrates soient relativement plus susceptibles de faire partie des plus touchés que des moins touchés, mais qu'ils soient sous-représentés dans ces deux groupes par rapport à la population générale. [Ce n'est pas le cas dans cette application, mais on ne peut pas le déduire à partir des tableaux seulement].

6. Pour $j = 0, 1$:

$$\widehat{\alpha}_j \in \arg\max_{\alpha} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{D_i = j\} (Y_i - X_i'\alpha)^2 + \lambda \|\alpha\|_1 ,$$

Le CATE peut alors être estimé par :

$$\widehat{\tau}(X) = X'(\widehat{\alpha}_1 - \widehat{\alpha}_0).$$

Le problème est que cet estimateur est composé de deux estimateurs qui sont calculés séparément et peuvent ne pas être optimaux pour estimer le CATE.

7. La solution proposée pour le CRF est de diviser selon un critère conjoint (pas un pour $D = 1$ et un autre pour $D = 0$) spécialement conçu pour cibler l'effet du traitement, et non le conditionnement du traitement sur D.

8. Ici, $\gamma = \beta - E(D)\delta$, $\tau(X) = X'\delta$, $\widehat{\tau}(X) = X'\widehat{\delta}$, et nous utilisons

$$\mathbb{E}\left[(Y - X'\beta - (D - \mathbb{E}[D])X'\delta) \begin{pmatrix} X \\ X(D - \mathbb{E}[D]) \end{pmatrix}\right] = 0.$$

9. β est un paramètre de nuisance. L'estimateur est immunisé car, d'après l'équation relative à δ_j,

$$\begin{aligned} \partial_\beta \mathbb{E}\left[(Y - X'\beta - (D - \mathbb{E}[D])X'\delta)X_j(D - \mathbb{E}[D])\right] &= -\mathbb{E}\left[XX_j(D - \mathbb{E}[D])\right] \\ &= -\mathbb{E}\left[XX_j\right]\mathbb{E}\left[(D - \mathbb{E}[D])\right] \\ &= 0. \end{aligned}$$

10. Ici, les deux coefficients sont estimés simultanément et sont susceptibles de donner une meilleure estimation de la CATE.

11. Cet estimateur est pertinent s'il y a de la rareté à la fois dans β_0 et δ_0, *c.-à-d.* que seules quelques composantes de X sont pertinentes pour prédire le résultat de base et l'hétérogénéité de l'effet du traitement. De plus, le CATE et la fonction de régression pour le résultat de base sont tous deux linéaires. Le CRF est plus adapté dans un contexte où la fonction de régression pour le résultat de base est constante par morceaux.

12. Nous remplaçons $\mathbb{E}[D]$ par $Z'\gamma$, et ajoutons une pénalité ainsi qu'une hypothèse de sparsité potentielle pour gérer la haute dimensionnalité potentielle de γ, qui est un paramètre de nuisance.

$$\begin{aligned} \left(\widehat{\beta}, \widehat{\gamma}, \widehat{\delta}\right) = \operatorname*{argmin}_{\beta,\delta} \frac{1}{n} \sum_{i=1}^{n} &(Y_i - X_i'\beta - (D_i - Z'\gamma)X_i'\delta)^2 + \\ &\lambda_\beta\|\beta\|_1 + \lambda_\gamma\|\gamma\|_1 + \lambda_\delta\|\delta\|_1. \end{aligned}$$

Elle est immunisée (au sens de la définition du cours) car d'après l'équation relative à δ_j,

$$\begin{aligned}\partial_\beta \mathbb{E}\left[(Y - X'\beta - (D - Z'\gamma)X'\delta)X_j(D - Z'\gamma)\right] &= -\mathbb{E}\left[XX_j(D - Z'\gamma)\right] \\ &= -\mathbb{E}\left[XX_j\right]\mathbb{E}\left[(D - Z'\gamma)\right] \\ &= 0.\end{aligned}$$

et

$$\partial_\gamma \mathbb{E}\left[(Y - X'\beta - (D - Z'\gamma)X'\delta)X_j(D - Z'\gamma)\right]$$

est la somme de deux termes : le premier

$$\mathbb{E}\left[Z(X'\delta)X_j(D - Z'\gamma)\right] = -\mathbb{E}\left[Z(X'\delta)X_j\right]\mathbb{E}\left[\zeta\right] = 0.$$

et le second

$$-\mathbb{E}\left[(Y - X'\beta - (D - Z'\gamma)X'\delta)X_jZ\right] = -\mathbb{E}\left[\epsilon X_jZ\right] = 0.$$

15.6 Contrôle synthétique et régularisation

1. Ce programme de minimisation trouve les paramètres qui permettent de reproduire au mieux le comportement de la série temporelle de l'unité 1 avant le traitement par une combinaison linéaire des unités non traitées, en espérant que $\hat{\mu} + \sum_{i=2}^{N+1} \hat{\omega}_i Y_{i,T+1}^{obs}$ sera un bon contrefactuel pour $Y_{1,T+1}(0)$.
2. (a) C'est un programme des moindres carrés.
   On utilisa la notation $\mathbf{Y}_t = \left(1, Y_{2,t}^{obs}, \ldots, Y_{N+1,t}^{obs}\right)$. Alors la solution $(\hat{\mu}, \hat{\boldsymbol{\omega}})$ est donnée par :

$$\left[\frac{1}{T}\sum_{t=1}^{'} \mathbf{Y}_t\mathbf{Y}_t'\right]^{-1}\left[\frac{1}{T}\sum_{t=1}^{'} \mathbf{Y}_t Y_{1,t}^{obs}\right],$$

   à condition que $\frac{1}{T}\sum_{t=1}^{'} \mathbf{Y}_t\mathbf{Y}_t'$ soit inversible. Cette condition ne peut pas être vérifiée par exemple si $N > T$, nous devons donc être dans un cas où nous avons des données de panel longues.

   (b) $\hat{\omega}_i$ est le poids donné à l'unité non traitée i lors de la reproduction de l'unité traitée. Il s'agit de la corrélation partielle de la série temporelle de cette unité particulière avec celle de l'unité traitée. Il peut être négatif s'ils sont corrélés négativement.

   (c) Plusieurs : elle ne peut pas toujours être calculée, elle permet l'extrapolation.
3. (a) $\hat{\omega}_1 = 1/N$.
   (b) $\hat{\mu} = (1/T)\sum_{t=1}^{'} Y_{1,t}^{obs} - \sum_{i=2}^{N+1} Y_{i,t}^{obs}/N$.

(c) En conséquence :

$$\hat{\theta} = \left[Y_{1,T+1}^{obs} - \frac{1}{T} \sum_{t=1}^{'} Y_{1,t}^{obs} \right] - \frac{1}{N} \left[\sum_{i=2}^{N+1} Y_{i,T+1}^{obs} - \frac{1}{T} \sum_{t=1}^{'} \sum_{i=2}^{N+1} Y_{i,t}^{obs} \right],$$

Il s'agit de l'estimateur par différence de différences.

4. Pour cette question, on ajoute à (OBJ) les trois contraintes : $\omega_i \geq 0$ pour $i = 2, \ldots, N+1$, $\sum_{i=2}^{N+1} \omega_i = 1$ et $\mu = 0$.
   (a) C'est l'estimateur du contrôle synthétique .
   (b) Non, il n'est pas unique en général.
5. (a) Cela aide à régulariser le problème.
   (b) $\alpha = 0$ correspond à l'estimateur Ridge :

$$\left[\frac{1}{T} \sum_{t=1}^{T} \mathbf{Y}_t \mathbf{Y}_t' + \lambda I_{N+1} \right]^{-1} \left[\frac{1}{T} \sum_{t=1}^{T} \mathbf{Y}_t Y_{1,t}^{obs} \right],$$

   mais le premier élément de I_{N+1} est zéro.
   Lorsque $\alpha = 1$ est la solution Lasso. Certains poids seront élevés et d'autres seront exactement nuls.
   (c) Une certaine forme de validation croisée utilisant la dimension temporelle ou la dimension transversale parmi les personnes non traitées.