Étude de cas 2 – Consignes de préparation

Estimation ciblée de l'effet d'un biomarqueur

Séance	7 – vendredi 26 février 2027
Chapitre	Tuglus, C. et van der Laan, M. J. (2011). Targeted methods for biomarker discovery. Chapitre 22 de M. J. van der Laan et S. Rose, *Targeted Learning : Causal Inference for Observational and Experimental Data*. New York : Springer.
Évaluation	**Aucune** : la préparation n'est pas notée, mais elle est indispensable pour profiter de la séance, pour le travail de synthèse et pour l'examen final.
Préalables	Séance 6 (équations d'estimation, fonction d'influence, variance sandwich ; au doctorat, théorie semi-paramétrique) ; séance 1 (estimateurs M)
Fichiers	Chapitre (PDF) ; `genes_select.RDS` ; jeu `golub` du module `multtest` (Bioconductor) ; fiche vierge (annexe)
Modules R	`multtest` (installer avec `BiocManager::install("multtest")`) ; `earth` (CRAN) ; `polspline` facultatif

1 Contexte et objectifs

Dans une étude de découverte de biomarqueurs, on veut mesurer l'effet propre de l'expression A d'un gène sur une issue Y en tenant compte de l'expression W de nombreux autres gènes. Plutôt que d'interpréter un coefficient de régression multiple dont la valeur dépend d'un modèle paramétrique possiblement faux, Tuglus et van der Laan définissent un paramètre d'intérêt – l'effet linéaire β du biomarqueur, ajusté pour W – et l'estiment par maximum de vraisemblance ciblé (TMLE). Les fonctions de nuisance (régression de Y et de A sur W) sont estimées de façon non paramétrique, puis l'estimation initiale est corrigée dans la direction de la fonction d'influence, ce qui fournit directement une erreur-type.

À la fin de la séance, vous devriez pouvoir :

1. définir le paramètre d'intérêt et distinguer ce paramètre des paramètres de nuisance ;
2. décrire les étapes du TMLE de β (estimation initiale, covariable astucieuse, mise à jour) et justifier le critère utilisé à chaque étape ;
3. utiliser la fonction d'influence pour obtenir une erreur-type et un test de Wald ;
4. implanter l'estimateur en R sur les données de Golub et coll. (1999).

2 Travail à faire avant la séance

1. **Lire le chapitre 22** en entier, en insistant sur la sous-section *Implementation* (deux encadrés gris, p. 370–371), la sous-section *Statistical inference* (bas de la p. 371) et l'application aux leucémies (section 22.5, tableau 22.4).
2. **Remplir la fiche des caractéristiques méthodologiques** (fiche vierge en annexe A, p. 5), en justifiant chaque choix par un renvoi au chapitre.
3. **Préparer les exercices** de la section 3 selon votre niveau et apporter dérivations et code R.
4. **Noter au moins deux questions** sur des passages du chapitre que vous n'avez pas compris.

3 Exercices

Les exercices sont repris du matériel de l'hiver 2026 (A. Bureau).

1. [Maîtrise et doctorat] Dans l'application aux données de leucémie de Golub et coll. (section 22.5), on utilise un modèle *linéaire* pour la réponse dichotomique leucémie lymphoblastique (ALL) ou myéloïde (AML). Justifiez l'emploi du critère des moindres carrés pour estimer le coefficient de la covariable astucieuse (point 2 du 2^{e} encadré, p. 371) avec une telle réponse. Indice : quelle équation d'estimation le critère des moindres carrés résout-il, et pourquoi suffit-elle ici ?
2. [Défi – recommandé au doctorat] Dérivez la fonction d'influence (« courbe d'influence ») du TMLE de l'effet linéaire β d'un biomarqueur A donnée au bas de la p. 371.
3. [Maîtrise et doctorat] Reproduisez l'estimation de l'effet linéaire β du gène TOP2B rapportée sur la première ligne du tableau 22.4 et obtenez une valeur p pour l'hypothèse nulle $\beta = 0$, en suivant les étapes ci-dessous.
   (a) Chargez le jeu `golub` du module `multtest` (`data(golub)`) et le fichier `genes_select.RDS` (`readRDS`), un vecteur d'indices qui identifie les 681 gènes retenus à un taux de fausses découvertes (FDR) inférieur à 0,05 par un test de Student de l'association marginale entre leur expression et le type de leucémie. Extrayez l'expression de ces 681 gènes de la matrice `golub` et leurs noms de la matrice `golub.gnames`. La réponse Y est `golub.cl` (0 = ALL, 1 = AML).
   (b) Calculez la corrélation entre l'expression de TOP2B (34^{e} des 681 gènes retenus) et celle des autres gènes retenus. Conservez comme covariables W les gènes dont la corrélation avec TOP2B est inférieure à 0,5. Indiquez si vous appliquez ce seuil à la corrélation ou à sa valeur absolue, et le nombre de gènes retenus ; si le temps le permet, comparez les deux choix.
   (c) Effectuez les étapes des deux encadrés gris de la sous-section *Implementation* (p. 370–371) pour obtenir le TMLE de β et la prédiction $r_n^1(W)$. Comme les auteurs (p. 378), prédisez les fonctions de nuisance de façon non paramétrique par splines de régression adaptatives : le module `earth` est plus stable et plus efficace que le module classique `polspline`. Tenez compte de la nature de la variable à prédire : variable réelle (expression de TOP2B, `earth`

avec réponse gaussienne) ou proportion (réponse binaire, `earth` avec l'argument `glm` réglé à la famille binomiale).

(d) Évaluez la fonction d'influence du TMLE de β à l'aide de la formule de la sous-section *Statistical inference* (bas de la p. 371) pour chaque observation ; déduisez-en une erreur-type (écart-type empirique de la fonction d'influence divisé par $\sqrt{n}$) et effectuez un test de Wald de $H_0 : \beta = 0$.

(e) Comparez votre estimation, votre erreur-type et votre valeur p à celles du tableau 22.4. Commentez les écarts éventuels (choix de W, méthode de lissage, réglages de `earth`).

Indications

— Dans `golub`, les lignes sont les gènes et les colonnes les échantillons : transposez la matrice pour obtenir une observation par ligne ($n = 38$).

— Avec $n = 38$ et plusieurs centaines de covariables, le choix des réglages de `earth` (p. ex. `degree`, `nk`) influence les résultats : notez ceux que vous utilisez.

4 Grille d'autoévaluation

Cette grille n'est pas utilisée pour noter : elle vous permet de juger vous-même si votre préparation est suffisante. Cochez une case par ligne avant la séance.

Critère	Je suis prêt ou prête si…	Prêt	Partiel	À revoir
Fiche méthodologique	Les six rubriques sont remplies ; je distingue clairement le paramètre d'intérêt β des fonctions de nuisance ; chaque choix renvoie au chapitre.	☐	☐	☐
Dérivations	J'ai une justification complète pour l'exercice 1 (et une tentative rédigée de l'exercice 2 au doctorat) que je pourrais présenter au tableau.	☐	☐	☐
Reproduction des résultats	Mon estimation de β pour TOP2B a le même signe et le même ordre de grandeur que celle du tableau 22.4, et ma conclusion au seuil de 5 % est la même ; je peux expliquer les écarts.	☐	☐	☐
Code R	Mon code s'exécute du début à la fin, est commenté, suit les étapes des encadrés et indique les réglages de `earth`.	☐	☐	☐
Compréhension	Je peux expliquer en deux minutes pourquoi la mise à jour ciblée permet d'utiliser la fonction d'influence pour l'inférence malgré l'estimation non paramétrique des nuisances.	☐	☐	☐
Questions	J'ai noté au moins deux questions précises pour la séance.	☐	☐	☐

5 Lien avec l'examen final et le travail de synthèse

— **Examen final.** On peut vous demander d'identifier l'objet d'intérêt et les paramètres de nuisance d'une méthode, de dériver ou d'utiliser une fonction d'influence pour obtenir une variance, ou de justifier un critère d'estimation. Les exercices 1 et 2 sont représentatifs de ce type de question.

— **Travail de synthèse.** Plusieurs articles suggérés (variance sandwich pour la g-computation, régression de Poisson robuste) reposent sur les mêmes outils : fonction d'influence et estimation des nuisances. La fiche remplie ici sert de modèle pour la fiche de lecture du travail de synthèse.

A Fiche des caractéristiques méthodologiques

Référence : Tuglus et van der Laan (2011), ch. 22 de *Targeted Learning*, Springer.

Élément	Choix	Justification
1. Objet statistique d'intérêt		
2. Paramètres nuisibles, données manquantes, sources de biais		
3. Approche pour traiter la nuisance ou éviter les biais		
4. Critère d'estimation ou de test		
5. Méthode d'optimisation ou de calcul		
6. Obtention de la distribution des statistiques		