Étude de cas 3 – Consignes de préparation

Échantillonnage à deux phases

Séance	10 – vendredi 2 avril 2027
Article	Breslow, N. E. et Chatterjee, N. (1999). Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 48(4), 457–468.
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	Séances 8 et 9 (données manquantes, vraisemblance des données observées, pondération par probabilité inverse) ; séance 6 (variance sandwich)
Fichiers	Article (PDF) ; `nwts.dat`, `NWTSc.dat`, `nwtsb.dat`, `nwtsb_indiv.csv` ; documentation de la fonction `tps` ; fiche vierge (annexe)
Modules R	`osDesign` (fonction `tps`) ; `stats::glm`

1 Contexte et objectifs

Dans l'étude nationale sur la tumeur de Wilms (NWTS), l'histologie déterminée par un laboratoire central – coûteuse – est la variable explicative d'intérêt, alors que l'histologie établie par l'hôpital d'origine, le stade et la rechute sont connus pour toute la cohorte. Un devis à deux phases mesure l'information coûteuse seulement dans un sous-échantillon dont la probabilité de sélection dépend de la réponse et de strates définies par des variables de la phase I. Breslow et Chatterjee (1999) comparent plusieurs devis (cas-témoins, équilibré) et trois critères d'estimation de la régression logistique qui tiennent compte de ce plan d'échantillonnage : la vraisemblance pondérée, la pseudo-vraisemblance et le maximum de vraisemblance semi-paramétrique.

À la fin de la séance, vous devriez pouvoir :

1. voir un devis à deux phases comme un problème de données manquantes *par devis* et en écrire la vraisemblance des données observées ;
2. expliquer les trois critères d'estimation, leurs hypothèses et leur efficacité relative ;
3. justifier la forme de l'estimateur sandwich lorsque les effectifs de phase I sont aléatoires ;
4. reproduire les analyses de l'article avec la fonction `tps`.

2 Travail à faire avant la séance

1. **Lire l'article** en entier, en insistant sur la présentation des trois critères, sur l'estimateur sandwich (formule 3), sur l'expression (5) et sur les tableaux 2 et 3.
2. **Remplir la fiche des caractéristiques méthodologiques** (annexe A, p. 5). Les rubriques 1 à 3 sont communes ; les rubriques 4 à 6 se remplissent *pour chacun des trois critères.*
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 de l'article que vous n'avez pas compris.

3 Exercices

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

1. [Doctorat] Démontrez la formule de la « viande » de l'estimateur sandwich (3) de Breslow et Chatterjee (1999). Indice : les effectifs N_{ij} de la phase I sont aléatoires ; calculez la variance théorique du score avant de substituer des estimations empiriques.
2. [Maîtrise et doctorat] Obtenez l'expression (5) à partir du modèle de risque et des probabilités d'échantillonnage.
3. [Maîtrise et doctorat] En vous basant sur l'équation de la p. 101 de Wu (2010), écrivez la fonction de vraisemblance des données observées d'un échantillon à deux phases de l'étude de la tumeur de Wilms comme une somme (ou une intégrale) de la vraisemblance des données complètes sur les données manquantes. Précisez quelles variables sont observées pour tous et lesquelles seulement en phase II.
4. [Maîtrise et doctorat] Écrivez les probabilités qui décrivent le mécanisme de génération des données manquantes (le *coarsening* des données). Ce mécanisme est-il MCAR, MAR ou MNAR ? Justifiez.
5. [Maîtrise et doctorat] Les fichiers fournis contiennent la cohorte entière (`nwts.dat`), un échantillon cas-témoins (`NWTSc.dat`) et un échantillon équilibré, en données groupées (`nwtsb.dat`) ou individuelles (`nwtsb_indiv.csv`). Reproduisez les analyses des tableaux 2 et 3 de Breslow et Chatterjee (1999) à l'aide de la fonction `tps` du module `osDesign` pour les trois critères (`method = "WL"`, `"PL"`, `"ML"`) et de la fonction `glm` pour la cohorte entière.
6. [Défi] Refaites les analyses en stratifiant sur l'histologie institutionnelle seulement (2 strates) et en traitant le stade comme une variable quantitative plutôt que comme un facteur. Comparez les coefficients obtenus par les trois critères aux estimations du maximum de vraisemblance dans la cohorte entière. Que se passe-t-il lorsque le stade n'est plus une variable de stratification ?

Indications pour l'exercice 5

— Les fichiers `.dat` ont un en-tête et des colonnes séparées par des espaces (`read.table(..., header = TRUE)`) ; le fichier `nwtsb_indiv.csv` utilise le point-

virgule comme séparateur (`read.csv2` ou `sep = ";"`). Sous Linux et macOS, respectez la casse du nom `NWTSc.dat`.

— Dans `tps`, `nn0` et `nn1` sont les effectifs de témoins et de cas *de la phase I* dans chaque strate (la documentation indique par erreur « phase II » pour `nn1`), et `group` donne la strate de chaque ligne des données de phase II. Vérifiez que l'ordre des strates est le même dans les trois arguments : l'exemple de la documentation de `tps` définit huit strates (stade × histologie institutionnelle).

— Si `osDesign` n'est plus offert sur CRAN pour votre version de R, installez-le à partir des archives de CRAN.

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 rubriques 1 à 3 sont remplies ; les rubriques 4 à 6 sont remplies pour les trois critères et je sais expliquer ce qui les distingue (poids, conditionnement, modélisation de la loi des covariables).	☐	☐	☐
Dérivations	J'ai rédigé les exercices 2 à 4 (et l'exercice 1 au doctorat) jusqu'au bout et je pourrais les présenter au tableau.	☐	☐	☐
Reproduction des résultats	Mes coefficients et erreurs-types des tableaux 2 et 3 coïncident avec ceux de l'article à l'arrondi près (écart d'au plus une unité sur la dernière décimale publiée) ; sinon, j'ai identifié la cause (strates, codage, échantillon).	☐	☐	☐
Code R	Mon code s'exécute du début à la fin, est commenté, lit tous les fichiers et produit les tableaux comparatifs.	☐	☐	☐
Compréhension	Je peux expliquer en deux minutes pourquoi la pseudo-vraisemblance et le maximum de vraisemblance sont plus efficaces que la vraisemblance pondérée, et à quelles conditions.	☐	☐	☐
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.** Les questions d'examen peuvent demander d'écrire la vraisemblance des données observées d'un devis à données manquantes, de décrire le mécanisme de non-réponse, de construire des poids de probabilité inverse ou de comparer des critères d'estimation selon leurs hypothèses et leur efficacité. Les exercices 2 à 4 sont représentatifs.

— **Travail de synthèse.** Comparer plusieurs critères pour un même objet d'intérêt, comme le fait la fiche à trois colonnes, est une excellente façon de structurer la présentation d'un

article qui compare des méthodes (p. ex. GEE et modèles mixtes, ou analyses avec données manquantes).

A Fiche des caractéristiques méthodologiques

Référence : Breslow et Chatterjee (1999), *Applied Statistics*, 48(4), 457–468.

Rubriques communes aux trois critères

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		

Rubriques propres à chaque critère

Élément	Vraisemblance pondérée (WL)	Pseudo-vraisemblance (PL)	Maximum de vraisemblance semi-paramétrique (ML)
4. Critère d'estimation : forme et justification			
5. Méthode d'optimisation ou de calcul			
6. Obtention de la distribution des statistiques			