Étude de cas 1 – Consignes de préparation

Effets intra et inter-grappes

Séance	5 – vendredi 12 février 2027
Article	Neuhaus, J. M. et McCulloch, C. E. (2006). Separating between- and within-cluster covariate effects by using conditional and partitioning methods. *Journal of the Royal Statistical Society, Series B*, 68(5), 859–872.
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 3 (modèles linéaires mixtes généralisés, effets individuels et marginaux) ; régression logistique conditionnelle (rappel fourni sur le site)
Fichiers	Article (PDF) ; `naisga.csv` ; fiche vierge (ci-dessous)

1 Contexte et objectifs

Dans un modèle mixte, on suppose habituellement que l'effet aléatoire de grappe est indépendant des covariables. Lorsque cette hypothèse est fausse – par exemple lorsque l'âge moyen d'une mère sur ses grossesses successives est associé à son risque de base de donner naissance à un bébé de faible poids – l'estimateur standard du coefficient d'une covariable variant à l'intérieur des grappes est biaisé. Neuhaus et McCulloch (2006) comparent deux remèdes : la vraisemblance *conditionnelle*, qui élimine l'effet de grappe, et la *décomposition* de la covariable en une composante inter-grappes (la moyenne de grappe $\bar{X}_i$) et une composante intra-grappe ($X_{ij} - \bar{X}_i$).

À la fin de la séance, vous devriez pouvoir :

1. expliquer pourquoi une corrélation entre l'effet aléatoire et une covariable biaise l'estimateur du modèle mixte standard, et ce que la vraisemblance conditionnelle et la décomposition intra-inter estiment exactement ;
2. ajuster les trois modèles en R et interpréter les coefficients intra et inter ;
3. quantifier par bootstrap l'incertitude d'une différence entre deux estimateurs.

2 Travail à faire avant la séance

1. **Lire l'article** en entier, en insistant sur les sections qui présentent les modèles, les résultats théoriques sur le biais et l'application aux naissances en Géorgie (tableau 2).
2. **Remplir la fiche des caractéristiques méthodologiques** (fiche vierge en annexe A, p. 4). Pour chaque rubrique, indiquez le choix fait par les auteurs *et* sa justification, avec un renvoi à

la section ou à l'équation de l'article.

3. **Préparer les exercices** de la section 3 selon votre niveau. Apportez vos dérivations (papier ou PDF) et votre code R : ils seront discutés en classe et certaines personnes seront invitées à présenter leur solution.

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). Le niveau indique ce qui est attendu de chacun ; les exercices de niveau supérieur sont fortement recommandés à tous.

1. [Doctorat] Démontrez l'équation (11) de Neuhaus et McCulloch (2006). Énoncez les hypothèses utilisées à chaque étape.

2. [Maîtrise et doctorat] On pose $\eta_{ij} = X_{ij} - \bar{X}_i$ et on définit l'effet aléatoire $b_i^* = b_i + \beta_B \bar{X}_i$. Montrez que la covariable peut s'écrire sous la forme $X_{ij} = h(b_i) + \eta_{ij}$ utilisée dans l'article, et donnez l'expression de la fonction h. Expliquez en une phrase pourquoi cette représentation rend explicite la corrélation entre l'effet aléatoire et la covariable.

3. [Maîtrise et doctorat] Le fichier `naisga.csv` contient un jeu de données de naissances en Géorgie plus complet que celui de l'article (variables `matid` : identifiant de la mère ; `poidsnais` : poids de naissance en grammes ; `mat_age` : âge de la mère ; ainsi que `dprenat`, `sexe`, `matrace`, `fumeuse`, non requises ici). Définissez le faible poids de naissance par `poidsnais` $\leq 2\,500$ g. Estimez, en fonction de l'âge maternel :

   (a) le modèle logistique mixte standard, avec ordonnée à l'origine aléatoire par mère (fonction `glmer` du module `lme4`) ;

   (b) le modèle logistique mixte avec décomposition intra-inter de l'âge (covariables $\bar{X}_i$ et $X_{ij} - \bar{X}_i$) ;

   (c) le modèle logistique conditionnel (p. ex. `survival::clogit` avec `strata(matid)`).

   Présentez les coefficients de l'âge et leurs erreurs-types dans un tableau semblable au tableau 2 de l'article et commentez. Les estimations devraient être *semblables* à celles de l'article, sans être identiques, puisque les données sont plus complètes.

4. [Défi] Refaites le bootstrap non paramétrique de Neuhaus et McCulloch (2006) pour estimer l'erreur-type de la différence entre le coefficient de l'âge du modèle mixte standard et le coefficient intra-grappe du modèle avec décomposition. Rééchantillonnez les *mères* (et non les naissances), utilisez au moins $B = 500$ réplications et fixez le germe. Concluez : la différence est-elle significative ?

Indications

— Retirez les naissances dont le poids ou l'âge est manquant et indiquez le nombre de mères et de naissances retenues ; les mères qui n'ont qu'une naissance ne contribuent pas à la vraisemblance conditionnelle.

— Dans le bootstrap, une même mère peut être tirée plusieurs fois : attribuez-lui un nouvel identifiant à chaque tirage pour que `glmer` la traite comme une grappe distincte.

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 ; chaque choix est justifié et renvoie à une section ou une équation de l'article ; je distingue clairement les deux approches.	☐	☐	☐
Dérivations	J'ai rédigé l'exercice 2 (et l'exercice 1 au doctorat) jusqu'au bout, chaque étape est justifiée et je pourrais l'expliquer au tableau.	☐	☐	☐
Reproduction des résultats	Mes trois coefficients de l'âge ont le même signe et le même ordre de grandeur que ceux du tableau 2 et s'en écartent de moins de deux erreurs-types ; je peux expliquer les écarts.	☐	☐	☐
Code R	Mon code s'exécute du début à la fin, est commenté et produit le tableau comparatif ; le germe du bootstrap est fixé.	☐	☐	☐
Compréhension	Je peux expliquer en deux minutes pourquoi le modèle mixte standard est biaisé ici et ce qu'estiment les coefficients intra et inter.	☐	☐	☐
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 demandent souvent de remplir une ou plusieurs rubriques de la grille méthodologique pour une situation nouvelle, de justifier un choix de modèle (mixte, conditionnel, marginal) ou de reconnaître une source de biais. L'opposition entre effets intra et inter-grappes et la régression logistique conditionnelle font partie de la matière d'examen.

— **Travail de synthèse.** La fiche que vous remplissez ici a la même structure que la fiche de lecture proposée pour le travail de synthèse ; l'exercice 3 est un modèle réduit de ce qui est attendu : reproduire une analyse publiée avec du code documenté.

A Fiche des caractéristiques méthodologiques

Référence : Neuhaus et McCulloch (2006), *JRSS B*, 68(5), 859–872. Remplissez une fiche distincte pour chacune des deux approches étudiées (vraisemblance conditionnelle et décomposition intra-inter) si leurs choix diffèrent pour une rubrique, ou indiquez les deux choix dans la même case.

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		