Devoir 1

Méthode delta, simulation et modèle logistique mixte

Pondération	20 % de la note finale (corrigé sur 100 points)
Matière	Séances 1 à 4 : inférence asymptotique, études de simulation, modèles linéaires mixtes généralisés, modèles additifs généralisés mixtes
Échéance	**vendredi 26 février 2027, 23 h 55**
Mode	Travail individuel
Remise	Boîte de dépôt « Devoir 1 » sur Brio : un rapport PDF et le code R
Données	`bsi.csv` (site du cours, section « Devoir 1 »)

1 Contexte et objectifs

Ce devoir comporte deux exercices indépendants. Le premier, plutôt théorique, porte sur la méthode delta et sur la validation d'estimateurs de variance par une étude de simulation ; il mobilise les séances 1 et 2. Le second est une analyse de données longitudinales binaires par modèle additif généralisé mixte puis par modèle logistique mixte ; il mobilise les séances 3 et 4.

Au terme du devoir, vous devriez être en mesure de :

1. appliquer la méthode delta d'ordre 2 lorsque la dérivée première s'annule ;
2. concevoir, réaliser et rapporter une étude de simulation qui valide un estimateur de variance (biais, taux de convergence, erreur de Monte Carlo) ;
3. décrire la forme d'un effet du temps par lissage dans un modèle additif mixte ;
4. sélectionner les effets aléatoires d'un modèle logistique mixte avec la bonne loi de référence (mélange de khi-deux) et tester ses effets fixes ;
5. écrire un modèle mixte sous forme matricielle en précisant les dimensions.

2 Exercice 1 – Variance de $\hat{\theta}(1-\hat{\theta})$ pour une loi de Bernoulli (45 points)

Soit $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\theta)$, $0 < \theta < 1$, et $\hat{\theta} = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Pour estimer $\text{Var}(\hat{\theta}) = \theta(1-\theta)/n$, on peut substituer $\hat{\theta}$ à θ dans la variance de la loi de Bernoulli,

$$\widehat{\text{Var}}(\hat{\theta}) = \frac{\hat{\theta}(1-\hat{\theta})}{n},$$

ou utiliser la variance échantillonnale :

$$v(\hat{\theta}) = \frac{s^2}{n}, \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2.$$

La quantité $\hat{\theta}(1-\hat{\theta})$ (qui divise par n) et la quantité s^2 (qui divise par $n-1$) sont donc deux estimateurs de la variance $\theta(1-\theta)$ d'une observation. Pour s^2, on utilisera l'estimateur de variance suivant :

$$v(s^2) = \frac{1}{n}\left(\hat{\mu}_4 - \frac{n-3}{n-1}\,s^4\right), \qquad \hat{\mu}_4 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^4.$$

a) **Dérivation. (12 points)**

(i) Posez $g(t) = t(1-t)$. Appliquez la méthode delta d'ordre 1 à $g(\hat{\theta})$ et montrez pourquoi elle ne fournit pas une approximation utile de $\mathrm{Var}\{\hat{\theta}(1-\hat{\theta})\}$ lorsque $\theta = 1/2$.

(ii) À l'aide d'un développement de Taylor d'ordre 2 de g autour de θ (méthode delta d'ordre 2), dérivez une approximation de $\mathrm{Var}\{\hat{\theta}(1-\hat{\theta})\}$ valable pour tout $\theta \in (0,1)$. Énoncez clairement les résultats de convergence utilisés et les termes négligés. Indiquez l'ordre de grandeur (en puissance de n) de cette variance lorsque $\theta \neq 1/2$ et lorsque $\theta = 1/2$.

(iii) Déduisez-en un estimateur de $\mathrm{Var}\{\hat{\theta}(1-\hat{\theta})\}$ en substituant $\hat{\theta}$ à θ.

b) **Étude de simulation. (13 points)** Validez l'estimateur de la variance de $\hat{\theta}(1-\hat{\theta})$ obtenu en a)(iii) et l'estimateur $v(s^2)$ de la variance de s^2. Plus précisément :

(i) considérez les six scénarios formés de $n \in \{100,\ 1\,000,\ 10\,000\}$ et $\theta \in \{0{,}2;\ 0{,}5\}$;

(ii) pour chaque scénario, générez $B = 5\,000$ échantillons indépendants de taille n (fixez le germe avec `set.seed()` et indiquez-le dans le rapport) ;

(iii) dans chaque échantillon, calculez $\hat{\theta}(1-\hat{\theta})$, s^2, l'estimateur de $\mathrm{Var}\{\hat{\theta}(1-\hat{\theta})\}$ de a)(iii), $v(s^2)$, ainsi que $\widehat{\mathrm{Var}}(\hat{\theta})$ et $v(\hat{\theta})$ (utiles en d) ;

(iv) présentez les résultats dans un tableau qui donne, pour chaque scénario et pour chacun des deux estimateurs $\hat{\theta}(1-\hat{\theta})$ et s^2 :

— la moyenne de l'estimateur sur les B réplications ;

— sa variance empirique (variance des B valeurs), qui sert de valeur de référence ;

— la moyenne de son estimateur de variance ;

— le biais relatif de l'estimateur de variance, (moyenne de l'estimateur de variance − variance empirique)/variance empirique ;

— l'erreur-type de Monte Carlo de la moyenne de l'estimateur de variance (écart-type des B valeurs divisé par $\sqrt{B}$).

Commentez : les estimateurs de variance sont-ils sans biais (à l'erreur de Monte Carlo près) ? Le comportement diffère-t-il entre $\theta = 0{,}2$ et $\theta = 0{,}5$?

c) **Taux de convergence. (10 points)** Comparez le taux de convergence de $\hat{\theta}(1-\hat{\theta})$ et de s^2 à $\theta = 0{,}2$ et à $\theta = 0{,}5$ en examinant comment évoluent, lorsque n est multiplié par 10, (1) leur

variance empirique et (2) la moyenne de leur estimateur de variance. Présentez les rapports successifs (de $n = 100$ à 1 000, puis de 1 000 à 10 000) dans un tableau ; un graphique log-log est un complément bienvenu. Les formules analytiques de a) et de $v(s^2)$ permettaient-elles de prédire les taux observés ? Justifiez, en particulier pour $\theta = 1/2$.

d) **Comparaison de $\widehat{\mathrm{Var}}(\hat{\theta})$ et de $v(\hat{\theta})$. (10 points)** À partir de la même simulation, comparez les deux estimateurs de $\mathrm{Var}(\hat{\theta})$ par rapport à la vraie valeur $\theta(1-\theta)/n$: biais, biais relatif, variance et erreur quadratique moyenne. Pour comparer des tailles d'échantillon différentes, il est commode de multiplier ces quantités par n (ou par n^2 pour la variance et l'erreur quadratique moyenne). Lequel recommanderiez-vous, et la différence a-t-elle une importance pratique ?

Indications pour l'exercice 1

- Il est attendu que vous généreriez les n observations de chaque échantillon. Si vous utilisez un raccourci (p. ex. générer directement $n\hat{\theta} \sim \mathrm{Binomiale}(n,\theta)$), démontrez d'abord qu'il donne exactement les mêmes estimateurs.
- Pour les calculs de vérification, rappelons que pour une loi de Bernoulli le quatrième moment centré vaut $\mu_4 = \theta(1-\theta)\{1-3\theta(1-\theta)\}$.
- Avec $B = 5\,000$, l'ensemble de la simulation s'exécute en quelques minutes ; si vous réduisez B pour des raisons de temps de calcul, justifiez-le à l'aide de l'erreur de Monte Carlo.

3 Exercice 2 – Drogues dures chez les parents et santé mentale des adolescents (45 points)

Contexte

Les enfants de parents infectés par le VIH vivent des situations stressantes qui peuvent affecter leur développement et leur santé mentale. L'effet de la maladie du parent sur l'enfant peut être aggravé par divers facteurs, dont la consommation de drogues « dures » par le parent. Rotheram-Borus et coll. (2004) ont suivi pendant six ans un échantillon de parents infectés par le VIH, résidant à New York et ayant un ou plusieurs adolescents (11 à 21 ans). Vous utiliserez ces données pour examiner l'association entre la consommation de drogues dures par le parent et la santé mentale de ses adolescents.

L'état mental des adolescents a été évalué par le *Brief Symptom Inventory* (BSI ; Derogatis et Melisaratos, 1983), un inventaire de 53 symptômes et états d'esprit négatifs couvrant un vaste ensemble de troubles psychologiques (dépression, anxiété, etc.). Chaque symptôme est coté de 0 à 4 et l'indice global de sévérité est la moyenne arithmétique des 53 items. Un indice supérieur ou égal à 1 indique la présence d'un trouble psychologique.

Le jeu de données contient les 215 parents dont la consommation de drogues dures est connue et leurs 323 adolescents. Les visites étaient prévues tous les 3 mois pendant au plus 7 ans ; en pratique, les intervalles entre visites sont souvent irréguliers. Certains parents sont décédés avant la fin du suivi : les données de leurs adolescents ne sont incluses que jusqu'au décès du parent. Le fichier `bsi.csv` contient les variables suivantes.

Variable	Description
`pid`	identifiant numérique de l'adolescent
`parent`	identifiant numérique du parent
`temps`	temps écoulé depuis le début de l'étude (mois)
`droguedure`	consommation de drogues dures par le parent au cours des 3 derniers mois : `"nonconsommateur"`, `"consommateur-courant"` ou `"consommateur-abst"` (consommateur qui s'est abstenu au cours des 3 derniers mois)
`dbsi_gsi`	indicatrice d'un indice BSI ≥ 1 (1 = trouble psychologique, 0 = pas de trouble)

Cadre de modélisation (pour toutes les sous-questions)

— La réponse est `dbsi_gsi` (binaire) ; utilisez le lien logit (lien canonique), ce qui facilite la convergence.

— Les adolescents sont nichés dans les parents, mais la variabilité au niveau des parents est négligeable : n'incluez des effets aléatoires **qu'au niveau de l'adolescent** (`pid`).

— Déclarez `droguedure` comme facteur et indiquez la catégorie de référence retenue (nous suggérons `"nonconsommateur"`).

a) **Forme de l'effet du temps (modèle additif mixte). (8 points)** Ajustez avec `mgcv::gam` (estimation par maximum de vraisemblance, `method = "ML"`) un modèle logistique additif mixte comprenant : un terme lisse du temps, `s(temps)` ; `droguedure` comme effet fixe ; une ordonnée à l'origine aléatoire par adolescent, `s(pid, bs = "re")` (avec `pid` déclaré comme facteur). N'incluez *ni* interaction entre `droguedure` et le temps, *ni* pente aléatoire pour le temps. Présentez le graphique de l'effet lisse estimé du temps sur la log-cote de trouble psychologique (avec bande de confiance) et décrivez sa forme. Rapportez le nombre effectif de degrés de liberté du terme lisse et commentez.

Pour la suite (b à d), quelles que soient vos observations en a), adoptez une spline linéaire avec changement de pente à 18 mois. Définissez la variable

```
bsi$apres18mois <- pmax(0, bsi$temps - 18)
```

et incluez `temps` et `apres18mois` dans le modèle logistique mixte (`lme4::glmer`). Les modèles sont donc désormais paramétriques.

Problèmes de convergence

Certains modèles risquent de produire des avertissements de convergence. Essayez des remèdes raisonnables (valeurs initiales tirées d'un modèle plus simple avec l'argument `start`, autre optimiseur via `glmerControl(optimizer = "bobyqa")`, centrage ou changement d'échelle du temps, p. ex. en années). Documentez ce que vous avez tenté. Si les avertissements persistent, poursuivez en supposant que les estimations sont correctes : l'expérience montre qu'elles sont alors raisonnables. Un avertissement documenté n'est pas pénalisé.

b) **Sélection des effets aléatoires. (12 points)** Dans tous les modèles comparés, incluez les mêmes effets fixes : `temps`, `apres18mois`, `droguedure` et les interactions de `droguedure` avec

`temps` et avec `apres18mois`. Comparez au minimum :

— un modèle avec ordonnée à l'origine aléatoire seulement ;

— un modèle avec ordonnée à l'origine et pentes aléatoires pour `temps` et `apres18mois` (matrice de covariance non structurée).

Effectuez un test du rapport de vraisemblances. Précisez les hypothèses testées, le nombre de paramètres de covariance ajoutés, la loi de référence appropriée (le paramètre testé est sur la frontière de l'espace des paramètres) et la valeur p correspondante ; expliquez pourquoi la valeur p produite par `anova()` n'est pas la bonne. Concluez sur la structure d'effets aléatoires retenue.

c) **Tests des effets fixes. (15 points)** Avec la structure d'effets aléatoires retenue en b), répondez aux questions suivantes en précisant chaque fois l'hypothèse nulle, la statistique (rapport de vraisemblances ou Wald), ses degrés de liberté et la valeur p :

(i) Y a-t-il une interaction entre la consommation de drogues dures par le parent et le temps ? (test conjoint de tous les termes d'interaction) ;

(ii) Y a-t-il une différence entre les statuts de consommation ? (test conjoint de l'effet de `droguedure`, dans le modèle sans interaction si l'interaction n'est pas retenue) ;

(iii) Si oui, quels statuts diffèrent ? Effectuez les comparaisons deux à deux avec un ajustement pour comparaisons multiples (p. ex. `emmeans::pairs`) et rapportez les rapports de cotes avec intervalles de confiance à 95 %.

Si le test du rapport de vraisemblances et le test de Wald mènent à des conclusions différentes, signalez-le et proposez une explication.

d) **Forme matricielle du modèle retenu. (10 points)** Écrivez le modèle retenu à l'issue de b) et c) sous la forme

$$\text{logit}\{\text{E}(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, \boldsymbol{b}_i)\} = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i, \qquad i = 1, \ldots, 323,$$

où i indice les adolescents et n_i est le nombre de visites de l'adolescent i. Donnez les dimensions de $\boldsymbol{Y}_i$, $\boldsymbol{X}_i$, $\boldsymbol{Z}_i$; écrivez explicitement une ligne type de $\boldsymbol{X}_i$ et de $\boldsymbol{Z}_i$ ainsi que les vecteurs $\boldsymbol{\beta}$ et $\boldsymbol{b}_i$; précisez la loi de $\boldsymbol{b}_i$. Dressez la liste complète des paramètres du modèle (effets fixes et paramètres de covariance) et leur nombre.

4 Livrables

1. **Rapport PDF** nommé `STT7360_D1_Nom_Prenom.pdf`, rédigé en français, d'au plus **15 pages** (annexes de code non comprises). Le rapport suit l'ordre des questions ; les dérivations de l'exercice 1 peuvent être tapées (LaTeX, R Markdown, Quarto) ou manuscrites et numérisées lisiblement. Chaque tableau et chaque figure est numéroté, possède un titre et est commenté dans le texte. Ne collez pas de sorties R brutes dans le corps du rapport : extrayez les nombres utiles dans des tableaux (arrondis à un nombre raisonnable de chiffres significatifs).

2. **Code R** nommé `STT7360_D1_Nom_Prenom.R` (ou `.Rmd`/`.qmd`), commenté, qui reproduit *tous* les résultats du rapport lorsqu'il est exécuté du début à la fin avec `bsi.csv` dans le répertoire

de travail. Indiquez en tête les modules utilisés et le germe aléatoire. Un fichier `.Rmd`/`.qmd` dont le rendu constitue le rapport est accepté s'il respecte les consignes de présentation.

3. Dépôt des deux fichiers dans la boîte de dépôt « Devoir 1 » de Brio au plus tard le **vendredi 26 février 2027 à 23 h 55**.

Retards, intégrité. La politique de retard et les règles d'intégrité (y compris l'usage déclaré des outils d'intelligence artificielle) sont celles du plan de cours. Vous pouvez discuter des idées avec vos collègues, mais le rapport et le code doivent être personnels.

5 Grille de correction (100 points)

Critère	Attendus et descripteurs	Pts
Exercice 1 – Variance de $\hat{\theta}(1-\hat{\theta})$		
a) Dérivation	(i) développement d'ordre 1 correct et constat que $g'(1/2) = 0$ (2 pts) ; (ii) développement d'ordre 2, loi limite de $\sqrt{n}(\hat{\theta}-\theta)$ et usage de $\mathrm{Var}(Z^2) = 2$ (ou calcul des moments) justifiés, formule finale exacte (7 pts) ; ordres de grandeur $O(n^{-1})$ et $O(n^{-2})$ (1 pt) ; (iii) estimateur par substitution (2 pts).	12
b) Simulation	Plan conforme : six scénarios, B indiqué, germe fixé (3 pts) ; calculs corrects de tous les estimateurs (4 pts) ; tableau complet et lisible avec biais relatif et erreur de Monte Carlo (4 pts) ; commentaire juste sur le biais selon θ et n (2 pts).	13
c) Taux de convergence	Rapports successifs calculés pour les deux estimateurs et les deux valeurs de θ (4 pts) ; taux identifiés ($1/n$ vs $1/n^2$ pour la variance) (3 pts) ; lien explicite avec les formules analytiques, y compris pour $v(s^2)$ à $\theta = 1/2$ (3 pts).	10
d) $\widehat{\mathrm{Var}}(\hat{\theta})$ vs $v(\hat{\theta})$	Biais, biais relatif, variance et EQM calculés par rapport à $\theta(1-\theta)/n$ (5 pts) ; mise à l'échelle appropriée selon n (2 pts) ; recommandation argumentée et importance pratique discutée (3 pts).	10
Sous-total exercice 1		**45**
Exercice 2 – Données BSI		
a) Modèle additif mixte	Modèle conforme (lissage du temps, effet fixe, effet aléatoire `re`, ML) (3 pts) ; graphique avec bande de confiance (2 pts) ; description juste de la forme et interprétation des degrés de liberté effectifs (3 pts).	8
b) Effets aléatoires	Modèles comparés conformes (mêmes effets fixes) (3 pts) ; hypothèses et nombre de paramètres ajoutés (2 pts) ; loi de référence en mélange de khi-deux correctement justifiée et valeur p calculée (5 pts) ; conclusion (2 pts).	12
c) Effets fixes	(i) test conjoint de l'interaction, d.l. corrects (5 pts) ; (ii) test de l'effet global de `droguedure` dans le modèle approprié (4 pts) ; (iii) comparaisons deux à deux ajustées, rapports de cotes et IC à 95 % interprétés (4 pts) ; discussion d'éventuels désaccords RV/Wald ou problèmes de convergence (2 pts).	15
d) Forme matricielle	Dimensions de $\boldsymbol{Y}_i$, $\boldsymbol{X}_i$, $\boldsymbol{Z}_i$ (3 pts) ; $\boldsymbol{\beta}$, $\boldsymbol{b}_i$ et lignes types écrits explicitement et cohérents avec le modèle retenu (4 pts) ; loi de $\boldsymbol{b}_i$ et liste complète des paramètres avec leur nombre (3 pts).	10
Sous-total exercice 2		**45**
Qualité du rapport et du code		
Rapport	*Excellent (5)* : structuré, rédaction claire et précise, tableaux et figures numérotés et commentés, conclusions répondant directement aux questions. *Satisfaisant (3)* : complet mais sorties brutes ou commentaires sommaires. *Insuffisant (0–1)* : désorganisé, réponses difficiles à repérer.	5
Code R	*Excellent (5)* : s'exécute sans erreur du début à la fin, reproduit les résultats, commenté, germe fixé. *Satisfaisant (3)* : fonctionne avec corrections mineures ou peu commenté. *Insuffisant (0–1)* : absent, incomplet ou ne reproduit pas les résultats.	5

Critère	Attendus et descripteurs	Pts
Sous-total qualité		**10**
Total		**100**

Références

— Derogatis, L. R. et Melisaratos, N. (1983). The Brief Symptom Inventory : an introductory report. *Psychological Medicine*, 13(3), 595–605.

— Rotheram-Borus, M. J., Lee, M. et coll. (2004). Six-year intervention outcomes for adolescent children of parents with the human immunodeficiency virus. *Archives of Pediatrics & Adolescent Medicine*, 158(8), 742–748.

— Wood, S. N. (2017). *Generalized Additive Models : An Introduction with R*, 2^{e} éd. Chapman & Hall/CRC.

— Wu, L. (2010). *Mixed Effects Models for Complex Data*. Chapman & Hall/CRC, ch. 2.