Devoir 2

GEE avec données manquantes et tests multiples

Pondération	20 % de la note finale (corrigé sur 100 points)
Matière	Séances 6 à 11 : équations d'estimation généralisées, données manquantes (imputation multiple, pondération par probabilité inverse), tests multiples
Échéance	**vendredi 23 avril 2027, 23 h 55**
Mode	Travail individuel
Remise	Boîte de dépôt « Devoir 2 » sur Brio : un rapport PDF et le code R
Fichiers fournis	`cbppw.csv` ; `pleuropneumonie.R` (programme du modèle mixte de la séance 3) ; le jeu `golub` est inclus dans le module `multtest` (Bioconductor)

1 Contexte et objectifs

Le premier exercice reprend l'exemple de la pleuropneumonie contagieuse bovine vu à la séance 3, mais sous l'angle d'un modèle *marginal* estimé par équations d'estimation généralisées (GEE), en présence de données manquantes. Le second exercice applique les procédures de contrôle de la FWER et de la FDR à un problème classique de génomique.

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

1. expliquer pourquoi une analyse des cas complets peut être biaisée et choisir une méthode de traitement des données manquantes en énonçant ses postulats ;
2. estimer un modèle log-binomial marginal par GEE avec imputation multiple ou avec pondération par l'inverse de la probabilité d'observation, et en tirer des rapports d'incidence avec intervalles de confiance ;
3. comparer un modèle marginal et un modèle mixte (interprétation des coefficients et erreurs-types) ;
4. appliquer et comparer des procédures de contrôle de la FWER et de la FDR.

2 Exercice 1 – Pleuropneumonie contagieuse bovine : modèle marginal (50 points)

Contexte

Des échantillons sanguins ont été prélevés tous les trois mois (quatre périodes) dans 15 troupeaux de zébus en Éthiopie afin de dénombrer les nouveaux cas sérologiques de pleuropneumonie contagieuse bovine (jeu `cbpp` du module `lme4`). À la séance 3, un modèle log-binomial *mixte*, avec ordonnée à l'origine aléatoire par troupeau, a été ajusté pour estimer les rapports d'incidence entre périodes (programme `pleuropneumonie.R`, fourni). Vous estimerez ici ces mêmes rapports d'incidence, avec intervalles de confiance à 95 %, sous un modèle log-binomial *marginal* estimé par GEE, le troupeau constituant la grappe.

Certains troupeaux n'ont pas été mesurés à certaines périodes. On suppose que **seule l'incidence est manquante** : la taille du troupeau est connue à toutes les périodes. Le fichier `cbppw.csv` présente les données en format large (une ligne par troupeau).

Variable	Description
`herd`	identifiant du troupeau (1 à 15)
`incidence.1` à `incidence.4`	nombre de nouveaux cas aux périodes 1 à 4 (`NA` si non mesuré)
`size.1` à `size.4`	taille du troupeau aux périodes 1 à 4 (toujours observée)
`p.1` à `p.4`	proportion d'incidence `incidence.k/size.k` (`NA` si non mesurée)

Modèle. Pour le troupeau i à la période j, soit Y_{ij} le nombre de nouveaux cas parmi m_{ij} animaux. Le modèle marginal est

$$\mathrm{E}(Y_{ij}) = m_{ij}\,\mu_{ij}, \qquad \log \mu_{ij} = \beta_0 + \beta_2\,\mathbf{1}(j=2) + \beta_3\,\mathbf{1}(j=3) + \beta_4\,\mathbf{1}(j=4),$$

de sorte que $\exp(\beta_j)$ est le rapport d'incidence de la période j par rapport à la période 1 (codage de traitement par défaut de R, comme à la séance 3). Si vous le souhaitez, présentez aussi les rapports entre périodes successives, $\exp(\beta_j - \beta_{j-1})$, avec leur intervalle de confiance.

a) **Biais de l'analyse naïve. (8 points)** Décrivez le patron des données manquantes (quels troupeaux, quelles périodes ; le patron est-il monotone ?). Expliquez pourquoi une analyse naïve portant uniquement sur les données observées pourrait produire des estimations biaisées des rapports d'incidence. Sous quel mécanisme l'analyse naïve par GEE resterait-elle sans biais ?

b) **Choix d'une méthode. (12 points)** Choisissez une méthode de traitement des données manquantes – *imputation multiple* ou *pondération par l'inverse de la probabilité d'observation* (IPCW) – et justifiez votre choix. Énoncez précisément les postulats requis pour que l'analyse soit valide : mécanisme de non-réponse (MCAR, MAR ou MNAR, et conditionnellement à quoi), modèle d'imputation ou modèle d'observation, et, pour l'IPCW, structure de corrélation de travail compatible avec la pondération. Discutez la plausibilité de ces postulats dans ce contexte.

c) **Estimation par GEE avec traitement des données manquantes. (18 points)** Ajustez le modèle log-binomial marginal par GEE (p. ex. `geepack::geeglm` avec `family = binomial(link = "log")` et `id = herd`) en appliquant la méthode choisie en b). Présentez dans un tableau les rapports d'incidence estimés et leurs intervalles de confiance à 95 %

(fondés sur l'erreur-type sandwich), ainsi que, à titre de comparaison, les résultats de l'analyse naïve des cas observés. Décrivez les étapes de façon que l'analyse soit reproductible (modèles auxiliaires, nombre d'imputations ou forme des poids, règle de combinaison).

d) **Comparaison avec le modèle mixte. (12 points)** Exécutez le modèle log-binomial mixte avec ordonnée à l'origine aléatoire par troupeau du programme `pleuropneumonie.R` (objet `gmlog1`). Comparez dans un tableau les estimations des coefficients et leurs erreurs-types entre ce modèle et votre modèle marginal de c). Que constatez-vous ? Interprétez les ressemblances et les différences en vous appuyant sur la relation entre paramètres marginaux et conditionnels sous le lien logarithmique, sur la façon dont chaque approche traite les données manquantes et sur la qualité de l'erreur-type sandwich avec seulement 15 grappes.

Indications pour l'exercice 1

- **Imputation multiple.** Le module `mice` ne propose pas de modèle binomial pour une proportion : modélisez `p.2` à `p.4` comme des variables continues (p. ex. `method = "norm"`), puis reconvertissez en nombres de cas entiers compris entre 0 et la taille du troupeau (`pmin(pmax(round(p * size), 0), size)`). Utilisez au moins $m = 20$ imputations, fixez le germe et combinez les estimations avec les règles de Rubin (`pool`).
- **IPCW.** Il n'y a pas de nouvelle sortie d'étude à la période 3 (un troupeau observé à la période 2 l'est aussi à la période 3). La probabilité d'être observé à la période 4 ne peut donc être modélisée que pour les troupeaux mesurés aux périodes 2 et 3. Le poids d'une observation est l'inverse du produit des probabilités d'observation jusqu'à cette période.
- Les données en format long nécessaires à `geeglm` s'obtiennent avec `tidyr::pivot_longer` ; triez-les par troupeau puis par période.
- Utilisez `scale.fix = TRUE` pour fixer le paramètre de dispersion à 1, comme dans le modèle binomial.

3 Exercice 2 – Tests multiples : leucémies ALL et AML (40 points)

Vous appliquerez des procédures de contrôle de la FWER et de la FDR à la comparaison du niveau d'expression des gènes entre la leucémie lymphoblastique aiguë (ALL) et la leucémie myéloïde aiguë (AML) dans le jeu de données `golub` du module `multtest` de Bioconductor (`data(golub)` : matrice `golub` de 3 051 gènes × 38 échantillons, vecteur `golub.cl` valant 0 pour les 27 cas ALL et 1 pour les 11 cas AML). Lisez au préalable le chapitre 3 de Farcomeni (2010), en particulier le tableau 3.5 et la figure 3.5.

1. **Tests individuels. (8 points)** Pour chacun des 3 051 gènes, testez l'hypothèse nulle d'égalité des moyennes de log-expression entre les 27 cas ALL et les 11 cas AML, contre l'alternative bilatérale, à l'aide de la statistique t à variances inégales (statistique de Welch). Calculez une valeur p brute soit par la loi de Student (degrés de liberté de Welch–Satterthwaite), soit par permutation ou bootstrap ; dans ce dernier cas, précisez le nombre de rééchantillonnages. Présentez l'histogramme des valeurs p brutes et commentez-le.

2. **Procédures d'ajustement. (12 points)** Appliquez au vecteur de valeurs p brutes, à l'aide de `mt.rawp2adjp`, les procédures de contrôle de la FWER et de la FDR qui figurent au

tableau 3.5 de Farcomeni (2010). Dans `mt.rawp2adjp`, les procédures pertinentes portent notamment les noms `"Bonferroni"`, `"Holm"`, `"Hochberg"`, `"SidakSS"`, `"SidakSD"`, `"BH"` et `"BY"` ; pour la méthode « plug-in », utilisez l'approche en deux étapes de Benjamini, Krieger et Yekutieli (2006), `"TSBH"`. Pour chaque procédure, indiquez en une phrase le critère contrôlé (FWER ou FDR), le type (une étape, descendante, ascendante) et les conditions de dépendance sous lesquelles le contrôle est garanti. Si une procédure du tableau n'est pas offerte par `mt.rawp2adjp`, signalez-le.

3. **Graphique. (8 points)** À l'aide de `mt.plot` (graphique de type `"rvsa"`), produisez un graphique du nombre d'hypothèses rejetées en fonction du seuil d'erreur de première espèce, pour les valeurs p brutes et pour les valeurs p ajustées de chaque procédure du point 2. Limitez l'axe horizontal à une plage informative (p. ex. de 0 à 0,2) et ajoutez une légende lisible. La courbe des valeurs p brutes devrait ressembler à celle de la figure 3.5 de Farcomeni (2010) ; commentez l'ordre des courbes.

4. **Tableau et comparaison. (12 points)** Présentez dans un tableau le nombre de gènes déclarés différentiellement exprimés par chaque procédure au niveau 5 % (valeur p ajustée $\leq 0{,}05$; les procédures FWER contrôlent aussi la FDR à ce niveau). Comparez vos résultats à ceux du tableau 3.5 de Farcomeni (2010) et expliquez les différences : vous utilisez les 38 cas, alors que Farcomeni utilisait des sous-échantillons de 9 ALL et 6 AML. Discutez enfin quelle famille de procédures est la plus appropriée pour une étape de criblage de gènes.

4 Livrables

1. **Rapport PDF** nommé `STT7360_D2_Nom_Prenom.pdf`, en français, d'au plus **12 pages** (annexes de code non comprises), qui suit l'ordre des questions. Tableaux et figures numérotés, titrés et commentés ; pas de sorties R brutes dans le corps du texte.

2. **Code R** nommé `STT7360_D2_Nom_Prenom.R` (ou `.Rmd`/`.qmd`), commenté, qui reproduit tous les résultats du rapport lorsqu'il est exécuté du début à la fin avec `cbppw.csv` dans le répertoire de travail. Indiquez les modules utilisés et les germes aléatoires (imputation, permutations).

3. Dépôt des deux fichiers dans la boîte de dépôt « Devoir 2 » de Brio au plus tard le **vendredi 23 avril 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. Le rapport et le code doivent être personnels.

5 Grille de correction (100 points)

Critère	Attendus et descripteurs	Pts
Exercice 1 – Pleuropneumonie		
a) Biais de l'analyse naïve	Patron des données manquantes décrit correctement (monotone, troupeaux et périodes en cause) (3 pts) ; mécanisme de biais expliqué (la non-réponse peut dépendre de l'incidence antérieure ; GEE des cas observés valides sous MCAR seulement) (5 pts).	8
b) Choix de la méthode	Méthode choisie et justifiée par rapport à l'autre (4 pts) ; postulat MAR énoncé avec les variables de conditionnement (4 pts) ; modèle d'imputation ou d'observation explicite et, pour l'IPCW, corrélation de travail indépendante justifiée (2 pts) ; discussion de la plausibilité (2 pts).	12
c) GEE avec traitement	Implantation correcte : modèles auxiliaires, poids ou imputations, reconversion en comptes, format long (8 pts) ; combinaison correcte (règles de Rubin) ou usage correct des poids et de l'erreur-type sandwich (4 pts) ; tableau des rapports d'incidence et IC à 95 %, comparé à l'analyse naïve (4 pts) ; interprétation (2 pts).	18
d) Comparaison au modèle mixte	Tableau comparatif des coefficients et erreurs-types (4 pts) ; interprétation marginal vs conditionnel sous le lien log (4 pts) ; discussion du traitement des données manquantes par chaque approche et de la fiabilité de l'erreur-type sandwich avec 15 grappes (4 pts).	12
Sous-total exercice 1		**50**
Exercice 2 – Tests multiples		
1. Tests individuels	Statistique de Welch correcte pour les 3 051 gènes (3 pts) ; valeurs p correctes et méthode précisée (3 pts) ; histogramme commenté (2 pts).	8
2. Procédures	Toutes les procédures pertinentes appliquées, dont TSBH pour le plug-in (6 pts) ; critère, type et conditions de dépendance correctement résumés (6 pts).	12
3. Graphique	Graphique `mt.plot` complet et lisible (légende, plage) (5 pts) ; commentaire sur l'ordre des courbes et le lien avec la figure 3.5 (3 pts).	8
4. Tableau et comparaison	Tableau des nombres de gènes détectés à 5 % exact (5 pts) ; comparaison avec le tableau 3.5 et explication des différences (taille d'échantillon, puissance) (4 pts) ; choix argumenté d'une procédure pour le criblage (3 pts).	12
Sous-total exercice 2		**40**
Qualité du rapport et du code		
Rapport	*Excellent (5)* : structuré, rédaction claire et précise, tableaux et figures numérotés et commentés, réponses directes 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, germes fixés. *Satisfaisant (3)* : fonctionne avec corrections mineures ou peu commenté. *Insuffisant (0–1)* : absent, incomplet ou ne reproduit pas les résultats.	5
Sous-total qualité		**10**
Total		**100**

Références

— Benjamini, Y., Krieger, A. M. et Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93(3), 491–507.

— Farcomeni, A. (2010). Multiple testing procedures. Ch. 3 de *Medical Biostatistics for Complex Diseases* (F. Emmert-Streib et M. Dehmer, dir.). Wiley.

— Golub, T. R. et coll. (1999). Molecular classification of cancer : class discovery and class prediction by gene expression monitoring. *Science*, 286, 531–537.

— Lesnoff, M. et coll. (2004). Within-herd spread of contagious bovine pleuropneumonia in Ethiopian highlands. *Preventive Veterinary Medicine*, 64, 27–40.

— Wu, L. (2010). *Mixed Effects Models for Complex Data*. Chapman & Hall/CRC, ch. 3 et § 9.2.