Plan de cours

STT-7360 Méthodes biostatistiques

Session	Hiver 2027 (11 janvier au 23 avril 2027)
Horaire	Vendredi, 8 h 30 à 11 h 20
Crédits	3
Enseignant	Julien Miron, chargé d'enseignement, Département de mathématiques et de statistique
Courriel	*à compléter*
Disponibilités	*à compléter* (sur rendez-vous)
Site du cours	Brio (version officielle)
Site web	`https://julienmiron.ca/STT-7360/`

Description du cours

Ce cours présente des méthodes statistiques avancées pour l'analyse de données biomédicales complexes : données corrélées (longitudinales ou groupées), modèles comportant des paramètres de nuisance de grande dimension, données manquantes ou incomplètes par devis, et problèmes de tests multiples. L'accent porte sur la compréhension des méthodes d'estimation et d'inférence, sur leur justification théorique et sur leur mise en œuvre en R. Le cours est formateur pour la recherche : il apprend à lire un article méthodologique, à en extraire les choix statistiques et à reproduire ses analyses.

Le cours s'adresse aux étudiantes et étudiants de maîtrise et de doctorat en statistique et en biostatistique. Un contenu supplémentaire, signalé dans les notes par un encadré « Niveau doctorat », porte sur la théorie de l'estimation semi-paramétrique ; il fait l'objet d'exercices et d'une question d'examen réservés au 3^{e} cycle.

Fil conducteur : la grille méthodologique

Chaque méthode vue en classe et chaque article étudié sont analysés selon la même grille, qui sert aussi de structure aux études de cas, au travail de synthèse et à l'examen final :

1. l'objet statistique d'intérêt ;
2. les paramètres nuisibles, les données manquantes et les sources de biais ;
3. l'approche retenue pour traiter la nuisance ou éviter le biais ;

4. le critère d'estimation ou de test ;
5. la méthode d'optimisation ou de calcul ;
6. l'obtention de la distribution des statistiques (inférence).

Objectifs d'apprentissage

À la fin du cours, l'étudiante ou l'étudiant sera en mesure de :

1. utiliser les outils asymptotiques (convergence, théorème de Slutsky, méthode delta, estimateurs M) pour obtenir la loi d'un estimateur ;
2. concevoir une étude de simulation qui valide un estimateur, l'estimateur de sa variance, un intervalle de confiance ou un test ;
3. formuler, estimer et interpréter un modèle linéaire mixte généralisé ou un modèle additif généralisé (mixte), et choisir la méthode d'approximation de la vraisemblance ;
4. distinguer les effets individuels des effets marginaux et choisir entre un modèle mixte et des équations d'estimation généralisées ;
5. dériver une fonction d'influence, en tirer un estimateur robuste de la variance et construire un estimateur robuste aux valeurs extrêmes ;
6. (doctorat) situer les équations d'estimation généralisées et l'estimation ciblée dans la théorie semi-paramétrique (espace tangent, score efficace, double robustesse) ;
7. reconnaître un mécanisme de données manquantes et appliquer la vraisemblance des données observées, l'imputation multiple ou la pondération par probabilité inverse ;
8. contrôler l'erreur de première espèce d'une famille de tests (FWER, FDR) ;
9. présenter un article méthodologique et implanter la méthode qu'il décrit.

Préalables

Inférence statistique de niveau avancé (maximum de vraisemblance, tests de Wald, du rapport de vraisemblances et du score), modèles linéaires généralisés, notions de modèle linéaire mixte et programmation en R. La première séance fait un rappel des outils asymptotiques.

Formule pédagogique

Chaque séance de trois heures combine un exposé théorique appuyé sur des notes de cours détaillées (déposées sur le site avant la séance) et des exemples traités en R. Les séances d'étude de cas portent sur un article : les étudiants le lisent à l'avance, remplissent la fiche des caractéristiques méthodologiques et préparent les exercices proposés, qui sont corrigés en classe. Les études de cas

ne sont pas évaluées, mais leur préparation est essentielle à la réussite du travail de synthèse et de l'examen final.

Calendrier

Séance	Date	Contenu	Lectures et échéances
1	15 janv.	Rappels d'inférence asymptotique : convergence, Slutsky, méthode delta ; maximum de vraisemblance ; tests de Wald, du score et du rapport de vraisemblances ; estimateurs M	Notes, section 1 ; Wu, § 9.2.2
2	22 janv.	Études de simulation : estimer une caractéristique d'une loi ; valider un estimateur, sa variance et un intervalle de confiance ; niveau et puissance d'un test ; génération de données	Notes, section 2
3	29 janv.	Modèles linéaires mixtes (BLUP, REML) et modèles linéaires mixtes généralisés : vraisemblance, quadrature, approximation de Laplace, PQL ; effets individuels et marginaux ; sélection des effets aléatoires	Notes, section 3 ; Wu, ch. 2
4	5 févr.	Modèles additifs généralisés et modèles additifs généralisés mixtes ; lissage comme composante de variance	Notes, section 4 ; Wu, § 2.5 ; Wood (2017)
5	12 févr.	**Étude de cas 1** : effets intra et inter-grappes, régression logistique conditionnelle	Neuhaus et McCulloch (2006)
6	19 févr.	Équations d'estimation généralisées ; fonction d'influence et variance sandwich ; robustesse. Doctorat : théorie semi-paramétrique	Notes, section 6 ; Wu, § 9.2.2 et p. 343 ; Tsiatis (2006), ch. 4 *Choix de l'article du travail de synthèse*
7	26 févr.	**Étude de cas 2** : estimation ciblée (TMLE) de l'effet d'un biomarqueur	Tuglus et van der Laan (2011) *Devoir 1 à remettre (23 h 55)*
–	5 mars	*Semaine de lecture*	
8	12 mars	Données manquantes I : mécanismes MCAR, MAR, MNAR ; vraisemblance des données observées ; imputation multiple	Notes, section 8 ; Wu, ch. 3
9	19 mars	Données manquantes II : pondération par probabilité inverse ; équations d'estimation généralisées pondérées	Notes, section 9 ; Wu, ch. 3
–	26 mars	*Vendredi saint (congé)*	
10	2 avril	**Étude de cas 3** : échantillonnage à deux phases et erreur de mesure	Breslow et Chatterjee (1999)
11	9 avril	Tests multiples : FWER, FDR, procédures marginales et par rééchantillonnage	Notes, section 11 ; Farcomeni (2010) ; Dudoit et van der Laan (2008)
12	16 avril	Présentations des travaux de synthèse (1)	
13	23 avril	Présentations des travaux de synthèse (2)	*Devoir 2 à remettre (23 h 55)*
–	26 avril au 2 mai	**Examen final** (date et local à confirmer)	

Évaluation

Évaluation	Pondération	Mode	Échéance
Devoir 1	20 %	Individuel	vendredi 26 février 2027, 23 h 55
Devoir 2	20 %	Individuel	vendredi 23 avril 2027, 23 h 55
Travail de synthèse et présentation	30 %	Individuel	16 ou 23 avril 2027
Examen final	30 %	Individuel	période d'examens

Devoirs. Les devoirs combinent une partie théorique et une analyse de données en R. Le devoir 1 porte sur les séances 1 à 4 (méthode delta, simulation, modèle logistique mixte et lissage) ; le devoir 2 porte sur les séances 6 à 11 (équations d'estimation généralisées avec données manquantes et tests multiples). Les consignes détaillées et les grilles de correction sont déposées sur le site du cours. Le code R est remis avec le rapport.

Travail de synthèse. Chaque étudiante ou étudiant présente oralement un article méthodologique et remet le code documenté de son implantation ou de son application de la méthode biostatistique décrite dans l'article. L'utilisation d'une fiche de lecture et la préparation d'exercices en lien avec l'article sont encouragées. Une liste d'articles suggérés est fournie ; un autre article peut être choisi avec l'accord de l'enseignant. Le choix est transmis au plus tard le 19 février 2027. Le code est déposé au plus tard la veille de la présentation à 23 h 55. Consignes et grille détaillées sur le site.

Examen final. Examen de résolution de problèmes de 2 h 50, sans ordinateur. Les questions demandent de formuler un modèle, d'identifier l'objet d'intérêt et les paramètres de nuisance, de choisir et de justifier une méthode d'estimation, de traiter des données manquantes, de contrôler l'erreur de première espèce sur plusieurs tests ou de dériver une fonction d'influence. Le barème diffère entre la maîtrise et le doctorat, et une question est réservée au doctorat. Des questions de pratique et des examens antérieurs avec solutionnaires sont disponibles.

Barème de conversion. Selon le barème en vigueur au Département pour les cours de 2^{e} cycle (*à compléter*).

Matériel

Manuel principal. Wu, L. (2010). *Mixed Effects Models for Complex Data*. Chapman & Hall/CRC.

Notes de cours. Notes détaillées pour chaque section, déposées sur le site du cours.

Ouvrages et articles complémentaires.

— Breslow, N. E. et Chatterjee, N. (1999). Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis. *Applied Statistics*, 48(4), 457–468.

— Dudoit, S. et van der Laan, M. J. (2008). *Multiple Testing Procedures with Applications to Genomics*. Springer.

— Farcomeni, A. (2010). Multiple testing procedures. Ch. 3 de *Medical Biostatistics for Complex Diseases*. Wiley.

— Fitzmaurice, G., Laird, N. et Ware, J. (2011). *Applied Longitudinal Analysis*, 2^{e} éd. Wiley.

— Jiang, J. et Nguyen, T. (2021). *Linear and Generalized Linear Mixed Models and Their Applications*, 2e éd. Springer.

— 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 B*, 68(5), 859–872.

— Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. Springer.

— van der Laan, M. J. et Rose, S. (2011). *Targeted Learning*. Springer (ch. 5 et ch. 22, Tuglus et van der Laan).

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

Logiciel. R, avec notamment les modules `lme4`, `mgcv`, `gee`, `geepack`, `mice`, `survival`, `osDesign` et `multtest` (Bioconductor).

Politiques

Retards. Un travail remis en retard sans motif valable reçoit une pénalité de 10 % par jour ouvrable de retard, jusqu'à concurrence de cinq jours ; au-delà, il reçoit la note zéro (*à ajuster selon la politique de l'enseignant*).

Plagiat et intelligence artificielle. Les règles de l'Université Laval en matière d'intégrité intellectuelle s'appliquent (Règlement disciplinaire à l'intention des étudiants). Le code et les rapports remis doivent être votre propre travail. L'usage d'outils d'intelligence artificielle générative est permis pour l'aide à la programmation, à condition d'être déclaré dans le travail remis (*à ajuster*).

Accommodements. Les étudiantes et étudiants ayant une lettre d'accommodements du Centre d'aide aux étudiants doivent communiquer avec l'enseignant au début de la session.