
Translations of Abstracts

BIOMETRIC METHODOLOGY**Ana-Maria Staicu, Ciprian M. Crainiceanu, Daniel S. Reich, and David Ruppert***Modeling Functional Data with Spatially Heterogeneous Shape*

331

Characteristics

Nous proposons une nouvelle classe de modèles pour les données fonctionnelles présentant une dissymétrie ou autre caractéristique de forme qui varie avec la localisation spatiale ou temporelle. Nous utilisons les copules de façon à ce que les distributions marginales et la structure de dépendance soient modélisées indépendamment. La dépendance est modélisée avec une Gaussienne ou t-copule, de manière à ce qu'il y ait un processus latent Gaussien sous-jacent. Nous modélisons les distributions marginales en utilisant la famille asymétrique t. La moyenne, variance et les paramètres de forme sont modélisés non-paramétriquement comme des fonctions de localisation. Un cadre d'inférence pour estimer les distributions marginales à queue lourde ou asymétrique hétérogène soluble de manière informatique est introduit. Le cadre fournit un nouveau jeu d'outils pour complexifier de façon croissante les données recueillies dans les études médicales et de santé publique. Nos méthodes sont motivées par et sont illustrées avec une étude d'état de l'art des voies neuronales chez des patients souffrant de sclérose en plaques et des contrôles sains. Utilisant les outils que nous avons développés, nous avons été capables de trouver les localisations le long de la voie les plus atteintes par la maladie. Cependant, nos méthodes sont générales et hautement pertinentes pour de nombreux jeux de données fonctionnelles. En plus de cette application à une voie unidimensionnelle illustrée ici, des extensions de la méthodologie à des dimensions plus hautes pourraient avoir des applications directes à d'autres données biologiques incluant les imageries par résonance magnétique fonctionnelles et structurelles.

Hulin Wu, Hongqi Xue, and Arun Kumar*Numerical Discretization-Based Estimation Methods for Ordinary Differential*

344

Equation Models via Penalized Spline Smoothing with Applications in Biomedical Research

Les équations différentielles sont largement utilisées pour modéliser la dynamique des processus physiques dans de nombreux domaines scientifiques comme l'ingénierie, la physique et les sciences biomédicales. L'estimation des paramètres des modèles d'équations différentielles est un problème difficile en raison de coût de calcul élevé et de la grande dimension de l'espace des paramètres. Dans ce papier, nous proposons une nouvelle classe de méthodes d'estimation des paramètres de modèles d'équation différentielle ordinaire (EDO), motivée par la modélisation de la dynamique du VIH. Les nouvelles méthodes exploitent la forme d'algorithmes de discrétisation numérique pour

un solveur d'EDO pour formuler des équations d'estimation. Une approche par spline pénalisée est d'abord utilisée pour estimer les variables d'état, ensuite les variables d'état estimées sont introduites dans une formule de discrétisation d'un solveur d'EDO pour obtenir les estimations des paramètres d'EDO par une approche de régression. Nous considérons trois méthodes ordonnées de discrétisation : Euler, trapèzes et Runge-Kutta. Un algorithme numérique d'ordre plus élevé réduit les erreurs numériques dans l'approximation de la dérivée, ce qui produit une estimation plus précise, mais avec un coût de calcul plus élevé. Pour trouver un compromis entre le coût de calcul et la précision de l'estimation, nous montrons par des études de simulation que l'estimation basée par la méthode des trapèzes est la meilleure et est recommandée pour une utilisation pratique. Les propriétés asymptotiques des estimateurs basés sur la méthode de discrétisation proposée sont établies. Les comparaisons entre les méthodes proposées et les méthodes existantes montrent un net avantage des méthodes proposées en ce qui concerne l'arbitrage entre coût de calcul et précision de l'estimation. Nous appliquons les méthodes proposées à une étude sur le VIH pour mieux illustrer l'utilité des approches proposées.

Lan Wang, Jianhui Zhou, and Annie Qu

Penalized Generalized Estimating Equations for High-Dimensional Longitudinal Data Analysis 353

Nous considérons les équations d'estimation généralisées pénalisées pour analyser des données longitudinales avec des covariables de grande dimension qui surviennent lors d'expérimentations par micro-puces et dans les études de santé à grande échelle. Les procédures existantes de régression de grande dimension supposent souvent l'indépendance des données et reposent sur la fonction de vraisemblance. La construction d'une fonction de vraisemblance faisable pour les données longitudinales de grande dimension est un défi, particulièrement pour des réponses discrètes corrélées. La procédure par équations d'estimation généralisées pénalisées requiert seulement de spécifier les deux premiers moments marginaux et une structure de corrélation de travail. Nous établissons la théorie asymptotique dans un cadre de grande dimension où le nombre de covariables p_n augmente lorsque le nombre de grappes n augmente et lorsque p_n peut atteindre le même ordre que n . Une caractéristique importante de la nouvelle procédure est que la consistance de la sélection du modèle est préservée même si la structure de corrélation de travail est mal spécifiée. Nous évaluons la performance de la méthode proposée en utilisant des simulations de Monte Carlo et démontrons son application en utilisant un jeu de données sur l'expression du cycle cellulaire d'une levure.

Dhiman Bhadra, Michael J. Daniels, Sungduk Kim, Malay Ghosh, and Bhramar Mukherjee

A Bayesian Semiparametric Approach for Incorporating Longitudinal Information on Exposure History for Inference in Case-Control Studies 361

Typiquement dans une étude cas-contrôle l'information sur l'exposition est obtenue à un moment unique pour les cas et les contrôles. Cependant les études cas-contrôle sont souvent incluses dans des études de cohortes existantes qui informent énormément sur l'histoire de l'exposition longitudinale des participants. Des études médicales récentes

ont montré que l'incorporation de l'histoire de l'exposition passée, ou d'une mesure construite de l'exposition cumulée dérivée de l'histoire de l'exposition passée, lorsqu'elle est disponible, peut produire des estimations du risque de maladie plus précises et cliniquement significatives. Dans cet article, nous proposons une approche Bayésienne semi-paramétrique souple pour modéliser les profils d'exposition longitudinaux des cas et des contrôles et ensuite utiliser des mesures d'exposition cumulée basées sur une intégrale pondérée de cette trajectoire dans le modèle final de risque de maladie. L'estimation est faite via une vraisemblance jointe. Dans la construction du résumé de l'exposition cumulée, nous introduisons une fonction d'influence, une fonction lisse du temps pour caractériser le schéma d'association du profil d'exposition et de l'état de la maladie avec différentes fenêtres temporelles ayant potentiellement des poids/influence différents. Cela nous permet d'analyser comment l'état actuel de la maladie d'un sujet est influencée par l'histoire de son exposition passée conditionnellement à l'exposition présente. La formulation en termes de vraisemblance jointe nous permet de prendre correctement en compte les incertitudes associées aux deux étapes du processus d'estimation de façon intégrée. L'analyse est conduite dans un cadre Bayésien hiérarchique utilisant des algorithmes de Monte Carlo avec chaîne de Markov à sauts réversibles (RJMCMC). La méthodologie proposée est motivée par, et appliquée à une étude cas-contrôle dans le cancer de la prostate, où une information longitudinale sur des biomarqueurs est disponible pour les cas et les contrôles.

G. Jones, W. O. Johnson, W. D. Vink, and N. French

A Framework for the Joint Modeling of Longitudinal Diagnostic Outcome Data and Latent Infection Status: Application to Investigating the Temporal Relationship between Infection and Disease 371

Pour beaucoup de maladies le statut infectieux des individus ne peut être observé directement, mais peut seulement être inféré à partir de biomarqueurs qui sont sujets à des erreurs de mesure. Le diagnostic d'infection, basé sur les symptômes observés peut lui-même être regardé comme un test imparfait du statut infectieux. La relation temporelle entre l'infection et le résultat des marqueurs peut être complexe, spécialement pour les maladies récurrentes où les individus peuvent subir plusieurs accès d'infection. Nous proposons une approche qui, d'abord modélise le statut infectieux longitudinal non observé d'individus, conditionnellement à des covariates pertinentes, et ensuite modélise la séquence longitudinale des résultats des biomarqueurs conditionnellement au statut infectieux et à l'information des covariates au cours du temps, de sorte qu'il en résulte un modèle joint pour l'infection longitudinale et la séquence de biomarqueurs. Ce modèle ne peut être utilisé pour examiner la dynamique temporelle de l'infection, et pour évaluer l'utilité des biomarqueurs dans un but de surveillance continue. Notre travail est motivé et illustré par une étude longitudinale de Dermatite Bovine Digitale (BDD) dans des fermes laitières au Nord Ouest de l'Angleterre et au Nord du Pays de Galles, dans lesquelles l'infection d'intérêt est le *Treponème spp.*, et les biomarqueurs sont des résultats de tests ELISA et une variable dichotomique, la présence de lésion aux pattes. BDD est connu comme une cause possible d'infection aux pattes des vaches.

Daniel Commenges, Benoit Liquet, and Cécile Proust-Lima

Choice of Prognostic Estimators in Joint Models by Estimating Differences of Expected Conditional Kullback–Leibler Risks 380

Les estimateurs pronostiques d'un événement clinique peuvent utiliser des mesures répétées de marqueurs en plus de covariables fixes. Ces mesures peuvent être liées à l'événement clinique par des modèles conjoints basés sur une structure latente. Lorsque l'objectif est de choisir entre des estimateurs pronostiques différents basés sur des modèles conjoints, le critère d'information d'Akaike (AIC) n'est pas bien adapté et le choix doit être fondé sur la précision de la prédiction. Nous définissons une fonction de risque adaptée, l'espérance de l'entropie croisée pronostique (EPCE). Nous définissons une autre fonction de risque pour le cas d'observations censurées à droite, l'espérance de l'entropie croisée pronostique observée (EPOCE). Ces risques peuvent être estimés par validation croisée « leave-one-out », pour laquelle nous donnons une formule approximée ainsi que la distributions asymptotique. L'estimateur basé sur cette formule de validation croisée approximée (CVPOLa) est étudié par simulation et appliqué à la comparaison de plusieurs modèles conjoints à classes latentes pour le pronostic de récurrence du cancer de la prostate, en utilisant des mesures de l'antigène prostatique spécifique (PSA).

Yingye Zheng, Tianxi Cai, Yuying Jin, and Ziding Feng

Evaluating Prognostic Accuracy of Biomarkers under Competing Risk

388

L'identification de marqueurs prédisant la survenue d'événements cliniques est un objectif de recherche important pour le développement de stratégies d'intervention ciblée. La caractérisation de la performance clinique d'un marqueur pour prédire différents types d'événements constitue une étape essentielle dans cette entreprise. Nous présentons des méthodes statistiques pour évaluer la performance d'un marqueur pronostique vis-à-vis de plusieurs événements compétitifs. Pour prendre en compte la variation potentielle de la performance prédictive en fonction du temps ainsi que l'existence de risques compétitifs, nous stratifions les cas par type d'événement et nous définissons des mesures de la valeur pronostique qui dépendent du temps et de l'événement. Cette approche permet d'évaluer la qualité pronostique d'un marqueur pour chaque type d'événement et d'effectuer des comparaisons inter-événements. En partant des estimateurs non paramétriques des courbes ROC spécifiques de chaque cause, proposés par Saha et Heagerty, nous développons des procédures d'inférence sur plusieurs mesures de prédictivité spécifiques de chaque type d'événement. Afin d'estimer ces valeurs pronostiques et d'apprécier dans quelle mesure des covariables peuvent les affecter en présence de risques compétitifs, nous considérons deux formes de modèles semi-paramétriques pour les risques spécifiques. Ces approches permettent une modélisation souple des relations entre le marqueur et les différents types d'événements, prenant en compte, si nécessaire, l'effet de covariables supplémentaires. Nous étudions les propriétés asymptotiques des estimateurs proposés et nous illustrons, à l'aide de simulations, leurs performances à distance finie. A titre d'exemple, les procédures proposées sont appliquées aux données d'une étude pronostique concernant le cancer de la prostate.

Pamela A. Shaw and Ross L. Prentice

Hazard Ratio Estimation for Biomarker-Calibrated Dietary Exposures

397

L'incertitude liée aux erreurs de mesure dans des enquêtes où les personnes interrogées reportent elles-mêmes leur régime alimentaire a de fortes implications sur la fiabilité des

rapports en épidémiologie nutritionnelle. Les biomarqueurs permettant de mesurer certains nutriments dans les urines peuvent fournir une mesure objective de la consommation de court terme de certains nutriments, et peuvent aussi servir à calibrer les consommations reportées par les individus interrogés lorsqu'ils sont utilisés dans le cadre d'une enquête de type cohorte. Un modèle non standard d'erreur de mesure qui prend en compte une erreur systématique et une erreur spécifique à l'individu en plus de l'habituelle erreur aléatoire indépendante est nécessaire dans ce cadre. Trois procédures d'estimation des paramètres du taux de hasard (modèle de Cox) sont adaptées à la structure d'erreurs complexe décrites ci dessus. Ces procédures sont : la calibration de régression par ensemble de risques, le score conditionnel et le score corrigé non paramétrique. Un estimateur de la fonction de hasard de base cumulé est également proposé. La performance de chaque méthode est évaluée dans une simulation comparative avant d'être appliquées à un exemple tiré du « Women's Health Initiative Dietary Modification Trial ».

Bin Dong and David E. Matthews

Empirical Likelihood for Cumulative Hazard Ratio Estimation with Covariate Adjustment 408

Dans les études médicales, il est souvent d'un intérêt scientifique d'évaluer l'effet d'un traitement par le rapport des risques cumulés, surtout lorsque ces risques ne sont pas proportionnels. Pour tenir compte de la non-proportionnalité dans le modèle de Cox, les investigateurs supposent souvent que l'effet du traitement a une certaine forme fonctionnelle. Toutefois, le fait de procéder ainsi peut créer un problème de mauvaise spécification du modèle, car il est généralement difficile de justifier le choix d'une forme paramétrique particulière pour modéliser l'effet du traitement. Dans cet article, nous utilisons la vraisemblance empirique (VE) pour obtenir un estimateur non paramétrique du rapport de risques cumulés avec prise en compte de covariables sous deux modèles de risques non proportionnels, dont un est un modèle stratifié, et un autre dans un cadre moins restrictif impliquant un ajustement spécifique à chaque groupe de sujets. Les propriétés asymptotiques de la statistique du rapport de vraisemblance empirique sont obtenues dans chacun des deux cas, et les propriétés en échantillons finis de l'estimateur par VE sont évaluées par simulation. Des bandes de confiance simultanées pour toutes les valeurs du rapport de risques cumulés dans un intervalle d'intérêt fixé à l'avance sont également développés. Les méthodes proposées sont illustrées en utilisant les données de deux études concernant la survie de patients atteints de lymphome non hodgkinien ou de cancer de l'ovaire.

Jun Yan and Jian Huang

Model Selection for Cox Models with Time-Varying Coefficients 419

Le modèle de Cox avec des coefficients dépendant du temps offre une très grande flexibilité pour décrire la dynamique de l'effet dans le temps d'une covariable en présence de données censurées à droite. Comme toutes les covariables ne présentent pas un tel effet dépendant du temps, la procédure de sélection des variables représente un nouveau challenge, qui consiste à distinguer les variables qui présentent un effet dépendant du temps de celles pour qui l'effet est indépendant du temps. Nous proposons une méthode adaptative par lasso groupé qui permet non seulement de sélectionner les

covariables importantes mais également de les spécifier comme ayant un effet dépendant ou indépendant du temps dans le modèle. L'effet de chaque covariable est ainsi réparti entre un effet indépendant du temps et un effet dépendant du temps, ce dernier étant caractérisé par un ensemble de coefficients sur la base de splines sans intercept. La procédure de sélection et d'estimation est réalisée à l'aide d'un algorithme de recherche groupée, rapide et itératif. Nous montrons que cette approche présente de bonnes propriétés dans une étude par simulation qui reproduit des situations réalistes incluant jusqu'à vingt covariables. Un exemple réel montre l'utilité de la méthode.

Zhangsheng Yu, Xihong Lin, and Wanzhu Tu

Semiparametric Frailty Models for Clustered Failure Time Data

429

Nous considérons des modèles de fragilité pour des données de durée classées, avec des effets de covariables semiparamétriques, additifs. Nous proposons une procédure de vraisemblance partielle doublement pénalisée (DPPL) pour estimer les fonctions nonparamétriques, en utilisant des splines de lissage. Nous montrons que les estimateurs DPPL peuvent être obtenus en ajustant un modèle augmenté de fragilité avec des effets de covariables paramétriques, dans lequel les fonctions nonparamétriques sont estimées comme combinaisons linéaires d'effets fixes et aléatoires et les paramètres de lissage sont estimés comme des composantes additionnelles de variance. Cette approche nous permet d'estimer tous les composants du modèle au sein d'un même contexte unifié de modèle de fragilité. Nous évaluons les performances de la méthode proposée pour un échantillon de taille finie par une étude de simulation, et nous appliquons la méthode à l'analyse des données d'une étude sur les infections sexuellement transmissibles (STI).

Heng Lian, Xin Chen, and Jian-Yi Yang

Identification of Partially Linear Structure in Additive Models with an Application to Gene Expression Prediction from Sequences

437

Sous le nom de modèle additif, on trouve une classe semiparamétrique de modèles, devenue très populaire en raison de sa plus grande flexibilité que le modèle linéaire, et de sa capacité à ajuster des données de grandes dimensions lorsque l'usage des modèles totalement nonparamétriques devient impossible. Nous envisageons le problème de la sélection de variables simultanément à l'identification des composantes paramétriques à l'aide d'approximations splines avec deux pénalités SCAD. L'avantage de notre approche est que le choix entre modèles additifs, modèles linéaires partiellement additifs et modèles linéaires peut être automatique dans une simple étape d'estimation. Des études de simulation sont utilisées pour illustrer notre méthode, et nous présentons également son application à un motif de régression.

Olivier Thas, Lieven Clement, John C.W. Rayner, Beatriz Carvalho, and Wim Van Criekinge

An Omnibus Consistent Adaptive Percentile Modified Wilcoxon Rank Sum Test with Applications in Gene Expression Studies

446

Nous proposons une modification du test de Wilcoxon, le test de Wilcoxon adaptatif par percentile, pour la comparaison de deux échantillons. Notre test est en fait un test de Wilcoxon appliqué sur une fraction des observations de l'échantillon et cette fraction est

déterminée de façon séquentielle par les observations de l'échantillon. La plupart de la théorie est développée pour un modèle de décalage, mais notre test est aussi pertinent pour les tests d'hypothèses alternatives plus générales. Notre test est particulièrement adapté à l'analyse de jeux de données de grande dimension pour lesquels des tests d'hypothèses quasi automatiques sont nécessaires. Nous évaluons la puissance de ce nouveau test à l'aide d'une étude de simulation et nous l'appliquons à une expérience de puces sur le cancer colorectal. Ces études empiriques montrent que notre test à globalement une puissance satisfaisante et qu'il est plus performant que d'autres tests populaires pour identifier les gènes différentiellement exprimés. Enfin, nous concluons que notre nouveau test non paramétrique est utilisable dans un grand nombre de situations et que sa puissance est comparable à celle du test de Baumgartner-Weiß-Schindler.

John A. Dawson and Christina Kendziorski

An Empirical Bayesian Approach for Identifying Differential Coexpression in High-Throughput Experiments 455

Un but commun aux puces à ADN et aux expériences de génomique à haut-débit est d'identifier des gènes qui varient en fonction de certaines conditions biologiques. Le plus souvent, cette tâche est effectuée en identifiant les gènes dont l'expression varie en moyenne, gènes également appelés différentiellement exprimés (DE). Un certain nombre de méthodes dédiées à l'identification de gènes DE ont été développés. Bien qu'utiles, ces approches ne s'accommodent pas d'autres types de régulation différentielle. Un exemple important concerne la recherche de variables différentiellement co-exprimées (DC). La recherche de gènes appartenant à cette classe est rendue ardue par la grande dimension de l'espace à parcourir ainsi que par l'influence de valeurs aberrantes. Ainsi, les méthodes d'identification de gènes DC existantes sont souvent peu puissantes, beaucoup trop promptes à identifier des faux positifs, et/ou trop demandeuse de ressources de calcul, même pour des nombres de paires modérés. Pour pallier à ce problème, nous avons développé une méthode bayésienne d'identification des gènes DC. Notre approche fournit une liste de gènes significativement DC dont le taux de faux positifs (TFP) est contrôlé, tout en ne sacrifiant pas à la puissance. Elle est applicable à des études simples, de même qu'à des études multiples. Les calculs sont grandement facilités par une modification de l'algorithme EM et l'utilisation d'une procédure heuristique. Nos simulations suggèrent que l'approche proposée surpasse les méthodes existantes en termes de temps de calcul, et les résultats d'une étude de cas pratique suggèrent que cette approche sera très probablement considérée comme complémentaire des méthodes d'identification de gènes DE couramment utilisées lors d'études génomiques à haut-débit.

Brisa N. Sánchez, Shan Kang, and Bhramar Mukherjee

A Latent Variable Approach to Study Gene-Environment Interactions in the Presence of Multiple Correlated Exposures 466

Beaucoup d'études de cohorte existantes initialement prévues pour étudier le risque de maladie comme une fonction des expositions environnementales ont recueillies des données génomiques ces dernières années avec l'objectif de tester les effets de l'interaction gène-environnement (GxE). En épidémiologie environnementale, l'intérêt

dans GxE se pose surtout après qu'un effet significatif de l'environnement ait été documenté. Les études de cohorte recueillent souvent des données d'exposition riches, comme un résultat, le fait d'évaluer les effets GxE en présence de marqueurs d'exposition multiple augmentent de plus le fardeau des tests multiples, un problème déjà présent à la fois dans les études génétiques et les études de santé environnementale. Les modèles à variables latentes (VL) ont été utilisés en épidémiologie environnementale pour réduire la dimension des données d'exposition, gain de puissance par la simplification des problèmes de multiplicité via la réduction des données d'exposition, et évite les problèmes de colinéarité due à la présence d'expositions multiples corrélées. Nous étendons le cadre VL pour caractériser l'interaction gène-environnement en présence d'expositions multiples corrélées et de catégories de génotype. De plus, de façon similaire à ce qui a été fait dans les études cas-témoins GxE, nous utilisons l'hypothèse d'indépendance gène-environnement (G-E) pour améliorer la puissance des tests pour l'interaction. Les conséquences de cette hypothèse, ou la question de comment modéliser explicitement l'association G-E n'a pas été étudiée auparavant dans les modèles VL. Nous postulons une hiérarchie d'hypothèses sur le modèle VL concernant les différentes formes de dépendance G-E et montrons que faire de telles hypothèses peut influencer les résultats inférentiels sur les paramètres G, E, GxE. Nous implémentons une classe d'estimateurs rétrécis aux données adaptativement compromis entre les formes les plus restrictives et les plus flexibles de l'hypothèse de dépendance G-E et notons qu'une telle classe d'estimateurs compromis peut servir de point de référence pour modéliser l'adéquation au modèle dans les modèles à variables latentes. Nous illustrons les méthodes avec un exemple basé sur les premières expositions, gènes du métabolisme du fer et poids de naissance, tiré de l'étude de l'exposition en début de vie aux neurotoxiques dans la ville de Mexico (Early Life Exposure in Mexico City to Neuro-Toxicants).

Jingyuan Yang and Shili Lin

*Likelihood Approach for Detecting Imprinting and In Utero Maternal Effects
Using General Pedigrees from Prospective Family-Based Association Studies*

477

L'empreinte génétique et les effets maternels in utero sont deux causes d'effet d'origine parentale mais ils sont le plus souvent indistinguables. Les tests développés pour détecter seulement l'un de ces deux effets auront de fait une sévère augmentation de l'erreur de type I si l'hypothèse d'absence de l'autre effet n'est pas vérifiée. Certaines méthodes existantes proposent de modéliser simultanément les deux effets. Cependant, ces méthodes ne sont pas adaptées à l'étude de familles étendues communément recrutées dans les études familiales. Ici, nous proposons une approche fondée sur la vraisemblance pour détecter l'empreinte génétique et les effets maternels (EGEM) dans des familles étendues recrutées lors d'études familiales prospectives. EGEM calcule la probabilité des génotypes familiaux sans supposer l'équilibre d'Hardy-Weinberg en introduisant un nouveau concept dit 'd'accouplement conditionnel' entre les fondateurs et leurs épouses non fondatrices. De plus, une fonction logit est utilisée pour modéliser la pénétrance. Pour prendre compte les données génotypiques incomplètes, EGEM impute implicitement les génotypes non observés en considérant tous ceux compatibles avec les génotypes observés. Nous comparons la puissance et l'erreur de type I d'EGEM avec deux autres méthodes existantes. Les résultats montrent que l'utilisation de généalogies étendues, même en présence d'informations incomplètes, sont bien plus puissantes que l'utilisation de seules familles nucléaires pour l'identification d'empreinte génétique et

d'effets maternels in utero sans inflation particulière de l'erreur de type I.

Oliver E. Lee and Thomas M. Braun

Permutation Tests for Random Effects in Linear Mixed Models

486

L'inférence concernant l'inclusion ou l'exclusion d'effets aléatoires dans les modèles linéaires mixtes est difficile parce que les composantes de la variance sont situées sur la frontière de leur espace paramétrique sous l'hypothèse nulle habituelle. Par conséquent la distribution nulle asymptotique de Wald, des scores, et des tests du rapport des vraisemblances n'est pas celle classique du χ^2 . Bien qu'il ait été prouvé que la distribution asymptotique correcte soit un mélange de distributions de χ^2 , la distribution appropriée du mélange est assez complexe et peu intuitive lorsque les hypothèses nulle et alternative diffèrent par plus d'un effet aléatoire. Comme alternative, nous présentons deux tests de permutation, l'un est fondé sur les meilleurs prédicteurs linéaires sans biais (BLUPs) et l'autre sur la statistique du test du rapport de vraisemblance restreint. Les deux méthodes impliquent l'utilisation de résidus pondérés, avec des poids déterminés par les composantes de la variance inter- et intra-sujet.

Sous l'hypothèse nulle, les distributions des statistiques de permutation sont calculées en permutant à la fois les résidus intra- et inter-sujets et elles sont valables autant asymptotiquement que pour des petits échantillons. Nous examinons la taille et la puissance de nos tests par simulation sous une variété de paramétrages, et nous appliquons notre test à un ensemble de données publiées de patients souffrant de leucémie myéloïde chronique.

J. D. Lebreton, R. Choquet and O. Gimenez

Simple Estimation and Test Procedures in Capture–Mark–Recapture Mixed Models

494

Le besoin de considérer dans des modèles de capture-recapture des effets aléatoires en sus d'effets fixes comme ceux de covariables environnementales a largement été reconnu ces dernières années. Toutefois, les approches formelles réclament des intégrations multiples lourdes, et des difficultés tant pratiques que conceptuelles ont ralenti la diffusion des modèles mixtes de capture recapture parmi les biologistes. Dans cet article, nous évaluons des procédures simples pour tester l'effet d'une covariable environnementales sur des paramètres tels que des probabilités de survie variant au cours du temps, en présence d'un effet aléatoire correspondant à une part de variation environnementale inexpliquée. Nous montrons que le test de rapport de vraisemblance usuel entre modèles fixes est fortement biaisé, et tend à détecter trop souvent l'effet d'une covariable. Nous montrons que les tests de permutation et d'analyse de déviance se comportent de façon adéquate, et recommandons donc leur utilisation. Les tests de permutation sont implémentés dans la version la plus récente du programme E-SURGE. Notre approche s'applique également aux modèles linéaires généralisés.

Hiroshi Okamura, Shingo Minamikawa, Hans J. Skaug, and Toshiya Kishiro

Abundance Estimation of Long-Diving Animals Using Line Transect Methods

504

L'échantillonnage par transects est la méthode la plus utilisée pour estimer la taille des populations animales dans la nature. Cette méthode suppose que la détection des

animaux le long du transect est infaillible. Cette hypothèse n'est pas toujours vérifiée par les animaux marins susceptibles de plonger et de refaire surface. On utilise souvent des détections en double à partir de deux plateformes d'échantillonnage parcourant le même transect pour estimer la probabilité de détection. Cependant, ces méthodes sont insuffisantes pour estimer l'abondance des animaux à longue durée de plongée car ces animaux peuvent être complètement indétectables pour les deux observateurs. Nous avons développé un modèle de probabilité de détection plus flexible en incorporant des informations sur les comportements typiques de plongée déduit de données télémétriques. Le modèle est basé sur un processus stochastique ponctuel et est analysable statistiquement. Une étude par simulation montre que le nouveau modèle fournit des estimations d'abondances très peu biaisées, alors que les modèles à taux ou probabilités de détection classiques ont un biais beaucoup plus important. Nous appliquons ce modèle à des relevés d'abondance de la bérardie de Baird (*Berardius bairdii*) dans le Pacifique Nord occidental à titre d'illustration.

Anton Grafström, Niklas L. P. Lundström, and Lina Schelin

Spatially Balanced Sampling through the Pivotal Method

514

Nous présentons une méthode simple pour sélectionner un échantillon spatialement équilibré en utilisant des probabilités d'inclusion égales ou inégales. Pour des populations présentant des tendances dans les variables analysées, l'estimation peut être considérablement améliorée en sélectionnant des échantillons bien répartis dans toute la population. La méthode est utilisable pour tout nombre de dimensions et peut donc aussi sélectionner des échantillons spatialement équilibrés dans un espace décrit par plusieurs variables auxiliaires. L'analyse comme les exemples concrets montrent que la méthode proposée produit un bon équilibre spatial et est donc efficace pour les populations à tendances marquées.

Kwun Chuen Gary Chan, and Mei-Cheng Wang

Estimating Incident Population Distribution from Prevalent Data

521

Un échantillon prévalent est constitué de personnes qui sont été atteintes d'une maladie incidente et en souffrent toujours au moment de l'enquête. Nous discutons les méthodes permettant d'estimer la fonction de distribution d'un vecteur aléatoire défini au départ pour une population incidente, alors que les données ont été recueillies par échantillonnage prévalent; L'échantillonnage de cas prévalents est souvent plus limité et plus économique qu'une étude de l'incidence afin d'étudier la distribution de survie d'une population atteinte par une maladie, mais l'on sait que les études de prévalence sont par essence biaisée compte tenu de leur schéma d'étude même. Les sujets ayant un temps de survie plus long ont en effet plus de chance d'être inclus dans une cohorte prévalente, et les paramètres initiaux qui sont corrélés à la survie sont également sujets au biais d'échantillonnage induit par le schéma de l'étude de prévalence. La méconnaissance de ce biais peut conduire à des biais très importants lorsque l'on applique une fonction de distribution empirique pour estimer la distribution des paramètres initiaux de la maladie au niveau de la population.

Dans cet article, des méthodes paramétriques et non paramétriques sont proposées pour estimer les distributions des paramètres initiaux dans les études de cas prévalents.

Cet article examine les procédures de test par groupe lorsque les unités au sein d'un groupe (ou d'un regroupement) peuvent être corrélées. Le nombre attendu de tests par unité (c'est à dire l'efficacité) de procédures hiérarchiques et matricielles est obtenu en se basant sur une classe de modèles à variables aléatoires binaires échangeables. On examine ensuite l'effet sur l'efficacité, de l'arrangement des unités corrélées au sein des regroupements. En général, lorsque les unités corrélées sont placées dans le même groupe, le nombre attendu de tests par unité décroît, parfois de façon substantielle, par rapport aux arrangements dans lesquels l'information sur la corrélation est ignorée.

L'estimation d'une effet de causalité pour une intervention sur une population recourt typiquement à la définition de paramètres dans un modèle nonparamétrique d'équation structurelle (Pearl, 2000, NPSEM) dans lequel le traitement (ou l'exposition) est posé de manière déterministe soit de façon statique soit de façon dynamique. Nous définissons un nouveau paramètre de causalité qui prend en compte le fait que les politiques d'intervention peuvent conduire à des expositions à caractère stochastique. Nous établissons le paramètre statistique qui identifie le paramètre de causalité auquel on s'intéresse. On développe les estimateurs selon la pondération par l'inverse de la probabilité de traitement (IPTW), selon l'IPTW augmentée (A-IPTW) et les estimateurs du maximum de vraisemblance ciblé (TMLE). Une étude par simulation est réalisée pour montrer les propriétés de ces estimateurs, qui incluent la double robustesse de l'A-IPTW et du TMLE. Un exemple d'application utilisant des données d'activité physique est présenté.

Nous présentons un modèle bayésien non paramétrique, élaboré pour un essai clinique de phase II chez des patients présentant différents sous-types de la maladie à l'étude. L'objectif est d'estimer la probabilité de succès associée à chaque sous-type, mais la faiblesse des effectifs des sous-groupes ainsi obtenus requiert une mise en commun de l'information. Parce qu'elles sont a priori spécifiques, et nullement interchangeables, les différentes déclinaisons de la maladie ne se prêtent guère à une modélisation par des modèles hiérarchiques traditionnels. Nous proposons donc un modèle à partition aléatoire, variante du modèle à partition par produit, de façon à combiner l'information. Il s'agit d'utiliser toutes les observations pour estimer les probabilités individuelles de succès, tout en faisant en sorte que le modèle considère bien les différentes formes de la maladie comme non interchangeables. Par conséquent, la probabilité de succès pour une déclinaison donnée est estimée en utilisant davantage l'information (sur le critère d'efficacité) issue des patients de même pronostic que l'information issue des autres patients.

Les données proviennent d'un essai clinique de phase II ayant inclus des patients atteints de sarcome, une forme rare de cancer. Chaque patient présente une déclinaison de la

maladie et ces déclinaisons sont catégorisées selon qu'elles sont associées à un pronostic bon, moyen ou mauvais. Le modèle a priori doit respecter la variabilité du pronostic selon les déclinaisons de la maladie. En raison des faibles effectifs, l'inférence nécessite, pour gagner en puissance, une mise en commun de l'information. Une étude de simulation compare le modèle proposé à un modèle plus simple, qui supposerait des probabilités de succès identiques entre les déclinaisons de même pronostic. Lorsque cette hypothèse est vérifiée, les deux modèles atteignent des niveaux de performance comparables ; en revanche, lorsqu'elle est incorrecte, le modèle que nous proposons s'avère plus performant que le modèle simplifié.

Bruce D. Spencer

When Do Latent Class Models Overstate Accuracy for Diagnostic and Other Classifiers in the Absence of a Gold Standard? 559

Les modèles à classes latentes sont de plus en plus utilisés pour établir la validité des tests de diagnostics médicaux ainsi que d'autres classifications lorsqu'il n'existe pas de référence et que l'état exact est inconnu. Lorsque la classe latente est traitée comme la classe vraie, les modèles à classes latentes donnent des mesures des composantes de la validité comme la spécificité, la sensibilité et les taux d'erreur de type I et II. Les taux d'erreur associés à un modèle à classe latente diffèrent cependant des véritables taux d'erreur et des comparaisons empiriques avec une référence suggèrent que les taux d'erreurs véritables sont souvent plus élevés. Nous étudions les conditions sous lesquelles les vrais taux d'erreur de type I et II sont plus grands que ceux obtenus par les modèles à classes latentes. Des résultats de Uebersax (1988) sont étendus pour prendre en compte des effets aléatoires et des covariables influant sur les réponses. Les résultats sont importants pour l'interprétation des résultats des analyses de modèles à classe latente. On présente une décomposition de l'erreur, qui incorpore une composante de l'erreur associée à la non-validité du modèle à classe latente.

Ruixin Guo, Hongtu Zhu, Sy-Miin Chow, and Joseph G. Ibrahim

Bayesian Lasso for Semiparametric Structural Equation Models 567

Le développement de modèles non linéaires à équations structurelles et des procédures associées d'inférence statistique, incluant les méthodes d'estimation et de sélection de modèles, ont rencontré un grand intérêt. Dans cet article, nous développons un modèle général semi-paramétrique d'équation structurelle (SSEM) dans lequel l'équation structurelle est composée à partir de fonctions non-paramétriques des variables exogènes latentes et de covariables sur un ensemble de variables latentes endogènes. Une représentation de base est utilisée pour approximer ces fonctions non-paramétriques, et une approche LASSO Bayésienne couplée à un algorithme de Monte-Carlo Chaîne de Markov (MCMC) est choisie pour réaliser simultanément l'estimation et la sélection du modèle. La méthode proposée est illustrée en utilisant une étude par simulation, ainsi que des données provenant de l'étude « Données Affectives et Différences individuelles » (ADID). Les résultats montrent que notre méthode permet une estimation précise des paramètres inconnus, et une identification correcte du modèle réel sous-jacent.

BIOMETRIC PRACTICE

Joseph G. Ibrahim, Ming-Hui Chen, H. Amy Xia, and Thomas Liu

Bayesian Meta-Experimental Design: Evaluating Cardiovascular Risk in New Antidiabetic Therapies to Treat Type 2 Diabetes 578

Des recommandations récentes de la FDA, portent sur l'évaluation des nouvelles thérapies dans le traitement du diabète de type II, appellent à ce que des méta-analyses des critères cardio-vasculaires soient planifiées au sein des programmes cliniques. Dans ce contexte, nous développons une nouvelle approche bayésienne de méta-analyse, utilisant des modèles de régression pour données de survie, afin d'évaluer si la taille du programme de développement clinique permet de caractériser de façon adéquate un critère donné de tolérance. A cet effet, nous proposons d'appliquer au schéma de méta-analyse des essais cliniques une méthodologie bayésienne de calcul d'effectif, qui porte une attention toute particulière au contrôle du risque de première espèce et de la puissance. Nous proposons également d'utiliser un a priori informatif que nous appelons le *partial borrowing power prior* pour intégrer les méta-données historiques de survie dans le modèle statistique. Nous examinons plusieurs propriétés de la méthodologie proposée et développons un algorithme efficace d'échantillonnage de type MCMC à partir des distributions a posteriori. De plus, nous développons un algorithme basé sur des simulations pour calculer plusieurs quantités, telles que la puissance et l'erreur de première espèce, dans le schéma bayésien de méta-analyse des essais cliniques. Pour finir, nous appliquons la méthodologie proposée pour élaborer un programme de phase II-III incluant notamment un essai clinique de non-infériorité, afin d'évaluer le risque cardio-vasculaire dans des études T2DM.

Swati Biswas and Shili Lin

Logistic Bayesian LASSO for Identifying Association with Rare Haplotypes and Application to Age-Related Macular Degeneration 587

Les variants rares ont été annoncés comme une clé pour découvrir «l'héritabilité manquante» dans des maladies complexes. Ces variants peuvent maintenant être génotypés en utilisant les technologies de séquençage de nouvelle génération; néanmoins, les haplotypes rares peuvent également résulter de la combinaison de SNP courants disponibles à partir des Etudes de l'Association « Genome-Wide » (GWAS). Les données de l'Institut National de la Vision « National Eye Institute » sur la dégénérescence maculaire liée à l'âge (DMLA) en sont un exemple. Des études sur AMD avaient identifié des variants rares potentiels, cependant, à cause du manque d'outils statistiques appropriés, les effets de différents haplotypes rares n'ont jamais été étudiés. Ici, nous développons une méthode pour identifier une association avec des haplotypes rares dans une expérience cas-témoins. Une régression logistique basée sur la vraisemblance rétrospective est formulée, elle est régularisée à l'aide d'une Logistique Bayésienne LASSO (LBL). En particulier, on pénalise les coefficients de régression en utilisant des *a-priori* appropriés pour éliminer les haplotypes non associés, rendant possible pour les plus rares associés de se démarquer. Nous avons appliqué la LBL aux données AMD et nous avons identifié les haplotypes communs et rares dans le gène de Facteur Complémentaire H, obtenant un aperçu des contributions des variants rares pour l'AMD au-delà de la littérature actuelle. Cette analyse montre également la richesse des données

GWAS pour cartographier des haplotypes rares, un potentiel largement inexploré. De plus, nous avons réalisé des simulations pour étudier la performance de LBL et le comparer au logiciel Hapassoc. Nos résultats montrent que LBL est beaucoup plus puissant pour identifier des haplotypes rares associés lorsque le taux de faux positifs pour les deux approches conserve les mêmes valeurs.

Nicholas G. Reich, Justin Lessler, Derek A. T. Cummings, and Ron Brookmeyer

Estimating Absolute and Relative Case Fatality Ratios from Infectious Disease Surveillance Data 598

Identifier les populations les plus à risque d'événements sévères dans une maladie infectieuse émergente est crucial pour décider d'une allocation optimale des ressources dans le plan d'action envers une épidémie. Le taux de létalité est la fraction des cas qui décèdent après avoir contracté la maladie. Le risque relatif de létalité est le facteur par lequel la létalité d'un groupe est multipliée ou divisée par rapport à celle du second groupe. La déclaration incomplète du nombre d'individus infectés, à la fois guéris et décédés, peut aboutir à des estimations biaisées du taux de létalité. Nous définissons des conditions sous lesquelles le taux de létalité et le risque relatif de létalité sont identifiables. De plus, nous proposons un estimateur du risque relatif de létalité prenant en compte des taux de déclaration variant dans le temps. Nous généralisons nos méthodes pour prendre en compte le temps passé entre l'infection et le décès. Pour illustrer la nouvelle méthodologie, nous utilisons les données de la pandémie de grippe de 1918 pour estimer le risque relatif de létalité entre les comtés du Maryland. Une étude de simulation évalue la performance des méthodes dans des scénarii d'épidémies. Un paquet pour le logiciel R rend les méthodes et les données présentées ici librement accessibles. Notre travail met en évidence les limitations et les enjeux associés avec l'estimation des risques absolus et relatifs de létalité en pratique. Cependant, dans certaines situations, les méthodes présentées ici peuvent aider à identifier des sous populations vulnérables précocement dans le développement d'une épidémie liée à un agent pathogène émergent comme dans la grippe pandémique.

Sebastian Meyer, Johannes Elias, and Michael Höhle

A Space-Time Conditional Intensity Model for Invasive Meningococcal Disease Occurrence 607

Nous proposons un nouveau modèle de processus ponctuel spatio-temporel continu est proposé pour quantifier les dynamiques de transmission des deux types de séquences antigéniques méningococciques les plus communément observées en Allemagne entre 2002 et 2008. La modélisation repose sur une fonction d'intensité conditionnelle (CIF) qui est décrite par une superposition de composantes additives et multiplicatives. Nous apportons une conclusion épidémiologique intéressante selon laquelle le comportement de diffusion dépend du type en plus de l'âge : les constantes de reproduction basales étant 0.25 (intervalle de confiance à 95% : 0.19 – 0.34) et 0.11 (intervalle de confiance à 95% : 0.07 – 0.17) respectivement pour les types B :P1.7-2,4F :1-5 et C :P1.5,2 :F3-3. Au total, la méthodologie proposée représente un cadre de régression, universel et explicite, pour la modélisation, la simulation et l'inférence de processus spatio-temporels continus auto-excités, basé sur les CIF. L'utilisation de cette modélisation en pratique biomédicale est facilitée par une implémentation dans le package R.

Yue Wang, Robin Mogg, and Jared Lunceford

Evaluating Correlation-Based Metric for Surrogate Marker Qualification within a Causal Correlation Framework 617

Les biomarqueurs jouent un rôle croissant dans le développement clinique de nouvelles thérapeutiques. Des décisions cliniques plus précoces facilitées par les biomarqueurs peuvent produire une réduction des coûts et des durées de développement des produits. Des associations entre des biomarqueurs et des critères cliniques sont souvent considérées comme des éléments initiaux à l'appui des objectifs envisagés. Le résultat est que, même si il est largement compris qu'une corrélation n'est pas une preuve de relation causale, la corrélation continue à être utilisée comme métrique pour qualifier un biomarqueur en pratique. Dans cet article, nous introduisons un cadre de corrélation causale, où deux types différents de corrélations sont définis au niveau individuel. Nous montrons que l'estimation de la corrélation est constituée de différentes composantes, et nécessite des précautions pour son interprétation lorsqu'on l'utilise pour qualifier un biomarqueur si l'on veut éviter des conclusions erronées. Autrement, on peut conclure à une corrélation significative même en l'absence d'une réelle association sous-jacente. Nous montrons également comment les quantités causales d'intérêt peuvent être testées dans un dispositif en crossover, et nous fournissons une discussion sur les difficultés à résoudre dans un contexte de groupes parallèles.

Dan Yang, Dylan S. Small, Jeffrey H. Silber, and Paul R. Rosenbaum

Optimal Matching with Minimal Deviation from Fine Balance in a Study of Obesity and Surgical Outcomes 628

Dans les appariements multivariés, un équilibre parfait contraint les distributions marginales d'une variable nominale dans les groupes appariés traités et contrôles à être identiques sans contrainte sur les appariements individuels. De ce fait, une contrainte d'équilibre parfait peut équilibrer une variable nominale avec de nombreuses modalités tout en portant les efforts sur les autres variables plus importantes quand on apparie les individus sur la minimisation de la distance totale des covariables à l'intérieur d'une paire. L'équilibre parfait n'est pas toujours possible ; en fait, il s'agit d'une contrainte dans un problème d'optimisation sans que celle-ci ne soit toujours satisfaite. Nous proposons un nouvel algorithme qui fournit un appariement avec distance minimale, parfaitement équilibré quand cela est faisable, et qui sinon, minimise la distance totale entre tous les échantillons appariés avec un écart minimal à l'équilibre parfait. Dans certains cas, nous pouvons être très proches de l'équilibre parfait quand celui-ci n'est pas accessible ; de plus, dans tous les cas, comme notre algorithme garantit d'obtenir la solution la plus proche de l'équilibre parfait, l'investigateur peut réaliser un appariement et sur cette base juger si l'équilibre optimal obtenu est adéquat ou pas. Nous montrons aussi comment incorporer une contrainte additionnelle. L'algorithme est implémenté de deux façons similaires, la première comme un problème d'allocation optimale avec une matrice de distance augmentée, la seconde comme un problème de minimisation de coût d'un flux dans un réseau. L'étude sur la chirurgie du genou dans l'obésité et de ses résultats chirurgicaux a motivé le développement de cet algorithme et est utilisée pour l'illustrer. Dans cet exemple, deux des 47 hôpitaux ont trop peu de patients non obèses pour obtenir un équilibre parfait pour la variable nominale à 47

niveaux représentant les hôpitaux, mais notre nouvel algorithme aboutit à une solution très proche de l'équilibre parfait. De plus, dans cet exemple, il existe un déficit de patients diabétiques non obèses ; l'incorporation d'une contrainte additionnelle a forcé à inclure dans l'appariement tous les patients non obèses et diabétiques, aboutissant de cette manière à un équilibre optimal sur cette covariable importante mais récalcitrante.

Dandan Liu, Douglas E. Schaubel, and John D. Kalbfleisch

Computationally Efficient Marginal Models for Clustered Recurrent Event Data 637

Les grandes bases de données observationnelles obtenues à partir des registres de maladies et d'études de cohorte rétrospectives se sont avérées très utiles pour étudier les services de santé. Cependant, l'utilisation de grandes bases de données peut introduire des difficultés de calcul, particulièrement lorsque l'événement auquel on s'intéresse est récurrent. Dans de telles situations, le regroupement des données d'événements récurrents dans des intervalles présélectionnés conduit à un modèle flexible de taux d'événement et à une contraction des données qui simplifie les aspects de calcul. Nous proposons un modèle marginal de taux proportionnels, stratifiable, avec un taux de base constant par morceaux pour des données d'événement récurrent. On considère aussi bien l'absence que la présence d'événement terminal. Pour de grands échantillons nous obtenons les distributions des estimateurs proposés. Nous réalisons des études de simulation sous diverses configurations de données, y compris les cas où le modèle est mal spécifié. Nous proposons des aides pour la sélection d'intervalle, que nous évaluons par des études numériques. Nous montrons ensuite que les procédures proposées peuvent être mises en œuvre avec des packages statistiques standard (tels que SAS, R). Nous donnons une application sur des données nationales d'hospitalisation de patients en insuffisance rénale terminale.

Alexander C. McLain, Kirsten J. Lum, and Rajeshwari Sundaram

A Joint Mixed Effects Dispersion Model for Menstrual Cycle Length and Time-to-Pregnancy 648

Les profils de cycle menstruel sont souvent utilisés comme indicateurs de la fécondité féminine et sont associés à des pathologies hormono-dépendantes comme le cancer du sein. L'identification des profils de cycle menstruel et leur association avec la fécondité représentent un important enjeu scientifique. Les enquêtes prospectives sur la grossesse qui recueillent des informations détaillées sur des femmes en âge de procréer permettent d'étudier ces questions. Cependant l'étude de l'association entre les deux processus soulève des problèmes méthodologiques liés au nombre relativement petit d'observations répétées du cycle menstruel, à leur censure informative par le délai de conception et à la forte asymétrie de la distribution des durées de cycles. Nous proposons une modélisation jointe associant un modèle à effets mixtes pour la dispersion des durées de cycles et un modèle de survie discret pour le délai de conception. Cela nous permet d'évaluer l'effet des caractéristiques du cycle menstruel sur la fécondité ainsi que l'effet des prédicteurs de la fécondité, comme l'âge aux premières règles, l'âge actuel ou la parité sur chacun des deux processus. Un avantage de cette approche est de pouvoir prédire le délai de conception et donc de pouvoir mesurer la valeur prédictive des caractéristiques du cycle menstruel. Nous illustrons cette méthode avec deux enquêtes prospectives sur la grossesse en ajustant un modèle aux données de l'Oxford Conception Study et en

évaluant sa valeur prédictive sur les données de la New York State Angler Prospective Pregnancy Cohort Study. Nos résultats sont applicables à l'évaluation de la fécondité.