Phd Position in Causal Inference

Title : Analyse de Survie en difference-en-difference (Survival Difference-in-Differences: SDiD)

Contexte En inférence causale, la présence de facteurs de confusion non mesurés constitue la principale limite des études observationnelles. Plusieurs stratégies ont été proposées pour pallier cette difficulté, parmi lesquelles l'approche en difference-in-differences (DiD). Le principe est de comparer l'évolution temporelle d'un critère de jugement entre un groupe traité et un groupe non traité. En oncologie, par exemple, de nouveaux traitements sont souvent approuvés sur la base d'essais non randomisés, à un seul bras et en ouvert. Pour déterminer si l'autorisation d'un nouveau traitement s'est traduite par une amélioration de la survie à l'échelle de la population, une stratégie consiste à comparer les tendances de survie avant et après la période d'autorisation dans les régions où les médicaments ont été introduits, à celles observées dans des régions ou pays où l'autorisation n'a pas été accordée. Bien que ce type d'expérience naturelle soit adapté au cadre DiD, il soulève d'importants défis méthodologiques. Si de nombreuses extensions ont été développées pour tenir compte de la complexité des designs d'étude, l'intégration de l'analyse de survie dans ce cadre reste encore largement inexplorée. Ce projet de thèse vise à combler un manque méthodologique important à l'interface entre analyse de survie et méthodes en DiD en proposant des outils robustes, flexibles et directement applicables aux données observationnelles en santé.

Projet 1 Dans un premier projet, l'étudiant s'attachera à définir les hypothèses nécessaires à l'identification de la quantité causale d'intérêt en présence de données de survie. Il mettra ensuite en place des stratégies d'estimation adaptées. Au-delà des estimateurs paramétriques classiques, il s'appuiera sur la théorie de l'estimation semi-paramétrique afin de développer des estimateurs robustes intégrant des méthodes d'apprentissage statistique, notamment le machine learning et les réseaux de neurones. Le travail sera dans un premier temps mené dans un cadre sans censure, avant d'être étendu à des situations plus réalistes intégrant des mécanismes de censure plus complexes. Les méthodes proposées seront évaluées à travers des études de simulation afin d'analyser leurs propriétés en échantillon fini, puis appliquées à des données issues du Système National des Données de Santé (SNDS). L'objectif sera d'évaluer l'impact de l'autorisation de mise sur le marché européenne des anti-PD1 comme traitement de première ligne des carcinomes cutanés épidermoïdes avancés, intervenue entre 2018 et 2019, sur la survie globale. Cette application permettra d'illustrer concrètement les approches développées et donnera lieu à la production d'outils reproductibles, sous forme de code et de tutoriels, à destination des épidémiologistes et des chercheurs en santé publique.

Projet 2 Dans un second projet, le travail portera sur les contraintes d'accès aux données individuelles, particulièrement fortes dans le domaine de la santé en raison de leur caractère sensible. L'étudiant développera des stratégies d'estimation adaptées à des contextes où les données ne peuvent pas être centralisées. Il s'agira notamment de proposer des procédures permettant d'estimer les paramètres d'intérêt à partir d'informations partielles ou distribuées, tout en préservant les propriétés statistiques des estimateurs.

Projet 3 Enfin, un troisième projet sera consacré à l'étude de l'hypothèse de tendances parallèles, qui constitue l'hypothèse clé des approches DiD. Cette hypothèse suppose qu'en l'absence de traitement, l'écart entre les groupes traité et contrôle resterait constant au cours du temps, ce qui est fondamentalement invérifiable. L'étudiant proposera un cadre d'analyse de sensibilité permettant de quantifier l'impact de violations de cette hypothèse sur les paramètres d'intérêt. En s'appuyant sur des modèles structuraux intégrant des facteurs de confusion non observés, il s'agira de dériver des expressions analytiques du biais induit et de caractériser les régimes de paramètres pour lesquels les conclusions causales sont robustes. Ces développements permettront de fournir des outils pratiques pour évaluer la crédibilité des hypothèses dans des applications réelles.

Environnement de recherche : Le doctorat se déroulera au sein de l'équipe de recherche Epilogy, affiliée à l'INSERM et basée à l'Université Paris-Est Créteil. Le projet de thèse sera dirigé par Dr. Tat-Thang Vo et Pr. Emilie Sbidian. L'étudiant profitera aussi de l'encadrement de Dr. Marie-Félicia Beclin et de Dr. Enrico Roma.

Compétences requises :

- Master (MSc) en statistique, biostatistique, data science ou dans un domaine connexe.
- Solides connaissances en statistique et en mathématiques requises. Un goût pour l'abstraction et la modélisation est attendu, ainsi qu'un intérêt pour les applications en santé.
- Maîtrise de R ou Python, avec une expérience des bibliothèques de statistique et de Machine-Learning.
- Bon niveau d'anglais.

Background In causal inference, the presence of unmeasured confounding is a major limitation of observational studies. Several strategies have been proposed to address this issue, among which the difference-in-differences (DiD) approach is widely used. The principle is to compare the temporal evolution of an outcome between a treated group and an untreated group. In oncology, for example, new treatments are often approved based on non-randomized, single-arm, open-label trials. To assess whether the approval of a new treatment has led to an improvement in survival at the population level, one strategy is to compare survival trends before and after the approval period in regions where the treatment was introduced with those observed in regions or countries where approval was not granted. Although such natural experiments are well-suited to the DiD framework, they raise important methodological challenges. While many extensions have been developed to account for complex study designs, the integration of survival analysis within this framework remains largely unexplored. This PhD project aims to fill an important methodological gap at the interface between survival analysis and DiD methods by developing robust, flexible tools that can be directly applied to observational health data.

Project 1 In a first project, the student will focus on defining the assumptions required for identifying the causal quantity of interest in the presence of survival data. Appropriate estimation strategies will then be developed. Beyond classical parametric estimators, the work will rely on semiparametric estimation theory to construct robust estimators incorporating modern statistical learning methods, including machine learning and neural networks. The initial developments will be carried out in a setting without censoring, before being extended to more realistic scenarios involving complex censoring mechanisms. The proposed methods will be evaluated through simulation studies to assess their finite-sample properties and then applied to data from the French National Health Data System (SNDS). The objective will be to evaluate the impact of the European approval of anti-PD1 therapies as first-line treatment for advanced cutaneous squamous cell carcinoma, which occurred between 2018 and 2019, on overall survival. This application will provide a concrete illustration of the proposed methods and will lead to the development of reproducible tools, including code and tutorials, for epidemiologists and public health researchers.

Project 2 In a second project, the focus will be on constraints related to access to individual-level data, which are particularly strong in the health domain due to their sensitive nature. The student will develop estimation strategies tailored to settings where data cannot be centralized. In particular, the goal will be to design procedures that allow the estimation of parameters of interest from partial or distributed information, while preserving the statistical properties of the estimators.

Project 3 Finally, a third project will address the parallel trends assumption, which is the key assumption underlying DiD approaches. This assumption states that, in the absence of treatment, the difference between treated and control groups would remain constant over time, which is fundamentally untestable. The student will develop a sensitivity analysis framework to quantify the impact of violations of this assumption on the parameters of interest. Relying on structural models incorporating unobserved confounding, the goal will be to derive analytical expressions for the induced bias and to characterize parameter regimes under which causal conclusions remain robust. These developments will provide practical tools for assessing the credibility of assumptions in real-world applications.

Research Environment: The PhD will be conducted within the Epilogy research team, affiliated with INSERM and based at Université Paris-Est Créteil. The doctoral project will be supervised by Dr. Tat-Thang Vo and Prof. Emilie Sbidian. The student will also benefit from the guidance of Dr. Marie-Félicia Beclin and Dr. Enrico Roma.

Required Skills:

- Enrolled in a Master's program (MSc) in Statistics, Biostatistics, Data Science, or a related field.
- Strong knowledge of statistics and mathematics is required. A taste for abstraction and modeling is expected, along with an interest in health-related applications.
- Proficiency in R or Python, including experience with statistical and/or machine learning libraries.
- Proficient in English. Proficiency in French is a plus but not mandatory.

Application Process:

- Applications should be submitted through Adum no later than May 8th. (Link of the offer)
- Contact : mariefelicia.beclin@gmail.com and tat-thang.vo@u-pec.fr