Synthèse de la Thèse de Doctorat

Contributions aux modèles de régression avec réponses manquantes : Risques concurrents et données longitudinales

Margarita Moreno Betancur

Introduction

La modélisation par régression avec des réponses manquantes

Les données manquantes sont fréquentes dans les études médicales, notamment dans celles impliquant des êtres humains. Par exemple, lorsqu'on demande aux participants d'une étude de remplir un questionnaire, certains d'entre eux laisseront souvent quelques items sans réponse. Dans les études médicales, il est commun de suivre un groupe d'individus au cours du temps, par exemple pour étudier l'évolution d'un marqueur biologique mesuré à plusieurs reprises ou le temps jusqu'à la survenue d'un évènement tel que le décès. L'observation prolongée des individus accentue l'occurrence de données manquantes. En effet, le recueil d'information se détériore avec le temps, entre autres parce que quelques individus n'assistent pas aux rendez-vous planifiés et d'autres sortent des études complètement par des raisons migratoires, médicales ou autres. Il est difficile d'éviter ces problèmes entièrement quelque soit le type d'étude, même lorsqu'il s'agit d'un essai clinique soigneusement planifié et suivant un protocole strict.

Dans la recherche médicale, la modélisation par régression est souvent utilisée pour étudier l'effet des covariables (ex. facteurs pronostiques, expositions) sur des réponses d'intérêt médical (ex. le niveau d'un marqueur biologique, le délai de survie). Dans ces études, les individus sont recrutés suivant des critères prédéfinis, et les réponses qui seront mesurées pour chacun d'entre eux, ainsi que les dates et horaires de ces mesures, sont déterminés à l'avance. L'intérêt médical porte sur les effets des covariables décrivant la distribution de la totalité des réponses prévues. Or, lorsque quelques réponses manquent pour certains individus, les réponses observées et les covariables représentent toute l'information disponible pour faire des inférences (c.-à-d. des estimations ponctuelles, des intervalles de confiance, des tests d'hypothèse) sur ces paramètres. Outre l'évidente perte de précision entrainée par l'information manquante, toute méthode pour faire des inférences à partir des données disponibles requière qu'une hypothèse sur le mécanisme de manquement soit vérifiée.

Rubin (1976) a proposé une taxonomie des mécanismes de manquement qui est clé pour comprendre la problématique de l'analyse statistique de données incomplètes. Dans le contexte des réponses manquantes, le mécanisme de manquement est appelé MCAR (pour les sigles en anglais de *manquant complètement au hasard*) si la probabilité qu'une réponse soit manquante est constante. Le mécanisme de manquement est appelé MAR (pour les sigles en anglais de *manquant au hasard*) si la probabilité qu'une réponse soit manquante ne dépende pas des réponses manquantes conditionnellement aux données observées (ex. des covariables, des réponses mesurées à d'autres moments). Enfin, le mécanisme de manquement est appelé MNAR (pour les sigles en anglais de *manquant non au hasard*) si la probabilité qu'une réponse soit manquante dépend des réponses manquantes même lorsqu'on conditionne par les données observées.

La taxonomie de Rubin a des implications importantes concernant les besoins de modélisation pour obtenir des inférences valides à partir des données disponibles. Sous l'hypothèse d'un mécanisme MCAR, les réponses observées sont un échantillon aléatoire des réponses prévues. Par conséquent, dans plusieurs contextes, un mécanisme MCAR garantit que l'on puisse obtenir des inférences, certes imprécises, mais tout de même valides en effectuant une analyse dite des « cas complets » (CC), c'est à dire, une analyse excluant les individus avec des réponses manquantes. Sous MAR, les réponses observées et manquantes ont la même distribution conditionnellement au reste des données observées. Ceci implique que des inférences valides peuvent être obtenues à partir des données disponibles sans avoir à faire des hypothèses supplémentaires sur la distribution des réponses manquantes. Sous MNAR, les distributions conditionnelles des réponses observées et manquantes sont différentes, et il est alors nécessaire de faire des hypothèses sur cette dernière, que ce soit de façon explicite ou implicite, pour résoudre les problèmes d'identifiabilité.

La taxonomie de Rubin relève des différences importantes entre les diverses méthodes existantes pour gérer les réponses manquantes. Or, comme Molenberghs et al. (2008) l'a démontré dans un contexte assez général, il n'est pas possible d'évaluer à partir des données observées si le mécanisme de manquement est MAR ou MNAR. En effet, sauf dans certaines situations où le schéma de l'étude prévoit le manquement de certaines données, les hypothèses de manquement requises par quelconque stratégie de modélisation avec des données manquantes ne sont pas vérifiables. C'est pourquoi il est indispensable d'effectuer des *analyses de sensibilité*, pour évaluer la robustesse des inférences obtenues dans une analyse principale aux écarts par rapport aux hypothèses de manquement sous-jacentes.

Objectifs de la thèse

L'analyse CC et d'autres méthodes ad-hoc étaient souvent utilisées auparavant, et elles sont toujours utilisées mais dans une moindre mesure. En effet, une meilleure compréhension des fortes hypothèses requises par ces méthodes, irréalistes dans la plupart des cas, a stimulé le développement, l'implémentation et l'utilisation de méthodes reposant sur des hypothèses moins fortes. Ceci est en particulier vrai dans le contexte des données multivariées incomplètes, c'est-à-dire, lorsque l'intérêt porte sur un vecteur de réponses $(Y_1, \ldots, Y_J)$ dont certaines composantes peuvent être manquantes. Pour ce type de données, plusieurs méthodes de modélisation reposant sur l'hypothèse MAR sont utilisées couramment et diverses méthodes sous MNAR ont été proposées aussi. En revanche, même si quelques méthodes pour effectuer des analyses de sensibilité ont été proposées, celui-ci est encore un domaine actif de recherche.

Le premier objectif de cette thèse était de développer une méthode d'analyse de sensibilité pour les données longitudinales continues avec des sorties d'étude. Ces données sont un type particulier de données multivariées incomplètes, où les réponses sont continues et ordonnées dans le temps. De plus, pour chaque individu, on observe soit toutes les réponses, soit toutes les réponses jusqu'à un certain moment, auquel l'individu sort de l'étude et ne revient jamais, de façon à ce que le reste de ses réponses sont manquantes.

Le second objectif de la thèse était de développer des méthodes pour la modélisation de risques concurrents avec des causes d'évènement manquantes en s'appuyant sur la théorie existante pour les données multivariées incomplètes. Le modèle de risques concurrents est une extension du modèle standard de l'analyse de survie où l'on distingue le type d'évènement ou la cause l'ayant entrainé. Certains auteurs ont proposé des méthodes de régression pour risques concurrents avec des causes d'évènement manquantes sous l'hypothèse MAR, notamment pour la modélisation semi-paramétrique du risque cause-spécifique. Mais d'autres modèles n'ont pas été considérés, de même que la modélisation sous MNAR et les analyses de sensibilité. Notre but était donc de combler certains de ces vides dans la littérature.

Modélisation de données longitudinales avec des sorties d'étude

Contexte et état de la question

On parle de *données longitudinales* lorsqu'une variable réponse d'intérêt est mesurée pour chaque individu à plusieurs reprises au cours du temps, donnant lieu à une séquence de mesures répétées ordonnées dans le temps. Ici, nous nous intéressons aux variables réponse continues et, pour faciliter la notation, nous nous concentrons sur le cas où tous les individus ont le même nombre J de mesures prévues, ayant lieu aux mêmes dates. On dénote par $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iJ})$ le vecteur de réponses de l'individu i ($i = 1, \ldots, n$). L'objectif d'une étude longitudinale est d'étudier l'évolution intra-individu de la variable réponse au cours du temps et de déterminer les facteurs qui influencent cette évolution.

Un exemple d'étude longitudinale est l'essai clinique randomisé qui a motivé les développements de cette partie de la thèse. Dans cet essai, on étudiait un traitement pour le trouble du maintien du sommeil. Les 962 patients recrutés dans cet essai ont eu à remplir quotidiennement un questionnaire pendant la durée de l'étude qui contenait des scores quantifiant la qualité du sommeil. Les scores étaient : la durée de réveil après l'induction du sommeil (WASO) ; le nombre de réveils (NAW) ; la qualité rafraichissante du sommeil (SLREF), prenant des valeurs de un (excellent) à quatre (pauvre) ; la sensation de somnolence (FEELC), prenant des valeurs de zéro (très somnolent) à neuf (pas du tout somnolent) ; le temps total de sommeil (TST) ; et la latence à l'endormissement (SOL), qui est la durée entre la fermeture des yeux et l'entrée effective en sommeil. Une décroissance dans les scores WASO, NAW, SLREF ou SOL, ou une croissance dans les scores FEELC ou TST indiquerait une amélioration dans la qualité du sommeil du patient. Les scores quotidiens n'étaient pas disponibles, seuls leurs moyennes sur six périodes communes à tous les individus. Donc, toutes les variables réponse considérées étaient continues.

Pour étudier l'effet des covariables sur l'évolution de la variable réponse, des méthodes de régression prenant compte de la corrélation entre les réponses d'un même individu au cours du temps sont nécessaires. Si toutes les réponses de tous les individus sont observées, on peut utiliser par exemple le modèle linéaire mixte (Laird and Ware, 1982).

Lorsqu'il y a des sorties d'étude, le vecteur de réponses s'écrit $\mathbf{Y}_i = (\mathbf{Y}_i^{\mathcal{O}}, \mathbf{Y}_i^{\mathcal{M}})$, où $\mathbf{Y}_i^{\mathcal{O}} = (Y_{i1}, \ldots, Y_{i(U_i-1)})$ et $\mathbf{Y}_i^{\mathcal{M}} = (Y_{iU_i}, \ldots, Y_{iJ})$ sont les parties observées et manquantes de $\mathbf{Y}_i$, respectivement, et U_i dénote le moment de la première réponse manquante, avec $U_i \leq J$ pour les individus qui sont sortis de l'étude et $U_i = J+1$ pour ceux qui complètent l'étude. On appelle U_i *l'indicateur de sortie d'étude*. Plusieurs méthodes sont disponibles pour l'analyse de données longitudinales avec des sorties d'étude sous

l'hypothèse MAR. Par exemple, Rubin (1976) a montré qu'une analyse par maximum de vraisemblance utilisant toutes les données disponibles permet d'obtenir des estimations non-biaisées et d'efficacité maximale à condition que (i) le mécanisme de manquement soit MAR et (ii) les paramètres des mécanismes de réponse et manquement soient distincts. Si ces deux conditions sont vérifiées on dit que le mécanisme de manquement est *ignorable*. En effet, on peut obtenir des inférences valides tout simplement en ajustant le modèle linéaire mixte aux réponses disponibles. Deux autres approches utiles sous MAR sont l'imputation multiple (Rubin, 1987) et les équations d'estimation pondérées par l'inverse de la probabilité d'observation (Robins et al., 1995).

Sous MNAR, il faut modéliser la distribution jointe du vecteur de réponses $\mathbf{Y}$ et de l'indicateur de sortie d'étude U, dénotée par $f(\mathbf{y}, u)$. Deux approches distinguées par Little and Rubin (1987, Chapitre 11) sont les *modèles de sélection* et les *modèles de mélange de profils*. Ces derniers, d'importance pour la suite, sont fondés sur la factorisation suivante :

$$f(\mathbf{y}, u) = f(\mathbf{y}|u) \times f(u).$$

Une troisième approche est celle des *modèles à paramètres partagés* (Wu and Carroll, 1988; Wu and Bailey, 1988, 1989; Little, 1995).

Une approche pour effectuer des analyses de sensibilité consiste à évaluer la discordance entre les inférences obtenues dans l'analyse principale et celles obtenues à partir d'une famille de modèles MNAR reposant sur des hypothèses de manquement, distributionnelles ou structurelles différentes (ex. Little and Yau, 1996; Kenward, 1998; Kenward and Molenberghs, 1999; Michiels et al., 2002). Si l'analyse principale repose sur l'hypothèse MAR, une version plus structurée de cette approche consiste à considérer une famille de modèles MNAR indexés par un paramètre qui quantifie le départ par rapport à l'hypothèse MAR (Little, 1994; Rotnitzky et al., 1998; Scharfstein et al., 1999; Daniels and Hogan, 2000; Molenberghs et al., 2001a). D'autres méthodes d'analyse de sensibilité reposant sur les idées d'influence de Cook (1977, 1986) ont été proposées (Thijs et al., 2000; Verbeke et al., 2001; Molenberghs et al., 2001b; Jansen et al., 2006).

Contributions

L'approche développée pour répondre au premier objectif de la thèse est fondée sur le principe d'une famille de modèles MNAR indexés par un paramètre quantifiant le départ par rapport à MAR. On utilise des modèles de mélange de profils, considérés par certains auteurs comme les plus appropriés pour les analyses de sensibilité (Daniels and Hogan, 2000; Daniels and Wang, 2009; Hogan, 2009). Plus précisément, soit $\boldsymbol{\varphi}$ le paramètre de $f(\mathbf{y}, u)$. Le paramètre d'intérêt $\boldsymbol{\theta}$ est en général une fonction h de $\boldsymbol{\varphi}$, c'est-à-dire $\boldsymbol{\theta} = h(\boldsymbol{\varphi})$. Les modèles de mélange de profils sont particulièrement utiles pour les analyses de sensibilité car ils peuvent être paramétrés de façon à ce que

$\boldsymbol{\varphi} = (\boldsymbol{\phi}, \boldsymbol{\kappa})$ où $\boldsymbol{\kappa}$ n'apparait pas dans le modèle pour les données observées, qui est donc indexé uniquement par $\boldsymbol{\phi}$. D'autre part, le modèle pour les réponses manquantes conditionnellement aux données observées, appelé le modèle d'extrapolation, est indexé par $(\boldsymbol{\phi}, \boldsymbol{\kappa})$:

$$f(\mathbf{y}^{\mathcal{O}}, \mathbf{y}^{\mathcal{M}}, u|\boldsymbol{\phi}, \boldsymbol{\kappa}) = f(\mathbf{y}^{\mathcal{O}}, u|\boldsymbol{\phi}) \times f(\mathbf{y}^{\mathcal{M}}|\mathbf{y}^{\mathcal{O}}, u, \boldsymbol{\phi}, \boldsymbol{\kappa}). \tag{1}$$

L'intérêt de cette paramétrisation découle des observations suivantes. Premièrement, pour une valeur fixe de $\boldsymbol{\phi}$, toute valeur de $\boldsymbol{\kappa}$ donne le même ajustement aux données observées. C'est-à-dire, la vraisemblance des données observées $\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\kappa}|\mathbf{y}^{\mathcal{O}}, u)$, considérée comme une fonction de $\boldsymbol{\kappa}$, est constante, impliquant que $\boldsymbol{\kappa}$ n'est pas identifiable. Deuxièmement, la vraisemblance $\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\kappa}|\mathbf{y}^{\mathcal{O}}, u)$, vue comme une fonction de $\boldsymbol{\phi}$, n'est pas constante, donc ce paramètre est identifiable. Enfin, le paramètre d'intérêt $\boldsymbol{\theta} = h(\boldsymbol{\phi}, \boldsymbol{\kappa})$ dépend en général de $\boldsymbol{\kappa}$. En conclusion, différentes valeurs de $\boldsymbol{\kappa}$ impliquent le même ajustement aux données observées, mais aussi des modèles d'extrapolation et des valeurs pour le paramètre d'intérêt différents. Un paramètre comme $\boldsymbol{\kappa}$ est appelé un *paramètre de sensibilité* car il représente la source des différences dans les inférences obtenues sous diverses hypothèses (non-vérifiables) sur le modèle d'extrapolation. Une analyse de sensibilité peut s'effectuer en faisant varier $\boldsymbol{\kappa}$ sur un ensemble de valeurs possibles, déterminé éventuellement avec l'aide d'experts, et en comparant les inférences obtenues sur cet ensemble et celles de l'analyse principale.

Nous considérons une famille de modèles de mélange de profils où l'on distingue uniquement entre les réponses observées et manquantes. Plus précisément, dénotant l'indicateur de manquement de Y_{ij} par $R_{ij} := I(U_i \leq j)$, on considère des modèles linéaires mixtes de la forme :

$$Y_{ij} = \mathbf{X}'_{ij}\boldsymbol{\beta} + \mathbf{Z}'_{ij}\mathbf{b}_i + \kappa R_{ij} + \varepsilon_{ij}, \qquad \varepsilon_{ij} \sim N(0, \sigma^2), \qquad \mathbf{b}_i \sim N(\mathbf{0}, \mathbf{G}), \tag{2}$$

pour $j = 1, \ldots, J$ et $i = 1, \ldots, n$, où : $\mathbf{X}_{ij}$ est un p-vecteur contenant des covariables fixes mesurées au début de l'étude et des polynômes de t_j, le temps de la mesure j ; $\boldsymbol{\beta}$ est le p-vecteur des effets fixes communs à la population ; le q-vecteur $\mathbf{Z}_{ij}$ contient les covariables dans $\mathbf{X}_{ij}$ à effet aléatoire ; les erreurs résiduelles ε_{ij} sont indépendantes et identiquement distribuées (i.i.d), suivant une distribution normale de moyenne zéro et variance σ^2 ; et les effets aléatoires, représentés par les q-vecteurs $\mathbf{b}_i$, sont i.i.d. suivant une distribution normale de moyenne zéro et variance $\mathbf{G} = \mathbf{G}(\boldsymbol{\alpha})$, où $\boldsymbol{\alpha}$ est un vecteur de paramètres inconnu. Les effets aléatoires sont supposés être indépendants des erreurs résiduelles et des covariables.

La première partie du prédicteur linéaire, $\mathbf{X}'_{ij}\boldsymbol{\beta} + \mathbf{Z}'_{ij}\mathbf{b}_i$, et les paramètres de la variance, σ^2 et $\mathbf{G}$, déterminent complètement la distribution des réponses observées, pour lesquelles $R_{ij} = 0$. Les paramètres correspondants peuvent être estimés à partir des données observées par maximum vraisemblance. D'autre part, la distribution des réponses manquantes est identifiée au paramètre κ près, qui est un paramètre de sensibilité. Donc, le modèle (2) correspond à une paramétrisation comme celle représentée dans

(1). Cette famille de modèles de mélange suppose que les réponses manquantes et observées ont la même distribution, sauf pour un décalage dans la valeur espérée quantifié par κ. L'hypothèse MAR est équivalente à l'hypothèse que ces deux distributions sont égales, ce qui se traduit par $\kappa = 0$. Donc, cette famille de modèles est 'centrée' en MAR, et κ quantifie le départ par rapport à cette hypothèse. En permettant que κ dépend des covariables, c'est-à-dire $\kappa = \kappa(\mathbf{X}_{ij})$, l'analyste peut intégrer des hypothèses plus détaillées concernant les différences entre les trajectoires des réponses observées et manquantes.

La performance de cette approche pour étudier la sensibilité des inférences a été évaluée dans une étude de simulations. Pour implémenter la méthode, une approche par imputation multiple a été développée. Dans cette approche, les réponses manquantes sont imputées en tirant des valeurs tirées directement du modèle (2), à l'aide des estimations des paramètres obtenus sous MAR et le paramètre $\kappa(\mathbf{X}_{ij})$ choisi. En prenant $\kappa = 0$, la procédure d'imputation développée offre une alternative d'analyse sous MAR par imputation multiple et a été validée dans ce contexte dans une autre étude de simulations.

L'approche proposée a été appliquée l'essai du trouble du maintien du sommeil. L'objectif de cet essai était de faire des inférences sur l'effet du traitement sur chaque score, définit comme la différence espérée entre les groupes de contrôle et de traitement dans le changement du score entre le début et la fin de l'étude. Or, 22% des individus étaient sortis de l'étude avant sa fin. Dans les essais cliniques, les sorties d'étude peuvent s'expliquer par des effets indésirables, un manque d'efficacité et des violations du protocole entre autres raisons (Molenberghs and Kenward, 2007). Nous avons utilisé la méthode proposée pour étudier la sensibilité des inférences obtenues sous MAR aux écarts par rapport à cette hypothèse. Les analyses ont confirmé un effet significatif du traitement sur les scores WASO et NAW, cette conclusion restant stable sous un large éventail de modèles d'une famille comme (2) et même lorsqu'on a considéré une autre définition de l'effet du traitement. Pour les autres scores, les résultats étaient très fragiles et sensibles aux hypothèses de manquement.

Par rapport à d'autres méthodes proposées dans la littérature (ex. Daniels and Hogan, 2000; Ratitch et al., 2013), l'approche proposée a comme avantage qu'elle peut être facilement appliquée à des données avec un grand nombre de mesures répétées, ayant lieu possiblement à des occasions différentes pour chaque individu. Celles-ci sont des avantages héritées du modèle linéaire mixte. De plus, les paramètres de sensibilité dans notre approche ont une interprétation intuitive car ils caractérisent les trajectoires des individus, ce qui facilite la formulation des hypothèses sur la distribution des réponses manquantes. Les résultats de cette partie de la thèse ont fait l'objet d'une publication (Moreno-Betancur and Chavance, 2013).

Modélisation de risques concurrents avec des causes d'évènement manquantes

Contexte et état de la question

Le modèle de *risques concurrents* est une extension du modèle standard de l'analyse de survie. Dans ce dernier, on étudie le délai de survenue d'un évènement, noté par T. Dans le modèle de risques concurrents, on distingue le type d'évènement ou la cause l'ayant entrainé, les différents types d'évènement étant mutuellement exclusifs. Les réponses d'intérêt dans ce contexte sont donc T et D, où D dénote la cause d'évènement. Sans perte de généralité, on peut supposer que D a deux catégories, une représentant la cause d'intérêt ($D = 1$) et l'autre regroupant toutes les autres causes ($D = 2$). En recherche médicale, le modèle de risques concurrents est typiquement employé lorsqu'on s'intéresse aux délais de survenue d'évènements tels que le décès par une cause spécifique, la rechute ou la réponse à un traitement. La modélisation par régression permet d'analyser l'effet des facteurs de risque ou des interventions sur la survenue d'évènements concurrents.

En général, T peut être *censuré à droite*, c'est-à-dire, pour certains individus on sait qu'ils n'ont pas subi d'évènement jusqu'à un certain moment, après lequel aucune information concernant la survenue d'un évènement n'est disponible. C'est cette particularité des délais de survenue des évènements qui a fait que l'analyse de survie se développe comme un domaine séparé de la statistique. Notamment, il existe de nombreuses méthodes pour la modélisation des risques concurrents en présence de censure à droite.

Soit $\mathbf{X}$ un p-vecteur de covariables mesurées au début de l'étude. Pour étudier l'impact de $\mathbf{X}$ sur la survenue des évènements entrainés par la cause j ($j = 1, 2$), on peut postuler des modèles de régression pour deux fonctions cause-spécifiques. La première fonction est le *risque cause-spécifique* (CSH), définit au temps t comme le taux auquel se produisent des évènements de type j parmi les individus à risque juste avant t : $\lambda_j(t) := \lim_{h\to 0} \frac{1}{h} P(t \leq T < t+h, D = j | T \geq t)$. Pour cette fonction, il est commun de considérer des modèles à risques proportionnels pour chaque cause, tel que le modèle de Cox (Cox, 1972), ou des modèles additifs tel que le modèle de Aalen (Aalen, 1980).

La deuxième fonction est appelée *la fonction d'incidence cumulée* (CIF), définie comme la probabilité d'observer un évènement de type j avant le temps t : $F_j(t) := P(T \leq t, D = j)$. Pour cette fonction, il est commun de postuler des modèles linéaires généralisés semi-paramétriques, comme le modèle de Fine et Gray (Fine and Gray, 1999) ou le modèle additif (Klein, 2006). Une modèle paramétrique a aussi été proposé (Jeong and Fine, 2007). L'incidence cumulée d'une cause donnée dépend des deux fonctions de risque, λ_1 et λ_2. Donc, une augmentation du risque d'une cause ne reflètera pas force-

ment une augmentation de l'incidence cumulée de cette cause (Beyersmann et al., 2007). C'est pourquoi, les deux fonctions doivent être étudiées pour comprendre le mécanisme de risques concurrents entièrement (Andersen et al., 2012; Latouche et al., 2013).

Les méthodes pour modéliser le CSH ou la CIF supposent en général que la cause d'évènement D est connue pour tous les individus ayant subi un évènement, ce qui n'est pas toujours le cas. Par exemple, si l'évènement étudié est le décès, quelques causes de décès peuvent manquer lorsque les certificats de décès ne sont pas bien remplis ou les patients meurent sans autopsie (Andersen et al., 1996; Manola and Gray, 2011). Certains auteurs ont proposé des méthodes de régression gérant les causes d'évènement manquantes sous l'hypothèse MAR, notamment pour la modélisation semi-paramétrique du CSH (Goetghebeur and Ryan, 1995; Andersen et al., 1996; Nicolaie et al., 2011; Lu and Tsiatis, 2001; Gao and Tsiatis, 2005; Gao, 2006; Lu and Liang, 2008). Cependant, d'autres modèles fréquemment employés n'ont été que peu ou pas du tout considérés. Par exemple, la modélisation semi-paramétrique de la CIF n'a été considéré que par Bakoyannis et al. (2010) qui a étudié le modèle de Fine et Gray avec une approche par imputation multiple. D'autre part, les modèles paramétriques pour le CSH et la CIF n'ont jamais été considérés dans ce contexte, et il est de même pour la modélisation sous MNAR et les analyses de sensibilité.

Dans la littérature, un point de départ pour le développement de méthodes de régression gérant les causes d'évènement manquantes a été un essai clinique randomisé de l'Eastern Cooperative Oncology Group (ECOG). Dans cet essai, 169 femmes âgées avaient été recrutées pour étudier un traitement du cancer du sein de stade II. Après un suivi médiane de 6.7 années, 79 femmes étaient décédées, dont 44 décès par cancer et 17 décès par autres causes. Pour les autres 18 femmes décédées (23%), la cause de décès était manquante, notamment parce que le schéma de l'étude n'était pas adapté à un suivi à long terme (Cummings et al., 1993). Les premières analyses de cet essai, effectuées par Cummings et al. (1985) et Cummings et al. (1986), n'ont pas montré d'effet significatif du traitement sur la survie, mais ont révélé des associations significatives entre la survie et deux facteurs pronostiques : le statut des récepteurs d'estrogène (positive ou négative) et le nombre de ganglions lymphatiques atteints par le cancer (inférieur à 4, ou supérieur ou égale à 4). Étant donné l'importance des causes de décès concurrentes dans cette étude et du haut pourcentage de causes manquantes, ces données on été de grande valeur dans la littérature pour illustrer de nouvelles méthodes pour gérer les causes manquantes (Goetghebeur and Ryan, 1995; Lu and Tsiatis, 2001; Gao, 2006; Nicolaie et al., 2011). Ces auteurs ont notamment étudié les effets des deux facteurs pronostiques mentionnés sur le décès par cancer. Dans cette thèse, nous avons utilisé ces données pour illustrer les méthodes proposées.

Contributions

Une première contribution, répondant au second objectif de la thèse, a été de proposer un cadre général pour la modélisation semi-paramétrique de la CIF sous l'hypothèse MAR. On considère le modèle linéaire généralisé semi-paramétrique suivant :

$$g\{F_j(t|\mathbf{X})\} = \beta_{j0}(t) + \boldsymbol{\beta}_j'\mathbf{X}, \qquad j = 1, 2, \tag{3}$$

où g est une fonction de lien monotone et différentiable et $\beta_{j0}(t)$ est un intercepte dépendant du temps non-spécifié. Le modèle (3) comprend des modèles tels que le modèle de Fine et Gray lorsque g est la fonction cloglog (*complementary log-log*) et le modèle additif lorsque g est la fonction identité. Lorsque toutes les causes d'évènement sont observées, ce modèle peut être ajusté avec la méthode Andersen-Klein (Andersen et al., 2003; Klein and Andersen, 2005). Cette méthode consiste à utiliser des pseudo-valeurs de jackknife de la CIF comme les réponses dans des équations d'estimation généralisées (Liang and Zeger, 1986). Nous avons proposé deux extensions de cette méthode pour gérer les causes d'évènement manquantes. La première extension consiste à utiliser des pseudo-valeurs pondérés par l'inverse de la probabilité que la cause d'évènement soit observée. Nous avons démontré que cette approche amène à des estimateurs consistants et asymptotiquement normaux sous l'hypothèse MAR. De plus, cette approche permet de prendre en compte toute l'information partielle disponible sur les individus avec cause manquante dans les équations d'estimation, impliquant des estimateurs efficaces. La deuxième extension s'agit de l'application de la procédure d'imputation multiple de Bakoyannis et al. (2010) pour imputer les causes manquantes, suivi de l'analyse par la méthode Andersen-Klein des jeux de données complétés. Ces deux méthodes ont été validées et comparées à l'analyse CC dans une étude de simulations, et ensuite appliqués à l'essai ECOG. Nous avons ainsi pu formuler des recommandations sur leur application dans la pratique. Ces travaux ont fait l'objet d'une publication (Moreno-Betancur and Latouche, 2013).

Une deuxième contribution pour la modélisation des risques concurrents avec causes d'évènement manquantes a été d'étudier formellement le concept de l'*ignorabilité* du mécanisme de manquement dans ce contexte. Notamment, nous avons déterminé des conditions suffisantes sur les mécanismes de manquement, de censure et des risques concurrents pour que le mécanisme manquement puisse être ignoré dans une analyse par maximum de vraisemblance. Nous avons ainsi déduit des expressions pour la vraisemblance en fonction du CSH et de la CIF permettant d'ajuster des modèles paramétriques sous ces conditions. Les modèles paramétriques n'avaient pas été considérés dans la littérature de causes manquantes auparavant. Pour le CSH, les expressions proposées permettent d'ajuster des modèles à risques proportionnels complètement paramétriques :

$$\lambda_j(t|\mathbf{X}) = \lambda_{j0}(\boldsymbol{\alpha}_j, t)\exp(\boldsymbol{\beta}_j'\mathbf{X}), \qquad j = 1, 2$$

où le risque de base λ_{j0} est connu au vecteur de paramètres $\boldsymbol{\alpha}_j$ près. Pour la CIF, le

modèle de Jeong and Fine (2007) peut être ajusté :

$$F_j(t|\mathbf{X}) = 1 - \left[1 + \alpha_j \exp(\boldsymbol{\beta}_j'\mathbf{X})\tau_j\{\exp(\rho_j t) - 1\}/\rho_j\right]^{-1/\alpha_j}, \qquad j = 1, 2.$$

La troisième contribution pour cette partie de la thèse a été le développement d'une méthodologie pour effectuer des analyses de sensibilité dans le contexte de causes manquantes. Nous avons repris les idées développées dans le cadre de données longitudinales avec des sorties d'étude. Plus précisément, nous avons considéré une famille de modèles de mélange de profils indexés par un paramètre de sensibilité. Pour modéliser sous MNAR dans ce contexte, on doit considérer la distribution jointe de (T, D, M) pour les individus non-censurés, où M est l'indicateur de manquement, c'est-à-dire, $M = 1$ si la cause est manquante et $M = 0$ sinon. Dénotant par U l'indicateur de censure, avec $U = 1$ si l'individu est censuré et $U = 0$ sinon, la factorisation par mélange de profils s'écrit

$$\begin{aligned} f(t, d, m \mid U = 0) &= f(t, d \mid M = m, U = 0) \times f(m \mid U = 0) \\ &= f(t, m \mid U = 0) \times f(d \mid T = t, M = m, U = 0). \end{aligned} \tag{4}$$

Ici, le modèle d'extrapolation est représenté par le dernier facteur dans (4), $f(d \mid T = t, M = m, U = 0)$. Nous avons considéré des modèles de mélange de profils où le modèle d'extrapolation a la forme suivant :

$$\text{logit}\{\Pi(\mathbf{X}, T, M)\} = \mathbf{h}(\mathbf{X}, T)'\boldsymbol{\gamma} + \kappa M, \tag{5}$$

où $\Pi(\mathbf{X}, T, M) := P(D = 1|\mathbf{X}, T, M, U = 0)$ et $\mathbf{h}(\mathbf{X}, T)$ est un vecteur incluant $\mathbf{X}$, T, et possiblement des interactions et des polynômes de T de plus grand ordre. La première partie du prédicteur linéaire, $\mathbf{h}(\mathbf{X}, T)'\boldsymbol{\gamma}$, détermine complètement la distribution des causes d'évènement pour les individus avec cause observée, pour qui $M_i = 0$. Le paramètre $\boldsymbol{\gamma}$ peut être estimé en ajustant le modèle logistique aux données de ces individus. D'autre part, la distribution des causes d'évènement pour les individus avec cause manquante est identifiée au paramètre κ près, qui est un paramètre de sensibilité. En effet, ce paramètre n'est pas identifiable à partir des données observées.

Le paramètre κ quantifie une différence entre les distributions des causes d'évènements des individus avec cause observée et manquante, ainsi que le départ par rapport à l'hypothèse MAR, celle-ci étant équivalent à la condition $\kappa = 0$. Plus précisément, κ est le logarithme du odds ratio ajusté, comparant les chances d'un évènement par la cause d'intérêt ($D = 1$) entre les individus non-censurés avec cause manquante et ceux avec cause observée. En permettant que κ varie sur un ensemble de valeurs plausibles, on peut évaluer la sensibilité des inférences aux écarts par rapport à MAR. Pour des hypothèses plus détaillées on peut permettre que κ dépend des covariables et du délai de survie, c'est-à-dire $\kappa = \kappa(\mathbf{X}, T)$. De même que pour le données longitudinales, nous proposons une implémentation de cette procédure par imputation multiple. D'abord, les

cause manquantes sont imputées en tirant des valeurs du modèle (5) en utilisant l'estimation de $\boldsymbol{\gamma}$ obtenue sous MAR et le $\kappa(\mathbf{X}, T)$ choisi. Ensuite, la méthode de régression souhaitée est utilisée pour analyser chaque jeu de données complété. Ainsi, cette méthode d'analyse de sensibilité est applicable à divers modèles pour risques concurrents, y compris des méthodes paramétriques ou semi-paramétriques pour le CSH et la CIF.

Les premiers résultats d'application de cette méthode à l'essai ECOG ont été encourageants. A la base, cette méthode avait été développée pour une étude épidémiologique sur les différentiels socioéconomiques dans la mortalité par suicide en France. On soupçonnait que la base de causes de décès de cette étude contenait de nombreux suicides codés comme décès à cause manquante suite à un problème de transmission des informations sur les suicides dans la région de l'étude. Ceci remettait en cause l'hypothèse MAR. Or, avec notre méthode nous avons trouvé que les résultats étaient peu sensibles aux hypothèses de manquement, sans doute grâce au faible pourcentage de causes manquantes dans cette étude (environ 10%).

Le projet sur les différentiels socioéconomiques dans la mortalité par suicide en France nous a amené à étudier en détail les indices utilisés dans la littérature pour mesurer ces différentiels à partir de données de survie et de risques concurrents. Un manuscrit sur cette problématique, externe à celle des données manquantes, est en préparation.

Références

Aalen, O. (1980). A model for nonparametric regression analysis of counting processes. In *Mathematical statistics and probability theory*, pages 1–25. Springer, New York.

Andersen, J., Goetghebeur, E., and Ryan, L. (1996). Missing cause of death information in the analysis of survival data. *Statistics in Medicine*, 15 :2191–2201.

Andersen, P. K., Geskus, R. B., de Witte, T., and Putter, H. (2012). Competing risks in epidemiology : possibilities and pitfalls. *International Journal of Epidemiology*, 41(3) :861–870.

Andersen, P. K., Klein, J. P., and Rosthoj, S. (2003). Generalised linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika*, 90(1) :15–27.

Bakoyannis, G., Siannis, F., and Touloumi, G. (2010). Modelling competing risks data with missing cause of failure. *Statistics in Medicine*, 29(30) :3172–3185.

Beyersmann, J., Dettenkofer, M., Bertz, H., and Schumacher, M. (2007). A competing risks analysis of bloodstream infection after stem-cell transplantation using subdistribution hazards and cause-specific hazards. *Statistics in Medicine*, 26(30) :5360–5369.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, 19(1) :15–18.

Cook, R. D. (1986). Assessment of local influence. *Journal of the Royal Statistical Society. Series B (Methodological)*, 48(2) :133–169.

Cox, D. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2) :187–220.

Cummings, F. J., Gray, R., Davis, T. E., Tormey, D. C., Harris, J. E., Falkson, G., and Arseneau, J. (1985). Adjuvant tamoxifen treatment of elderly women with stage II breast cancer : A double-blind comparison with placebo. *Annals of Internal Medicine*, 103(3) :324–329.

Cummings, F. J., Gray, R., Davis, T. E., Tormey, D. C., Harris, J. E., Falkson, G. G., and Arseneau, J. (1986). Tamoxifen versus placebo : double-blind adjuvant trial in elderly women with stage II breast cancer. *NCI monographs : a publication of the National Cancer Institute*, (1) :119–123.

Cummings, F. J., Gray, R., Tormey, D. C., Davis, T. E., Volk, H., Harris, J., Falkson, G., and Bennett, J. M. (1993). Adjuvant tamoxifen versus placebo in elderly women with node-positive breast cancer : long-term follow-up and causes of death. *Journal of Clinical Oncology*, 11(1) :29–35.

Daniels, M. J. and Hogan, J. W. (2000). Reparameterizing the pattern mixture model for sensitivity analyses under informative dropout. *Biometrics*, 56(4) :1241–1248.

Daniels, M. J. and Wang, C. (2009). Comments on : Missing data methods in longitudinal studies : a review. *TEST : An Official Journal of the Spanish Society of Statistics and Operations Research*, 18(1) :51–58.

Fine, J. P. and Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94(446) :496–509.

Gao, G. Z. (2006). *Semiparametric Estimators for the Regression Coefficients in the Linear Transformation Competing Risks Models with Missing Cause of Failure*. PhD thesis, North Carolina State University, Raleigh.

Gao, G. Z. and Tsiatis, A. A. (2005). Semiparametric estimators for the regression coefficients in the linear transformation competing risks model with missing cause of failure. *Biometrika*, 92(4) :875–891.

Goetghebeur, E. and Ryan, L. (1995). Analysis of competing risks survival data when some failure types are missing. *Biometrika*, 82(4) :821–833.

Hogan, J. (2009). Comments on : Missing data methods in longitudinal studies : a review. *TEST : An Official Journal of the Spanish Society of Statistics and Operations Research*, 18(1) :59–64.

Jansen, I., Hens, N., Molenberghs, G., Aerts, M., Verbeke, G., and Kenward, M. G. (2006). The nature of sensitivity in monotone missing not at random models. *Computational Statistics & Data Analysis*, 50(3) :830–858.

Jeong, J. H. and Fine, J. P. (2007). Parametric regression on cumulative incidence function. *Biostatistics*, 8(2) :184–196.

Kenward, M. G. (1998). Selection models for repeated measurements with non-random dropout : an illustration of sensitivity. *Statistics in Medicine*, 17(23) :2723–2732.

Kenward, M. G. and Molenberghs, G. (1999). Parametric models for incomplete continuous and categorical longitudinal data. *Statistical Methods in Medical Research*, 8 :51–83.

Klein, J. P. (2006). Modelling competing risks in cancer studies. *Statistics in Medicine*, 25 :1015–1034.

Klein, J. P. and Andersen, P. K. (2005). Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. *Biometrics*, 61 :223–229.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4) :963–974.

Latouche, A., Allignol, A., Beyersmann, J., Labopin, M., and Fine, J. P. (2013). A competing risks analysis should report results on all cause-specific hazards and cumulative incidence functions. *Journal of Clinical Epidemiology*, 66(6) :648–653.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1) :13–22.

Little, R. J. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika*, 81(3) :471–483.

Little, R. J. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90(431) :1112–1121.

Little, R. J. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics. Wiley, New York, 1st edition.

Little, R. J. and Yau, L. (1996). Intent-to-treat analysis for longitudinal studies with drop-outs. *Biometrics*, 52(4) :1324–1333.

Lu, K. F. and Tsiatis, A. A. (2001). Multiple imputation methods for estimating regression coefficients in the competing risks model with missing cause of failure. *Biometrics*, 57(4) :1191–1197.

Lu, W. and Liang, Y. (2008). Analysis of competing risks data with missing cause of failure under additive hazards model. *Statistica Sinica*, 18 :219–234.

Manola, J. B. and Gray, R. J. (2011). When bad things happen to good studies. *Journal of Clinical Oncology*, 29(26) :3497–3499.

Michiels, B., Molenberghs, G., Bijnens, L., Vangeneugden, T., and Thijs, H. (2002). Selection models and pattern-mixture models to analyse longitudinal quality of life data subject to drop-out. *Statistics in Medicine*, 21 :1023–1041.

Molenberghs, G., Beunckens, C., Sotto, C., and Kenward, M. G. (2008). Every missingness not at random model has a missingness at random counterpart with equal fit. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 70(2) :371–388.

Molenberghs, G. and Kenward, M. G. (2007). *Missing data in clinical studies*. Wiley, Chichester.

Molenberghs, G., Kenward, M. G., and Goetghebeur, E. (2001a). Sensitivity analysis for incomplete contingency tables : the Slovenian plebiscite case. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 50(1) :15–29.

Molenberghs, G., Verbeke, G., Thijs, H., Lesaffre, E., and Kenward, M. G. (2001b). Influence analysis to assess sensitivity of the dropout process. *Computational Statistics & Data Analysis*, 37 :93–113.

Moreno-Betancur, M. and Chavance, M. (2013). Sensitivity analysis of incomplete longitudinal data departing from the missing at random assumption : Methodology and application in a clinical trial with drop-outs. *Statistical Methods in Medical Research [Epub ahead of print on May 22, 2013]*.

Moreno-Betancur, M. and Latouche, A. (2013). Regression modeling of the cumulative incidence function with missing causes of failure using pseudo-values. *Statistics in Medicine*, 32(18) :3206–3223.

Nicolaie, M. A., van Houwelingen, H. C., and Putter, H. (2011). Vertical modeling : Analysis of competing risks data with missing causes of failure. *Statistical Methods in Medical Research [Epub ahead of print on December 16, 2011]*.

Ratitch, B., O'Kelly, M., and Tosiello, R. (2013). Missing data in clinical trials : from clinical assumptions to statistical analysis using pattern mixture models. *Pharmaceutical statistics [Epub ahead of print on January 4, 2013]*.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429) :106–121.

Rotnitzky, A., Robins, J. M., and Scharfstein, D. O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association*, 93(444) :1321–1339.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3) :581–592.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, Ltd., New York.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448) :1096–1120.

Thijs, H., Molenberghs, G., and Verbeke, G. (2000). The milk protein trial : influence analysis of the dropout process. *Biometrical Journal*, 42(5) :617–646.

Verbeke, G., Molenberghs, G., Thijs, H., Lesaffre, E., and Kenward, M. G. (2001). Sensitivity analysis for nonrandom dropout : a local influence approach. *Biometrics*, 57(1) :7–14.

Wu, M. C. and Bailey, K. R. (1988). Analysing changes in the presence of informative right censoring caused by death and withdrawal. *Statistics in Medicine*, 7 :337–346.

Wu, M. C. and Bailey, K. R. (1989). Estimation and comparison of changes in the presence of informative right censoring : conditional linear model. *Biometrics*, 45(3) :939–955.

Wu, M. C. and Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, 44(1) :175–188.