Inférence statistique dans le modèle de mélange à risques proportionnels

Rim Ben Elouefi

Thèse

UNIVERSITE BRETAGNE LOIRE

THESE INSA Rennes
sous le sceau de l'Université Bretagne Loire
pour obtenir le titre de
DOCTEUR DE L'INSA RENNES
Spécialité : Mathématiques et leurs intéractionstions

présentée par

RIM BEN ELOUEFI

ECOLE DOCTORALE : *Matstic / EDSTI*
LABORATOIRE : *IRMAR*

Inférence statistique dans des modèles de mélange à risques proportionnels

Thèse soutenue le 05.09.2017
devant le jury composé de :

Célestin KOKONENDJI
Professeur, université de Besançon (France) / Président
Natacha HEUTTE
Professeur, université de Rouen (France)/ Rapporteure
Sophie DABO
Professeur, université de Lille 3 (France) / Rapporteure
Pierrette CHAGNEAU
Maitre de conférence, INSA de Rennes (France) / Examinatrice
Jean François DUPUY
Professeur, INSA de Rennes (France) / Co-directeur de thèse
Said ZARATI
Professeur, université de Tunis Manar (France) / Co-directeur de thèse

Inférence statistique dans des modèles de mélange à risques proportionnels

Rim BEN ELOUEFI

En partenariat avec

THESE DE DOCTORAT

Institut National des Sciences Appliquées de Rennes, France

Et Faculté des Sciences de Monastir, Tunisie

Discipline : Mathématiques et leurs intéractions

Spécialité : Statistique

présentée par

Rim BEN ELOUEFI

Inférence statistique dans des modèles de mélange à risques proportionnels

---

Soutenue devant le jury composé de Mesdames et Messieurs les Professeurs :

KOKONENDJI Célestin	Université de Besançon	Président
DABO Sophie	Université de Lille 3	Rapporteure
HEUTTE Natacha	Université de Rouen	Rapporteure
CHAGNEAU Pierrette	INSA de Rennes	Examinatrice
DUPUY Jean-François	INSA de Rennes	Co-directeur
ZARATI Said	Université de Tunis El Manar	Co-directeur

Dédicace

À ma chère maman.
À mes chers frères.
À mes anges Adem et Mariem.
À l'âme de mon père.

Remerciements

Je tiens en premier lieu à remercier Jean-François DUPUY de m'avoir offert la possibilité de commencer cette thèse et qui a accepté de superviser mes premiers pas dans la recherche. Mes chaleureux remerciements à Samir BEN AMMOU de m'avoir accompagnée au cours de ces quatre années, aussi pour ses conseils et ses encouragements.

Je désire témoigner de ma reconnaissance envers Said ZARATI, pour la constance de son soutien et la justesse des orientations qu'il a voulu me suggérer.

Je tiens également à remercier Sophie DABO et Natacha HEUTTE qui ont accepté d'être les rapporteures de ce travail de thèse. Je remercie également Célestin KOKONENDJI et Pierrette CHAGNEAU d'avoir accepté d'examiner la thèse et d'être parti du jury.

Je remercie aussi tous les membres et personels du département de mathématiques de L'INSA.

Je souhaite enfin exprimer toute ma reconnaissance à ma mère et mes frères pour m'avoir aider si patiemment à réaliser ce rêve, je pense aussi à Ã mes amis Houssem, Foued SAADAOUI et oncle Rafik qui ont été d'un grand secours pour venir à bout de certains obstacles.

Abréviations & Notations

Abréviation et notations

- Les variables aléatoires seront définies sur un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$.
- $\mathbb{N}$: ensemble des entiers naturels.
- $\mathbb{R}$: ensemble des nombres réels.
- $\mathbb{R}^+$: ensemble des nombres réels positifs.
- $\mathbb{R}^p$: espace vectoriel de dimension p sur $\mathbb{R}$.
- $\mathbb{P}$: probabilité.
- $\mathbb{E}$: espérance.
- $\mathbb{V}$ar: variance.
- $\mathbb{C}$ov: covariance.
- $\mathbb{E}(X|Y)$: espérance conditionnelle de X sachant Y.
- $a \wedge b$: minimum de a et b.
- $a \vee b$: maximum de a et b.
- $1_{\{A\}}$: indicatrice de A.
- v.a: variable aléatoire.
- p.s: presque sûrement.
- $\mathcal{F}_t$: filtration à l'instant t.
- $^\top$: transposé du vecteur ligne $(x_1, \cdots, x_n)$.
- $\otimes$: produit direct.
- Si a est un vecteur de dimension p, $a^{\otimes 0} = 1$, $a^{\otimes 1} = a$ et $a^{\otimes 2} = aa^\top$.
- $\rightarrow^{\mathbb{P}}$: convergence en probabilité.
- $\rightarrow^{\mathcal{L}}$: convergence en loi.
- i.i.d: identiquement et indépendamment distribuées.
- càdlàg: continue à droite avec une limite à gauche.

Table des matières

Publications et Conférences

Articles

- Ben Elouefi, R. A goodness-of-fit test for the stratified proportional hazards model for survival data. Soumis.
- Ben Elouefi, R., Dupuy, J.-F. Nonparametric inference in stratified survival samples with missing data. Soumis.

Conférences

- Ben Elouefi R., Dupuy J.-F. Two problems in the stratified proportional hazards model: goodness-of-fit testing and stratification according to an unknown thresold. 6th International Conference on Accelerated Life Testing and Degradation Models, Troyes, 2016.
- Ben Elouefi R., Dupuy J.-F. A goodness-of-fit test for the stratified proportional hazards model for survival data. 8th International Conference of the ERCIM Working Group on Computational and Methodological Statistics, Londres, 2015.
- Ben Elouefi R., Dupuy J.-F. Inférence statistique dans le modèle stratifié de risque proportionnel: test d'ajustement. 22ème colloque de la société mathématique de la Tunisie, Mahdia, 2017.
- Ben Elouefi R., Dupuy J.-F. Nonparametric inference in stratified survival samples with missing data. Colloque International de Statistique et Econométrie Mahdia, Tunisie 2017.

Table des figures

Liste des tableaux

Introduction générale

L'Analyse statistique des durées de vie (ou durées de survie) s'est initialement développée dans le domaine des sciences médicales et biologiques. Elle est également largement utilisée en sciences économiques et sociales, ainsi que dans l'industrie. L'analyse des durées de vie restera longtemps un problème étudié par les démographes et les actuaires. Elle remonte à l'école anglaise au **XVII**^{e} siècle. L'objectif des analystes de ce siècle est l'étude de la mortalité ainsi que des autres caractéristiques de la population. À partir du **XIX**^{e} siècle, et avec l'apparition des catégorisations suivant des variables telles que le sexe, la nationalité, les catégories professionnelles, apparaissent les premières modélisations de la probabilité d'un événement, désignée sous le terme de "fonction de risque".

Durant le **XX**^{e} siècle, l'analyse des durées de vie commence à irriguer de nombreuses autres disciplines: l'actuariat, la physique, l'industrie, la médecine.
En 1951, Weibull s'intéresse à l'une de particularités importantes des durées de vie, à savoir la présence de données tronquées ou censurées. De plus, il propose une nouvelle distribution de probabilité qui sera par la suite fréquemment utilisée en analyse de survie: la "loi de Weibull".
Deux autres dates importantes doivent êtres citées: en 1958 Kaplan et Meier ont proposé d'utiliser un estimateur non-paramétrique de la fonction de survie en présence de censure.
En 1972, Cox a publié un article posant les bases d'un cas particulier important du modèle à "risques proportionnels". Ce modèle de référence a donné lieu à des nombreux développements et extensions: introduction de coefficient de régression dépendant du temps, prise en compte de dépendances entre les variables observées, stratification de l'effet de covariables, etc.

Dans cette thèse, nous nous intéressons à une généralisation du modèle à risques proportionnels, dite modèle à risques proportionnels stratifié. Ce modèle permet à dif-

férents groupes de la population, appelés strates, de posséder des fonctions d'intensité de base distinctes tandis que la valeur du paramètre de régression est commune. Ce modèle est plus en plus utilisé dans de nombreux domaines, tels que l'économie, la médecine et la santé publique. Par exemple, Kiefer et Larson (2015, [32]) ont utilisé ce modèles pour évaluer les prêts hypothécaires résidentiels au Etats-Unis. Haenlein (2013, [23]) l'a utilisé pour identifier les facteurs de risque d'attrition parmi les clients d'un opérateur de téléphonie mobile... Face à l'utilisation de plus en plus répondre de ce modèle, la construction de tests d'ajustement qui lui soient spécifiques est devenue un problème important. Ce domaine de recherche a suscité une littérature abondante dans le cas du modèle de Cox non stratifié. Citons par exemple: Gandy et Jensen (2009, [19]), Lin, Wei et Ying (1993, [37]), Marzec et Marzec (1997, [43]), Verweij, Houwelingen et Stijnen (1998, [66]). La construction d'un test pour le modèle à risques proportionnels stratifié est l'objet de la première partie de ce travail. On propose une statistique de test d'ajustement pour le modèle à risques proportionnels stratifié défini par la fonction de risque

$$\alpha_j(t) = \alpha_{0,j}(t)\exp(\beta^\top \mathbf{Z}), \tag{1}$$

où $\mathbf{Z}$ est un vecteur de variables explicatives (covariables), β est le vecteur de paramètre de régression inconnu et $\{\alpha_{0,j}(t); t \geq 0, j = 1, \cdots, J\}$ sont les fonctions de risque de base dans les J strates. Pour cela, nous considérons le processus suivant:

$$Q^j_{n,\mathbf{z}}(t, \widehat{\beta}_n) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widehat{M}_i(t, \widehat{\beta}_n) 1_{\{S_i = j\}} 1_{\{\mathbf{Z}_i \leq \mathbf{z}\}}, \tag{2}$$

où $\widehat{M}_i$ sont les résidus de martingale obtenus à partir d'un échantillon d'observations indépendantes $(X_i, \Delta_i, \mathbf{Z}_i, S_i)$, où $X = \min(T, C)$ désigne la durée observée, C le temps de censure, $\Delta = 1_{\{T \leq \min(C,\tau)\}}$ l'indicatrice de censure et S l'indicatrice des strates. Nous établissons rigoureusement la loi limite de la statistique de test proposée, sous l'hypothèse nulle d'un ajustement correct du modèle aux données. Nous réalisons une étude de simulations pour évaluser les propriétés (niveau, puissance) de cette statistique pour différentes alternatives, tailles d'échantillon, proportions de censure. L'ensemble de ce travail est décrit dans le chapitre 3 de ce manuscrit.

A partir du test d'ajustement proposé, nous proposons une procédure numérique permettant de stratifier le modèle à risques proportionnels suivant le seuil inconnu ω d'un variable quantitative, dite de stratification. Nous consiérons le modèle:

$$\alpha_{0,1}(t)\exp(\beta^\top \mathbf{Z}) 1_{\{W \leq \omega\}} + \alpha_{0,2}(t)\exp(\beta^\top \mathbf{Z}) 1_{\{W > \omega\}}, \tag{3}$$

où $\alpha_{0,1}(t)$ et $\alpha_{0,2}(t)$ sont des fonctions de risque de base inconnues et β le paramètre de régression inconnu. Nous évaluons cette procédure au moyen de simulations. Ce travail est décrit dans le chapitre 3 de ce manuscrit.

Dans des nombreuses situations, la strate S ne peut être observée chez tous les sujets suivis. L'inférence statistique dans le modèle à risques proportionnels startifié se heurte alors à un problème de données manquantes. Parmi les approches qui ont été proposées pour résoudre ce problème, citons à titre d'exemple Dupuy et Leconte (2009, [16]) qui ont proposé une version modifiée de la vraisemblance partielle basée sur la méthode " Régression Calibration" et Dupuy et Détais (2008, [14]) qui ont utilisé l'algorithme Espérance-Maximisation (EM). Afin d'estimer les paramètres d'intérêt de ce modèle à données manquantes, une autre solution consiste à mener une analyse en "cas complets" ("CC" par la suite) c'est-à-dire, à: i) retirer de l'échantillon les individus pour lesquels S est manquante, ii) réaliser l'inférence sur les individus restants. Cette solution n'est pas satisfaisante puisqu'elle entraîne une perte d'information. Robin, Lipshitz, Harrington et Pugh (1980, [50]) ont proposé la méthode d'estimation dite "Pondération par probabilité inverse" (ou "Inverse Probability Weighting", "IPW" par la suite). Nous nous appuyons sur ce principe pour construire un nouveau test du logrank stratifié en présence de données manquantes. Ce travail est décrit dans le chapitre 4 de ce manuscrit.
Plus précisément, nous proposons une version modifiée de l'estimateur de Nelson-Aalen, puis nous construisons un nouveau test du logrank stratifié, adapté au cas de données manquantes, en utilisant le principe IPW.
Nous en étudions la distribution asymptotique de la statistique sous l'hypothèse nulle de l'égalité inter-groupes des distributions de survie. Nous comparons cette nouvelle statistique à celle basée sur l'analyse en CC, au moyen de simulations. Nous réalisons une application sur des données médicales réelles.

1

Analyse statistique des durées de survie

Résumé

Dans ce chapitre, nous présentons les différents outils de modélisation utilisés dans l'analyse des durées de vie. Nous rappelons également quelques notions probabilistes utile dans ce domaine.

Sommaire

1.1 Motivation

LE terme "durées de vie" est employé de manière générale pour désigner le temps écoulé jusqu'à la survenue d'un événement précis. L'événement étudié (communément appelé "décès") est le passage irréversible entre deux états (communément nommés "vivant" et "décès"), survenant après une durée appelée "durée de survie".
L'événement terminal n'est pas forcément la mort: il peut s'agir de l'apparition d'une maladie, le temps avant une rechute ou le rejet d'une greffe, la panne d'une machine (durée de fonctionnement des composantes éléctroniques en fiabilité industrielle) ou la survenue d'un sinistre (temps entre deux sinistres, en actuariat).
L'analyse des durées (données) de survie est l'étude du délai de la survenue de cet événement. Dans le domaine biomédical, on étudie ces durées dans le contexte des études longitudinales comme les enquêtes de cohorte (suivi de patient dans le temps) ou les essais thérapeutiques (pour tester l'efficacité d'un médicament). Nous cherchons alors à estimer les fonctions de survie de plusieurs groupes ou à analyser la manière dont les variables explicatives modifient les fonctions de survie.

Le premier chapitre expose, les définitions, les notations et les outils de modélisation utilisés dans la théorie des durées de vie, tels que les processus stochastiques, les martingales, les processus de comptage et les processus empiriques.

1.2 Modèles de durée de vie

Quelques définitions sont couramment utilisées dans les études des durées de survie:

- Date d'origine qui correspond à l'origine de la durée étudiée. Elle peut être la date de naissance pour la mesure d'un âge, date d'une opération chirurgicale (l'origine du temps diffère d'une personne à une autre).

- Date de point ou fin d'étude: c'est la date au delà de laquelle on arrêtera l'étude et on ne tiendra plus compte des informations sur les sujets.

- Date des dernières nouvelles: c'est la date la plus récente où des informations sur un sujet ont été récueillies.

1.2.1 Définitions et propriétés

Nous donnons ci-dessous les définitions des principaux outils classiques de modélisation utilisés en analyse de survie.
Nous désignons par T une variable aléatoire positive absolument continue définie sur un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$ et représentant une durée jusqu'à la survenue d'un événement d'intérêt. Par la suite, la durée T sera appelée durée de survie. L'interprétation d'une variable aléatoire en terme de durée de survie conduit à définir d'autres fonctions que la fonction de répartition, définie par $F(t) = \mathbb{P}(T \leq t)$, si $t \geq 0$ et $F(t) = 0$ si $t < 0$. La fonction de répartition F est croissante, continue à droite limitée à gauche (cadlàg), avec $\lim_{t\to\infty} F(t) = 1$. Elle représente la probabilité pour que l'événement d'intérêt survienne pendant l'intervalle de temps $[0, t]$.

Notons que chaque individu est susceptible de subir une seule fois l'événement d'intérêt. De plus, l'observation de la survenue ou non de cet événement constitue la donnée fondamentale pour une modélisation de durée de survie. La loi de T peut également être caractérisée par d'autres fonctions à interpréter facilement.

Définition 1 *On appelle fonction de survie S la probabilité de survivre au-delà de l'instant t:*

$$\forall t \geq 0, S(t) = \mathbb{P}(T > t) = 1 - F(t).$$

Par définition, S est décroissante, continue à droite et limitée à gauche.

Si T admet une densité de probabilité $f(t) \geq 0$ par rapport à la mesure de Lebesgue sur $\mathbb{R}^+$ alors

$$\forall t \geq 0, F(t) = \int_0^t f(u)du.$$

Si la fonction de répartition F admet une dérivée au point t alors

$$f(t) = F'(t) = -S'(t).$$

Définition 2 *On appelle fonction de risque instantané λ la fonction définie pour t dans $\mathbb{R}^+$ par*

$$\forall t \geq 0, \lambda(t) = \lim_{h\to 0} \frac{1}{h}\mathbb{P}(t < T \leq t + h | T \geq t).$$

Supposons que T soit une variable continue, on a alors

$$\forall t \in \mathbb{R}, \lambda(t) = \frac{f(t)}{S(t)} = -\frac{\partial}{\partial t} \ln(S(t)).$$

Par convention $c/0 = +\infty, \forall c \geq 0$.

De façon heuristique, la fonction de risque instantané représente la probabilité de subir l'événement à l'instant t sachant que l'on n'a pas encore subi cet événement à l'instant $t-$ i.e. juste avant l'instant t. Par conséquent, la fonction de risque instantané traduit l'évolution dans le temps du risque de survenue de l'événement. C'est ainsi que l'allure de cette fonction apporte une information immédiate sur les caractéristiques du sujet étudié.

Définition 3 *On appelle fonction de risque cumulé Λ la fonction définie pour t dans $\mathbb{R}^+$ par*

$$\Lambda(t) = \int_0^t \lambda(s)ds = -\ln(S(t)),$$

qui vaut $+\infty$ quand $S(t) = 0$;

On peut déduire de cette équation une expression de la fonction de survie à l'aide de la fonction de risque cumulé (ou risque instantané). La définition de la distribution de probabilité de T repose de manière équivalente sur sa densité, sa fonction de risque cumulé, sa fonction de risque instantané.
En effet, les relations suivantes entre les différentes fonctions peuvent être établies, pour $t \geq 0$

$$\begin{aligned}
\lambda(t) &= \frac{f(t)}{S(t)} = -\frac{d}{dt} \log(S(t)), \\
\Lambda(t) &= \int_0^t \lambda(u)du = -\log(S(t)), \\
S(t) &= \exp(-\Lambda(t)).
\end{aligned}$$

Bien que toutes les fonctions caractérisent la loi de l'événement d'intérêt, la fonction de risque instantané λ est la plus intéressante, puisqu'elle donne une description probabiliste du futur immédiat du sujet observé et permettra le plus souvent de guider le choix du modèle pour des données de durée.

1.2.2 Covariables

Les individus ont des carractéristiques observables différentes, qui peuvent être modélisées dans la plupart des domaines d'applications par des variables explicatives, qui donnent des informations supplémentaires. Ces covariables peuvent être

fixes dans le temps (sexe, groupe sanguin,...) ou bien dépendantes du temps (mesure d'une quantité biologique,...).

Nous considérons dans la suite une durée de survie aléatoire T et $Z = (Z_1, ..., Z_p)^\top$ un vecteur de p variables explicatives réelles associées à chaque sujet. Les éléments $Z_1, ..., Z_p$ sont appelées des covariables. En analyse statistique des durées de survie, la famille des lois possibles pour le couple (T, Z) est généralement determinée en choisissant une loi conditionnelle pour T sachant Z. Cette loi est caractérisée par la fonction de risque instantané conditionnelle $\lambda_{T|Z}$ définie, sous réserve de conditions générales de régularité et si Z est fixe dans le temps, par

$$\lambda_{T|Z}(t) = \lim_{\Delta_t \downarrow 0} \frac{1}{\Delta_t} \mathbb{P}(t < T < t + \Delta_t | \varepsilon_t),$$

où $\varepsilon_t = \sigma(Z; T > t)$ est la σ-algèbre engendrée par Z et l'événement $\{T > t\}$. Si Z dépend du temps, cette fonction est donnée par

$$\lambda_{T|Z}(t) = \lim_{\Delta_t \downarrow 0} \frac{1}{\Delta_t} \mathbb{P}(t < T < t + \Delta_t | \mathcal{F}_t),$$

avec $\mathcal{F}_t = \sigma(Z(s) : 0 \leq s \leq t; T > t)$ la σ-algèbre engendrée par $\{Z(s) : 0 \leq s \leq t\}$ et $\{T > t\}$.

Pour prendre en compte la présence des covariables dans l'analyse statistique des durée de survie, il existe notamment deux grands types de modèles: les modèles à risques additifs et les modèles à risques multiplicatifs. Ces derniers sont définis par une fonction de risque instantané $\lambda_{T|Z}$ conditionnelle au vecteur de covariables Z de la forme

$$\lambda_{T|Z}(t) = \lambda(t) h(Z; \beta),$$

où λ est la fonction de risque de base, $h(Z; \beta)$ est une fonction de régression sur Z, et β est un vecteur de paramétre inconnu. La fonction de lien $h(Z; \beta)$ doit être positive, on choisit généralement $h(Z; \beta) = e^{\beta^\top Z}$. Ains, le modèle multiplicatif permettant d'inclure des covariables dans le modèle de Weibull est défini par

$$\lambda_{T|Z}(t) = \alpha \lambda t^{\alpha - 1} e^{\beta^\top Z}, \quad t > 0, \alpha > 0, \lambda > 0,$$

où $\beta, Z \in \mathbb{R}^p$. La partie $\alpha \lambda t^{\alpha-1}$ s'interprète comme la fonction de risque de base, c'est-à-dire le risque instantané d'occurrence de l'événement pour un individu associé à des covariables telles que $Z = 0$, et (α, λ, β) est le paramètre à estimer.
Les fonctions conditionnelles de densité, de répartition, de survie et de risque cumulé

qui se déduisent de $\lambda_{T|Z}$ sont respectivement, pour $t > 0$,

$$f_{T|Z}(t) = \alpha\lambda t^{\alpha-1} e^{\beta^\top Z} e^{-\lambda t^\alpha \exp(\beta^\top Z)}, \qquad F_{T|Z}(t) = 1 - e^{-\lambda t^\alpha \exp(\beta^\top Z)},$$

$$S_{T|Z}(t) = e^{-\lambda t^\alpha \exp(\beta^\top Z)}, \qquad \Lambda_{T|Z}(t) = \lambda t^\alpha e^{\beta^\top Z}.$$

1.2.3 Censure et troncature

L'une des caractéristiques des durées de survie est l'incomplétude des données, c'est-à-dire une perte d'information. En effet, les données sont souvent recueillies partiellement, notamment, à cause des processus de censure. Les données censurées proviennent du fait qu'on n'a pas accès à toutes les informations. Ce mécanisme est supposé indépendant de l'événement étudié.

Définition 4 *La variable de censure C est définie par la possible non-observation de l'événement. Si l'on observe C, et non T, et que l'on sait que $T > C$, on dit qu'il y a une censure à droite. Ce modèle est adapté au cas où l'événement considéré est la date de fin d'étude qui est préalablement fixée.*

Définition 5 *Lorsque nous observons la censure C, et non la durée de survie T, et que l'on sait que $T < C$, un phénomène symétrique au précédent se produit et on dit qu'il s'agit d'une censure à gauche. Ce modèle est adapté au cas où l'incorporation de l'individu dans une étude est conditionnée par un évènement initial.*

Définition 6 *Lorsque la censure à droite et la censure à gauche sont conjuguées, dans ce cas l'information apportée par l'expérience se traduit par une appartenance de la durée de survie à un intervalle de temps $(C_1 < T < C_2)$. Cette censure est appelée censure par intervalle. Ce modèle est adapté au cas de suivi périodique de sujets.*

Les modèles de censure peuvent être affinés suivant le mécanisme de censure existant. Les différents modes de censure, décrits dans le cadre d'une censure à droite, sont les suivants,

Définition 7 *La censure est dite non aléatoire de type I si, étant donné un nombre positif fixé c et un échantillon $T_1, ..., T_n$, les observations consistent en (X_i, Δ_i), où*

$$\begin{cases} X_i = T_i \wedge c, \\ \Delta_i = 1_{\{T_i \leq c\}}. \end{cases}$$

Définition 8 *La censure est dite aléatoire de type I si, étant donné un échantillon $T_1, ..., T_n$, il existe une v.a. n-dimentionnelle $(C_1, ..., C_n)$ de $(\mathbb{R}^+)^n$ telle que les observations consistent en (X_i, Δ_i), où*

$$\begin{cases} X_i = T_i \wedge C_i, \\ \Delta_i = 1_{\{T_i \leq C_i\}}. \end{cases}$$

Ce modèle est typiquement utilisé pour les essais thérapeutiques.

Remarque 1 *Désormais, nous nous plaçons dans le cadre d'une censure à droite, aléatoire de type I et indépendante de la date de survenue de l'événement d'intérêt. Ce cadre correspond à un modèle fréquemment employé dans la pratique.*

Les troncatures diffèrent des censures au sens où elles concernent l'échantillonnage lui-même. Ainsi, une variable X est tronquée par un sous ensemble éventuellement aléatoire A de $\mathbb{R}^+$ si au lieu de X, on observe X uniquement si $X \in A$. Les points de l'échantillon "tronqué" appartiennent tous à A, et suivent donc la loi de T conditionnée par l'appartenance à A. Il ne faut pas confondre censure et troncature. S'il y a troncature, une partie des individus (donc des X_i) ne sont pas observables et on n'étudie qu'un sous-échantillon (problème d'échantillonnage). Le biais de séléction est un cas particulier de troncature.

- La troncature à gauche

  Soit Z une variable aléatoire indépendante de X, on dit qu'il y a troncature à gauche lorsque X n'est observable que si $X > Z$. On observe le couple (X, Z), avec $X > Z$. Par exemple, si la durée de vie d'une population est étudiée à partir d'une cohorte tirée au sort dans cette population, seule la survie des sujets vivants à l'inclusion pourra être étudiée (il y a troncature à gauche car seuls les sujets qui ont survécu jusqu'à la date d'inclusion dans la cohorte sont observables).

- La troncature à droite

  De même, il y a troncature à droite lorsque X n'est observable que si $X < Z$.

- La troncature par intervalle

  Quand une durée est tronquée à droite et à gauche, on dit qu'elle est tronquée par intervalle. Par exemple, on rencontre ce type de troncature lors de l'étude des patients d'un registre: les patients diagnostiqués avant la mise en place du registre ou répertoriés après la consultation du registre ne seront pas inclus dans l'étude.

1.2.4 Mécanisme des données manquantes

En statistique, on parle de donnée manquante lorsque l'on n'observe pas la valeur d'une variable donnée pour un individu. Le problème de la gestion des données manquantes est un sujet très complexe et aux enjeux multiples. Ces données ne peuvent pas être ignorées lors d'une analyse statistique. Mais selon leur proportion et leur type, des solutions différentes vont être choisies. On pourra soit retirer les individus présentant des données manquantes, soit imputer ces variables à ces données ou encore développer des méthodes plus complexes permettant de mener à bien les analyses en présence de ce type de données.

Différents types de données manquantes

— Données manquantes complètement au hasard (MCAR: Missing Completely At Random): La probabilité de ne pas observer une valeur ne dépend ni des données observées ni des données non observées.
— Données manquantes au hasard (MAR: Missing At Random): La probabilité pour qu'une donnée soit manquante dépend uniquement des données observées.
— Données manquantes non ignorables (NMAR: No Missing At Random): La probabilité qu'une donnée soit manquante dépend de cette donnée.

Plus de détails et d'exemples sur ces mécanismes de données manquantes peuvent être trouvés dans Little et Rubin (1987, [38]), Gill et al. (1997, [21]) et Tsiatis (2006, [63]).

Méthodes de traitement des données manquantes au hasard (MAR)

Ils existe plusieures méthodes permettant de réaliser l'inférence statistique avec ce type de données. Parmi ces méthodes, citons la méthode "cas complets" et la "pondération par la probabilité inverse".

(a) La méthode "cas complets": elle consiste à éliminer les individus présentant des données manquantes et à garder ceux pour lesquels les données complètes. Le résultat est une perte d'information. Cette méthode entraîne que la taille de l'échantillon diminue, la puissance de ce test se réduit et les estimateurs obtenus peuvent être biaisés.

(b) La pondération par la probabilité inverse: dans cette méthode, on corrige l'inférence statistique en pondérant habilement les individus entièrement observés.

1.2.5 Vraisemblance dans un modèle de survie censuré

Dans un modèle de régression de durées, l'inférence statistique consiste à estimer le paramètre d'intérêt β du modèle qui permet de quantifier de l'influence des covariables. Ce problème repose souvent sur la méthode du maximum de vraisemblance. Dans la suite, nous rappelons brièvement cette méthode. Pour plus de détails le lecteur pourra se reporter aux ouvrages de Bagdonavičius et Nikulin (2002, [5]), Kalbfleisch et Prentice (1980, [31]), Lawless (2003, [36]).

Soit T^0 une durée de survie aléatoire. On suppose que la loi P_{T^0} de T^0 appartient à une famille de lois de probabilité $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$; avec $\Theta \subseteq \mathbb{R}^p$. On note par P_{θ_0} la loi de T^0 ou $\theta_0 \in \Theta$.

On note $f_{T^0;\theta}(\cdot), F_{T^0;\theta}(\cdot), S_{T^0;\theta}(\cdot), \lambda_{T^0;\theta}(\cdot), \Lambda_{T^0;\theta}(\cdot)$ les densité, fonction de répartition, fonction de survie, fonction de risque instantané et fonction de risque cumulé de la variable durée de vie T^0, sous la loi P_θ.

On définit le temps de censure C sur le même espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$ que T, et on suppose que C et T sont indépendantes et la loi de C ne dépend pas du paramètre θ (la loi de C est "non informative"). On désigne par $f_C(\cdot), F_C(\cdot), S_C(\cdot)$ les densité, fonction de répartiton et fonction de survie de C avec $F_C(t) = \mathbb{P}[C \leq t]$, $S_C(t) = 1 - F_C(t)$ et $f_C(t) = \frac{dF_C(t)}{dt}$.

Les observations correspondent alors à des réalisations de $T = \min(T^0, C)$ et de l'indicatrice de censure $\Delta = 1_{\{T^0 \leq C\}}$. Notons $(T_i, \Delta_i)_{i \in \{1,\ldots,n\}}$ un échantillon des variables (T, C). À partir de ces observations, on peut utiliser la méthode de maximum du vraisemblance pour estimer θ_0.

La vraisemblance associée à l'échantillon $(T_i, \Delta_i)_{i \in \{1,\ldots,n\}}$ s'écrit sous la forme

$$\begin{aligned} L_n(\theta) &= \prod_{i=1}^{n} (f_{T^0;\theta}(T_i) S_C(T_i))^{\Delta_i} (S_{T^0;\theta}(T_i) f_C(T_i))^{1-\Delta_i} \\ &= \prod_{i=1}^{n} \lambda_{T^0;\theta}(T_i)^{\Delta_i} S_{T^0;\theta}(T_i) S_C(T_i)^{\Delta_i} f_C(T_i))^{1-\Delta_i}, \end{aligned}$$

en utilisant les relations liant les densité, fonction de survie et fonction de risque instantané.

On remarque qu'il est équivalent de chercher l'estimateur du maximum de vraisemblance de θ_0, sous l'hypothèse de censure non informative, en maximisant l'expression

$$\prod_{i=1}^{n} \lambda_{T^0;\theta}(T_i)^{\Delta_i} S_{T^0;\theta}(T_i).$$

1.2.6 Vraisemblance dans un modèle de survie censuré avec covariables

Supposons que l'on soit dans le cadre d'un modèle à risque multiplicatif avec des covariables et que la loi conditionnelle $P_{T^0|Z}$ appartienne à une famille de lois de probabilité $\mathcal{P}_Z = \{P_{\theta,Z} : \theta \in \Theta \subseteq \mathbb{R}^p\}$, avec $\Theta \in \mathbb{R}^p$. On note par $P_{\theta_0,Z}$ la loi de T^0 sachant Z, avec $\theta_0 \in \Theta$.

On note $f_{T^0|Z;\theta}(\cdot), F_{T^0|Z;\theta}(\cdot), S_{T^0;\theta}(\cdot), \lambda_{T^0|Z;\theta}(\cdot), \Lambda_{T^0|Z;\theta}(\cdot)$ les densité, fonction de répartition, fonction de survie, fonction de risque instantané et fonction de risque cumulé de la variable durée de survie T^0, sous la loi $P_{\theta,Z}$. Nous supposons que la densité $f_Z(Z;\theta)$ de Z ne dépend pas du paramètre θ et que la loi de T^0 sachant Z est indépendante de la loi de C sachant Z. On suppose que la loi de C est non-informative. Avec les mêmes notations que précédemment, la vraisemblance associée à l'échantillon $(X_i, \Delta_i, Z_i)_{i\in\{1,\cdots,n\}}$ s'écrit sous la forme

$$L_n(\theta) = \prod_{i=1}^{n} (f_{T^0|Z;\theta}(T_i) S_{C|Z}(T_i))^{\Delta_i} (S_{T^0|Z;\theta}(T_i) f_{C|Z}(T_i))^{1-\Delta_i} f_Z(Z_i).$$

Sous l'hypothèse de censure non informative, il est équivalent, pour obtenir l'estimateur du maximum de vraisemblance, de maximiser l'expression

$$\prod_{i=1}^{n} \lambda_{T^0|Z;\theta}(T_i)^{\Delta_i} S_{T^0|Z;\theta}(T_i).$$

1.3 Rappels sur les processus

Nous commençons ici par rappeler des notions sur les processus, qui seront nécessaires par la suite.

Définition 9 *Soit un espace probabilisé* $(\Omega, \mathcal{A}, \mathbb{P})$ *où* $\mathbb{P}$ *est la mesure de probablité sur* $(\Omega, \mathcal{A})$. *Un processus aléatoire est une fonction de deux variables,* t *et* ω*, notée* $X(t,\omega)$*, définie par:*

$$\begin{aligned} X : (T \times \Omega) &\rightarrow \mathbb{R} \\ (t, \omega) &\mapsto X(t, \omega). \end{aligned}$$

Les fonctions définies pour $\omega \in \Omega$ par $X(., \omega)$ sont appelées trajectoires de X. À t fixé, la fonction $X_t \mapsto X(t, \omega)$ est appelée coordonnée à l'instant t.
Un processus est dit continu à droite, à variation bornée, croissant, ayant une limite à droite si l'ensemble des trajectoires ayant la propriété correspondante est de probabilité 1.

Définition 10 *Une filtration $\{\mathcal{F}_t, t \in \mathbb{R}^+\}$ est une famille croissante et continue à droite de sous tribus de $\mathcal{A}$. C'est-à-dire vérifiant pour tout $s \leq t, \quad \mathcal{F}_s \subseteq \mathcal{F}_t$.*

- *Si $(\mathcal{F}_t)_{t \in \mathbb{R}^+}$ est une filtration, $\bigcap_{h>0} \mathcal{F}_{t+h}$ est une tribu et on la note $\mathcal{F}_{t^+}$. De la même façon, $\mathcal{F}_{t^-}$ est la tribu engendrée par $\bigcap_{h>0} \mathcal{F}_{t-h}$*
- *Une filtration $(\mathcal{F}_t)_{t \in \mathbb{R}^+}$ est dite continue à droite si pour tout $t \in \mathbb{R}^+$, $\mathcal{F}_{t^+} = \mathcal{F}_t$.*

Remarque 2
- *Une filtration permet de définir ce qui appartient au passé du processus (qui est observé et connu) et ce qui appartient au futur (pas encore accessible). Les filtrations naturelles sont les familles $(\mathcal{F}_t)_{t \in \mathbb{R}^+}$ telles que $\mathcal{F}_t = \sigma(X(s); 0 \leq s \leq t)$ qui est la plus petite tribu rendant pour tout $(0 \leq s \leq t)$ les variables $X(s)$ mesurables, c'est-à-dire $\mathcal{F}_t$ contient l'information engendrée par le processus X sur $[0, t]$.*
- *On dispose souvent de plus d'information que la seule histoire de processus. À chaque instant, on observe également d'autres processus qui peuvent apporter de l'information sur le processus étudié. Dans ce cas, la filtration est élargie pour intégrer cette information (par exemple prise en compte de covariables pour améliorer la prévision).*

Définition 11 *Le processus $(X_t)_{t \in \mathbb{R}^+}$ est dit $\mathcal{F}_t$-adapté si $\forall t \in \mathbb{R}^+$, $X(t)$ est $\mathcal{F}_t$-mesurable.*

Définition 12 *Soient $(\Omega, \mathcal{A}, \mathbb{P})$ un espace probabilisé et $(\mathcal{F}_t)_{t \in \mathbb{R}^+}$ une filtration. Un processus stochastique $(X_t)_{t \in \mathbb{R}^+}$ est dit*

(i) *intégrable si $\sup_{t \in \mathbb{R}^+} \mathbb{E}(|X_t|) < \infty$,*

(ii) *de carré intégrable si $\sup_{t \in \mathbb{R}^+} \mathbb{E}(X_t^2) < \infty$,*

(iii) *borné s'il existe une constante $C \in \mathbb{R}^+$ telle que*

$$\mathbb{P}(\sup_{t \in \mathbb{R}^+} |X(t)| \leq C) = 1.$$

1.3.1 Martingales

Dans le cadre de la modélisation de durées de survie censurées, la théorie des martingales présente un outil indispensable permettant l'étude des estimateurs. Dans la suite, nous donnons la définition des martingales à temps continu ainsi que quelques propriétés. On peut se reporter aux ouvrages de Flemming et Harrington (1991, [18]) ou Bickel et al. (1993, [6]).

Soit $M = (M(t))_{t\geq 0}$ un processus stochastique continu à droite avec une limite à gauche et $(\mathcal{F}_t)_{t\in\mathbb{R}^+}$ une filtration.

Définition 13 *Soit un espace probabilisé* $(\Omega, \mathcal{A}, \mathbb{P})$. $M = (M(t))_{t\geq 0}$ *est une martingale adaptée à la filtration* $(\mathcal{F}_t)_{t\in\mathbb{R}^+}$ *si*

— $\forall t, M_t \in \mathrm{L}^1$, *(i.e. est intégrable) et* $M = (M(t))_{t\geq 0}$ *est* $\mathcal{F}_t$*-adaptée*

— $\forall s, t, 0 \leq s \leq t, \quad \mathbb{E}(M_t|\mathcal{F}_s) = M_s \quad p.s.$

Si la dernière condition est remplacée par $\mathbb{E}(M_t|\mathcal{F}_s) \geq M_s \quad p.s$, M *est appelée sous-martingale.*

Si la dernière condition est remplacée par $\mathbb{E}(M_t|\mathcal{F}_s) \leq M_s \quad p.s$, M *est appelée sur-martingale.*

Proposition 1.3.1 *Soit* $M_t = (M(t))_{t\geq 0}$ *une martingale. Alors* $M_t^2 - \langle M\rangle_t$ *est une martingale, avec* $\langle M\rangle_t$ *est défini par*

$$\langle M\rangle_t = \lim_{|\delta|\to 0} \sum_i \mathbb{E}[(M_{t_{i+1}} - M_{t_i})^2|\mathcal{F}_{t_i}].$$

$\langle M\rangle_t$ *est appelé le processus prévisible croissant associé à* M_t.

Définition 14 *Soit* $H = \{H(t), t \in T\}$ *un processus stochastique prévisible, intuitivement ça signifie que pour tout* t*, la valeur de* $H(t)$ *est connue juste avant* t*. On note que la condition suffisante pour que* H *soit prévisible:*

(i) H *est adapté par rapport à* $(\mathcal{F}_t)_t \geq 0$.

(ii) *les trajectoires de* H *sont continues à gauche.*

On introduit l'intégrale stochastique par rapport aux martingales $I(t) = \int_0^t H(s)dM(s)$ qui sont des martingales à moyenne nulle par rapport à $\mathcal{F}_t$. Par conséquent, les propriétés des martingales sont conservées pour les intégrales stochastiques. On a donc:

$$\langle \int HdM\rangle = \int H^2 d\langle M\rangle, \quad \langle \int H_1 dM_1, \int H_2 dM_2\rangle = \int H_1 H_2 d\langle M_1, M_2\rangle.$$

On va voir comment un processus stochastique peut être décomposé en la somme d'un processus prévisible et d'une martingale.

Théorème 1.3.1 ***Décompostion de Doob-Meyer***
Soit X une sous-martingale positive continue à droite et adaptée à la filtration $\mathcal{F}_t$. Il existe une martingale continue à droite M telle que presque sûrement on a:

$$X(t) = M(t) + A(t) \quad pour \quad tout \quad t \geq 0,$$

avec A est un processus prévisible croissant continu à droite vérifiant $\mathbb{E}(A_t) < \infty$.

Définition 15 *Le processus prévisible A est appelé compensateur de X.*

Pour l'obtention des résultats théoriques, comme le calcul des variances, on considère souvent la décomposition de Doob-Meyer associée aux carrés des martingales. En particulier, comme le carré d'une martingale est une sous-martingale (inégalité de Jensen), la décomposition de Doob-Meyer permet d'obtenir le résultat suivant.
Soit M une martingale adaptée à $\mathcal{F}_t$, continue à droite, telle que $\mathbb{E}(M_t^2) < \infty$ pour tout t. Alors il existe un unique processus continue à droite prévisible noté: $\langle M \rangle = \langle M, M \rangle$ tel que $\langle M \rangle(0) = 0$ presque sûrement, $\mathbb{E}(\langle M \rangle(t)) < 0$ pour tout t et

$$M^2(t) - \langle M \rangle(t), \quad t \geq 0$$

est une martingale continue à droite. $\langle M \rangle$ est appelé le compensateur de M^2.

1.3.2 Processus de Poisson

Un processus de Poisson homogène décrit la distribution des événements qui se produisent entièrement indépendemment l'un de l'autre. On désigne par λ l'intensité du ce processus et λdt la probabilité qu'un événement se produise dans l'intervalle $[t, t + dt)$. Voici quelques propriétés de processus de Poisson:

(i) Le temps entre les événements est exponentiellement distribué avec la densité de probabilité $\lambda e^{-\lambda t}$.

(ii) L'espérance et la variance du nombre d'événements dans un intervalle de temps de longueur h sont toutes les deux égales à λh.

(iii) Le nombre d'événements dans un intervalle de temps de longueur h suit la loi de Poisson; la probabilité de survenue de k événements est $(\lambda h)^k e^{-\lambda h}/k!$.

(iv) Le processus de Poisson est à accroissements indépendants.

Soit $N(t)$ le nombre d'événements dans $[0, t]$, on introduit le processus:

$$M(t) = N(t) - \lambda t.$$

On désigne par $\mathcal{F}_t$ les informations concernant tous les événements qui se produisent dans l'intervalle de temps $[0, t]$. Le résultat suivant découle du fait que le processus de Poissson est à accroissements indépendants, $t > s$:

$$E(M(t) - M(s)|\mathcal{F}_s) = E(M(t) - M(s)) = E(N(t) - N(s)) - \lambda(t - s) = 0,$$

ce qui donne

$$E(M(t)|\mathcal{F}_s) = M(s).$$

C'est la définition d'une martingale, et par conséquent le processus $M(t) = N(t) - \lambda t$ est une martingale. Celà montre que λt est le compensateur de processus de Poisson $N(t)$.
Par un argument similaire

$$E(M^2(t) - \lambda t|\mathcal{F}_s) = M^2(s) - \lambda s.$$

1.3.3 Processus de comptage

Un processus de comptage peut être vu comme l'enregistrement des occurences dans le temps d'événements disjoints et discrets. Ce type de processus intervient de manière naturelle dans la modélisation des durées de vie.

Définition 16 *Un processus de comptage $N = \{N(t), t \geq 0\}$ est un processus continu à droite avec une limite à gauche, $(\mathcal{F}_t)_{t\in\mathbb{R}^+}$-adapté, nul à zéro, croissant, à trajectoire constante par morceaux, ayant des sauts d'amplitude 1 et tel que $N(t)$ est presque sûrement fini pour tout $t \in \mathbb{R}^+$.*

On considère la filtration naturelle engendrée par N, $\mathcal{F} = \sigma(N(s), s \leq t)$. On pourra également ajouter les ensembles négligeables. Comme un processus de comptage (croissant, nul en zéro et borné) est une sous martingale positive, on déduit la proposition suivante de la décomposition de Doob Meyer.

Proposition 1.3.2 *Soit $N = \{N(t), t \geq 0\}$ un processus de comptage adapté à $(\mathcal{F}_t)_{t\in\mathbb{R}^+}$ tel que $\mathbb{E}(N(t)) < +\infty$ pour tout t. Alors il existe un unique processus $\lambda(.)$ $\mathcal{F}_t$-prévisible, croissant, càdlàg et nul en zéro presque sûrement vérifiant $\mathbb{E}(\lambda(t)) < +\infty$ pour tout t tel que*

$$M(t) = N(t) - \Lambda(t), \quad t \in \mathcal{T}$$

soit une martingale continue à droite de moyenne nulle.

Proposition 1.3.3 *Soient N un processus ponctuel de dimension 1, et Λ son compensateur. Si N est absolument continu, alors N possède une intensité λ, i.e. il existe un*

processus prévisible λ tel que

$$\Lambda(t) = \int_0^t \lambda(s)ds,$$

pour tout t.

Ça donne une définition formelle d'intensité λ du processus de comptage

$$\lambda(s) = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \mathbb{P}(N(s+\varepsilon) - N(s) \geq 1 | \mathcal{F}_s);$$

où $\mathcal{F}_s, s \in S$ est la filtration naturelle. On a encore

$$M(t) = N(t) - \int_0^t \lambda(s)ds.$$

Par conséquent et par unicité de la décomposition on obtient

$$\langle M \rangle(t) = \langle M, M \rangle(t) = \int_0^t \lambda(s)ds.$$

Dans le contexte du processus du comptage, l'intégrale stochastique

$$I(t) = \int_0^t H(s)dM(s)$$

peut être divisée simplement en deux comme suit:

$$I(t) = \int_0^t H(s)dNs - \int_0^t H(s)\lambda(s)ds,$$

la dernière étant une intégrale de Riemann.
Nous obtenons l'expression suivante

$$\langle \int HdM \rangle(t) = \int H^2(s)\lambda(s)ds.$$

1.3.4 Théorie asymptotique des martingales

Parfois on considère les martingales comme processus de bruit, contenant des déviations aléatoires de la moyenne. Ainsi cela vaut pour les martingales. En fait il existe des théorèmes des limites centraux pour les martingales qui sont étroitement analogues à ceux connus pour la somme de variables aléatoires indépendantes. Pour une lecture appropfondie sur cette théorie, on peut se reporter au livre de Aalen, Borgan et Gjessing (2008, [1]).

Si on a une suite des processus de comptage, où le nombre des sauts croit et devient de plus en plus dense, la martingale normalisée associée (ou l'intégrale stochas-

tique par rapport à ces martingales) converge vers une martingale limite à trajectoire continue. La martingale limite est étroitement reliée à un processus de Wiener, ou un mouvement brownien. En effet, si le processus à variation prévisible de la martingale limite est égale à une fonction deterministe $V(t)$, alors elle est gaussienne. Cela découle du fait qu'une martingale à trajectoire continu est déterminée uniquement par son processus à variation.
Donc il y a deux conditions à vérifier pour qu'une suite de martingales converge vers une martingale gaussienne:

(i) le processus à variation prévisible doit converger vers une fonction déterministe,

(ii) les tailles des sauts des martingales tendent vers zéro.

Il est important de noter que ces deux hypothèses concernent des différents aspects de la suite des processus. La première hypothèse implique la stabilisation de l'espace d'échantillon de processus, alors que la seconde est une exigence sur les trajectoires des processus.

1.3.5 Formulation mathématique

Il existe plusieures versions du théorème de la limite central pour les martingales, qui reformulent les deux hypothèses données dans la section précédente. Une version très générale est formulée par Rebolledo (1980, [52]). On peut se reporter aussi à l'ouvrage Anderson et al (1993, [3]). Helland (1982) a montré aussi comment le théorème de Rebolledo à temps continu peut être déduit du théorème de limite centrale pour les martingale à temps discret. Cette version du théorème est inspirée de l'ouvrage Aalen, Borgan et Gjessing (2008, [1]).

Théorème 1.3.2 *Soit $\widetilde{M}^{(n)}, n \geq 1$ une suite des martingales de moyenne nulle définies sur $[0, \tau]$, et soit $\widetilde{M}_{\varepsilon}^{(n)}$ la martingale qui contient tous les sauts de $\widetilde{M}^{(n)}$ supérieure à ε donné. Considérons les conditions suivantes:*

(i) $\langle \widetilde{M}^{(n)} \rangle \to^{P} V(t)$, *pour tout $t \in [0, \tau]$ lorsque $n \to \infty$, où V est une fonction croissante continue vérifiant $V(0) = 0$.*

(ii) $\langle \widetilde{M}_{\varepsilon}^{(n)} \rangle \to^{P} 0$, *pour tout $t \in [0, \tau]$ et tout $\varepsilon > 0$ lorsque $n \to \infty$.*

Alors, si $n \to \infty$, la suite des martingales $\widetilde{M}^{(n)}$ converge en distribution vers une martingale gaussienne de moyenne nulle et de variance U donnée par $U(t) = W(V(t))$.

Ainsi, sous des hypothèses très générales, il y a une convergence en distribution vers une martingale gaussienne. On va utiliser le théorème de la limite centrale pour

les martingales pour obtenir le comportement asymptotique d'une suite d'intégrales stochastiques de la forme $\int_0^t H^{(n)}(s)dM^{(n)}(s)$, où $H^{(n)}(s)$ est le processus prévisible et $N^{(n)}(t) = M^{(n)}(t) + \int_0^t \lambda^{(n)}(s)ds$ est le processus de comptage. Plus généralement, on considère la somme des intégrales stochastiques

$$\sum_{j=1}^{k} \int_0^t H_j^{(n)}(s)dM_j^{(n)},$$

où $H_j^{(n)}(t)$ est le processus prévisible et

$$M_j^{(n)} = N_j^{(n)} - \int_0^t \lambda_j^{(n)}(s)ds,$$

est la décomposition de la martingale en fonction de processus de comptage pour $j = 1, ..., k$. Dans cette situation les conditions (i) et (ii) du théorème 1.3.2 prennent la forme:

- $\sum_{j=1}^{k} \int_0^t (H_j^{(n)}(s))^2 \lambda_j^{(n)}(s)ds \to^P V(t)$, pour tout $t \in [0, \tau]$,
- $\sum_{j=1}^{k} \int_0^t (H_j^{(n)}(s))^2 1_{\{|H_j^{(n)}|>\varepsilon\}} \lambda_j^{(n)}(s)ds \to^P 0$, pour tout $t \in [0, \tau]$.

Remarque 3 *L'utilité de théorème de la limite centrale (TLC) consiste à montrer la convergence en distribution des estimateurs ou des statistiques en vérifiant les deux conditions (i) et (ii).*

1.3.6 Rappel sur les processus empiriques

La théorie des processus empiriques joue un rôle central en statistique, puisqu'elle concerne l'ensemble des résultats limites généraux se rapportant aux échantillons aléatoires. Parmi ces plus importants résultats, on trouve ceux liés aux théorèmes de Glivenko-Cantelli (1933) et Donsker (1951).
Un processus empirique est un processus stochastique basé sur un échantillon aléatoire. Soit $X_1, X_2, \cdots$ une suite de variables aléatoires réelles i.i.d. de loi P, définies sur l'espace probabilisé $(\Omega, \mathcal{C}, \mathbb{P})$. On note F la fonction de répartition de P définie par

$$F(t) = P(]-\infty, t]) = \mathbb{P}(\{\omega : X(\omega) \leq t\}).$$

Définition 17 *On définit la mesure empirique $\mathbb{P}_n$ associée à l'échantillon $X_1, \cdots, X_n$ par*

$$\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}, \tag{1.1}$$

où δ_x est la mesure de Dirac au point x. et une fonction de répartition empirique $\mathbb{F}_n(\omega)(\cdot) : \mathbb{R} \to [0,1]$ définie par:

$$\mathbb{F}_n(\omega)(t) = \mathbb{P}_n(\omega)(]-\infty, t]) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i(\omega)\leq t\}}$$

Notons que $\mathbb{P}_n$ est une variable aléatoire à valeurs dans l'ensemble des lois de probabilité sur $\mathbb{R}$ et que $\mathbb{F}_n$ est une variable aléatoire à valeurs dans l'ensemble des fonctions de répartition sur $\mathbb{R}$. Pour tout $\omega \in \Omega$, $\mathbb{P}_n(\omega)$ est donc une probabilité sur $\mathbb{R}$, de fonction de répartition $\mathbb{F}_n(\omega)(\cdot)$.
Par la suite, on omettra ω dans les notations. Ainsi, on notera

$$\begin{aligned}
\mathbb{P}_n(A) &= \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}(A) = card\{i \leq n : X_i \in A\}/n \\
\mathbb{F}_n(t) &= \mathbb{P}_n(]-\infty, t]) = \frac{1}{n}\sum_{i=1}^{n}\{X_i \leq t\}.
\end{aligned}$$

On appelle processus empirique réel la quantité $\alpha_n = \sqrt{n}(\mathbb{F}_n - F)$. Une réalisation $\alpha_n(\omega)$ de la fonction aléatoire α_n est une fonction $t \mapsto \alpha_n(t)$ appelée trajectoire du processus.
Pour tout t, les variables aléatoires $1_{\{X_t \leq t\}}$ sont des Bernouillis $\mathcal{B}(F(t))$ indépendantes donc $n\mathbb{F}_n(t)$ est binomiale $\mathcal{B}(n, F(t))$ de sorte que:

$$\mathbb{E}(\mathbb{F}_n(t)) = F(t) \quad et \quad \mathbb{V}ar(\mathbb{F}_n(t)) = \frac{F(t)(1-F(t))}{n}.$$

De plus, $\forall t \in \mathbb{R}$,

$$\mathbb{F}_n(t) \to^{p.s} F(t) \quad et \quad \alpha_n(t) \to \mathcal{N}(0, F(t)(1-F(t))).$$

Le théorème de la limite centrale multivarié assure de plus que pour tout $(t_1, \cdots, t_k) \in \mathbb{R}^k$, $(\alpha_n(t_1), \cdots, \alpha_n(t_k))$ converge en loi vers un vecteur gaussien centré, de matrice de variance-covariance $(V_{i,j})$, où

$$\begin{aligned}
V_{i,j} &= F(t_i)F(t_j)(V_{i,j} = \mathbb{C}ov(1_{\{X\leq t_i\}}, 1_{\{X\leq t_j\}}) \\
&= \mathbb{P}(X \leq t_i, X \leq t_j) - F(t_i)f(t_j).
\end{aligned}$$

La théorie des processus empiriques s'intéresse au cas plus général de la convergence uniforme de ces processus sur des classes de fonctions. La loi forte des grands nombres permet de donner, en tout point de $\mathbb{R}$, la convergence presque sûre de la fonction de répartition empirique vers la fonction de répartition des observations. Les théorèmes de Glivenko-Cantelli et Donsker donnent une première extension uni-

forme sur les classes de fonctions pour les variables aléatoires à valeurs réelles indépendantes. On note $D[-\infty,\infty]$ l'espace des fonctions càdlàg (espace de Skorohod) muni de la norme de supremum.

Théorème 1.3.3 ***(Théorème de Glivenko-Cantelli, 1933)***
Soit $X_1, X_2, \cdots$, une suite de variables aléatoires réelles indépendantes de même loi de probabilité de fonction de répartition F. Alors

$$|\mathbb{F}_n - F_X| \longrightarrow_{p.s} 0. \tag{1.2}$$

Théorème 1.3.4 ***(Théorème de Donsker, 1951)***
Soit $X_1, X_2, \cdots$, une suite de variables aléatoires réelles indépendantes de même loi de probabilité de fonction de répartition F. Alors la suite de processus empiriques α_n converge en loi dans $D[-\infty,\infty]$ vers un processus gaussien G_F centré et de fonction de covariance

$$\mathbb{C}ov(G_F(s), G_F(t)) = \mathbb{E}(G_F(s)G_F(t)) = F(s \wedge t) - F(s)F(t),$$

Les théorèmes de Glivenko-Cantelli et Donsker pour les processus empiriques réels peuvent être vus comme des cas particuliers des résultats généraux pour des processus empiriques indexés par des classes de fonctions.

Processus empiriques indexés par des fonctions

Considérons de nouveau $X_1, X_2, \cdots$, une suite de variables aléatoires réelles indépendantes de même loi de probabilité P sur $(\mathcal{X}, \mathcal{A})$ et soit $\mathcal{F} \subset L_1(P)$ une classe de fonctions mesurables de $\mathcal{X}$ dans $\mathbb{R}$. Notons:

$$\forall f \in \mathcal{F}, Pf = \mathbb{E}f(X) = \int f dP$$

et

$$\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^{n} f(X_i) = \int f d\mathbb{P}_n.$$

Le processus stochastique

$$\sqrt{n}(\mathbb{P}_n - P)(f), \quad f \in \mathcal{F},$$

où $\sqrt{n}(\mathbb{P}_n - P)(f) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(f(X_i) - Pf)$, s'appelle le processus empirique (centré normalisé) indexé par $\mathcal{F}$. On le notera par la suite $G_n = \sqrt{n}(\mathbb{P}_n - P)$, ou $\{\mathbb{G}_n f, f \in \mathcal{F}\}$. C'est une fonction aléatoire de $\mathcal{F}$ dans $\mathbb{R}$.
Si $\mathcal{X} = \mathbb{R}$, le processus $\{\alpha_n(t), t \in \mathbb{R}\}$ peut être ré-exprimé comme $\{\mathbb{G}_n f, f \in \mathcal{F}\}$, où $\mathcal{F} = \{1_{\{x \leq t, t \in \mathbb{R}\}}$. Ainsi, on peut avoir le processus empirique réel comme indexé

par $t \in \mathbb{R}$ ou par $f \in \mathcal{F}$. Pour $f \in \mathcal{F}$, la loi forte des grands nombres assure que $\mathbb{P}_n f \to^{p.s} Pf$, et une classe de fonctions $\mathcal{F}$ pour laquelle une version uniforme de ce résultat existe, appelée classe de Glivenko-Cantelli.

Définition 18 ***(Classe de Glivenko-Cantelli)***
Une classe $\mathcal{F} \subset L_1(P)$ de fonctions mesurables $f : \mathcal{X} \to \mathbb{R}$ est dite P-Glivenko-Cantelli si

$$\|\mathbb{P}_n - P\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |\mathbb{P}_n f - Pf| \to^{p.s} 0.$$

Remarque 4 *Une classe de fonction Glivenko-Cantelli fournit une loi des grands nombres uniforme, car*

$$\lim_{n \to \infty} \sup_{f \in \mathcal{F}} |\frac{1}{n} \sum_{i=1}^{n} f(x_i) - \mathbb{E}f(x)| = 0 \quad p.s.$$

contrôle une infinité de lois de grands nombres simultanément.

Remarque 5 *La distance aléatoire $\|\mathbb{P}_n - P\|_{\mathcal{F}}$ n'est pas forcément mesurable, d'où l'utilisation, pour contourner cette difficulté, de la convergence p.s.. $\mathcal{F}$ est aussi dite P-Glivenko-Cantelli lorsque la convergence a lieu en P-probabilité.*
Par le théorème de la limite centrale (TLC), on a $\mathbb{G}_n f \to^L \mathcal{N}(0, P(f - Pf)^2)$ (si $Pf^2 \leq \infty$), et d'après le TLC multidimensionnel, on a pour tout ensemble fini $(f_1, \cdots, f_k)$ la fonction de $\mathcal{F}$ telles que $Pf_i^2 \leq \infty$

$$(\mathbb{G}_n f_1, \cdots, \mathbb{G}_n f_k) \to (Gf_1, \cdots, Gf_k),$$

où $(Gf_1, \cdots, Gf_k)$ est un vecteur gaussien sur $\mathbb{R}^k$, d'espérance nulle et de covariances $P(f_i f_j) - Pf_i p f_j$.

Nous supposerons par la suite que

$$\sup_{f \in \mathcal{F}} |f(x) - Pf| < \infty, \quad \forall x \in \mathcal{X},$$

de sorte que le processus $\mathbb{G}_n$ soit à valeurs dans $l^\infty(\mathcal{F})$, que l'on munira de la norme $\|H\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |H(f)|$.

Définition 19 ***(Classe de Donsker)***
Une classe $\mathcal{F} \subset L_2(P)$ de fonctions mesurables $f : \mathcal{X} \to \mathbb{R}$ est dite P-Donsker si la suite de processus $\{\mathbb{G}_n f, f \in \mathcal{F}\}$ converge en loi dans l'espace $l^\infty(\mathcal{F})$ vers le processus $\{Gf, f \in \mathcal{F}\}$. Le processus limite G est un processus gaussien centré de fonction de covariance

$$\mathbb{C}ov(Gf_1, Gf_2) = P(f_1 f_2) - Pf_1 Pf_2$$

appelé P-pont brownien.

Remarque 6 *Une classe de Donsker fournit un TLC uniforme car le TLC usuel*

$$\sqrt{n}(\frac{1}{n}\sum_{i=1}^{n} f(X_i) - Pf) \rightarrow^{\mathcal{L}} \mathcal{N}(0, \mathbb{V}ar(f(X)))$$

est vérifié "conjointement" pour tous les $f \in \mathcal{F}$.

2

Modèle semi-paramétrique de Cox

Résumé

Le modèle semi-paramétrique de Cox introduit par Cox (1972, [11]) est le plus répandu parmi les modèles à risques multiplicatifs. C'est un modèle semi-paramétrique qui permet la modélisation de l'influence des covariables sur des durées de survie. Ce modèle définit une famille de lois conditionnelles de la donnée de survie, sachant un vecteur de variables explicatives. On pourra se reporter à plusieurs ouvrages tels que: Martinussen et Scheike (2002, [42]); Fleming et Harrington (1991, [18]); Therneau et Grambesch (2000, [61]).

Dans ce chapitre, nous introduisons le modèle de Cox puis nous présentons quelques estimateurs et leurs propriétés asymptotiques. Enfin, nous présentons le modèle de Cox stratifié.

Sommaire

2.1 Présentation générale

Concernant les modèles statistiques proprement dits, trois approches sont possibles: paramétrique, non paramétrique et semi-paramétrique.

Approche paramétrique: Elle nécessite l'appartenance de la loi de probabilité réelle des observations à une classe particulière de lois, qui dépend d'un nombre fini de paramètres. Cette approche permet de faciliter l'estimation des paramètres ainsi que d'obtenir des intervalles de confiance et construire des tests statistiques. Cependant, le choix d'un modèle paramétrique particulier n'est pas toujours aisé en pratique.

Approche non- paramétrique: Elle ne nécessite aucune hypothèse quant à la loi de probabilité réelle des observations, mais cette approche nécessite la disposition d'un nombre important d'observations.

Approche semi-paramétrique: Elle représente le cas intermédiaire entre les deux approches précédentes. La loi de probabilité réelle des observations est supposée appartenir à une classe ayant une partie dépendante de paramètres et l'autre partie s'ecrivant sous la forme de fonctions non paramétriques. Cette approche est la plus utilisée pour analyser des durées de survie, notamment au travers du modèle de régression de Cox.

2.2 Modèle semi-paramétrique de Cox

On introduit une fonction de risque de base qui donne la forme générale du risque et qui est commune à tous les individus. Les modèles à risques proportionnels se caractérisent par la relation suivante, pour tout $t > 0$

$$\alpha(t|Z) = \alpha_0(t)h(\beta^\top Z), \tag{2.1}$$

où Z est un vecteur de covariables, β le paramètre d'intérêt et h une fonction positive. En général, on suppose que l'effet des covariables se résume à une quantité réelle $\beta' Z$.

Remarque 7 *Si α_0 et ou h ont une forme inconnue le modèle est dit semi-paramétrique.*

2.2.1 Hypothèse de proportionnalité

Le modèle (2.1) est dit à risques proportionnels car, quelques soient deux individus i et j qui ont pour covariables Z_i et Z_j, le rapport des fonctions de risque ne varie pas au cours du temps,

$$\frac{\alpha(t|Z_i)}{\alpha(t|Z_j)} = \frac{h(\beta^\top Z_i)}{h(\beta^\top Z_j)}.$$

Les fonctions de risque sont donc proportionnelles, c'est une conséquence du modèle mais c'est aussi une hypothèse qu'il faudra vérifier. Ce rapport est un risque relatif à l'instant t des sujets de caractéristiques Z_i par rapport aux sujets de caractéristiques Z_j . Vérifier l'hypothèse de proportionnalité est primordiale lorsque l'on vient d'ajuster au modèle de Cox à un jeu de données puisque ce dernier repose essentiellement sur cette hypothèse. En cas de non proportionnalité, les tests statistiques et les seuils observés sont invalides et les rapports de risque ne sont pas interprétables correctement. Ainsi, différentes méthodes ont été mises au point pour permettre de vérifier cette hypothèse importante.

Le modèle de Cox est le plus utilisé parmi les modèles multivariés d'analyse des durées de survie. Il permet de calculer une courbe de survie avec un ajustement sur l'influence de plusieurs variables. Ces covariables, qualitatives ou quantitatives, peuvent être choisies ou séléctionnées par des procédures dites "pas-à-pas". Pour chacune des variables présentes dans le modèle final, on obtient une estimation de risque relatif ajusté (hazard ratio) de survenue de l'événement d'intérêt en présence de la variable. Nous définissons la fonction de risque instantané conditionnelle aux covariables Z pour tout $t \in \mathbb{R}^+$ par:

$$\alpha(t|Z) = \alpha_0(t)\exp(\beta^\top Z),$$

où Z est un vecteur de covariables de dimension $p \times 1$ et β un vecteur $(p \times 1)$ de coefficients de régression. Définissons la fonction de risque cumulé pour tout $t \in \mathbb{R}^+$ par $A(t) = \int_0^t \alpha(s)ds$ et notons qu'elle peut être écrite, dans le modèle de Cox, par

$$A(t) = \exp(\beta^\top Z)A_0(t),$$

où $A_0(t) = \int_0^t \alpha_0(s)ds$ est la fonction de risque de base.
Ainsi, la fonction de densité de la durée de survie T s'écrit

$$f(t) = A_0(t)e^{\beta^\top Z}\exp(-e^{\beta^\top Z}A_0(t)).$$

L'estimation du paramètre d'intérêt β, permettant d'expliquer la nature de l'influence des covariables, ainsi que la fonction de risque cumulé A, considérée comme un paramètre de nuisance, est abordée dans la section suivante.

2.2.2 Notations

On considère un échantillon contenant n individus. Soit, pour $i \in \{1, ..., n\}$:

— τ date de fin d'observation;
— T_i la date de survenue de l'évènement chez l'individu i;
— C_i la date de censure correspondante;
— $X_i = \min(T_i, \min(C_i, \tau))$;
— $\Delta_i = 1_{\{T_i < \min(C_i, \tau)\}}$;
— $Z_i = (Z_{i,1}, ..., Z_{i,p})^\top \in \mathbb{R}^p$ le vecteur de covariables;
— $Y_i(t) = 1_{\{X_i \geq t\}}$ l'indicatrice de risque;
— $N_i(t)$ le processus de comptage;
— $\Lambda_i(t)$ le compensateur de $N_i(t)$ par rapport à la filtration $\mathcal{F}_t$.

2.2.3 Méthode du maximum de vraisemblance

Pour mener l'inférence statistique dans ce modèle, Cox (1972, [11]) a proposé de s'appuyer sur une vraisemblance partielle dans laquelle le paramètre de nuisance α_0 n'intervient pas. Nous construisons la vraisemblance partielle de Cox pour un échantillon de durées de survie et de covariables en appliquant les techniques de vraisemblance classique dans le modèle de Cox à risques proportionnels.
Nous rappelons tout d'abord le principe de cette méthode et nous l'appliquons ensuite pour construire la vraisemblance partielle dite de Cox, pour un échantillon de durées et de covariables. Pour un exposé plus complet, le lecteur pourra se reporter à Bagdonavičius et Nikulin (1996, [5]), Cox (1975, [12]), Fleming et Harrington (1991, [18]) et Kalbfleisch et Prentice (1980, [31]).

On considère ici un vecteur X de densité $f_X(x, \theta)$, où $\theta = (\phi, \beta)$ est un vecteur de paramètres. Supposons que l'on s'intéresse à β, paramètre d'intérêt, et non à ϕ, paramètre de nuisance. Par exemple dans le modèle de Cox, qui définit une famille de lois paramétrées par un vecteur β et une fonction $\alpha(t), \beta$ est le paramètre d'intérêt et $\alpha(t)$ est le paramètre de nuisance.
Pour faire de l'inférence statistique sur le paramètre β, il suffit de maximiser la vraisemblance en $\theta = (\phi, \beta)$ conjointement, mais de n'utiliser de la matrice de covariance asymptotique de l'estimateur de θ que la sous-matrice relative à β. Cependant, l'uti-

lisation de cette démarche est rendue difficile lorsque le paramètre de nuisance ϕ est fonctionnel. La maximisation jointe de la vraisemblance peut s'avèrer difficile.

On suppose qu'il est possible de décomposer X en une paire (V, W) telle que la densité de X peut être donnée par:

$$f_X(x, \theta) = f_{W|V}(w|v, \theta) f_V(v, \theta). \tag{2.2}$$

Un exemple classique de décomposition telle que (2.2), est l'écriture d'un vecteur d'observation comme un vecteur de statistique d'ordre et les rangs des observations originales. Une telle décomposition est fournie pour le vecteur V des valeurs de X ordonnées par ordre croissant et W le vecteur des rangs de ces valeurs. On peut baser l'inférence statistique sur le paramètre β sur le facteur qui ne dépend pas de ϕ. Ceci entraîne une perte d'information si le terme ignoré dépend de β, mais le gain en simplicité est susceptible de compenser une certaine perte d'efficacité. Le principe proposé par Cox est basé sur cette idée de décomposition.
Étant donnée la décomposition plus générale du vecteur X

$$(V_1, W_1, V_2, W_2, ..., V_N, W_N)$$

d'une suite de paires (V_i, W_i), $i = 1, .., N$. La vraisemblance de θ peut alors s'écrire

$$\begin{aligned} f_X(x, \theta) &= f_{V_1, W_1, ..., V_n, W_n}(v_1, w_1, ..., v_n, w_n; \theta) \\ &= \prod_{i=1}^{n} f_{W_i|V_1, W_1, ..., V_i}(w_1|v_1, w_1, ..., v_i; \theta) \\ &\times f_{V_1|V_1, W_1, ..., V_{i-1}, W_{i-1}}(v_i|v_1, w_1, ..., v_{i-1}, w_{i-1}; \theta) \\ &= [\prod_{i=1}^{N} f_{W_i|Q_i}(w_i|q_i); \theta].[\prod_{i=1}^{N} f_{V_i|P_i}(v_i|p_i); \theta] \end{aligned} \tag{2.3}$$

où $P_1 = \emptyset, Q_1 = V_1$ et pour $i = 2, ..., N$,

$$P_i = (V_1, W_1, ..., V_{i-1}, W_{i-1}) \tag{2.4}$$

et

$$Q_i = (V_1, W_1, ..., W_{i-1}, V_i). \tag{2.5}$$

Si le premier facteur de (2.3) dépend seulement de β, Cox l'appelle vraisemblance partielle de β basée sur W. Son intérêt pratique consiste alors dans sa simplicité par rapport à celle de $F_X(x, \theta)$.

2.2.4 La vraisemblance partielle de Cox

L'utilisation de la méthode de vraisemblance partielle met en oeuvre des problèmes à étudier tels que la consistance et la normalité asymptotique des estimateurs obtenus. Dans le cas de modèle de Cox, Tsiatis (1981, [62]) et Anderson et Gill (1982, [4]) ont montré que l'estimateur de maximum de vraisemblance partielle possède les propriétés asymptotiques usuelles d'un estimateur du maximum de vraisemblance totale. On considère le triplet de l'observation d'un n-échantillon $(X_i, \Delta_i, Z_i)_{1\leq i\leq n}$ du vecteur aléatoire (T, Δ, Z), où $T = X \wedge C, \Delta = 1_{\{X\leq C\}}$ et Z est un p-vecteur ou un processus p-dimensionnel de covariables. Nous notons S_C la fonction de survie de C, f_C la densité de C et f_Z la densité de Z.
Nous supposons que X et C sont indépendants conditionnellement à Z. Nous supposons que la fonction de risque instantané est $\alpha(t)e^{\beta^\top Z}$ si Z est fixé.

On se place sous l'hypothèse de censure non informative et on suppose que la distribution de Z est indépendante de β et A. La vraisemblance obtenue en ignorant les termes expliquant les distributions de C et Z pour estimer $\theta = (\beta, A)$ peut s'écrire sous la forme:

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} [\alpha(X_i)e^{\beta^\top Z_i} \exp(-e^{\beta^\top Z_i} A(X_i))]^{\Delta_i} [\exp(-e^{\beta^\top Z_i} A(X_i))]^{1-\Delta_i} \tag{2.6}$$

Estimer le paramètre $A(t) = \int_0^t \alpha(s)ds$ relève d'un problème d'estimation fonctionnelle.

Cox propose d'ignorer le terme $\prod_{i=1}^n f_{V_i|P_i}(v_i|p_i)$ apportant peu d'information sur β et de baser l'inférence pour β sur la vraisemblance partielle. Ainsi, en maximisant la vraisemblance partielle de Cox, l'estimation du paramètre β peut se faire sans spécifier ni estimer la fonction A puisque son estimation pose problème. Il n'existe pas de maximum de (2.6) lorsque α appartient à l'ensemble des fonctions positives sur $\mathbb{R}^+$, alors on ne peut pas appliquer cette méthode d'estimation. Pour cela Cox a proposé d'estimer β sans spécifier ni estimer la fonction A_0 à partir d'une vraisemblance partielle obtenue par le principe décrit précédemment.

Nous considérons un n-échantillon $(X_i, \Delta_i, Z_i)_{1\leq i\leq n}$ du triplet (X, Δ, Z). Supposons que l'on observe L instants d'occurrence de l'événement d'intérêt (ces instants sont supposés distincts). Nous les réordonnons par ordre croissant: $T_{(1)} < \cdots < T_{(L)}$, où $(l)(l = 1, ..., L)$ désigne l'indice de l'observation $i(i = 1, \cdots, n)$ telle que $T_i = T_{(L)}$. Notons $Z_{(l)}$ le vecteur de covariables associées à cette observation.
Supposons que m_i censures se produisent dans l'intervalle $[T_{(i)}, T_{(i+1)})$ et notons $X_{(i,1)}, ..., X_{(i,m_i)}$

ces instants, auxquels sont associés les vecteurs de covariables $Z_{(i,1)}, ..., Z_{(i,m_i)}$. Reprenant les notations de l'introduction de cette section, nous définissons, pour $i = 1, \cdots, L$:

$$V_i = \{Z_l; l = 1, ..., n; T_{(i)}, X_{(i-1,j)} : j = 1, ..., m_{i-1}\}$$

et

$$W_i = \{(i)\} = \{j : T_j = T_{(i)}\}.$$

À partir de la décomposition (2.4), nous pouvons définir une suite P_i, qui contient tous les instants de censure jusqu'à $T_{(i-1)}$, tous les instants d'événement jusquà $T_{(i-1)}$ inclus, et les indices des observations correspondant à ces censures et événements. P_i contient également l'information sur les covariables de tous les sujets. La suite Q_i obtenue à partir de la décomposition (2.4) contient en plus l'événement $T_{(i)}$, ainsi que les instants de censure précédant $T_{(i)}$ et les indices des observations correspondantes à ces censures.
Pour faire de l'inférence sur β nous ignorons les termes $f_{V_i|P_i}(v_i|p_i; \beta, \alpha_0)$ dans (2.3) qui revient à ignorer, pour chaque i, la probabilité qu'aucun événement d'intérêt ne survienne dans l'intervalle $(T_{(i-1)}, T_{(i)})$, qu'un événement survienne en $T_{(i)}$, et que m_{i-1} censures se produisent en $X_{(i-1,j)}(j = 1, ..., m_{i-1})$, sachant tous les événements survenus jusqu'à $T_{(i-1)}$ inclus, toutes les censures jusqu'à $T_{(i-1)}$ et connaissant Z_l : $l = 1, ..., n$.

On peut supposer que sous certaines conditions il n'y a pas d'information perdue sur β. En effet, le fait qu'un événement se produise en $T_{(i)}$ mais qu'aucun ne survienne entre $T_{(i-1)}$ et $T_{(i)}$ pourrait être expliqué seulement par une fonction de risque α_0 voisine de 0 sur $(T_{(i-1)}, T_{(i)})$ et prenant une valeur élevée en $T_{(i)}$. Sous l'hypothèse de censure non-informative, Cox a proposé de baser l'inférence sur β sur la vraisemblance partielle

$$\prod_{i=1}^{L} P[W_i = (i)|q_i] \tag{2.7}$$

Ce terme s'interprète comme la probabilité que ce soit pour le sujet (i) que se produise l'événement d'intérêt à l'instant $T_{(i)} = t_i$, sachant qu'un événement se produit en t_i et connaissant l'ensemble des sujets à risque à cet instant.
$P[W_i = (i)|q_i]$ peut être donné comme suit:

$$P[W_i = (i)|q_i] = \frac{e^{\beta^\top Z_{(i)}}}{\sum_{j \in R(t_i)} e^{\beta^\top Z_j}}, \tag{2.8}$$

avec $R(t)$ est l'ensemble des sujets pour lesquels ni l'événement d'intérêt ni une censure ne se sont produits à l'instatnt t, appelé aussi ensemble des sujets à risque à

l'instant t.
D'après (2.8), la vraisemblance partielle de Cox est donnée par

$$\prod_{i=1}^{n} \frac{e^{\beta^\top Z_{(i)}}}{\sum_{j\in R(t_i)} e^{\beta^\top Z_{(i)}}} = \prod_{i=1}^{n} \Big[\frac{e^{\beta^\top Z_{(i)}}}{\sum_{j=1}^{n} e^{\beta^\top Z_j 1_{\{X_i\leq X_j\}}}}\Big]^{\Delta_i},$$

2.2.5 Estimation de la fonction de risque cumulé de base

La maximisation de la vraisemblance partielle ne permet pas d'estimer la fonction de risque cumulé de base A_0, puisqu'elle ne figure pas dans sa formule. Parmi les méthodes d'estimation les plus reconnues, celle proposée par Breslow (1972 [7], 1974 [8]) qui généralise l'estimateur de Nelson-Aalen. Après avoir estimé β_0 par $\widehat{\beta}_n$, l'estimateur de Breslow $\widehat{A}_0$ est donné par:

$$\begin{aligned}
\widehat{A}_0(t) &= \sum_{i=1}^{n} \frac{\Delta_i 1_{\{T_i\leq t\}}}{\sum_{j=1}^{n} e^{\widehat{\beta}_n^\top Z_i} 1_{\{T_i\leq T_j\}}} \\
&= \int_0^t \sum_{i=1}^{n} \frac{J(s)}{\sum_{j=1}^{n} e^{\widehat{\beta}_n^\top Z_i} Y_j(s)} dN_\cdot(s) \\
&= \int_0^t \frac{J(s)}{\sum_{k=1}^{m} S^{(0)}(s,\widehat{\beta}_n)} dN_\cdot(s)
\end{aligned}$$

avec $J(s) = 1_{\{Y_\cdot(s)>0\}}$. Dans le cas où il n'y a pas de covariables dans le modèle, l'estimateur de Breslow devient l'estimateur de Nelson-Aalen de A_0, donné par:

$$\widehat{A}_0(t) = \sum_{i=1}^{n} \frac{\Delta_i 1_{\{T_i\leq t\}}}{\sum_{j=1}^{n} 1_{\{T_i\leq T_j\}}}.$$

La vraisemblance partielle associée à un processus de comptage N est donnée par:

$$\begin{aligned}
L(\beta) &= \prod_{t\leq\tau}\prod_{i=1}^{n} \Big\{(\alpha_i(t)Y_i(t))^{\Delta N_i(t)}\Big\} \times \exp\Big[-\sum_i Y_i(t)A_i(\tau)\Big] \qquad (2.9)\\
&= \prod_{t\leq\tau}\prod_{i=1}^{n} \Big\{(\alpha_0(t)Y_i(t))\exp(\beta^\top Z_i)^{\Delta N_i(t)}\Big\} \times \exp\Big[-\int_0^\tau S^{(0)}(\beta,u)\alpha_0(u)du\Big]
\end{aligned}$$

avec

$$S^{(0)}(\beta,t) = \sum_j Y_j(t)\exp(\beta^\top Z_j).$$

En remplaçant dans (2.9) A_0 par son estimateur de Breslow, nous obtenons

$$\begin{aligned}
\mathcal{L}(\beta) &= \prod_t \prod_i \Big\{ (d\widehat{A}_0(t) Y_i(t) \exp(\beta^\top Z_i))^{\Delta N_i(t)} \Big\} \times \Big[- \int_0^\tau S^{(0)}(\beta, u) d\widehat{A}_0(u) \Big]; \\
&= \prod_t \prod_i \Big\{ (Y_i(t) \exp(\beta^\top Z_i))^{\Delta N_i(t)} (d\widehat{A}_0(t))^{\Delta N_i(t)} \Big\} \times \Big[- \int_0^\tau S^{(0)}(\beta, u) \frac{J(u) dN.(u)}{S^{(0)}(\beta, u)} \Big]; \\
&= \prod_t \prod_i \Big\{ [Y_i(t) \exp(\beta^\top Z_i)]^{\Delta N_i(t)} \times \Big[\frac{J(t) dN.(t)}{S^{(0)}(\beta, t)} \Big]^{\Delta N_i(t)} \Big\} \times \Big[- \int_0^\tau J(u) dN.(u) \Big]; \\
&= \prod_t \prod_i \Big\{ \Big[\frac{Y_i(t) \exp(\beta^\top Z_i)}{S^{(0)}(\beta, t)} \Big]^{\Delta N_i(t)} \times \Big[J(t) dN.(t) \Big]^{\Delta N_i(t)} \times \Big[- \int_0^\tau J(u) dN.(u) \Big]; \\
&= L(\beta) \times \prod_t \prod_i \Big\{ J(t) dN.(t) \Big\}^{\Delta N_i(t)} \times \Big[- \int_0^\tau J(u) dN.(u) \Big]
\end{aligned}$$

avec

$$L(\beta) = \prod_t \prod_{i=1} (Y_i(t) \frac{e^{\beta^\top Z_i}}{S^{(0)}(\beta, t)})^{\Delta N_i(t)},$$

dépendant de β (le reste de la vraisemblance étant indépendant de β). Par définition, $L(\beta)$ est la vraisemblance partielle de Cox.

2.2.6 Propriétés asymptotiques

Nous présentons dans cette section les propriétés asymptotiques de l'estimateur de maximum du vraisemblance partielle de β et de l'estimateur de Breslow de $A_0(.)$ définis précédemment dans le modèle de Cox. L'étude de la consistance et de la distribution asymptotique est faite par Andersen et Gill (1982, [4]). Le lecteur pourra se reporter aussi à l'exposé de Bagdonavicùis et Nikulin (2002, [5]).
Nous désignons par $\widehat{\beta}_n$ l'estimateur de maximum de vraisemblance partielle de β, $\widehat{A}_0(.)$ l'estimateur de $A_0(.)$ associé calculé pour $\beta = \widehat{\beta}_n$. Notons pour $i = 1, ..., n$, β et t donnés:

$$s^{(k)}(\beta, t) = \mathbb{E}\Big[Y_i(s) Z_i^{\otimes^k} e^{\beta^\top Z_i} \Big].$$

Notons également la matrice:

$$\sum(\beta) = \int_0^\tau \Big[s^{(2)}(\beta, t) - \frac{\{s^{(1)}(\beta, t)\}^{\otimes^2}}{s^{(0)}(\beta, t)} \Big] dA(t),$$

avec

$$\begin{aligned}
s^{(1)}(\beta, t) &= \frac{\partial s^{(0)}(\beta, t)}{\partial \beta} \\
s^{(2)}(\beta, t) &= \frac{\partial s^{(1)}(\beta, t)}{\partial \beta}.
\end{aligned}$$

L'étude asymptotique des estimateurs est basée sur les conditions suivantes:

C1 la fonction A_0 vérifie $A_0(\tau) < +\infty$,

C2 le paramètre β appartient à un ensemble borné $\mathcal{B}$ de $\mathbb{R}^p$,

C3 $\mathbb{P}(T \geq \tau) > 0$,

C4 $\sum(\beta)$ est définie positive,

C5 la covariable Z est bornée presque sûrement et sa matrice de covariance est définie positive.

Andersen et Gill (1982, [4]) ont montré le résultat suivant sous les hypothèses données ci-dessus:

Théorème 2.2.1 *L'estimateur du maximum de vraisemblance partielle $\widehat{\beta}_n$ et l'estimateur de Breslow $\widehat{A}_n$ possèdent les propriétés suivantes:*

- *$\widehat{\beta}_n$ converge en probabilité vers β_0,*
- *$\sqrt{n}(\widehat{\beta}_n - \beta_0)$ converge en loi vers un vecteur aléatoire gaussien centré de matrice de covariance $\sum(\beta_0)^{-1}$,*
- *$\sqrt{n}(\widehat{A}_0 - A_0)$ converge en loi vers un processus gaussien centré G de fonction de covariance*

$$\begin{aligned}\mathbb{C}ov(G(s), G(t)) &= \int_0^{\min(s,t)} \frac{\alpha_0(u)}{s^{(0)}(\beta, u)} du \\ &+ \int_0^t \frac{s^{(1)}(\beta, u)'}{s^{(0)}(\beta, u)} \alpha_0(u) du. \sum(\beta)^{-1} \frac{s^{(1)}(\beta, u)}{s^{(0)}(\beta, u)} \alpha_0(u) du.\end{aligned}$$

À t fixé, la variable aléatoire $\sqrt{n}(\widehat{A}_n(t) - A_0(t))$ converge vers une gaussienne centrée de variance donnée par $\mathbb{C}ov(G(s), G(t))$.

2.2.7 Résidus de martingales

Une fois que l'on a estimé le paramètre β et la fonction de base A_0, on peut se demander si le modèle est adéquat. Pour cela, on considère des résidus, et en particulier les résidus des martingales. Ils ont une moyenne égale à 0 et une distribution plutôt asymétrique même si le modèle ajusté est adéquat. Ces résidus des martingales peuvent être interprétés comme la différence au cours du temps entre le nombre d'événements observés et le nombre d'événements prédits par le modèle de Cox ajusté (Klein et Moeschberger (2003, [34])). Ces résidus peuvent être utilisés pour évaluer:

(i) la forme fonctionnelle de l'influence d'une covariable, dans un modèle qui tient compte déjà des autres covariables,

(ii) l'adéquation du modèle relativement à l'hypothèse de risques proportionnels,

(iii) l'efficacité du modèle pour prédire ce qui attend un nouveau sujet,

(iv) l'influence de chacun de ces sujets de l'étude sur l'estimation des paramètres.

On prend comme base des résidus la différence entre le processus de comptage et son compensateur:

$$M_i(t) = N_i(t) - \int_0^t Y_i(s) e^{\beta' Z_i} dA_0(s) \quad (i = 1, ..., n),$$

$M_i(t)$ est la martingale résiduelle associée au sujet i. On en a une estimation en y remplaçant β et A_0 par leurs estimateurs

$$\widehat{M}_i(t) = N_i(t) - \int_0^t Y_i(s) e^{\widehat{\beta}' Z_i} d\widehat{A}_0(s) \quad (i = 1, ..., n).$$

Notons $\widehat{M}_i(\infty)$ simplement $\widehat{M}_i$. Les résidus peuvent être interprétés comme, à chaque instant t, la différence sur $[0, t]$ entre le nombre d'événements et son espérence conditionnelle. Les résidus ont quelques une des propriétés des résidus du modèle linéaire:

(i) $\sum \widehat{M}_i(t) = 0 \quad \forall t.$

(ii) $\mathbb{E}(\widehat{M}_i) = \mathbb{C}ov(\widehat{M}_i, \widehat{M}_i) = 0$, asymptotiquement.

2.3 Modèle de Cox stratifié

Le modèle de Cox stratifié est une généralisation du modèle de Cox usuel. À partir de ce modèle, on peut séparer l'échantillon en J groupes ou strates selon les catégories d'une variable explicative. L'hypothèse de proportionalité est toujours vérifiée. À l'intérieur de ces strates, la fonction de risque de base diffère mais les covariables Z agissent de la même manière pour les différentes fonctions de risque instantané:

$$\alpha_j(t) = \alpha_{0,j}(t) \exp(\beta_0^\top Z), \quad j = 1, ..., J,$$

où β_0 est le paramètre de régression commun à toutes les strates et $\alpha_{0,j}$ est la fonction de risque de base spécifique à la strate j. On note par S la variable aléatoire qui renseigne sur la strate de l'individu. Alors de la même façon que dans le cas du modèle de Cox non stratifié, la vraisemblance partielle du n-échantillon $(T_i, \Delta_i, X_i, S_i)_{1 \leq i \leq n}$ peut être obtenue grâce au produit des vraisemblances partielles à l'intérieur de chaque strate:

$$L^{(p)}(\beta) = \prod_{j=1}^{J} \prod_{i=1}^{n} \Big(\frac{e^{\beta^\top Z_i}}{\sum_{k=1}^{n} Y_k(X_i) e^{\beta^\top Z_k} 1_{\{S_k = j\}}} \Big)^{\Delta_i 1_{\{S_i = j\}}}.$$

L'estimateur du maximum de vraisemblance partielle de β est obtenu en maximisant l'expression précédente, ou son logarithme, donné par:

$$\ln(L_n(\beta)) = \sum_{i=1}^{n}\sum_{j=1}^{J} 1_{\{S_i=j\}}\Big(\beta^\top Z_i - \ln(\sum_{k=1}^{n} Y_k(X_i)e^{\beta^\top Z_i}1_{\{S_k=j\}})\Big).$$

Si l'on note

$$S_n^{(p)}(\beta) = (\frac{\partial \ln(L_n^{(p)}(\beta))}{\partial \beta_1}, \cdots, \frac{\partial \ln(L_n^{(p)}(\beta))}{\partial \beta_p})$$

le score dérivé de la log-vraisemblance partielle, alors l'estimateur du maximum de vraisemblance partielle $\widehat{\beta}_n$ peut être obtenu en résolvant numériquement l'équation du score $S_n^{(p)}(\widehat{\beta}_n) = 0$. Les fonctions de risque cumulé de base $(A_{0,j})_{1\leq j\leq J}$ peuvent être estimées comme dans le cas du modèle non stratifié grâce à l'estimateur de Breslow appliqué à chaque strate, après avoir obtenu l'estimateur de maximum de vraisemblance partielle $\widehat{\beta}_n$ de β_0. Ces estimateurs sont donnés par

$$\widehat{A}_{0,j}(t) = \int_0^t \frac{\sum_{i=1}^n 1_{\{S_i=j\}}dN_i(s)}{\sum_{k=1}^n Y_k(s)e^{\widehat{\beta}_n^\top Z_k}1_{\{S_k=j\}}}, \quad j = 1, ..., J.$$

2.3.1 Le modèle de Cox stratifié avec strate aléatoirement manquante

Parmi les problèmes relatifs aux covariables dans le modèle de Cox, citons le problème des valeurs manquantes, qui survient le plus souvent pour des raisons pratiques, liées à la difficulté d'obtenir régulièrement des mesures de ces covariables sur tous les sujets suivis. Cette situation se rencontre lorsque des contraintes techniques inattendues viennent limiter les possibilités de recueil des données ou interrompre une partie d'un essai en cours. Par exemples dans certaines applications, l'étude du stade histologique du patient nécessite une biopsie qui ne peut pas, parfois, être effectuée sur chaque individu. Le lecteur intéressé par cette problématique pourra notamment se reporter à Paik (1997, [46]), Paik et Tsai (1997, [47]), Chen et Little (1999, [9]), Martinussen (1999, [41]) ou encore Pons (2002, [48]). Cependant, peu de travaux se sont intéressés au cas des strates manquantes.

Dans ce cadre, l'inférence dans le modèle à risques proportionnels stratifié, basée sur la vraisemblance partielle, ne peut pas être appliquée directement. En effet, la vraisemblance partielle

$$L_n^{(p)}(\beta) = \prod_{i=1}^{n}\prod_{k=1}^{K}\left(\frac{e^{\beta^\top Z_i}}{\sum_{j=1}^n e^{\beta^\top Z_j 1_{\{X_i\leq X_j\}}}}\right)^{1_{\{S_i=k\}}} \tag{2.10}$$

est obtenue à partir du produit sur toutes les strates du modèle, mais l'indicatrice $1_{\{S_i=k\}}$ n'est pas connue pour tous les individus i. Ce problème d'estimation dans le modèle de Cox stratifié avec strates manquantes a été étudié par Dupuy et Leconte (2008, [16]). Leur approche est de recourir à la méthode de Régression-Calibration (méthode RC par la suite). Détais et Dupuy (2008, [14]) ont de leur côté élaboré l'estimation avec l'algorithme Estimation-Maximisation (algorithme EM par la suite). Nous décrivons bièvement ces deux approches d'estimation dans le modèle de Cox avec strates manquantes.

2.3.2 Approche "Régression-Calibration"

La Régression-Calibration (RC) est une méthode simple pour estimer les paramètres d'un modèle de régression lorsqu'une covariable est manquante pour certains sujets d'étude. Elle consiste à remplacer la covariable non observée par un estimateur de son espérance conditionnelle aux covariables disponibles. Cette méthode a été récemment étudiée dans divers modèles de régression, tels que les modèles linéaires Huang (2005, [29]), le modèle de régression logistique Rosner et al (1989, [54]) et (1990, [53]), Spiegelman et al (2001, [57]), Weller et al (2007, [68]), les modèles généralisés Hardin et al (2003, [25]) et les modèles à risques proportionnels Prentice (1982, [49]) et Tsiatis (2006, [63]).

Notations et méthode RC

Pour simplifier les notations, on considère le cas de deux strates ($S = 1, 2$). Soit l'indicatrice de strate $D = 1_{\{S=1\}}$ égale à 1 si l'individu appartient à la strate 1 et 0 si l'individu appartient à la strate 2. Avec ces notations, la fonction de risque instantané peut être écrite comme:

$$\lambda_{Z,D} = \lambda_1 D \exp(\beta^\top Z) + \lambda_2 \overline{D} \exp(\beta^\top Z), \tag{2.11}$$

où $\overline{D} = 1 - D$ et β est le paramètre de régression commun à toutes les strates.
La log-vraisemblance partielle de β pour n copies indépendantes $(X_i, \Delta_i, Z_i, S_i)$, $i = 1, \cdots, n$ de (X, Δ, Z, S) est:

$$\sum_{i=1}^{n} \int_0^\tau \Big(\beta^\top Z_i - \overline{D}_i \log\{n^{-1} \sum_{j=1}^{n} Y_j(s) \overline{D}_j e^{\beta^\top Z_j}\} \\ - D_i \log\{n^{-1} \sum_{j=1}^{n} Y_j(s) D_j e^{\beta^\top Z_j}\} \Big) dN_i(s) \tag{2.12}$$

L'estimateur du maximum de vraisemblance de β est obtenu en maximisant (2.12).
Nous considérons maintenant la situation dans laquelle la valeur de S, ou de manière équivalente l'indicatrice de la strate D, est manquante pour certains individus. Ainsi, un sous échantillon est disponible où toutes les variables (T, Z, D) sont observées, tandis que pour les autres sujets, seules les données (T, Z) sont observées.
Dans le cas où on peut appliquer la méthode RC, on suppose qu'on observe une covariable $W \in \mathbb{R}^q$ qui fournit des informations partielles sur D. Cette covariables auxiliaire peut contenir à la fois des composantes discrètes et continues et elle est supposée indépendante de T. On note par R l'indicatrice qui vaut 1 si D est observée et 0 sinon et $\overline{R} = 1 - R$. On se place sous le régime des données aléatoirement manquantes (MAR) qui stipule que R, D et Z sont indépendantes de W. La relation entre ces données peut être exprimée de la manière suivante:

$$\mathbb{P}[D = 1|R = 0, Z, W] = \mathbb{P}[D = 1|R = 1, Z, W] = \mathbb{P}[D = 1|Z, W] \tag{2.13}$$

Procédure "Régression Calibration"

Soit $(X_i, Z_i, W_i, D_i, R_i), i = 1, \cdots, n$, n copies indépendantes de (X, Z, W, D, R). D_i est non observée lorsque $R_i = 0$. L'approche consiste alors à: (i) remplacer les variables non observées D_i par leur espérance conditionnelle $\mathbb{E}(D_i|Z_i, W_i)$ puis (ii) estimer le paramètre de régression en maximisant une version approchée de la log-vraisemblance partielle, obtenue en remplaçant les D_i manquantes par leurs estimateurs:

$$\begin{aligned}\sum_{i=1}^{n} \int_0^{\tau} \Big(\beta^\top Z_i - \overline{\xi}_i \log\{n^{-1} \sum_{j=1}^{n} Y_j(s) \overline{\xi}_j e^{\beta^\top Z_j}\} \\ -\xi_i \log\{n^{-1} \sum_{j=1}^{n} Y_j(s) \xi_j e^{\beta^\top Z_j}\}\Big) dN_i(s),\end{aligned} \tag{2.14}$$

où $\xi = R_i D_i + \overline{R}_i \mathbb{E}[D_i|Z_i, W_i]$ et $\overline{\xi} = R_i \overline{D}_i + \overline{R}_i (1 - \mathbb{E}[D_i|Z_i, W_i])$.
Ensuite, si on suppose que $\mathbb{E}[D_i|Z_i, W_i] = \mu(Z_i, W_i; \gamma)$, où μ est une fonction connue, on peut remplacer γ par l'estimateur consistant $\widehat{\gamma}$ dans l'expression de μ et on définit: $\widehat{\xi}_i = R_i D_i + \overline{R}_i \mu(Z_i, W_i; \widehat{\gamma})$ et $\widehat{\overline{\xi}}_i = 1 - \widehat{\xi}_i$. Puis on remplace ξ_i par $\overline{\xi}_i$ et $\widehat{\xi}_i$ par $\widehat{\overline{\xi}}_i$, on obtient alors la version approchée de la log-vraisemblance partielle:

$$\begin{aligned}\widehat{l}_n(\beta) = \sum_{i=1}^{n} \int_0^{\tau} \Big(\beta^\top Z_i - \widehat{\overline{\xi}}_i \log\{n^{-1} \sum_{j=1}^{n} Y_j(s) \widehat{\overline{\xi}}_j e^{\beta^\top Z_j}\} \\ -\widehat{\xi}_i \log\{n^{-1} \sum_{j=1}^{n} Y_j(s) \widehat{\xi}_j e^{\beta^\top Z_j}\}\Big) dN_i(s).\end{aligned} \tag{2.15}$$

L'estimateur de RC $\widehat{\beta}_n$ de β est défini comme la solution de l'équation

$$\widehat{U}_n(\beta) = \frac{\partial \widehat{l}_n(\beta)}{\partial \beta} = 0. \tag{2.16}$$

Théorie asymptotique

Afin d'étudier les propriétés asymptotiques de l'estimateur RC, nous introduisons quelques notations. Soit $\beta_0 = (\beta_{0,1}, \cdots, \beta_{0,d})$, $\lambda_{0,1}(\cdot), \lambda_{0,2}(\cdot)$, les vraies valeurs des paramètres. Le processus du comptage $N_i(t)$ admet l'intensité $Y_i(t)[\lambda_{0,1}(\cdot)D_i + \lambda_{0,2}(\cdot)\overline{D}_i]\exp(\beta_0^\top Z_i)$. Si les D_i sont non-observées, on a:

$$\lambda_{Z_i,W_i}(t) = (\lambda_{0,1}(t)\mathbb{E}[D_i|\mathcal{G}^i_{t-}] + \lambda_{0,2}(t)\mathbb{E}[\overline{D}_i|\mathcal{G}^i_{t-}])\exp(\beta_0^\top Z_i)$$

avec $\mathcal{G}^i_t = \sigma(1_{\{T_i \le s\}}, 1_{\{C_i \le s\}}, Z_i, W_i : 0 \le s \le t)$ et l'intensité de processus de comptage est égale à $Y_i(t)\widetilde{\lambda}_i(t)$ avec $\widetilde{\lambda}_i(t) = R_i\lambda_{Z_i,D_i}(t) + \overline{R}_i\lambda_{Z_i,W_i}(t)$. On définit aussi:

$$s^{(k)}(s,\beta) = \mathbb{E}[Y(s)Z^{\otimes k}\xi e^{\beta^\top Z}];$$

$$\overline{s}^{(k)}(s,\beta) = \mathbb{E}[Y(s)Z^{\otimes k}\overline{\xi} e^{\beta^\top Z}];$$

$$r^{(k)}(s) = \mathbb{E}[Y(s)Z^{\otimes k}\xi\widetilde{\lambda}(s)];$$

$$\overline{r}^{(k)}(s) = \mathbb{E}[Y(s)Z^{\otimes k}\overline{\xi}\widetilde{\lambda}(s)];$$

et soit $m^{(k)}(s) = r^{(k)}(s) + \overline{r}^{(k)}(s)$.

$$\widehat{S}^{(k)}_n(s,\beta) = n^{-1}\sum_{j=1}^n Y_j(s)\widehat{\xi}_j Z_j^{\otimes k} e^{\beta^\top Z};$$

$$\widehat{\overline{S}}^{(k)}_n(s,\beta) = n^{-1}\sum_{j=1}^n Y_j(s)\widehat{\overline{\xi}}_j Z_j^{\otimes k} e^{\beta^\top Z};$$

$$\overline{S}^{(k)}_n(s,\beta) = n^{-1}\sum_{j=1}^n Y_j(s)\overline{\xi}_j Z_j^{\otimes k} e^{\beta^\top Z};$$

$$S^{(k)}_n(s,\beta) = n^{-1}\sum_{j=1}^n Y_j(s)\xi_j Z_j^{\otimes k} e^{\beta^\top Z};$$

Finalement, pour $\beta \in \mathbb{R}^d$, on définit:
$\overline{I}(\beta) = \int_0^\tau \left[\frac{\overline{s}^{(2)}(s,\beta)}{\overline{s}^{(2)}(s,\beta)} - \{\frac{\overline{s}^{(1)}(s,\beta)}{\overline{s}^{(0)}(s,\beta)}\}^{\otimes 2}\right]\overline{r}^{(0)}(s)ds,$
$I(\beta) = \int_0^\tau \left[\frac{s^{(2)}(s,\beta)}{s^{(2)}(s,\beta)} - \{\frac{s^{(1)}(s,\beta)}{s^{(0)}(s,\beta)}\}^{\otimes 2}\right]r^{(0)}(s)ds,$
et $\Gamma(\beta) = \overline{I}(\beta) + I(\beta)$.
Pour établir la théorie asymptotique, nous supposons les conditions de régularité sui-

vantes:

C1 Le paramètre de régression appartient à un espace compact convexe de $\mathbb{R}^d$.

C2 Les fonctions de risque de base satisfont $\int_0^\tau \lambda_{0,1}(s)ds < \infty$ et $\int_0^\tau \lambda_{0,1}(s)ds < \infty$.

C3 Le vecteur des covariables Z est borné.

Théorème 2.3.1 ***(Dupuy et Leconte(2013, [16]))***
Sous les conditions données précédemment, $\widehat{\beta}_n$ converge en probabilité vers β^* solution du système d'équations $h(\beta) = 0$, avec

$$h(\beta) = \int_0^\tau m^{(1)}(s)ds - \int_0^\tau \frac{s^{(1)}(s,\beta)}{s^{(0)}(s,\beta)} r^{(0)}(s)ds - \int_0^\tau \frac{\overline{s}^{(1)}(s,\beta)}{\overline{s}^{(0)}(s,\beta)} \overline{r}^{(0)}(s).$$

Nous passons maintenant à la distributon asymptotique de l'estimateur RC. Rappelons que $\widehat{\beta}_n$ est la solution de l'équation $\widehat{U}_n(\widehat{\beta}_n) = 0$, où:

$$\widehat{U}_n(\beta) = \sum_{i=1}^n \int_0^\tau \left(Z_i - \widehat{\overline{\xi}}_i \frac{\widehat{\overline{S}}_n^{(1)}(s,\beta)}{\widehat{\overline{S}}_n^{(0)}(s,\beta)} - \widehat{\xi}_i \frac{\widehat{S}_n^{(1)}(s,\beta)}{\widehat{S}_n^{(0)}(s,\beta)} \right)$$

est la version approchée de:

$$U_n(\beta) = \sum_{i=1}^n \int_0^\tau \left(Z_i - \overline{\xi}_i \frac{\overline{S}_n^{(1)}(s,\beta)}{\overline{S}_n^{(0)}(s,\beta)} - \xi_i \frac{S_n^{(1)}(s,\beta)}{S_n^{(0)}(s,\beta)} \right).$$

On se propose d'exprimer $n^{-\frac{1}{2}} U_n$ comme suit:

$$n^{-\frac{1}{2}} U_n(\beta) = n^{-\frac{1}{2}} \sum_{i=1}^n V_i(\beta) + o_p(1),$$

où

$$\begin{aligned} V_i(\beta) &= \int_0^\tau \left\{ Z_i - \xi_i \frac{s^{(1)}(s,\beta)}{s^{(0)}(s,\beta)} - \overline{\xi}_i \frac{\overline{s}^{(1)}(s,\beta)}{\overline{s}^{(0)}(s,\beta)} \right\} dN_i(s) - \int_0^\tau \frac{Y_i(s)\xi e^{\beta^\top Z_i}}{s^{(0)}(s,\beta)} \left\{ Z_i - \frac{s^{(1)}(s,\beta)}{s^{(0)}(s,\beta)} \right\} \\ &\times \; r^{(0)}(s)ds - \int_0^\tau \frac{Y_i(s)\overline{\xi} e^{\beta^\top Z_i}}{\overline{s}^{(0)}(s,\beta)} \left\{ Z_i - \frac{\overline{s}^{(1)}(s,\beta)}{\overline{s}^{(0)}(s,\beta)} \right\} \overline{r}^{(0)}(s)ds. \end{aligned}$$

Théorème 2.3.2 ***(Dupuy et Leconte(2013, [16]))***
Sous les conditions de régularité données précédemment, le vecteur aléatoire $n^{-\frac{1}{2}} \widehat{U}_n(\beta^*)$ admet une distribution asymptotique normale de moyenne nulle et de matrice de covariance $W(\beta^*) = \mathbb{E}[V_1(\beta^*)^{\otimes 2}]$.

Le théorème suivant établit la normalité asymptotique de $n^{-\frac{1}{2}}(\widehat{\beta}_n - \beta^*)$.

Théorème 2.3.3 ***(Dupuy et Leconte(2013, [16]))***
Sous les mêmes conditions de régularité, $n^{-\frac{1}{2}}(\widehat{\beta}_n - \beta^*)$ admet une distribution asymptotique normale de moyenne nulle et de matrice de covariance:

$$\Gamma^{-1}(\beta^*)W(\beta^*)\Gamma^{-1}(\beta^*).$$

2.3.3 Estimation par l'algorithme EM

Nous supposons que la loi de T conditionnelle aux variables explicatives est donnée par:

$$\lambda(t) = \lambda_k(t)\exp(\beta^\top X), \tag{2.17}$$

où β est le paramètre de régression commun à toutes les strates $\lambda_k, k = 1, \cdots, K$ sont des fonctions de risque instantané de base associées aux K strates et $\Lambda_k = \int \lambda_k$ sont les fonctions de risque cumulé de base. L'idée est de fournir, dans le modèle de Cox stratifié avec strates manquantes, un estimateur du paramètre de régression qui soit à la fois consistant et asymptotiquement gaussien. Cette approche est traitée par Détais et Dupuy (2008, [14]). Soit $W \in \mathbb{R}^m$ une covariable auxiliaire qui donne des informations partielles sur S lorsqu'elle est manquante. La probabilité d'appartenir à la strate $k \in \mathcal{K} = \{1, \cdots, K\}$ connaissant les covariables W est ainsi donnée par

$$\mathbb{P}(S = k|W) = \frac{\exp(\gamma_k^\top W)}{\sum_{j=1}^k \exp(\gamma_k^\top W)},$$

où $\gamma_k \in \mathbb{R}^m$ est un vecteur de régression. Dans la suite, par souci d'identifiabilité, nous posons $\gamma_K = 0$ et notons $\gamma = (\gamma_1^\top, \cdots, \gamma_{K-1}^\top)$ le vecteur concaténé des vecteurs de régression associés à chaque strate. On désigne par R l'indicatrice qui est égale à 1 si S est observée et 0 si non. Notons $\pi_{k,\gamma}(W) = \mathbb{P}(S = k|W)$ pour $k \in \mathcal{K} = \{1, \cdots, K\}$. Notre n-échantillon est donc composé des observations $\mathcal{O}_i = (X_i, \Delta_i, Z_i, W_i, R_i, R_iS_i), i \in \{1, \cdots, n\}$ qui sont supposées être n répliques indépendantes et de même loi que (X, Δ, Z, W, R, RS).

Le paramètre inconnu est $\theta = (\beta, \gamma, \Lambda_k; 1 \leq k \leq K)$ et la vraie valeur du modèle est notée $\theta_0 = (\beta_0, \gamma_0, \Lambda_{k,0}; 1 \leq k \leq K)$. Le paramètre d'intérêt pour l'inférence sur ce modèle est le paramètre de régression β, les paramètres γ et Λ_k sont considérés comme des paramètres de nuisance. Les estimateurs de $(\beta_0, \gamma_0, \Lambda_{k,0}; k \in \mathcal{K})$ sont obtenus en maximisant une version modifiée de la log-vraisemblance. S'ils existent, ces estimateurs sont des estimateurs non paramétriques de maximum de vraisemblance (NPMLE, par la suite).

Comme les strates ne sont pas complètement observées, nous ne pouvons pas utiliser la formule de la vraisemblance partielle pour l'estimation de β. Par conséquent, nous calculons la vraisemblance observée relative au paramètre θ. La vraisemblance est donnée par:

$$\begin{aligned} L_n^{(obs)}(\theta) &= \prod_{i=1}^{n}\Big(\prod_{k=1}^{K}(\lambda_k(X_i)^{\Delta_i}\exp(\Delta_i\beta^\top Z_i-\Lambda_k(X_i)e^{\beta^\top Z_i})\pi_{k,\gamma}(W_i))^{1_{\{S_i=k\}}}\Big)^{R_i} \\ &\times \Big(\sum_{k=1}^{K}\lambda_k(X_i)^{\Delta_i}\exp(\Delta_i\beta^\top Z_i-\Lambda_k(X_i)e^{\beta^\top Z_i})\pi_{k,\gamma}(W_i))^{1_{\{S_i=k\}}}\Big)^{1-R_i}. \end{aligned}$$

Cette expression pose des difficultés lors de la recherche de l'estimateur non-paramétrique du maximum de vraisemblance (NPMLE). Cette problématique peut être résolue par l'algorithme Espérance Maximisation (EM), Dempster et al.(1977, [13]). Cette approche consiste à résoudre le problème de maximisation en procédant itérativement sur la log-vraisemblance des données complètes. La covariable non observée est remplacée par son espérance conditionnée par les observations. Cela conduit à envisager un algorithme itératif qui se résume en deux étapes principales:

- Étape de calcul de l'espérance de la log-vraisemblance des données complètes conditionnée par les données réellement observées.
- Étape de maximisation: actualisation du paramètre courant comme une des valeurs maximisant la quantité calculée dans l'étape précédente.

Pour $1 \leq i \leq n$ et $k \in K$, soit $w_i(k,\theta)$ la probabilité d'appartenir à la k-ième strate sachant $(T_i, \Delta_i, X_i, W_i)$ et la valeur du paramètre θ. Soit $\mathcal{Q}(\mathcal{O}_i, k, \theta)$ l'espérance de $1_{\{S_i=k\}}$ sachant $\mathcal{O}_i$ et θ. Alors $\mathcal{Q}(\mathcal{O}_i, k, \theta)$ peut s'écrire sous la forme

$$\mathcal{Q}(\mathcal{O}_i, k, \theta) = R_i 1_{\{S_i=k\}} + (1-R_i)w_i(k,\theta)$$

Théorème 2.3.4 ***(Détais et Dupuy (2008, [14]))***
Sous certaines conditions de régularité, le NPMLE $\widehat{\theta}_n = (\widehat{\beta}_n, \widehat{\gamma}_n, \widehat{\Lambda}_{k,n}; k \in \mathcal{K})$ de θ_0 existe et l'estimateur NPMLE $\widehat{\theta}_n$ satisfait l'équation pour tout $k \in \mathcal{K}$:

$$\widehat{\Lambda}_{k,n}(t) = \int_0^t \sum_{i=1}^{n} \frac{\mathcal{Q}(\mathcal{O}_i, k, \theta)}{\mathcal{Q}(\mathcal{O}_j, k, \theta)\exp(\widehat{\beta}_n^\top X_j)Y_j(s)} dN_i(s).$$

Voici quelques résultats asymptotiques des estimateurs proposés:

Théorème 2.3.5 ***(Détais et Dupuy (2008, [14]))***
Sous certaines conditions de régularité, $\|\widehat{\beta}_n - \beta_0\|$ et $\sup_{t\in[0,\tau]} |\widehat{\Lambda}_{k,n}(t) - \Lambda_{k,0}(t)|$ convergent vers 0 presque sûrement lorsque n tend vers l'infini.

Afin de donner la convergence asymptotique de $\widehat{\beta}_n$, on définit $\{e_1, \cdots, e_p\}$ la base canonique de $\mathbb{R}^p$, $\widetilde{\sigma}_\beta^{-1} = \sigma_\beta^{-1}(u, 0, 0, k \in K)$, $\widetilde{\sigma}_\gamma^{-1} = \sigma_\beta^{-1}(0, u, 0, k \in K)$ et $\widetilde{\sigma}_\Lambda^{-1} = \sigma_\beta^{-1}(0, 0, u, k \in K)$ sont des opérateurs linéaires.

Théorème 2.3.6 ***(Détais et Dupuy (2008, [14]))***
Sous certaines conditions de régularité, $\sqrt{n}(\widehat{\gamma}_n - \gamma_0)$ admet une distribution asymptotique normale $N(0, \Sigma_\gamma)$. De plus, pour tout $t \in [0, \tau]$ et $j \in K$, $\sqrt{n}(\widehat{\Lambda}_{j,n} - \Lambda_{j,0}(t)$ est asymptotiquement distribuée comme une loi $N(0, v_j^2(t))$, où

$$v_j^2(t) = \int_0^t \sigma_{\Lambda_j}^{-1}(h_{j,t})(u) d\Lambda_{j,0}(u)$$

avec h une application linéaire donnée. Les quantités $v_j^2(t)$ et σ^{-1} n'ont pas des expression explicites.

Le lecteur intéressé peut se reporter à Détais et Dupuy (2008, [14]) dans lequel les auteurs ont traité ce problème en donnant explicitement les conditions de régularité pour montrer les propriétés asymptotiques.

3

A goodness-of-fit test for the stratified proportional hazards model for survival data and the problem of stratification with unknown threshold

Abstract

Goodness-of-fit testing is addressed in the stratified proportional hazards model for survival data. A test statistic based on within-strata cumulative sums of martingale residuals over covariates is proposed and its asymptotic distribution is derived under the null hypothesis of model adequacy. A Monte Carlo procedure is proposed to approximate the critical value of the test. Simulation studies are conducted to examine finite-sample performance of the proposed statistic. The stratified proportional hazards model with unknown threshold is also studied. An estimator of the regression parameter β is proposed, based on the partial likelihood and its behaviour is also examined. A two-steps estimation method of the variance of regression parameter is proposed and comparisons are made with an oracle method. Important gains in term of variance and mean square error can be made with the proposed method.

Keywords: asymptotic distribution, martingale, residuals, simulations, survival analysis.

Sommaire

Ce chapitre fait l'objet d'un article soumis pour publication à "Communication in Statistics:Theory and methods".

3.1 A goodness-of-fit test for the stratified proportional hazards model for survival data

THe stratified proportional hazards model (see for example [5, 42]) generalizes the usual Cox proportional hazards regression model (see [11]) for survival data by allowing different groups (called strata) of the population under study to have distinct baseline hazard functions. Precisely, in the stratified model, the strata divide the sample individuals into J disjoint groups, each having a distinct baseline hazard function $\alpha_{0,j}$ but a common value for the regression parameter. The hazard function of an individual in stratum j thus takes the form

$$\alpha_j(t) = \alpha_{0,j}(t)e^{\beta_0^\top \mathbf{Z}}, \quad j = 1, \ldots, J, \tag{3.1}$$

where $\mathbf{Z}$ is a p-vector of covariates, β_0 is a p-vector of unknown regression parameters of interest, $\top$ denotes the transpose and $\{\alpha_{0,j}(t) : t \geq 0, j = 1, \ldots, J\}$ are J unknown baseline hazard functions. A consistent and asymptotically normal estimator of β_0 is obtained by maximizing the so-called partial likelihood function (see [12]). The partial likelihood for the stratified model (3.1) is the product over strata of the within-stratum partial likelihoods (see [5]). Further discussion of this model, including estimation of the cumulative baseline hazard functions $A_{0,j}(t) = \int_0^t \alpha_{0,j}(u)\, du$,

can be found in [3] and [42].

The stratified proportional hazards model recently began to be increasingly used in various fields, such as economy, marketing, medicine and public health. For example, in [32], authors use this model to assess the relative performance of piggyback loans and insured loans with respect to residential mortgage lifetimes in USA. In [23], the author uses a stratified proportional hazards model to identify risk factors of churn among customers of a mobile phone operator. In [55], a stratified proportional hazards model is used to evaluate mortality after radical prostatectomy. In [45], association between exposure to antibiotics in fetal and early life and asthma in childhood is also investigated by using a stratified model.

Despite this recent interest, methodological developments for the stratified proportional hazards model are still a few. In [16] and [22], authors investigate estimation in the stratified model with covariate missing values and covariate measurement error respectively. The prognostic ability of the stratified model is assessed in [10] and [44]. In [26], authors propose a measure to quantify the partial dependence between a survival time and a covariate in the stratified model. Confidence intervals for the difference of median survival times in the stratified model are investigated in [33]. Finally, a stratified proportional hazards model with spatio-temporal heterogeneity is developed in [24]. In the present paper, we consider goodness-of-fit testing for the stratified proportional hazards model.

Many goodness-of-fit tests have been proposed for the usual (unstratified) proportional hazards model. In particular, several authors discussed goodness-of-fit tests based on weighted sums of martingale residuals. A non-exhaustive list of references includes [20, 37, 43, 66], see also [5] for a detailed account of this topic. However, to the best of our knowledge, no goodness-of-fit tests have been proposed yet for the stratified proportional hazards model (3.1). Our paper aims at filling this gap.

We propose a goodness-of-fit test statistic for model (3.1) based on within-stratum cumulative sums of residuals. We establish rigorously the asymptotic distribution of this statistic under the null hypothesis that model (3.12) is correct. A Monte Carlo procedure is proposed to approximate the critical value of the test for a given asymptotic level. Finite-sample performance of the test are investigated via simulations.

The rest of the paper is organized as follows. In Section 3.2, we construct our test statistic, we investigate its null asymptotic distribution and we discuss Monte Carlo approximation of the critical value. In Section 3.3, we conduct a simulation study to assess level and power properties of our test for various numbers of strata, sample sizes and proportions of censored observations. Section 3.6 concludes the paper with some perspectives. Proofs of some intermediate technical results are given in an appendix.

3.2 The proposed test statistic and decision rule

3.2.1 Preliminaries and notations

All random variables are defined on the same probability space $(\Omega, \mathcal{C}, \mathbb{P})$. Let T be a random failure time whose distribution depends on a vector of covariates $\mathbf{Z} = (Z_1, \ldots, Z_p)^\top \in \mathbb{R}^p$ and on a stratum indicator $S \in \{1, \ldots, J\}$. We assume that conditionally on $\mathbf{Z}$ and $S = j$, the hazard function of T is given by model (3.12). We suppose that T is randomly right-censored by a positive random variable C and that T and C are independent conditionally on $\mathbf{Z}$ and S. The analysis is restricted to the time interval $\mathcal{T} := [0, \tau]$, where $\tau < \infty$ denotes the end of the study. Therefore, we actually observe the duration $X = \min(T, \min(C, \tau))$ and a censoring indicator $\Delta = 1_{\{T \leq \min(C,\tau)\}}$.

Available data consist of n independent and identically distributed copies $(X_i, \Delta_i, \mathbf{Z}_i, S_i)$ of the random vector $(X, \Delta, \mathbf{Z}, S)$. For every $i = 1, \ldots, n$ and $t \in \mathcal{T}$, we denote by $N_i(t) = 1_{\{X_i \leq t\}}\Delta_i$ and $Y_i(t) = 1_{\{X_i \geq t\}}$ the failure counting and at-risk processes of the i-th individual respectively. The process $N_i(t)$ has intensity $Y_i(t) \sum_{j=1}^J \alpha_{0,j}(t) e^{\beta_0^\top \mathbf{Z}_i} 1_{\{S_i = j\}}$ with respect to the filtration $(\mathcal{F}_{t,i})_{t \geq 0}$ defined by $\mathcal{F}_{t,i} = \sigma\{\mathbf{Z}_i, S_i, N_i(s), Y_i(s) \colon 0 \leq s \leq t\}$, so that processes

$$M_i(t) = N_i(t) - \int_0^t Y_i(s) \sum_{j=1}^J \alpha_{0,j}(s) e^{\beta_0^\top \mathbf{Z}_i} 1_{\{S_i = j\}}\, ds, i = 1, \ldots, n$$

are martingales.

As mentioned above, a consistent and asymptotically normal estimator $\widehat{\beta}_n$ of β_0 in model (3.1) is obtained by maximizing the partial likelihood

$$\prod_{j=1}^J \prod_{i=1}^n \left(\frac{e^{\beta^\top \mathbf{Z}_i}}{\sum_{k=1}^n Y_k(X_i) e^{\beta^\top \mathbf{Z}_k} 1_{\{S_k = j\}}} \right)^{\Delta_i 1_{\{S_i = j\}}},$$

(see [5]) while cumulative baseline hazard functions $A_{0,j}(t) = \int_0^t \alpha_{0,j}(u)\, du$ can be estimated by Breslow-type estimators

$$\widehat{A}_{0,j}(t) = \int_0^t \frac{\sum_{i=1}^n 1_{\{S_i = j\}} dN_i(s)}{\sum_{k=1}^n Y_k(s) e^{\widehat{\beta}_n^\top \mathbf{Z}_k} 1_{\{S_k = j\}}}, \quad j = 1, \ldots, J.$$

Martingale residuals in model (3.1) can be defined similarly as in the classical unstratified Cox model. Precisely, we define the residual martingale for the i-th individual

as

$$\widehat{M}_i(t,\widehat{\beta}_n) = N_i(t) - \sum_{j=1}^{J} \int_0^t \frac{Y_i(s)e^{\widehat{\beta}_n^\top \mathbf{Z}_i} 1_{\{S_i=j\}}}{S_j^{(0)}(s,\widehat{\beta}_n)}\, d\bar{N}_j(s), \tag{3.2}$$

where $\bar{N}_j(s) = \sum_{i=1}^n N_i(s) 1_{\{S_i=j\}}$ and $S_j^{(0)}(s,\widehat{\beta}_n) = \sum_{i=1}^n Y_i(s) e^{\widehat{\beta}_n^\top \mathbf{Z}_i} 1_{\{S_i=j\}}$.

3.2.2 The proposed test statistic and its asymptotic distribution under H_0

We construct a test statistic for the null hypothesis that the stratified proportional hazards model (3.1) is true, that is, for

$$H_0 : \alpha_j(t) = \alpha_{0,j}(t) e^{\beta_0^\top \mathbf{Z}}, \quad \text{for } j = 1, \ldots, J.$$

Our test statistic is based on within-stratum partial cumulative sums of the martingale residuals (3.2). So far, model-checking techniques based on cumulative sums of residuals have been developed for various models such as the generalized linear model (*e.g.*, [58, 59]), Cox's model (*e.g.*, [5, 20, 37, 43, 66]) and Aalen's additive risk model ([19]). For the stratified proportional hazards model (3.1), we propose to consider the process

$$Q_{n,\mathbf{z}}^j(t,\widehat{\beta}_n) := \frac{1}{\sqrt{n}} \sum_{i=1}^n \widehat{M}_i(t,\widehat{\beta}_n) 1_{\{S_i=j\}} 1_{\{\mathbf{Z}_i \le \mathbf{z}\}}, \tag{3.3}$$

where $\mathbf{z} = (z_1, \ldots, z_p)^\top \in \mathbb{R}^p$ and the event $\{\mathbf{Z}_i \le \mathbf{z}\}$ means that all components $Z_{1,i}, \ldots, Z_{p,i}$ of $\mathbf{Z}_i$ are less than or equal to the corresponding components of $\mathbf{z}$. Under model (3.12), we have

$$\sum_{i=1}^n \widehat{M}_i(t,\widehat{\beta}_n) 1_{\{S_i=j\}} = 0$$

for every $j = 1, \ldots, J$. Under H_0, one would thus expect that the partial-sum process $Q_{n,\mathbf{z}}^j(t,\widehat{\beta}_n)$ fluctuates around 0 and a large value of $Q_{n,\mathbf{z}}^j(t,\widehat{\beta}_n)$ should lead to the conclusion that model (3.12) is misspecified for the j-th stratum. Note that the function $\mathbf{z} \mapsto Q_{n,\mathbf{z}}^j(t,\widehat{\beta}_n)$ has possible jumps at the distinct values of the $\mathbf{Z}_i$, $i = 1, \ldots, n$. Therefore it is sufficient to consider a finite number Q of values for the vector $\mathbf{z}$ (this issue is discussed in Section **??**). If $\mathbf{z}_1, \ldots, \mathbf{z}_Q$ are Q such values, we consider the within-stratum statistics $\max_{q \in \mathcal{Q}} |Q_{n,\mathbf{z}_q}^j(t,\widehat{\beta}_n)|$ (for $j = 1, \ldots, J$), where $\mathcal{Q} = \{1, \ldots, Q\}$

and we combine all within-stratum statistics in the overall test statistic

$$\max_{1\le j\le J}\max_{q\in\mathcal{Q}}|Q^j_{n,\mathbf{z}_q}(t,\widehat{\beta}_n)|.$$

We consider this statistic at $t=\tau$ in order to use the whole data information and finally, we define our goodness-of-fit test statistic as:

$$\mathcal{S}_n := \max_{1\le j\le J}\max_{q\in\mathcal{Q}}|Q^j_{n,\mathbf{z}_q}(\tau,\widehat{\beta}_n)|. \tag{3.4}$$

We now establish the asymptotic distribution of $\mathcal{S}_n$ under the null hypothesis H_0. We first introduce some further notations and some regularity conditions. If u is a vector in $\mathbb{R}^p$, let $u^{\otimes 0}=1$ and $u^{\otimes 1}=u$. For $k=0,1$, $j=1,\dots,J$, $s\in\mathcal{T}$ and $\beta\in\mathbb{R}^p$, we define

$$\begin{aligned}
S^{(k)}_j(s,\beta) &= \sum_{i=1}^n Y_i(s)\mathbf{Z}_i^{\otimes k}e^{\beta^\top\mathbf{Z}_i}1_{\{S_i=j\}},\\
S^{(k)}_{j,\mathbf{z}}(s,\beta) &= \sum_{i=1}^n Y_i(s)\mathbf{Z}_i^{\otimes k}e^{\beta^\top\mathbf{Z}_i}1_{\{S_i=j\}}1_{\{\mathbf{Z}_i\le\mathbf{z}\}},\\
s^{(k)}_j(s,\beta) &= \mathbb{E}\left(Y_i(s)\mathbf{Z}_i^{\otimes k}e^{\beta^\top\mathbf{Z}_i}1_{\{S_i=j\}}\right),\\
s^{(k)}_{j,\mathbf{z}}(s,\beta) &= \mathbb{E}\left(Y_i(s)\mathbf{Z}_i^{\otimes k}e^{\beta^\top\mathbf{Z}_i}1_{\{S_i=j\}}1_{\{\mathbf{Z}_i\le\mathbf{z}\}}\right).
\end{aligned}$$

Finally, $\|\cdot\|_{\mathcal{T}}$ and $\|\cdot\|_{\mathcal{T}\times\mathcal{B}}$ will denote uniform norms on $\mathcal{T}$ and $\mathcal{T}\times\mathcal{B}$ respectively. The asymptotic distribution of our goodness-of-fit statistic $\mathcal{S}_n$ will be established under the following conditions:

C1 β_0 belongs to the interior of a compact subset $\mathcal{B}$ of $\mathbb{R}^p$.

C2 The time τ is such that $\int_0^\tau \alpha_{0,j}(u)\,du<\infty$ for all $j=1,\dots,J$.

C3 The covariate $\mathbf{Z}$ is bounded.

C4 For $j=1,\dots,J$ and every $\mathbf{z}\in\mathbb{R}^p$, the families $\{s^{(k)}_{j,\mathbf{z}}(s,\cdot);s\in\mathcal{T}\}$ and $\{s^{(k)}_j(s,\cdot);s\in\mathcal{T}\}$ $(k=0,1)$ of functions of β are equicontinuous at β_0.

C5 For $j=1,\dots,J$, there exists a constant $c>0$ such that $\inf_{s\in\mathcal{T}} s^{(0)}_j(s,\beta_0)>c$.

Our result is as follows:

Theorem 3.1 *Let $Q\in\mathbb{N}$ and $\mathbf{z}_1,\dots,\mathbf{z}_Q$ be Q vectors in $\mathbb{R}^p$. Let $\mathbb{W}$ be a $(J\cdot Q)$-dimensional Gaussian vector with components denoted by $\mathbb{W}:=(\mathbb{W}^{1\top},\dots,\mathbb{W}^{J\top})^\top$, where $\mathbb{W}^j:=(W^j_1,\dots,W^j_Q)^\top$ for $j=1,\dots,J$. Assume that $\mathbb{W}$ has mean zero and variance-covariance matrix $\mathbb{S}$, where $\mathbb{S}$ is the $(J\cdot Q)\times(J\cdot Q)$ block-matrix*

$$\begin{pmatrix}\Sigma^1 & \dots & 0_{Q\times Q}\\ \vdots & \ddots & \vdots\\ 0_{Q\times Q} & \dots & \Sigma^J\end{pmatrix}$$

and $\Sigma^j = (\Sigma^j_{k,\ell})_{1\le k,\ell\le Q}$ *is the* $Q\times Q$ *matrix with components*

$$\Sigma^j_{k,\ell} := \mathbb{E}\left[\int_0^\tau \alpha_{0,j}(s)Y_i(s)e^{\beta_0^\top \mathbf{Z}_i}1_{\{S_i=j\}}1_{\{\mathbf{Z}_i\le \mathbf{z}_k\}}1_{\{\mathbf{Z}_i\le \mathbf{z}_\ell\}}\,ds\right] - \int_0^\tau \alpha_{0,j}(s)\frac{s^{(0)}_{j,\mathbf{z}_k}(s,\beta_0)\cdot s^{(0)}_{j,\mathbf{z}_\ell}(s,\beta_0)}{s^{(0)}_j(s,\beta_0)}\,ds$$

and $0_{Q\times Q}$ *is a* $Q\times Q$ *matrix with all components zero. Assume that conditions C1-C5 hold. Then under* H_0*, the test statistic* $\mathcal{S}_n$ *converges in distribution, as* n *tends to infinity, to* $\mathcal{S} := \max_{1\le j\le J}\max_{q\in\mathcal{Q}}|W^j_q|$.

Proof. We need two intermediate lemmas. Their proofs are postponed to an appendix.

Lemma 3.1 *Let* $\mathbf{z}_1,\dots,\mathbf{z}_Q$ *be* Q *vectors in* $\mathbb{R}^p$*. Assume that conditions C1-C5 hold and for* $j=1,\dots,J$*, let* $\mathbb{Q}^j_n(\cdot,\beta) := (Q^j_{n,\mathbf{z}_1}(\cdot,\beta),\dots,Q^j_{n,\mathbf{z}_Q}(\cdot,\beta))^\top$*. Then under* H_0*, the process* $(\mathbb{Q}^1_n(\cdot,\beta_0)^\top,\dots,\mathbb{Q}^J_n(\cdot,\beta_0)^\top)^\top$ *converges weakly in* $(D(\mathcal{T}))^{J\cdot Q}$ *to a zero-mean Gaussian process* $(\mathbb{W}^{1\top},\dots,\mathbb{W}^{J\top})^\top$ *(with* $\mathbb{W}^j := (W^j_1,\dots,W^j_Q)^\top$*) with covariance function*

$$cov(W^j_k(t_1),W^j_\ell(t_2)) = \mathbb{E}\left[\int_0^{\min(t_1,t_2)} \alpha_{0,j}(s)Y_i(s)e^{\beta_0^\top \mathbf{Z}_i}1_{\{S_i=j\}}1_{\{\mathbf{Z}_i\le \mathbf{z}_k\}}1_{\{\mathbf{Z}_i\le \mathbf{z}_\ell\}}\,ds\right] - \int_0^{\min(t_1,t_2)} \alpha_{0,j}(s)\frac{s^{(0)}_{j,\mathbf{z}_k}(s,\beta_0)\cdot s^{(0)}_{j,\mathbf{z}_\ell}(s,\beta_0)}{s^{(0)}_j(s,\beta_0)}\,ds$$

and $cov(W^j_k(t_1),W^{j'}_\ell(t_2)) = 0$ *if* $j\ne j'$ *(with* $1\le k,\ell\le Q$ *and* $t_1,t_2\in\mathcal{T}$*).*

Lemma 3.2 *Assume that conditions C1-C5 hold. Then under* H_0*,* $\mathbb{Q}^j_n(\cdot,\widehat{\beta}_n) = \mathbb{Q}^j_n(\cdot,\beta_0) + o_{\mathbb{P}}(1)$ *where the* $o_{\mathbb{P}}(1)$ *is uniform on* $\mathcal{T}$*.*

We now prove Theorem 3.1. It follows from Lemma 3.1 and Lemma 3.2 that under H_0, the process $(\mathbb{Q}^1_n(\cdot,\widehat{\beta}_n)^\top,\dots,\mathbb{Q}^J_n(\cdot,\widehat{\beta}_n)^\top)^\top$ converges weakly to a zero-mean Gaussian process on $\mathcal{T}^{J\cdot Q}$, with covariance function as in Lemma 3.1. Hence $(\mathbb{Q}^1_n(\tau,\widehat{\beta}_n)^\top,\dots,\mathbb{Q}^J_n(\tau,\widehat{\beta}_n)^\top)^\top$ converges in distribution to a Gaussian vector with mean zero and variance-covariance matrix given by $\mathbb{S}$ in Theorem 3.1 (this vector is denoted by $\mathbb{W} := (\mathbb{W}^{1\top},\dots,\mathbb{W}^{J\top})^\top$, with $\mathbb{W}^{j\top} := (W^j_1,\dots,W^j_Q)$, $j=1,\dots,J$). By the continuous mapping theorem, $\mathcal{S}_n$ converges in distribution to $\max_{1\le j\le J}\max_{q\in\mathcal{Q}}|W^j_q|$.

□

Based on Theorem 3.1, a natural decision rule is as follows: reject H_0 (at the asymptotic level $\alpha\in(0,1)$) if $\mathcal{S}_n$ is greater than the quantile $q_{1-\alpha}(\mathcal{S})$ of order $1-\alpha$ of the distribution of $\mathcal{S}$. However, this quantile - or critical value of the test - is unknown. In the next section, we propose to estimate it via Monte Carlo simulations.

3.2.3 Monte Carlo estimation of the critical value and decision rule

In order to estimate $q_{1-\alpha}(\mathcal{S})$, we propose to: i) simulate a large number (say M) of realizations of $\mathbb{W}$, ii) calculate the corresponding values $s_1, \ldots, s_M$ of $\mathcal{S}$, iii) approximate $q_{1-\alpha}(\mathcal{S})$ by the empirical quantile of order $1-\alpha$ of $(s_1, \ldots, s_M)$. However, the matrix $\mathbb{S}$ needed to simulate data from $\mathbb{W} \sim \mathcal{N}_{(J \cdot Q)}(0, \mathbb{S})$ is unknown. Under H_0, this matrix can be estimated by replacing unknown terms in $\Sigma^j_{k,\ell}$ by their empirical counterparts. Precisely, we define:

$$\widehat{\Sigma}^j_{k,\ell} := \int_0^\tau \left\{ \frac{S^{(0)}_{j,\check{\mathbf{z}}_{k,\ell}}(s, \widehat{\beta}_n)}{S^{(0)}_j(s, \widehat{\beta}_n)} - \frac{S^{(0)}_{j,\mathbf{z}_k}(s, \widehat{\beta}_n) \cdot S^{(0)}_{j,\mathbf{z}_l}(s, \widehat{\beta}_n)}{(S^{(0)}_j(s, \widehat{\beta}_n))^2} \right\} \frac{d\bar{N}_j(s)}{n},$$

where $\check{\mathbf{z}}_{k,\ell}$ denotes the vector in $\mathbb{R}^p$ whose j-th component ($j = 1, \ldots, p$) is the minimum of the j-th components of $\mathbf{z}_k$ and $\mathbf{z}_\ell$. This estimator is consistent, as stated in Proposition 3.2.1.

Proposition 3.2.1 *Assume that conditions C1-C5 hold. Then, under H_0, $\widehat{\Sigma}^j_{k,\ell}$ converges in probability to $\Sigma^j_{k,\ell}$ as n tends to infinity.*

Proof. Let

$$V^j_{k,\ell}(s, \beta) := \frac{S^{(0)}_{j,\check{\mathbf{z}}_{k,\ell}}(s, \beta)}{S^{(0)}_j(s, \beta)} - \frac{S^{(0)}_{j,\mathbf{z}_k}(s, \beta) \cdot S^{(0)}_{j,\mathbf{z}_\ell}(s, \beta)}{(S^{(0)}_j(s, \beta))^2}$$

and

$$v^j_{k,\ell}(s, \beta) := \frac{s^{(0)}_{j,\check{\mathbf{z}}_{k,\ell}}(s, \beta)}{s^{(0)}_j(s, \beta)} - \frac{s^{(0)}_{j,\mathbf{z}_k}(s, \beta) \cdot s^{(0)}_{j,\mathbf{z}_\ell}(s, \beta)}{(s^{(0)}_j(s, \beta))^2}.$$

We have:

$$\begin{aligned}
\left|\widehat{\Sigma}^j_{k,\ell} - \Sigma^j_{k,\ell}\right| &= \left| \int_0^\tau V^j_{k,\ell}(s, \widehat{\beta}_n) \frac{d\bar{N}_j(s)}{n} - \int_0^\tau v^j_{k,\ell}(s, \beta_0) s^{(0)}_j(s, \beta_0) \alpha_{0,j}(s)\, ds \right| \\
&\leq \left| \int_0^\tau \left\{ V^j_{k,\ell}(s, \widehat{\beta}_n) - v^j_{k,\ell}(s, \widehat{\beta}_n) \right\} \frac{d\bar{N}_j(s)}{n} \right| \\
&\quad + \left| \int_0^\tau \left\{ v^j_{k,\ell}(s, \widehat{\beta}_n) - v^j_{k,\ell}(s, \beta_0) \right\} \frac{d\bar{N}_j(s)}{n} \right| \\
&\quad + \left| \int_0^\tau v^j_{k,\ell}(s, \beta_0) \left\{ \frac{d\bar{N}_j(s)}{n} - \frac{1}{n} S^{(0)}_j(s, \beta_0) \alpha_{0,j}(s)\, ds \right\} \right| \\
&\quad + \left| \int_0^\tau v^j_{k,\ell}(s, \beta_0) \left\{ \frac{1}{n} S^{(0)}_j(s, \beta_0) - s^{(0)}_j(s, \beta_0) \right\} \alpha_{0,j}(s)\, ds \right| \qquad (3.5)
\end{aligned}$$

Under conditions C1-C5, $\|V^j_{k,\ell} - v^j_{k,\ell}\|_{\mathcal{T}\times\mathcal{B}}$ converges to zero in probability. Moreover, $\bar{N}_j(\tau)/n$ tends to $\mathbb{E}[N(\tau)1_{\{S=j\}}] < \infty$. Therefore, the first term in (3.5) converges to zero in probability. Condition C4 and the convergence of $\bar{N}_j(\tau)/n$ to $\mathbb{E}[N(\tau)1_{\{S=j\}}] < \infty$ imply that the second term in (3.5) also converges to zero. Consider the third term in (3.5). Let $\delta > 0, \eta > 0$. By Lenglart inequality (*e.g.*, [3]), we have

$$\mathbb{P}\left(\left|\int_0^\tau v^j_{k,\ell}(s,\beta_0)\frac{1}{n}d\bar{M}_j(s)\right| > \eta\right) \\ \leq \frac{\delta}{\eta^2} + \mathbb{P}\left(\int_0^\tau (v^j_{k,\ell}(s,\beta_0))^2\frac{1}{n}S^{(0)}_j(s,\beta_0)\alpha_{0,j}(s)\,ds > n\delta\right).$$

For any $\delta > 0$, the probability on the right-hand side of this inequality converges to zero as n tends to infinity, and since δ is arbitrary, $\mathbb{P}(|\int_0^\tau v^j_{k,\ell}(s,\beta_0)\frac{1}{n}d\bar{M}_j(s)| > \eta)$ must converge to zero. Thus the third term in (3.5) converges to zero in probability. Finally,

$$\left|\int_0^\tau v^j_{k,\ell}(s,\beta_0)\left\{\frac{1}{n}S^{(0)}_j(s,\beta_0) - s^{(0)}_j(s,\beta_0)\right\}\alpha_{0,j}(s)\,ds\right| \\ \leq \int_0^\tau \left|v^j_{k,\ell}(s,\beta_0)\right|\alpha_{0,j}(s)\,ds \times \left\|\frac{1}{n}S^{(0)}_j(\cdot,\beta_0) - s^{(0)}_j(\cdot,\beta_0)\right\|_{\mathcal{T}}.$$

The first term on the right-hand side of this inequality is bounded under conditions C1-C5 and the second term converges to zero in probability by Glivenko-Cantelli theorem. Thus the fourth term in (3.5) converges to zero and the proof is complete.

□

In what follows, we denote by $\widehat{\mathbb{S}}$ the estimator of $\mathbb{S}$ obtained by replacing the $\Sigma^j_{k,\ell}$ by their estimators $(j = 1,\ldots,J; k,\ell = 1,\ldots,Q)$.
We now apply a parametric bootstrap procedure: i) we simulate M vectors $\mathbb{W}^*_1,\ldots\mathbb{W}^*_M$ from $\mathbb{W}^* \sim \mathcal{N}_{(J\cdot Q)}(0,\widehat{\mathbb{S}})$, ii) we calculate the corresponding values $s^*_i = \max_{1\leq j\leq J}\max_{q\in\mathcal{Q}}|W^{*j}_{i,q}|$, $i = 1,\ldots,M$, iii) we estimate $q_{1-\alpha}(\mathcal{S})$ by the empirical quantile $q^*_{1-\alpha}$ of order $1-\alpha$ of $s^*_1,\ldots,s^*_M$. Our decision rule is finally: "reject H_0 at the asymptotic level $\alpha \in (0,1)$ if $\mathcal{S}_n \geq q^*_{1-\alpha}$".

In the next section, we investigate this decision rule via simulations. Several issues are discussed, such as the choice of Q and $\mathbf{z}_1,\ldots,\mathbf{z}_Q$.

3.3 Simulation study

In this simulation study, we investigate the finite sample behaviour of our test statistic for various numbers of strata typically encountered in practice ($J = 3, 5$). Precisely, we assess level and power of our test against various alternatives, for several sample sizes and censoring fractions. Simulations are run using the statistical language R (see [51]) under a Linux Ubuntu Server 14.04 LTS 64Bits with two pro-

cessors Intel E5 2640v2 running at 1600MHZ, with 256Go RAM.

First, we simulate $N = 10^5$ samples of size n of observations from model (3.12) with $J = 3$. The sample sizes in the 3 strata are denoted by (n_1, n_2, n_3), with $n = \sum_{j=1}^{3} n_j$. We consider several values for (n_1, n_2, n_3), namely $(100, 110, 80)$, $(150, 175, 120)$ and $(200, 225, 190)$. The baseline hazard function in stratum j ($j = 1, 2, 3$) is $\alpha_{0,j}(t) = \lambda_j \alpha_j t^{\alpha_j - 1}$ with $(\alpha_1, \lambda_1) = (2.1, 1)$, $(\alpha_2, \lambda_2) = (1.2, 0.75)$, $(\alpha_3, \lambda_3) = (1.8, 1.5)$. We consider a two-dimensional covariate $\mathbf{Z} = (Z_1, Z_2)^\top$, where Z_1 and Z_2 are independent and distributed as a Gaussian $\mathcal{N}(0, 1)$ and a uniform $\mathcal{U}(1, 3)$ respectively. We take $\beta_0 = (0.2, 0.7)^\top$. The model used for simulating data is thus $\alpha_j(t) = \lambda_j \alpha_j t^{\alpha_j - 1} e^{0.2Z_1 + 0.7Z_2}$, $j = 1, 2, 3$. Censoring times are simulated from an exponential distribution with parameter $\mu > 0$, where μ is chosen to yield some pre-specified proportion c of censored observations (we consider $c = 0.1, 0.2, 0.4$).

As mentioned in Section 3.2.2, the function $\mathbf{z} \mapsto Q^j_{n,\mathbf{z}}(t, \widehat{\beta}_n)$ has possible jumps at the distinct values of the $\mathbf{Z}_i$, $i = 1, \ldots, n$. Therefore it is sufficient to consider a finite number Q of values for the vector $\mathbf{z}$. One may consider all distinct $\mathbf{Z}_i$ but our numerical experiments showed that no major change affects the outcome of the test when Q is smaller than n, provided that Q stays large enough. Thus in this simulation study, we take $Q = 250$ when $(n_1, n_2, n_3) = (100, 110, 80)$, $Q = 400$ when $(n_1, n_2, n_3) = (150, 175, 120)$, $Q = 600$ when $(n_1, n_2, n_3) = (200, 225, 190)$ and we use a regular grid of values $\mathbf{z}_q$ over the range of the $\mathbf{Z}_i$.

Under the setting described above, the null hypothesis H_0 holds and we investigate level of our test. For each of the N simulated samples, we calculate the test statistic $\mathcal{S}_n$ and approximate the critical value (for an asymptotic level $\alpha = 0.05$) by using the Monte Carlo procedure described in Section 3.2.3 (with $M = 5000$). Finally, we apply the proposed decision rule. Table 1 provides empirical level of the test for the various configurations of the simulation design parameters (see the row labeled "H_0").

Next, we investigate power of our test. For $J = 3$, we consider the following alternatives (respectively denoted by $H_{1,a}, H_{1,b}, H_{1,c}$ and $H_{1,d}$):

— *non-proportional hazards model*: the hazard function in stratum j is chosen as $\alpha_j(t) = \alpha_j e^{1.2Z_1 + 1.5Z_2 \times t}$, $j = 1, 2, 3$. We take $\alpha_1 = 0.01, \alpha_2 = 0.1, \alpha_3 = 0.25$. Hazard ratios are not constant in time and thus, the model is not a proportional hazards model.

— *covariate threshold effect*: the hazard function in stratum j is chosen as $\alpha_j(t) = \lambda_j \alpha_j t^{\alpha_j - 1} e^{1.7Z_1 1_{\{Z_1 > \xi_j\}} + 0.5Z_2}$, $j = 1, 2, 3$. We take $(\alpha_1, \lambda_1, \xi_1) = (1.5, 1, 0.6)$, $(\alpha_2, \lambda_2, \xi_2) = (0.5, 0.75, 1)$ and $(\alpha_3, \lambda_3, \xi_3) = (1, 1.25, 0.8)$.

— *distinct regression parameters across strata*: the hazard function in stratum j is chosen as $\alpha_j(t) = \lambda_j \alpha_j t^{\alpha_j - 1} e^{\beta_{1,j} Z_1 + \beta_{2,j} Z_2}$, $j = 1, 2, 3$. We take $(\alpha_1, \lambda_1, \beta_{1,1}, \beta_{2,1}) = (2.1, 1, 0.2, 0.7)$, $(\alpha_2, \lambda_2, \beta_{1,2}, \beta_{2,2}) = (1.2, 0.75, 1, 1)$ and $(\alpha_3, \lambda_3, \beta_{1,3}, \beta_{2,3}) = (1.8, 1.5, 0.2, 0.2)$.

— *omitted covariates*: the hazard function in stratum j is taken as $\alpha_j(t) = \lambda_j \alpha_j t^{\alpha_j - 1} e^{Z_1 - 0.7Z_2 + 0.75}$ $(j = 1, 2, 3)$ where the additional covariate Z_3 is distributed as a $\mathcal{N}(1, 0.25)$. We take $(\alpha_1, \lambda_1) = (2.1, 1)$, $(\alpha_2, \lambda_2) = (1.2, 0.75)$ and $(\alpha_3, \lambda_3) = (1.8, 1.5)$.

We simulate $N = 10^5$ samples of observations under each of these alternatives. The simulation design (within-stratum sample sizes, censoring proportion) is the same as for evaluating level. For each sample and each alternative hypothesis, we calculate $\mathcal{S}_n$ (based on the fitted model $\alpha_{0,j}(t)\exp(\beta_1 Z_1 + \beta_2 Z_2)$, $j = 1, 2, 3$), we approximate the critical value (for an asymptotic level $\alpha = 0.05$) by using the Monte Carlo procedure of Section 3.2.3 (with $M = 5000$) and finally, we apply the proposed decision rule. Empirical powers of the test under each alternative are reported in Table 3.1 (see rows labeled $H_{1,a}, H_{1,b}, H_{1,c}, H_{1,d}$).

	$(100, 110, 80)$			$(150, 175, 120)$			$(200, 225, 190)$		
c	0.1	0.2	0.4	0.1	0.2	0.4	0.1	0.2	0.4
H_0	0.0569	0.0679	0.0629	0.0436	0.0612	0.0629	0.0686	0.0505	0.0650
$H_{1,a}$	0.6762	0.5408	0.4187	0.8177	0.6982	0.4683	0.9705	0.8443	0.5790
$H_{1,b}$	0.6287	0.5320	0.4635	0.7577	0.7370	0.7216	0.9149	0.8600	0.8561
$H_{1,c}$	0.7469	0.7239	0.6648	0.9178	0.9130	0.8563	0.9818	0.9779	0.9659
$H_{1,d}$	0.6763	0.6483	0.4889	0.8423	0.8217	0.7897	0.9357	0.9191	0.8931

Tableau 3.1: Empirical size and power of the proposed test for various censoring proportions and sample sizes, with $J = 3$. All results are based on 10^5 simulated samples.

Then, we conduct a similar study as above, with $J = 5$. For investigating level, we simulate data from the stratified model $\alpha_{0,j}(t) = \lambda_j \alpha_j t^{\alpha_j - 1} e^{0.2Z_1 + 0.7Z_2}$ $(j = 1, \ldots, 5)$ with the same (α_i, λ_i), $i = 1, 2, 3$ as above and $(\alpha_4, \lambda_4) = (1, 1)$, $(\alpha_5, \lambda_5) = (1.2, 0.5)$. We consider following sample sizes: $(100, 110, 80, 110, 70)$, $(150, 175, 120, 80, 110)$ and $(200, 225, 190, 150, 120)$. A similar procedure as for $J = 3$ yields results in Table 2 (see the row labeled "H_0"). The power of the test is investigated under the same alternatives as above, with two additional strata. For conciseness, we postpone description of parameters values for strata $j = 4, 5$ to an appendix. Empirical powers of the test under each alternative are reported in Table 3.2 (see rows labeled $H_{1,a}, H_{1,b}, H_{1,c}, H_{1,d}$).

From these results, the proposed test statistic performs well under a variety of conditions. The empirical level is close to the nominal level even when censoring is large (40%, say). As expected, power of the test increases when within-strata sample

	$(100, 110, 80, 110, 70)$			$(150, 175, 120, 80, 110)$			$(200, 225, 190, 150, 120)$		
c	0.1	0.2	0.4	0.1	0.2	0.4	0.1	0.2	0.4
H_0	0.0638	0.0612	0.0428	0.0401	0.0584	0.0502	0.0524	0.0624	0.0609
$H_{1,a}$	0.7365	0.5967	0.2355	0.9341	0.8517	0.5581	0.9782	0.9263	0.6105
$H_{1,b}$	0.4502	0.4335	0.3837	0.8079	0.7283	0.6823	0.8876	0.8794	0.8627
$H_{1,c}$	0.8167	0.6616	0.5869	0.9195	0.9003	0.8830	0.9785	0.9718	0.9282
$H_{1,d}$	0.5898	0.5233	0.5211	0.8018	0.7937	0.7955	0.9657	0.9541	0.9210

Tableau 3.2: Empirical size and power of the proposed test for various censoring proportions and sample sizes, with $J = 5$. All results are based on 10^5 simulated samples.

sizes increase and decreases when censoring increases. However, when censoring is low to moderate (20%, say), the test maintains satisfactory power against every alternative provided that within-strata sample sizes are all sufficiently large (greater than 100, say). When censoring is heavy (40%, say), the test still maintains good power if all within-strata sample sizes are sufficiently large, say 200. Overall, our test appears to provide an efficient tool for assessing adequacy of the stratified proportional hazards model under usual conditions of sample size and censoring.

Appendix A. Proofs of Lemma 3.1 and Lemma 3.2.

Proof of Lemma 3.1 We use the martingale central limit theorem (*e.g.*, Theorem 5.3.5 in [18]) to establish weak convergence of the process $(\mathbb{Q}_n^1(\cdot, \beta_0)^\top, \ldots, \mathbb{Q}_n^J(\cdot, \beta_0)^\top)$. First, we provide an alternative expression for $Q_{n,\mathbf{z}}^j(t, \beta_0)$. For $j = 1, \ldots, J$, let $\bar{N}_j(s) = \sum_{i=1}^n 1_{\{S_i=j\}} N_i(s)$. From (3.3), we have:

$$\begin{aligned}
Q_{n,\mathbf{z}}^j(t, \beta_0) &:= \frac{1}{\sqrt{n}} \sum_{i=1}^n \widehat{M}_i(t, \beta_0) 1_{\{S_i=j\}} 1_{\{\mathbf{Z}_i \le \mathbf{z}\}}, \\
&= \sum_{i=1}^n \int_0^t \frac{1}{\sqrt{n}} 1_{\{S_i=j\}} 1_{\{\mathbf{Z}_i \le \mathbf{z}\}} \left\{ dN_i(s) - \sum_{\ell=1}^J \frac{Y_i(s) e^{\beta_0^\top \mathbf{Z}_i} 1_{\{S_i=\ell\}}}{S_\ell^{(0)}(s, \beta_0)} d\bar{N}_\ell(s) \right\} \\
&= \sum_{i=1}^n \int_0^t \frac{1}{\sqrt{n}} 1_{\{\mathbf{Z}_i \le \mathbf{z}\}} \left\{ 1_{\{S_i=j\}} dN_i(s) - \frac{Y_i(s) e^{\beta_0^\top \mathbf{Z}_i} 1_{\{S_i=j\}}}{S_j^{(0)}(s, \beta_0)} d\bar{N}_j(s) \right\} \\
&= \sum_{i=1}^n \int_0^t \frac{1}{\sqrt{n}} \left\{ 1_{\{\mathbf{Z}_i \le \mathbf{z}\}} - \frac{S_{j,\mathbf{z}}^{(0)}(s, \beta_0)}{S_j^{(0)}(s, \beta_0)} \right\} 1_{\{S_i=j\}} \, dN_i(s).
\end{aligned}$$

Now, recall that $N_i(t) = \int_0^t Y_i(s) \sum_{j=1}^J \alpha_{0,j}(s) e^{\beta_0^\top \mathbf{Z}_i} 1_{\{S_i=j\}}\, ds + M_i(t)$ and thus:

$$Q^j_{n,\mathbf{z}}(t,\beta_0) = \sum_{i=1}^n \int_0^t \frac{1}{\sqrt{n}} \left\{ 1_{\{\mathbf{Z}_i \le \mathbf{z}\}} - \frac{S^{(0)}_{j,\mathbf{z}}(s,\beta_0)}{S^{(0)}_j(s,\beta_0)} \right\} 1_{\{S_i=j\}} Y_i(s) \alpha_{0,j}(s) e^{\beta_0^\top \mathbf{Z}_i} ds$$
$$+ \sum_{i=1}^n \int_0^t \frac{1}{\sqrt{n}} \left\{ 1_{\{\mathbf{Z}_i \le \mathbf{z}\}} - \frac{S^{(0)}_{j,\mathbf{z}}(s,\beta_0)}{S^{(0)}_j(s,\beta_0)} \right\} 1_{\{S_i=j\}}\, dM_i(s).$$

The first term in the right-hand side of this equality is 0 (straightforward calculations are omitted), which yields:

$$Q^j_{n,\mathbf{z}}(t,\beta_0) := \sum_{i=1}^n \int_0^t \frac{1}{\sqrt{n}} \left\{ 1_{\{\mathbf{Z}_i \le \mathbf{z}\}} - \frac{S^{(0)}_{j,\mathbf{z}}(s,\beta_0)}{S^{(0)}_j(s,\beta_0)} \right\} 1_{\{S_i=j\}}\, dM_i(s). \tag{3.6}$$

The process $Q^j_{n,\mathbf{z}} := (Q^j_{n,\mathbf{z}}(\cdot,\beta_0))$ is then a martingale with respect to the filtration $\bigvee_{i\le n} \mathcal{F}_{t,i}$. Let $k,\ell \in \{1,\dots,Q\}$ and $t \in \mathcal{T}$. Using expression (3.6), we have:

$$\langle Q^j_{n,\mathbf{z}_k}, Q^j_{n,\mathbf{z}_\ell} \rangle(t) = \sum_{i=1}^n \int_0^t \frac{1}{n} \left\{ 1_{\{\mathbf{Z}_i \le \mathbf{z}_k\}} - \frac{S^{(0)}_{j,\mathbf{z}_k}(s,\beta_0)}{S^{(0)}_j(s,\beta_0)} \right\}$$
$$\times \left\{ 1_{\{\mathbf{Z}_i \le \mathbf{z}_\ell\}} - \frac{S^{(0)}_{j,\mathbf{z}_\ell}(s,\beta_0)}{S^{(0)}_j(s,\beta_0)} \right\} 1_{\{S_i=j\}} \alpha_{0,j}(s) Y_i(s) e^{\beta_0^\top \mathbf{Z}_i}\, ds.$$

As n tends to infinity, $\langle Q^j_{n,\mathbf{z}_k}, Q^j_{n,\mathbf{z}_\ell} \rangle(t)$ converges in probability to

$$\begin{aligned}
C^j_{\mathbf{z}_k,\mathbf{z}_\ell}(t) \;:=\; & \mathbb{E}\left[\int_0^t 1_{\{\mathbf{Z} \le \mathbf{z}_k\}} 1_{\{\mathbf{Z} \le \mathbf{z}_\ell\}} 1_{\{S=j\}} \alpha_{0,j}(s) Y(s) e^{\beta_0^\top \mathbf{Z}}\, ds \right] \\
& - \mathbb{E}\left[\int_0^t 1_{\{\mathbf{Z} \le \mathbf{z}_k\}} \frac{s^{(0)}_{j,\mathbf{z}_\ell}(s,\beta_0)}{s^{(0)}_j(s,\beta_0)} 1_{\{S=j\}} \alpha_{0,j}(s) Y(s) e^{\beta_0^\top \mathbf{Z}}\, ds \right] \\
& - \mathbb{E}\left[\int_0^t 1_{\{\mathbf{Z} \le \mathbf{z}_\ell\}} \frac{s^{(0)}_{j,\mathbf{z}_k}(s,\beta_0)}{s^{(0)}_j(s,\beta_0)} 1_{\{S=j\}} \alpha_{0,j}(s) Y(s) e^{\beta_0^\top \mathbf{Z}}\, ds \right] \\
& + \mathbb{E}\left[\int_0^t \frac{s^{(0)}_{j,\mathbf{z}_k}(s,\beta_0) s^{(0)}_{j,\mathbf{z}_\ell}(s,\beta_0)}{(s^{(0)}_j(s,\beta_0))^2} 1_{\{S=j\}} \alpha_{0,j}(s) Y(s) e^{\beta_0^\top \mathbf{Z}}\, ds \right] \\
= \; & \mathbb{E}\left[\int_0^t 1_{\{\mathbf{Z} \le \mathbf{z}_k\}} 1_{\{\mathbf{Z} \le \mathbf{z}_\ell\}} 1_{\{S=j\}} \alpha_{0,j}(s) Y(s) e^{\beta_0^\top \mathbf{Z}}\, ds \right] \\
& - \int_0^t \frac{s^{(0)}_{j,\mathbf{z}_k}(s,\beta_0) s^{(0)}_{j,\mathbf{z}_\ell}(s,\beta_0)}{s^{(0)}_j(s,\beta_0)} \alpha_{0,j}(s)\, ds.
\end{aligned}$$

Moreover, when $j \neq j'$, we have:

$$\begin{aligned}\langle Q^j_{n,\mathbf{z}_k}, Q^{j'}_{n,\mathbf{z}_\ell}\rangle(t) &= \sum_{i=1}^n \int_0^t \frac{1}{n}\left\{1_{\{\mathbf{Z}_i\leq \mathbf{z}_k\}} - \frac{S^{(0)}_{j,\mathbf{z}_k}(s,\beta_0)}{S^{(0)}_j(s,\beta_0)}\right\}\left\{1_{\{\mathbf{Z}_i\leq \mathbf{z}_\ell\}} - \frac{S^{(0)}_{j',\mathbf{z}_\ell}(s,\beta_0)}{S^{(0)}_{j'}(s,\beta_0)}\right\} \\ &\qquad \times 1_{\{S_i=j\}}1_{\{S_i=j'\}}Y_i(s)\sum_{m=1}^J \alpha_{0,m}(s)e^{\beta_0^\top \mathbf{Z}_i}1_{\{S_i=m\}}\,ds \\ &= 0\end{aligned}$$

since $1_{\{S_i=j\}}1_{\{S_i=j'\}} = 0$. Next, we verify Lindeberg condition (*e.g.*, condition (3.18) of Theorem 5.3.5 in [18]). Let $\varepsilon > 0$, $E^{(0)}_{j,\mathbf{z}_k}(s,\beta_0) := S^{(0)}_{j,\mathbf{z}_k}(s,\beta_0)/S^{(0)}_j(s,\beta_0)$ and define the jump process $Q^j_{n,\mathbf{z}_k,\varepsilon}$ by $Q^j_{n,\mathbf{z}_k,\varepsilon}(t,\beta_0) =$

$$\sum_{i=1}^n \int_0^t \frac{1}{\sqrt{n}}\left\{1_{\{\mathbf{Z}_i\leq \mathbf{z}_k\}} - E^{(0)}_{j,\mathbf{z}_k}(s,\beta_0)\right\}1_{\left\{\left|\frac{1}{\sqrt{n}}\left(1_{\{\mathbf{Z}_i\leq \mathbf{z}_k\}} - E^{(0)}_{j,\mathbf{z}_k}(s,\beta_0)\right)\right|>\varepsilon\right\}}1_{\{S_i=j\}}\,dM_i(s).$$

We have:

$$\begin{aligned}\langle Q^j_{n,\mathbf{z}_k,\varepsilon}, Q^j_{n,\mathbf{z}_k,\varepsilon}\rangle(t) = \sum_{i=1}^n \int_0^t &\left\{\frac{1}{\sqrt{n}}\left(1_{\{\mathbf{Z}_i\leq \mathbf{z}_k\}} - E^{(0)}_{j,\mathbf{z}_k}(s,\beta_0)\right)\right\}^2 1_{\{S_i=j\}} \\ &\times \alpha_{0,j}(s)Y_i(s)e^{\beta_0^\top \mathbf{Z}_i}1_{\left\{\left|\frac{1}{\sqrt{n}}\left(1_{\{\mathbf{Z}_i\leq \mathbf{z}_k\}} - E^{(0)}_{j,\mathbf{z}_k}(s,\beta_0)\right)\right|>\varepsilon\right\}}\,ds.\end{aligned}$$

By using the inequality $|u-v|^2 1_{\{|u-v|>\varepsilon\}} \leq 4u^2 1_{\{|u|>\varepsilon/2\}} + 4v^2 1_{\{|v|>\varepsilon/2\}}$, we obtain that $\langle Q^j_{n,\mathbf{z}_k,\varepsilon}, Q^j_{n,\mathbf{z}_k,\varepsilon}\rangle(t)$ is bounded above by

$$\begin{aligned}&\frac{4}{n}\sum_{i=1}^n \int_0^t 1_{\{\mathbf{Z}_i\leq \mathbf{z}_k\}}1_{\{1_{\{\mathbf{Z}_i\leq \mathbf{z}_k\}}>\varepsilon\sqrt{n}/2\}}Y_i(s)e^{\beta_0^\top \mathbf{Z}_i}1_{\{S_i=j\}}\alpha_{0,j}(s)\,ds \\ &\quad + \frac{4}{n}\int_0^t \left(E^{(0)}_{j,\mathbf{z}_k}(s,\beta_0)\right)^2 1_{\{|E^{(0)}_{j,\mathbf{z}_k}(s,\beta_0)|>\varepsilon\sqrt{n}/2\}}S^{(0)}_j(s,\beta_0)\alpha_{0,j}(s)\,ds.\end{aligned}$$

When n is sufficiently large, $1_{\{1_{\{\mathbf{Z}_i\leq \mathbf{z}_k\}}>\varepsilon\sqrt{n}/2\}} = 0$ for every $i = 1,\ldots,n$ and the first term is zero. The second term converges to zero in probability, as the event $|E^{(0)}_{j,\mathbf{z}_k}(s,\beta_0)| > \varepsilon\sqrt{n}/2$ cannot be true for large n under conditions C1-C5. By Theorem 5.3.5 in [18], the process $(\mathbb{Q}^1_n(\cdot,\beta_0)^\top,\ldots,\mathbb{Q}^J_n(\cdot,\beta_0)^\top)$ converges weakly in $(D(\mathcal{T}))^{J\cdot Q}$ to a zero-mean Gaussian process $(\mathbb{W}^{1\top},\ldots,\mathbb{W}^{J\top})$ (with $\mathbb{W}^j := (W^j_1,\ldots,W^j_Q)^\top$) with covariance function $\mathrm{cov}(W^j_k(t_1), W^j_\ell(t_2)) = C^j_{\mathbf{z}_k,\mathbf{z}_\ell}(\min(t_1,t_2))$ and $\mathrm{cov}(W^j_k(t_1), W^{j'}_\ell(t_2)) = 0$ if $j \neq j'$, for $k,\ell = 1,\ldots,Q$.

□

Proof of Lemma 3.2 By a first-order Taylor expansion of $\beta \mapsto Q^j_{n,\mathbf{z}}(t,\beta)$ around β_0,

we have:

$$Q^j_{n,\mathbf{z}}(t,\widehat{\beta}_n) - Q^j_{n,\mathbf{z}}(t,\beta_0) = (\widehat{\beta}_n - \beta_0)^\top \frac{\partial}{\partial\beta} Q^j_{n,\mathbf{z}}(t,\widetilde{\beta}_n),$$

where $\widetilde{\beta}_n$ is on the line segment between $\widetilde{\beta}_n$ and β_0. The derivative $\partial Q^j_{n,\mathbf{z}}(t,\beta)/\partial\beta$ is given by:

$$\frac{\partial}{\partial\beta} Q^j_{n,\mathbf{z}}(t,\beta) = -\frac{1}{\sqrt{n}} \int_0^t H_{j,\mathbf{z}}(s,\beta)\, d\bar{M}_j(s),$$

where $\bar{M}_j(s) = \sum_{i=1}^n 1_{\{S_i=j\}} M_i(s)$ and

$$H_{j,\mathbf{z}}(s,\beta) := \frac{S^{(1)}_{j,\mathbf{z}}(s,\beta)}{S^{(0)}_j(s,\beta)} - \frac{S^{(0)}_{j,\mathbf{z}}(s,\beta) S^{(1)}_j(s,\beta)}{(S^{(0)}_j(s,\beta))^2}.$$

Then, by Cauchy-Schwarz inequality,

$$\left\| Q^j_{n,\mathbf{z}}(\cdot,\widehat{\beta}_n) - Q^j_{n,\mathbf{z}}(\cdot,\beta_0) \right\|_{\mathcal{T}} \leq \|\widehat{\beta}_n - \beta_0\| \left\| \frac{\partial}{\partial\beta} Q^j_{n,\mathbf{z}}(\cdot,\widetilde{\beta}_n) \right\|_{\mathcal{T}}.$$

Under H_0, $\widehat{\beta}_n$ converges in probability to β_0 as n tends to infinity and thus, $\|\widehat{\beta}_n - \beta_0\|$ converges to zero. We prove that $\left\| \frac{\partial}{\partial\beta} Q^j_{n,\mathbf{z}}(\cdot,\widetilde{\beta}_n) \right\|_{\mathcal{T}}$ is bounded in probability. Let

$$h_{j,\mathbf{z}}(s,\beta) := \frac{s^{(1)}_{j,\mathbf{z}}(s,\beta)}{s^{(0)}_j(s,\beta)} - \frac{s^{(0)}_{j,\mathbf{z}}(s,\beta) s^{(1)}_j(s,\beta)}{(s^{(0)}_j(s,\beta))^2}.$$

We prove that $\|H_{j,\mathbf{z}}(\cdot,\widetilde{\beta}_n) - h_{j,\mathbf{z}}(\cdot,\beta_0)\|_{\mathcal{T}}$ converges in probability to zero as n tends to infinity. We have:

$$\begin{aligned}
\|H_{j,\mathbf{z}}(\cdot,\widetilde{\beta}_n) - h_{j,\mathbf{z}}(\cdot,\beta_0)\|_{\mathcal{T}} &\leq \left\| \frac{n^{-1} S^{(1)}_{j,\mathbf{z}}(\cdot,\widetilde{\beta}_n)}{n^{-1} S^{(0)}_j(\cdot,\widetilde{\beta}_n)} - \frac{s^{(1)}_{j,\mathbf{z}}(\cdot,\beta_0)}{s^{(0)}_j(\cdot,\beta_0)} \right\|_{\mathcal{T}} \\
&\quad + \left\| \frac{s^{(0)}_{j,\mathbf{z}}(\cdot,\beta_0) s^{(1)}_j(\cdot,\beta_0)}{(s^{(0)}_j(\cdot,\beta_0))^2} - \frac{S^{(0)}_{j,\mathbf{z}}(\cdot,\widetilde{\beta}_n) S^{(1)}_j(\cdot,\widetilde{\beta}_n)}{(S^{(0)}_j(\cdot,\widetilde{\beta}_n))^2} \right\|_{\mathcal{T}} \\
&:= U_{n,1} + U_{n,2}.
\end{aligned}$$

We prove that $U_{n,1}$ converges in probability to zero as n tends to infinity (arguments are similar for $U_{n,2}$ and are thus omitted). Using the elementary equality $\frac{u_n}{v_n} - \frac{u}{v} =$

$(\frac{u_n}{v} - \frac{uv_n}{v^2})\frac{v}{v_n}$, we obtain:

$$\begin{aligned}
U_{n,1} &= \left\| \left(\frac{n^{-1}S^{(1)}_{j,\mathbf{z}}(\cdot,\widetilde{\beta}_n)}{s^{(0)}_j(\cdot,\beta_0)} - \frac{s^{(1)}_{j,\mathbf{z}}(\cdot,\beta_0)n^{-1}S^{(0)}_j(\cdot,\widetilde{\beta}_n)}{(s^{(0)}_j(\cdot,\beta_0))^2} \right) \frac{s^{(0)}_j(\cdot,\beta_0)}{n^{-1}S^{(0)}_j(\cdot,\widetilde{\beta}_n)} \right\|_{\mathcal{T}} \\
&\leq \left\| \frac{n^{-1}S^{(1)}_{j,\mathbf{z}}(\cdot,\widetilde{\beta}_n)}{s^{(0)}_j(\cdot,\beta_0)} - \frac{s^{(1)}_{j,\mathbf{z}}(\cdot,\beta_0)n^{-1}S^{(0)}_j(\cdot,\widetilde{\beta}_n)}{(s^{(0)}_j(\cdot,\beta_0))^2} \right\|_{\mathcal{T}} \left\| \frac{s^{(0)}_j(\cdot,\beta_0)}{n^{-1}S^{(0)}_j(\cdot,\widetilde{\beta}_n)} \right\|_{\mathcal{T}} \\
&\leq \frac{1}{c^2} \left\| n^{-1}S^{(1)}_{j,\mathbf{z}}(\cdot,\widetilde{\beta}_n)s^{(0)}_j(\cdot,\beta_0) - s^{(1)}_{j,\mathbf{z}}(\cdot,\beta_0)n^{-1}S^{(0)}_j(\cdot,\widetilde{\beta}_n) \right\|_{\mathcal{T}} \left\| \frac{s^{(0)}_j(\cdot,\beta_0)}{n^{-1}S^{(0)}_j(\cdot,\widetilde{\beta}_n)} \right\|_{\mathcal{T}}
\end{aligned}$$

where the second to third line follows by condition C5. Now,

$$\begin{aligned}
&\left\| n^{-1}S^{(1)}_{j,\mathbf{z}}(\cdot,\widetilde{\beta}_n)s^{(0)}_j(\cdot,\beta_0) - s^{(1)}_{j,\mathbf{z}}(\cdot,\beta_0)n^{-1}S^{(0)}_j(\cdot,\widetilde{\beta}_n) \right\|_{\mathcal{T}} \\
&\qquad \leq \left\| n^{-1}S^{(1)}_{j,\mathbf{z}}(\cdot,\widetilde{\beta}_n)s^{(0)}_j(\cdot,\beta_0) - s^{(1)}_{j,\mathbf{z}}(\cdot,\widetilde{\beta}_n)s^{(0)}_j(\cdot,\beta_0) \right\|_{\mathcal{T}} \\
&\qquad + \left\| s^{(1)}_{j,\mathbf{z}}(\cdot,\widetilde{\beta}_n)s^{(0)}_j(\cdot,\beta_0) - s^{(1)}_{j,\mathbf{z}}(\cdot,\beta_0)s^{(0)}_j(\cdot,\beta_0) \right\|_{\mathcal{T}} \\
&\qquad + \left\| s^{(1)}_{j,\mathbf{z}}(\cdot,\beta_0)s^{(0)}_j(\cdot,\beta_0) - s^{(1)}_{j,\mathbf{z}}(\cdot,\beta_0)s^{(0)}_j(\cdot,\widetilde{\beta}_n) \right\|_{\mathcal{T}} \\
&\qquad + \left\| s^{(1)}_{j,\mathbf{z}}(\cdot,\beta_0)s^{(0)}_j(\cdot,\widetilde{\beta}_n) - s^{(1)}_{j,\mathbf{z}}(\cdot,\beta_0)n^{-1}S^{(0)}_j(\cdot,\widetilde{\beta}_n) \right\|_{\mathcal{T}}
\end{aligned}$$

and thus

$$\begin{aligned}
&\left\| n^{-1}S^{(1)}_{j,\mathbf{z}}(\cdot,\widetilde{\beta}_n)s^{(0)}_j(\cdot,\beta_0) - s^{(1)}_{j,\mathbf{z}}(\cdot,\beta_0)n^{-1}S^{(0)}_j(\cdot,\widetilde{\beta}_n) \right\|_{\mathcal{T}} \\
&\quad \leq \left\| s^{(0)}_j(\cdot,\beta_0) \right\|_{\mathcal{T}} \left(\left\| n^{-1}S^{(1)}_{j,\mathbf{z}} - s^{(1)}_{j,\mathbf{z}} \right\|_{\mathcal{T}\times\mathcal{B}} + \left\| s^{(1)}_{j,\mathbf{z}}(\cdot,\widetilde{\beta}_n) - s^{(1)}_{j,\mathbf{z}}(\cdot,\beta_0) \right\|_{\mathcal{T}} \right) \\
&\quad + \left\| s^{(1)}_{j,\mathbf{z}}(\cdot,\beta_0) \right\|_{\mathcal{T}} \left(\left\| s^{(0)}_j(\cdot,\beta_0) - s^{(0)}_j(\cdot,\widetilde{\beta}_n) \right\|_{\mathcal{T}} + \left\| s^{(0)}_j - n^{-1}S^{(0)}_j \right\|_{\mathcal{T}\times\mathcal{B}} \right).
\end{aligned}$$

From Glivenko-Cantelli theorem, $\|n^{-1}S^{(1)}_{j,\mathbf{z}} - s^{(1)}_{j,\mathbf{z}}\|_{\mathcal{T}\times\mathcal{B}}$ and $\|n^{-1}S^{(0)}_j - s^{(0)}_j\|_{\mathcal{T}\times\mathcal{B}}$ converge to zero. By condition C4, $\|s^{(1)}_{j,\mathbf{z}}(\cdot,\widetilde{\beta}_n) - s^{(1)}_{j,\mathbf{z}}(\cdot,\beta_0)\|_{\mathcal{T}}$ and $\|s^{(0)}_j(\cdot,\beta_0) - s^{(0)}_j(\cdot,\widetilde{\beta}_n)\|_{\mathcal{T}}$ converge to zero. Moreover, $\|s^{(0)}_j(\cdot,\beta_0)\|_{\mathcal{T}}$ and $\|s^{(1)}_{j,\mathbf{z}}(\cdot,\beta_0)\|_{\mathcal{T}}$ are bounded by conditions C1 and C3. Thus, the quantity $\|n^{-1}S^{(1)}_{j,\mathbf{z}}(\cdot,\widetilde{\beta}_n)s^{(0)}_j(\cdot,\beta_0) - s^{(1)}_{j,\mathbf{z}}(\cdot,\beta_0)n^{-1}S^{(0)}_j(\cdot,\widetilde{\beta}_n)\|_{\mathcal{T}}$ tends to zero. Using similar arguments, one can show that $\|s^{(0)}_j(\cdot,\beta_0)/n^{-1}S^{(0)}_j(\cdot,\widetilde{\beta}_n)\|_{\mathcal{T}}$ converges to 1 and finally, $U_{n,1}$ converges to zero. Convergence to zero of $U_{n,1}$ and $U_{n,2}$ implies that $\|H_{j,\mathbf{z}}(\cdot,\widetilde{\beta}_n) - h_{j,\mathbf{z}}(\cdot,\beta_0)\|_{\mathcal{T}}$ converges in probability to zero. Now, we have:

$$\|H_{j,\mathbf{z}}(\cdot,\widetilde{\beta}_n)\|_{\mathcal{T}} \leq \|H_{j,\mathbf{z}}(\cdot,\widetilde{\beta}_n) - h_{j,\mathbf{z}}(\cdot,\beta_0)\|_{\mathcal{T}} + \|h_{j,\mathbf{z}}(\cdot,\beta_0)\|_{\mathcal{T}}.$$

Under conditions C1, C3, C5, $\|h_{j,\mathbf{z}}(\cdot,\beta_0)\|_{\mathcal{T}}$ is bounded and thus $\|H_{j,\mathbf{z}}(\cdot,\widetilde{\beta}_n)\|_{\mathcal{T}}$ and then $\|\frac{\partial}{\partial\beta}Q^j_{n,\mathbf{z}}(\cdot,\widetilde{\beta}_n)\|_{\mathcal{T}}$ are bounded in probability. Finally, $Q^j_{n,\mathbf{z}}(\cdot,\widehat{\beta}_n) = Q^j_{n,\mathbf{z}}(\cdot,\beta_0) + o_{\mathbb{P}}(1)$ where the $o_{\mathbb{P}}(1)$ is uniform on $\mathcal{T}$. This concludes the proof.

□

Appendix B. Parameters values in the simulation study for $J = 5$.

For $J = 5$, we consider the following alternatives (respectively denoted by $H_{1,a}, H_{1,b}, H_{1,c}$ and $H_{1,d}$ in Table 3.2):

— *non-proportional hazards model*: the hazard function in stratum j is chosen as $\alpha_j(t) = \alpha_j e^{1.2Z_1+1.5Z_2\times t}$, $j = 1, \ldots, 5$. We take $\alpha_1 = 0.01, \alpha_2 = 0.1, \alpha_3 = 0.25, \alpha_4 = 0.3, \alpha_5 = 0.2$.

— *covariate threshold effect*: the hazard function in stratum j is chosen as $\alpha_j(t) = \lambda_j\alpha_j t^{\alpha_j-1}e^{1.7Z_1 1_{\{Z_1>\xi_j\}}+0.5Z_2}$, $j = 1, \ldots, 5$. We take $(\alpha_1, \lambda_1, \xi_1) = (1.5, 1, 0.6)$, $(\alpha_2, \lambda_2, \xi_2) = (0.5, 0.75, 1)$, $(\alpha_3, \lambda_3, \xi_3) = (1, 1.25, 0.8)$, $(\alpha_4, \lambda_4,$
$\xi_4) = (0.75, 0.8, 1.2)$ and $(\alpha_5, \lambda_5, \xi_5) = (1, 0.8, 0.75)$.

— *distinct regression parameters across strata*: the hazard function in stratum j is chosen as $\alpha_j(t) = \lambda_j\alpha_j t^{\alpha_j-1}e^{\beta_{1,j}Z_1+\beta_{2,j}Z_2}$, $j = 1, \ldots, 5$. We take $(\alpha_1, \lambda_1, \beta_{1,1}, \beta_{2,1}) = (2.1, 1, 0.2, 0.7)$, $(\alpha_2, \lambda_2, \beta_{1,2}, \beta_{2,2}) = (1.2, 0.75, 1, 1)$, $(\alpha_3, \lambda_3, \beta_{1,3}, \beta_{2,3}) = (1.8, 1.5, 0.2, 0.2)$, $(\alpha_4, \lambda_4, \beta_{1,4}, \beta_{2,4}) = (1, 1, 1.3, 0.5)$ and $(\alpha_5, \lambda_5, \beta_{1,5}, \beta_{2,5}) = (1.5, 0.5, 0.15, 0.55)$.

— *omitted covariates*: the hazard function in stratum j is taken as $\alpha_j(t) = \lambda_j\alpha_j t^{\alpha_j-1}e^{Z_1-0.7Z_2+0.75}$ $(j = 1, \ldots, 5)$ where the additional covariate Z_3 is distributed as a $\mathcal{N}(1, 0.25)$. We take $(\alpha_1, \lambda_1) = (2.1, 1)$, $(\alpha_2, \lambda_2) = (1.2, 0.75)$, $(\alpha_3, \lambda_3) = (1.8, 1.5)$, $(\alpha_4, \lambda_4) = (1.2, 1.25)$ and $(\alpha_5, \lambda_5) = (0.5, 0.8)$.

3.4 Stratified proportional hazard model with unknown threshold

In the first part of this chapter, we proposed a goodness-of-fit test statistic for the stratified proportional hazard model and we established its asymptotic distribution under the null hypothesis that a stratified model holds. Our simulation study suggests that the proposed test performs well under a wide range of conditions (sample size, censoring fraction, alternative hypothesis).

Now, several issues deserve attention. For example, in the stratified proportional hazards model, the variable used for stratifying the population under study is often discrete (*e.g.*, gender, disease stage, socio-professional category...). Sometimes, it may

be relevant to stratify according to a threshold ω in a continuous variable W (such as tumour size, level of salary. . .). To the best of our knowledge, no procedure was proposed yet to estimate w^*. Our test statistic may be used for that purpose. For example, one may test goodness of fit of the model

$$\alpha_{0,1}(t)\exp(\beta^\top \mathbf{Z})1_{\{W\leq\omega\}} + \alpha_{0,2}(t)\exp(\beta^\top \mathbf{Z})1_{\{W>\omega\}} \tag{3.7}$$

for several values of ω (using our test statistic $\mathcal{S}_n$) and retain the value $\hat{\omega}$ yielding the less significant result. We are currently exploring this original application of our test statistic.

In what follows, we provide a method allowing to stratify the proportional hazards model according to an unknown threshold.

3.4.1 Model and notations

Let T be a random failure time whose distribution depends on a vector of covariates $\mathbf{Z} = (Z_1, \ldots, Z_p)^\top \in \mathbb{R}^p$. We assume that the hazard function for T takes the form of a stratified Cox model such as model (3.7). We consider the case of 2 strata (labeled $k = 0, 1$ for convenience). Let define the stratum indicator S equal to say, 0 if the individual belongs to stratum 0, and 1 if the individual belongs to stratum 1. We suppose that T may be right-censored at a non-informative censoring time C such that C is independent of T conditionally on $\mathbf{Z}$ and S. The observed time variables are therefore the possibly censored time $X = \min(T, \min(C, \tau))$ and a censoring indicator $\Delta = 1_{\{T\leq\min(C,\tau)\}}$. Let denote by τ the maximum observation time i.e. the analysis is restricted to the time interval $\mathcal{T} := [0, \tau]$, where $\tau < \infty$.

Available data consist of n independent and identically distributed copies $(X_i, \Delta_i, \mathbf{Z}_i, S_i)$ of the random vector $(X, \Delta, \mathbf{Z}, S)$. For every $i = 1, \ldots, n$ and $t \in \mathcal{T}$, we denote by $N_i(t) = 1_{\{X_i\leq t\}}\Delta_i$ and $Y_i(t) = 1_{\{X_i\geq t\}}$ the failure counting and at-risk processes of the i-th individual respectively. We assume that The process $N_i(t)$ has intensity :

$$Y_i(t)[\alpha_{0,1}(t)e^{\beta_0^\top \mathbf{Z}}1_{\{W\leq\omega^*\}} + \alpha_{0,2}(t)e^{\beta_0^\top \mathbf{Z}}1_{\{W>\omega^*\}}]. \tag{3.8}$$

with respect to the filtration $(\mathcal{F}_{t,i})_{t\geq 0}$ defined by $\mathcal{F}_{t,i} = \sigma\{\mathbf{Z}_i, S_i, N_i(s), Y_i(s)\colon 0 \leq s \leq t\}$ Then under model (3.8), the process defined by:

$$M_i(t) = \delta - \int_0^t 1_{\{X_i\leq X_j\}}[\alpha_{0,1}(t)e^{\beta_0^\top \mathbf{Z}}1_{\{W\leq\omega^*\}} + \alpha_{0,2}(t)e^{\beta_0^\top \mathbf{Z}}1_{\{W>\omega^*\}}]ds, \quad i = 1, ..., n$$

is a martingale.

If ω^* were known, we could estimate β_0 by using maximum partial likelihood estimation:

$$L_n(\beta,\omega^*) = \prod_{i:W_i\le\omega^*}\left(\frac{e^{\beta^\top \mathbf{Z}_i}}{\sum_{j:W_j\le w^*} e^{\beta^\top \mathbf{Z}_j} Y_j(X_i)}\right)^{\delta_i} \times \prod_{i:W_i>\omega^*}\left(\frac{e^{\beta^\top \mathbf{Z}_i}}{\sum_{j:W_j>w^*} e^{\beta^\top \mathbf{Z}_j} Y_j(X_i)}\right)^{\delta_i}.$$

The log-partial likelihood for β is

$$\begin{aligned}\sum_{i=1}^n \int_0^\tau \Big(\beta^\top Z_i \quad &- \quad 1_{\{W_j\le w^*\}} \ln \sum_{j=1}^n Y_j(s) 1_{\{W_j\le w^*\}} e^{\beta^\top Z_j} \\ &- \quad 1_{\{W_j> w^*\}} \ln \sum_{j=1}^n Y_j(s) 1_{\{W_j> w^*\}} e^{\beta^\top Z_j}\Big) dN_i(s)\end{aligned} \tag{3.9}$$

The maximum partial likelihood estimator of the regression parameter is defined as the value of β maximizing (3.9), or equivalently as the solution to the estimating equation

$$\begin{aligned}\sum_{i=1}^n \int_0^\tau \Big(Z_i \quad &- \quad 1_{\{W_j\le w^*\}} \frac{\sum_{j=1}^n Y_j(s) 1_{\{W_j\le w^*\}} Z_j e^{\beta^\top Z_j}}{\sum_{j=1}^n Y_j(s) 1_{\{W_j\le w^*\}} e^{\beta^\top Z_j}} \\ &- \quad 1_{\{W_j> w^*\}} \frac{\sum_{j=1}^n Y_j(s) 1_{\{W_j> w^*\}} Z_j e^{\beta^\top Z_j}}{\sum_{j=1}^n Y_j(s) 1_{\{W_j> w^*\}} e^{\beta^\top Z_j}}\Big) dN_i(s) = 0,\end{aligned}$$

while cumulative baseline hazard functions $A_{0,1,\omega}(t) = \int_0^t \alpha_{0,1} ds$ and $A_{0,2,\omega}(t) = \int_0^t \alpha_{0,2} ds$ can be estimated by Breslow-type estimators

$$\widehat{A}_{0,1,\omega}(t) = \sum_{i:W_i\le\omega^*} \frac{\delta_i 1_{\{X_i\le t\}}}{\sum_{k:W_k\le\omega^*} e^{\widehat{\beta}_n^\top \mathbf{Z}_k} Y_k(X_i)},$$

$$\widehat{A}_{0,2,\omega}(t) = \sum_{i:W_i>\omega^*} \frac{\delta_i 1_{\{X_i\le t\}}}{\sum_{k:W_k>\omega^*} e^{\widehat{\beta}_n^\top \mathbf{Z}_k} Y_k(X_i)}.$$

As in the first part of this chapter, model adequacy can be assessed by using the martingale

$$M_{i,\omega}(t) = \begin{cases} N_i(t) - \int_0^t Y_i(s) e^{\beta_0^\top \mathbf{Z}_i}\, dA_{0,1,\omega}(s) & \text{if } W \le \omega^* \\ N_i(t) - \int_0^t Y_i(s) e^{\beta_0^\top \mathbf{Z}_i}\, dA_{0,2,\omega}(s) & \text{if } W > \omega^*, \end{cases} \tag{3.10}$$

and the martingale residual process for the ith observation is defined as

$$\widehat{M}_{i,\omega}(t) = \begin{cases} N_i(t) - \int_0^t Y_i(s) e^{\beta_0^\top \mathbf{Z}_i} \, d\widehat{A}_{0,1,\omega}(s) & \text{if } W \leq \omega^* \\ N_i(t) - \int_0^t Y_i(s) e^{\beta_0^\top \mathbf{Z}_i} \, d\widehat{A}_{0,2,\omega}(s) & \text{if } W > \omega^*. \end{cases} \tag{3.11}$$

and it verifies the property

$$\sum_{i=1}^{n} \widehat{M}_{i,\omega} = 0$$

3.4.2 The proposed test statistic

We consider the simple model

$$H_0 : \alpha(t|\mathbf{Z}, W) = \begin{cases} \alpha_{0,1}(t) e^{\beta_0^\top \mathbf{Z}} & \text{si } W \leq \omega^* \\ \alpha_{0,2}(t) e^{\beta_0^\top \mathbf{Z}} & \text{si } W > \omega^* \end{cases} \tag{3.12}$$

where ω^* is an unknown threshold, $\alpha_{0,1}(\cdot)$ and $\alpha_{0,2}(\cdot)$ are unknown baseline hazard functions and β_0 is an unknown regression parameter. We propose the test statistic based on an idea similar on within-stratum partial cumulative sum of the martingale residuals (3.10):

$$Q_r^{(1)}(\widehat{\beta}_{n,\omega}, \omega) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widehat{M}_{i,\omega} 1_{\{W_i \leq \omega\}} 1_{\{\widehat{\beta}_{n,\omega}^\top \mathbf{Z}_i \leq r\}}$$

and

$$Q_r^{(2)}(\widehat{\beta}_{n,\omega}, \omega) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widehat{M}_{i,\omega} 1_{\{W_i > \omega\}} 1_{\{\widehat{\beta}_{n,\omega}^\top \mathbf{Z}_i \leq r\}},$$

where r ranges over some suitable interval $\mathcal{R} \subset \mathbb{R}$.
The quantity $Q_r^{(\ell)}(\widehat{\beta}_{n,\omega}, \omega)$ measures adequacy (or conversely, the lack-of-fit) of model (3.12) within stratum ℓ, $\ell = 0, 1$, for a threshold value equal to ω. Under model(3.12), one would expect the partial-sum process (also referred to as cusum process for cumulative sum process)

$$Q_r^{(\ell)}(\widehat{\beta}_{n,\omega}, \omega) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widehat{M}_{i,\omega} 1_{\{W_i [\leq or >] \omega\}} 1_{\{\widehat{\beta}_{n,\omega}^\top \mathbf{Z}_i \leq r\}}$$

to fluctuate about 0.

Remark 3.1 *Residual cumulative sum processes have been proposed for testing lack-of-fit in various models: generalized linear models, for example Su and Wei (1991,*

[59]); Stute and Zhu (2002, [58]), proportional hazards model Lin et al. (1993, [37]) Marzec and Marzec (1997, [43]); Gandy and Jensen (2009, [20])...

Remark 3.2 — *Some authors propose to index the cumulative sum residual process by* $1_{\{\mathbf{Z}_i \leq r\}}$ *instead of* $1_{\{\widehat{\beta}_{n,\omega}^\top \mathbf{Z}_i \leq r\}}$*, where the event* $\{\mathbf{Z}_i \leq r\}$ *indicates that all the components of* $\mathbf{Z}_i$ *are less than or equal to those of* r*,*

— *When dimension of* $\mathbf{Z}_i$ *is large, sparseness of data points in a high-dimensional space may be problematic, hence the alternative of "reducing dimensionality" by projecting* $\mathbf{Z}_i$ *on* $\widehat{\beta}_{n,\omega}$*.*

We consider the statistic at $t = \tau$ in order to use the whole data information and we define our statistic as:

$$Q^{(\ell)}(\widehat{\beta}_{n,\omega}, \omega) = \sup_{r \in \mathcal{R}} \left| Q_r^{(\ell)}(\widehat{\beta}_{n,\omega}, \omega) \right|, \quad \ell = 0, 1.$$

We estimate ω^* by the value $\hat{\omega}$ which achieves the best global model fit in the following sense, we define:

$$\hat{\omega} = \arg\min_{\omega \in \mathcal{G}} \left(Q^{(1)}(\widehat{\beta}_{n,\omega}, \omega) \vee Q^{(2)}(\widehat{\beta}_{n,\omega}, \omega) \right)$$

in order to estimate β with

$$\widehat{\beta}_{n,\hat{\omega}} = \arg\max_{\beta} L_n(\beta, \hat{\omega}).$$

3.5 A simulation study

A variance estimate of $\widehat{\beta}_{n,\widehat{\omega}}$ is necessary to obtain confidence intervals and hypothesis tests but we cannot simply plug $\widehat{\beta}_{n,\widehat{\omega}}$ in the Fisher information matrix for calculating variance estimates: we have to account for uncertainty in the estimation of ω^*. The failure to do this result in variance under- estimation in term of the bootstrap method. A simulation study was carried out to investigate the performance of the proposed bootstrap method to estimate the variance of $\widehat{\beta}_{n,\omega}$

In order to complete our investigations of the performance of the bootstrap proposed method in the context of the stratified Cox model with unknown threshold defining the strata, we carried out a simulation study under various scenario. We

consider the stratified Cox model of Eq. (3.12) with Weibull baseline hazard function in both strata: $\alpha_{0,\ell}(t) = \lambda_\ell \alpha_\ell t^{\alpha_\ell - 1}$ with $(\alpha_1, \lambda_1) = (1, 3)$ et $(\alpha_2, \lambda_2) = (1, 1)$.
The set of the regression parameter is considered : $\beta_0 = (1.3, 0.75, -0.5, 0.25, 0)^\top$. **Z** was generated from discrete and continuous covariates (uniform, Gaussian, Bernouilli). Censoring time is generated from the exponential distribution with parameter p where p is chosen to yield censoring percentage approximately equal to 10%, 20% and 40%. We note n_1 and n_2 the number of observations in the strata 1 and 2 respectively, with $n_1 = n_2 = n$. Three sample sizes are considered (n=425, 225 and 155). For each simulation run, 1000 data sets are generated using an **R** program.
For comparison, the method which applied a bootstrap estimation is evaluated: the procedure generates B bootstrap samples of size n from the initial data set $(X_i, \delta_i, Z_i, W_i)$,$i = 1, \cdots, n$, for each bootstrap sample $(X_i^{(b)}, \delta_i^{(b)}, Z_i^{(b)}, W_i^{(b)})$, $i = 1, \cdots, n$, we compute

$$\widehat{\beta}_{n,\widehat{\omega}} = \arg\max L_n^{(b)}(\beta, \widehat{\omega}), \quad pour \quad b = 1, \cdots, B.$$

Finally, we estimate the variance $\mathbb{V}ar(\widehat{\beta}_{n,\widehat{\omega}})$ by the empirical variance of $\widehat{\beta}_{n,\widehat{\omega}}^{(b)}$, $b = 1, \cdots, B$. . The results of the Oracle method, i.e. when we assume that the true ω^* is known, are also obtained. For each combination of different parameters of study (sample size, censored percentage) we estimate $\widehat{\beta}_{n,\hat{\omega},j}^{(1)}, \ldots, \widehat{\beta}_{n,\hat{\omega},j}^{(N)}$ which applied bootstrap and $\widetilde{\beta}_{n,j}^{(1)}, \ldots, \widetilde{\beta}_{n,j}^{(N)}, j = 1, \cdots, 5$ which proceed the oracle method. We also determined for each value the empirical root mean square error (RMSE),the empirical standard deviation (SD), as well as the empirical coverage probability (CP) of 95%-level confidence intervals and average length of 95%-level confidence intervals (ℓ), the table "Tableau 3.3" summarise the key results of these simulations (table 2 and 3 are given in Appendix). Which allows us to conclude that the value obtained appears to be similar for both methods. The oracle method is used to evaluate the efficiency of our proposed method. We note that our method has good performance in term of bias and appears to be efficient similarly than Oracle. The figure 3.3 , 3.7 and 3.11 of Boxplots clearly displays the comparative of N two-step estimates $\widehat{\beta}_{n,\hat{\omega},j}^{(1)}, \ldots, \widehat{\beta}_{n,\hat{\omega},j}^{(N)}$ and oracle estimates $\widetilde{\beta}_{n,j}^{(1)}, \ldots, \widetilde{\beta}_{n,j}^{(N)}, j = 1, \cdots, 5$ and censoring percentage respectively 40%, 20% and 10% (the other figures for the different numbers of observations are given in Appendix B and C) . We construct histograms, quantile-quantile plot of N normalized estimates $\widehat{\beta}_{n,\hat{\omega},j}^{(1)}, \ldots, \widehat{\beta}_{n,\hat{\omega},j}^{(N)}$, j=1,...,5, $(n_1, n_2) = 1(00, 125)$ and censoring percentage 40%, 20% and 10% respectively, the figure 3.1, 3.2, 3.5, 3.6, 3.9 and 3.10 confirms that the asymptotic normality of the estimators. The figure 3.4, 3.8 and 3.12 ensures that the density of the two methods are almost identical for different censoring percentage.

		proposed method					oracle				
c		β_1	β_2	β_3	β_4	β_5	β_1	β_2	β_3	β_4	β_5
10%	RMSE	0.0171	0.0208	0.0261	0.0033	0.0745	0.0117	0.0180	0.0237	0.0030	0.0694
	bias	-0.0581	-0.0348	0.0284	-0.0112	-0.0034	0.0220	0.0136	-0.0043	0.0042	-0.0098
	SE	0.1105	0.1433	0.1627	0.0550	0.2729	0.1069	0.1355	0.1542	0.0519	0.2576
	SD	0.1173	0.1400	0.1590	0.0561	0.2730	0.1061	0.1334	0.1538	0.0547	0.2633
	CP	0.8870	0.9500	0.9500	0.9500	0.9390	0.9560	0.9610	0.9470	0.9440	0.9440
	ℓ	0.4333	0.5617	0.6379	0.2154	1.0697	0.4192	0.5311	0.6044	0.2036	1.0100
20%	RMSE	0.0166	0.0213	0.0295	0.0036	0.0805	0.0136	0.0204	0.0271	0.0032	0.0753
	bias	-0.0508	-0.0210	0.0297	-0.0089	0.0042	0.0266	0.0207	0.0004	0.0064	0.0043
	SE	0.1159	0.1513	0.1728	0.0581	0.2903	0.1118	0.1431	0.1634	0.0548	0.2723
	SD	0.1186	0.1445	0.1691	0.0592	0.2838	0.1135	0.1415	0.1646	0.0559	0.2745
	CP	0.9080	0.9660	0.9530	0.9410	0.9510	0.9460	0.9510	0.9540	0.9390	0.9540
	ℓ	0.4542	0.5931	0.6775	0.2276	1.1379	0.4381	0.5608	0.6404	0.2150	1.0673
40%	RMSE	0.0195	0.0306	0.0421	0.0044	0.1052	0.0172	0.0279	0.0392	0.0042	0.0998
	bias	-0.0380	-0.0311	0.0167	-0.0050	-0.0012	0.0335	0.0084	-0.0100	0.0077	-0.0018
	SE	0.1323	0.1770	0.2053	0.0682	0.3393	0.1253	0.1651	0.1913	0.0635	0.3164
	SD	0.1346	0.1722	0.2045	0.0659	0.3245	0.1267	0.1669	0.1978	0.0641	0.3161
	CP	0.9220	0.9470	0.9470	0.9540	0.9590	0.9430	0.9450	0.9520	0.9530	0.9480
	ℓ	0.5186	0.6937	0.8047	0.2674	1.3301	0.4911	0.6471	0.7500	0.2488	1.2404

Tableau 3.3: Simulation results with $(n_1, n_2) = (110, 125)$. RMSE: empirical root mean square error. SD: empirical standard deviation. CP: empirical coverage probability of 95%-level confidence intervals. ℓ: average length of 95%-level confidence intervals. All results are based on N = 1000 simulated samples.

3.6 Conclusion

In this paper, we propose a goodness-of-fit test statistic for the stratified proportional hazards model and we establish its asymptotic distribution under the null hypothesis that a stratified model holds. Our simulation study suggests that the proposed test performs well under a wide range of conditions (sample size, censoring fraction, alternative hypothesis).

Now, several issues deserve attention. First, in the stratified proportional hazards model, the variable used for stratifying the population under study is often discrete (*e.g.*, gender, disease stage, socio-professional category...). Very first step towards an objective method for categorising a stratifying variable in proportional hazards regression $\Rightarrow$ still much to be done, *e.g.* in the case of three strata, the resulting model may be written as $\alpha_{0,1}(t)\exp(\beta^\top \mathbf{Z})1_{\{W\leq\omega_1\}} + \alpha_{0,2}(t)\exp(\beta^\top \mathbf{Z})1_{\{\omega_1<W<\omega_2\}} + \alpha_{0.3}(t)\exp(\beta^\top \mathbf{Z})1_{\{\omega_2<W\}}$, where ω_1, ω_2 and ω_3 are some unkown thresholds. To the best of our knowledge, no procedure was proposed yet to estimate ω_i, $i = 1, 2, 3$. We are currently exploring this original application of our test statistic.

Second, our test is feasible when the number of covariates stays moderate. A large number of covariates will raise some computational issues. For example, calculating (3.3) over a fine grid of a high-dimensional space will eventually be time-consuming.

We are currently exploring some directions to reduce this computational burden.

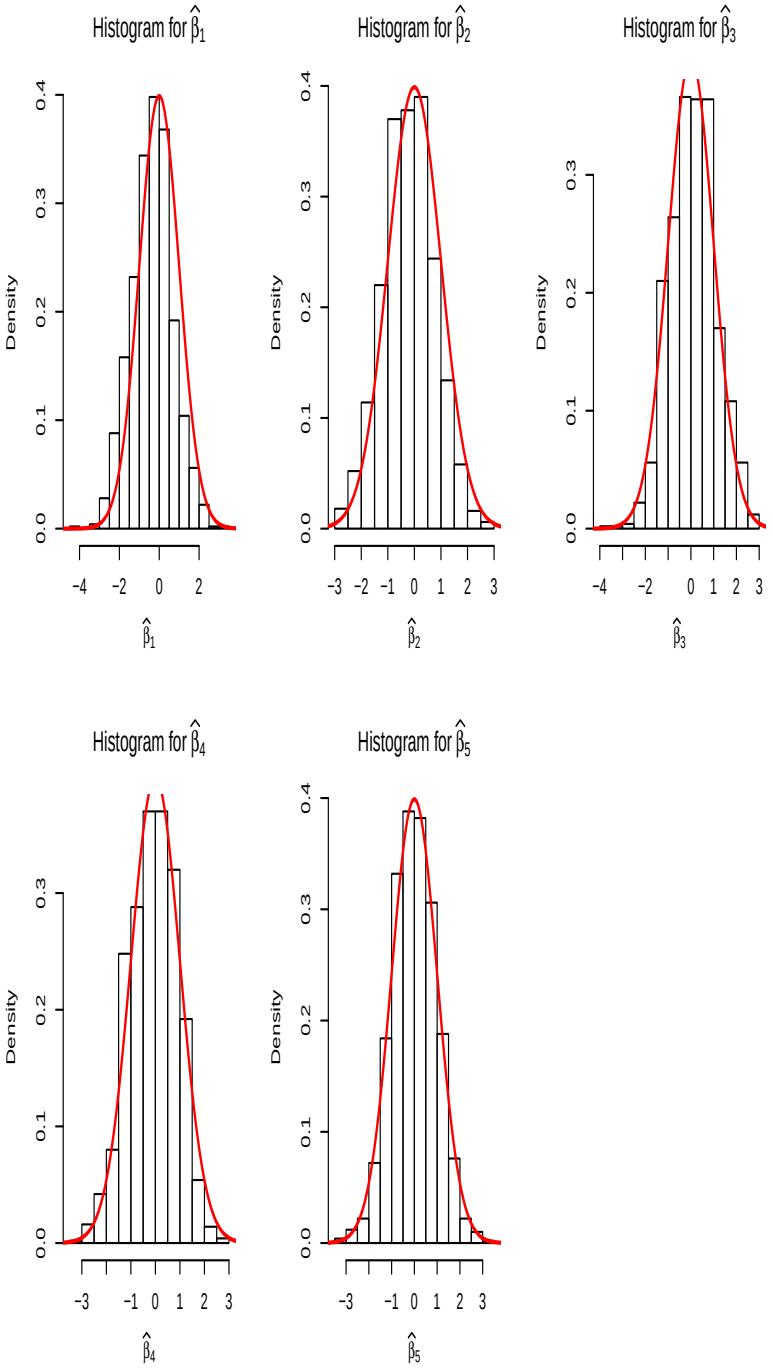

FIGURE 3.1: Histogram of the N normalized estimates $\widehat{\beta}^{(1)}_{n,\hat{\omega},j}, \ldots, \widehat{\beta}^{(N)}_{n,\hat{\omega},j}$, $j = 1, \ldots, 5$, $(n_1, n_2) = (110, 125)$ and censoring percentage = 40%. In red: density of the $\mathcal{N}(0, 1)$.

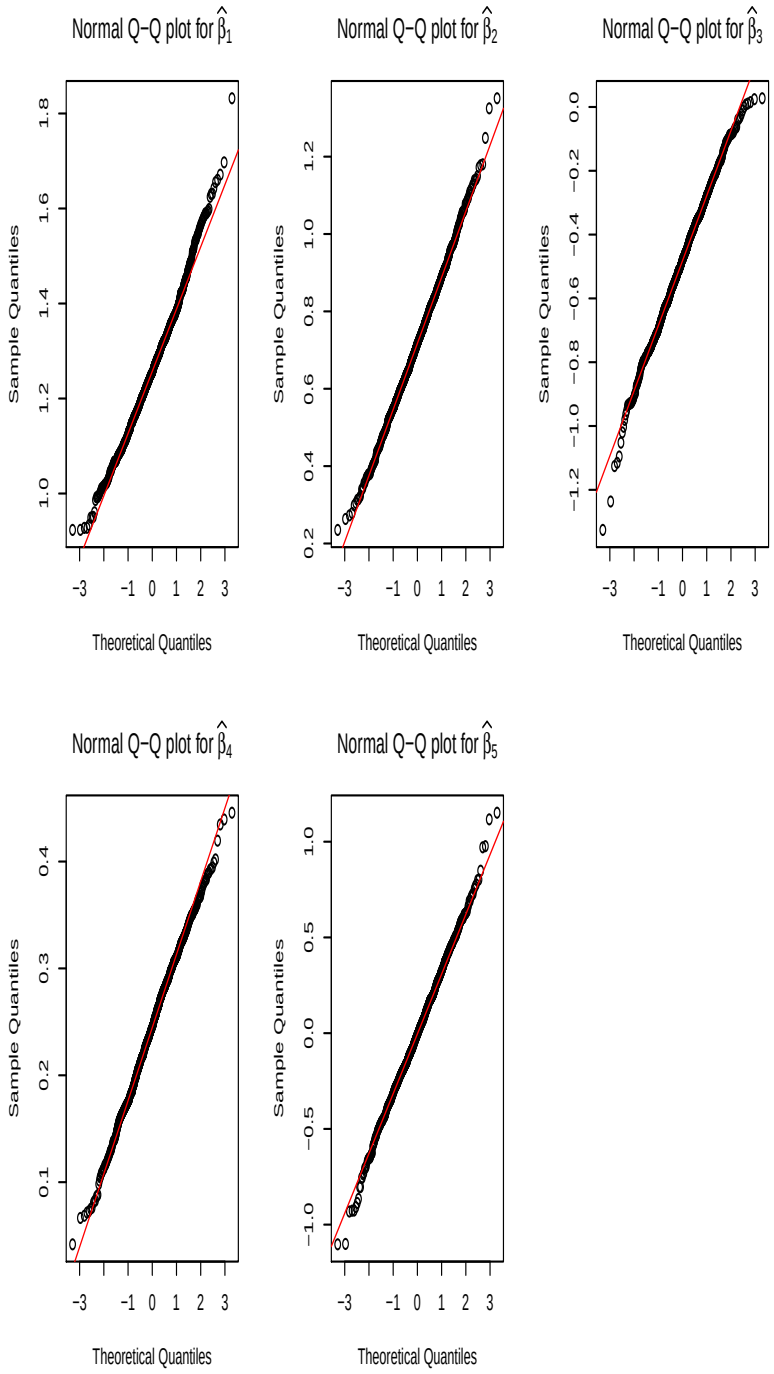

FIGURE 3.2: QQ-plots of the N estimates $\widehat{\beta}^{(1)}_{n,\hat{\omega},j}, \ldots, \widehat{\beta}^{(N)}_{n,\hat{\omega},j}$, $j = 1, \ldots, 5$, $(n_1, n_2) = (110, 125)$ and censoring percentage = 40%.

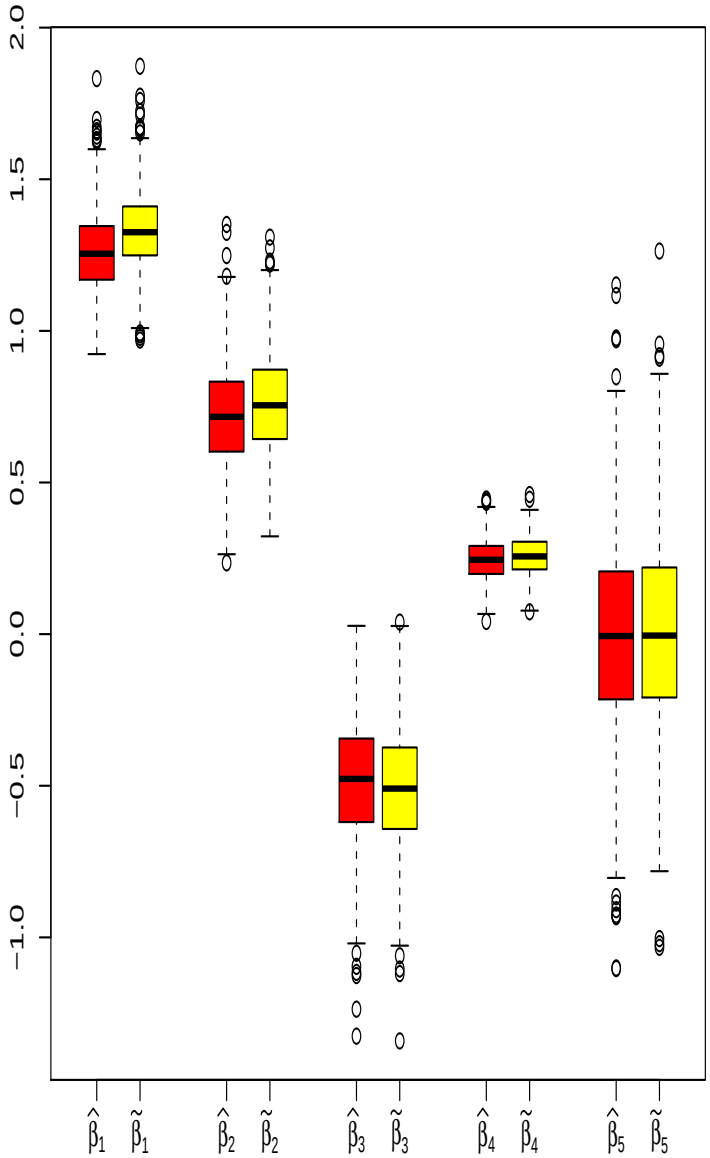

FIGURE 3.3: Boxplots-plots of the N **two-step estimates** $\widehat{\beta}^{(1)}_{n,\hat{\omega},j}, \ldots, \widehat{\beta}^{(N)}_{n,\hat{\omega},j}$ and **oracle estimates** $\widetilde{\beta}^{(1)}_{n,j}, \ldots, \widetilde{\beta}^{(N)}_{n,j}$, $j = 1, \ldots, 5$, $(n_1, n_2) = (110, 125)$ and censoring percentage = 40%.

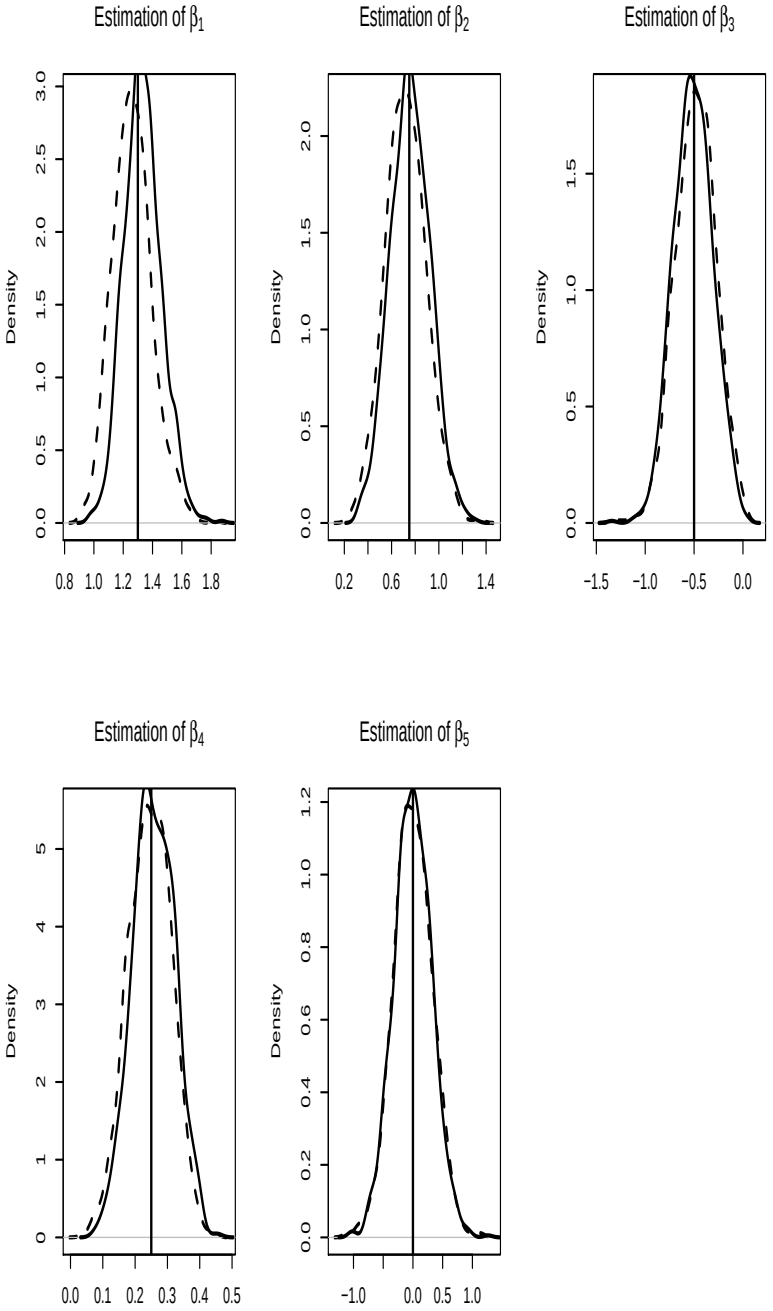

FIGURE 3.4: Density estimation from the N two-step estimates (dashed line) and oracle estimates (solid line), $(n_1, n_2) = (110, 125)$ and censoring percentage = 40%. The true value is indicated by a vertical line.

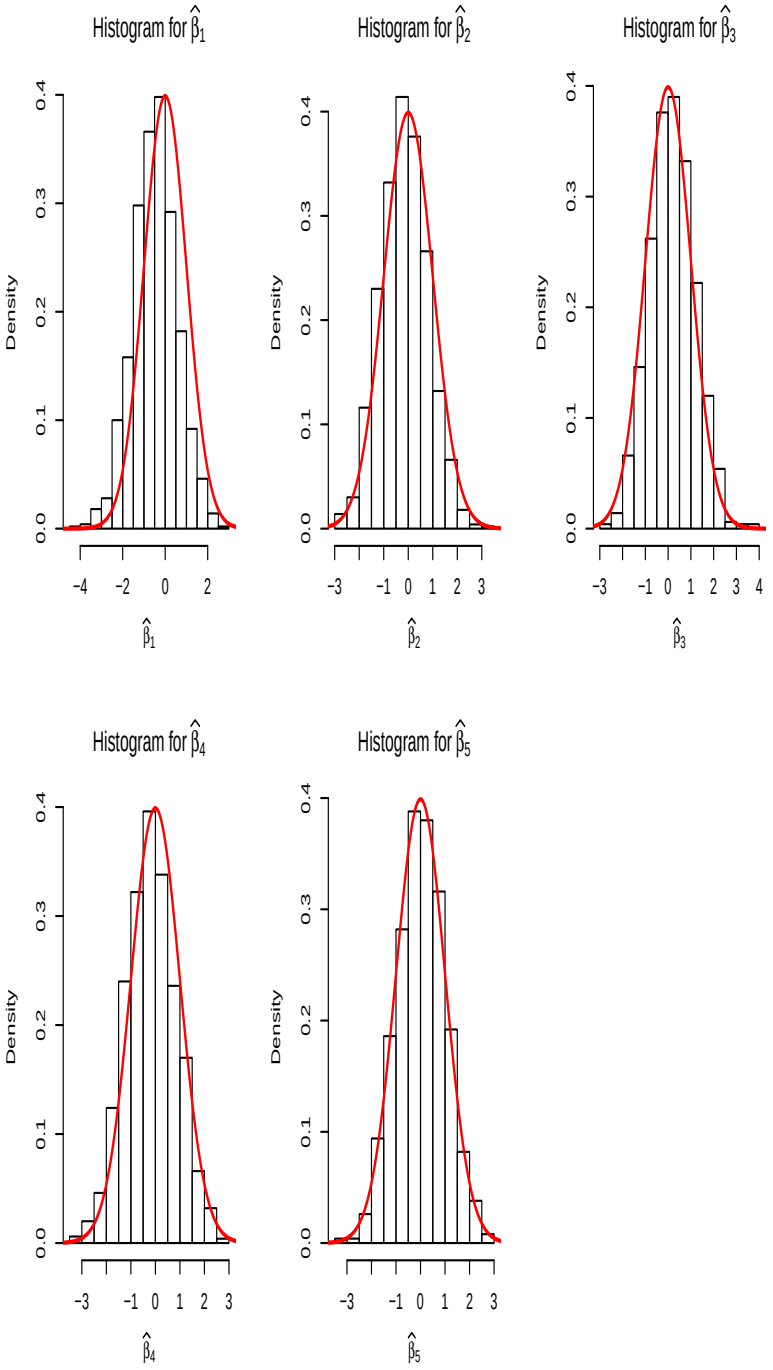

FIGURE 3.5: Histogram of the N normalized estimates $\widehat{\beta}^{(1)}_{n,\hat{\omega},j}, \ldots, \widehat{\beta}^{(N)}_{n,\hat{\omega},j}$, $j = 1, \ldots, 5$, $(n_1, n_2) = (110, 125)$ and censoring percentage = 20%. In red: density of the $\mathcal{N}(0, 1)$.

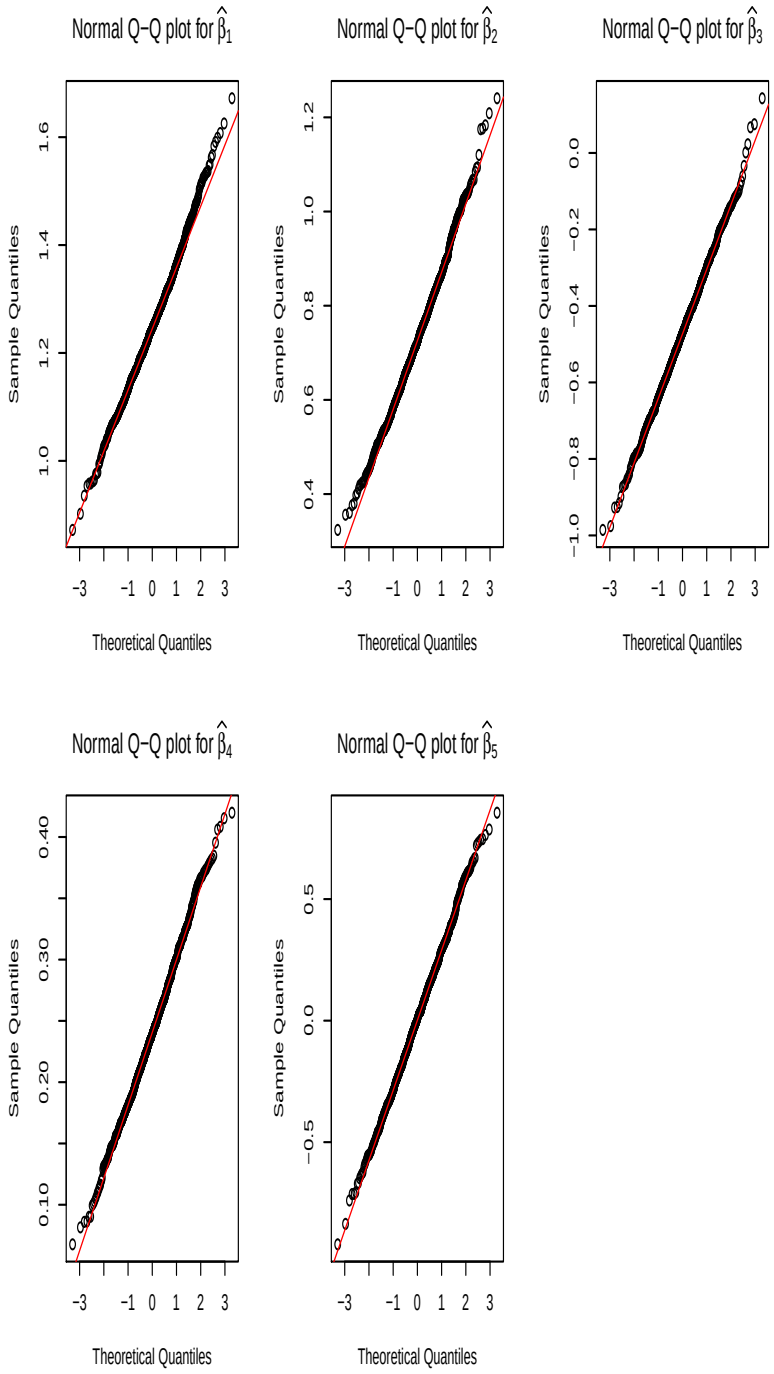

FIGURE 3.6: QQ-plots of the N estimates $\widehat{\beta}^{(1)}_{n,\hat{\omega},j}, \ldots, \widehat{\beta}^{(N)}_{n,\hat{\omega},j}$, $j = 1, \ldots, 5$, $(n_1, n_2) = (110, 125)$ and censoring percentage = 20%.

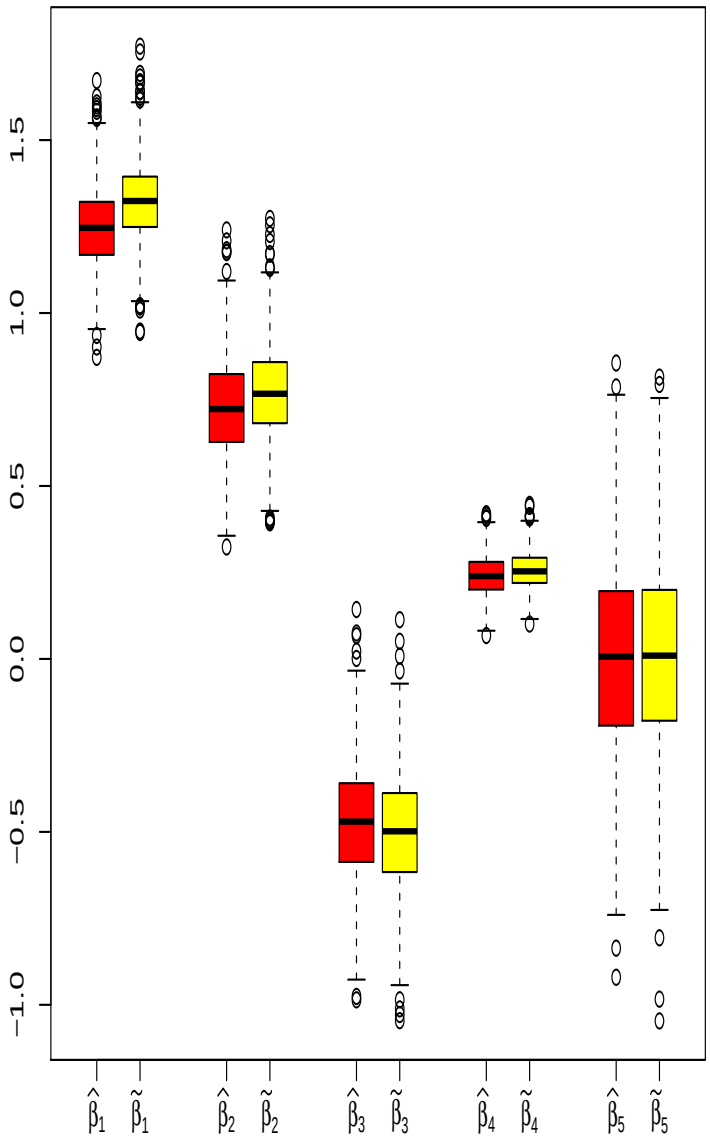

FIGURE 3.7: Boxplots-plots of the N **two-step estimates** $\widehat{\beta}^{(1)}_{n,\hat{\omega},j}, \ldots, \widehat{\beta}^{(N)}_{n,\hat{\omega},j}$ and **oracle estimates** $\widetilde{\beta}^{(1)}_{n,j}, \ldots, \widetilde{\beta}^{(N)}_{n,j}$, $j = 1, \ldots, 5$, $(n_1, n_2) = (110, 125)$ and censoring percentage = 20%.

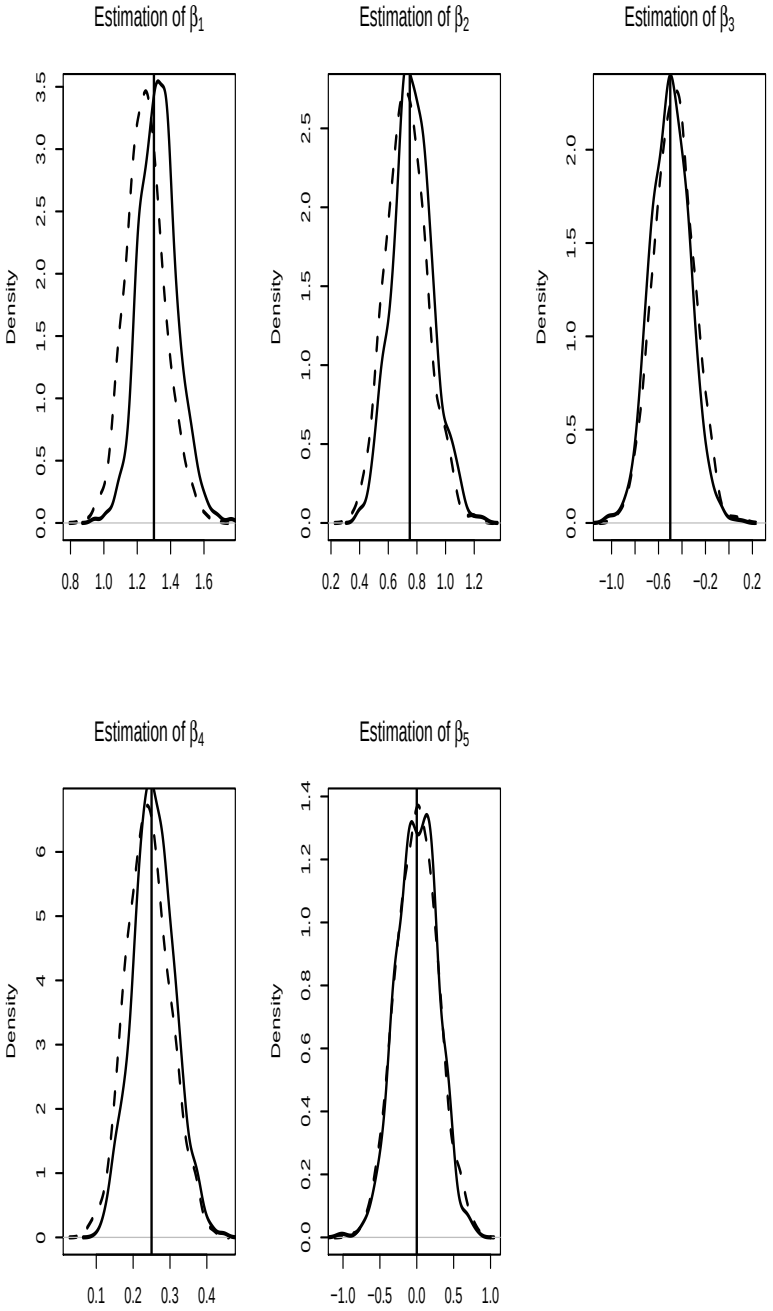

FIGURE 3.8: Density estimation from the N two-step estimates (dashed line) and oracle estimates (solid line), $n = 100, 125$ and censoring percentage = 20%. The true value is indicated by a vertical line.

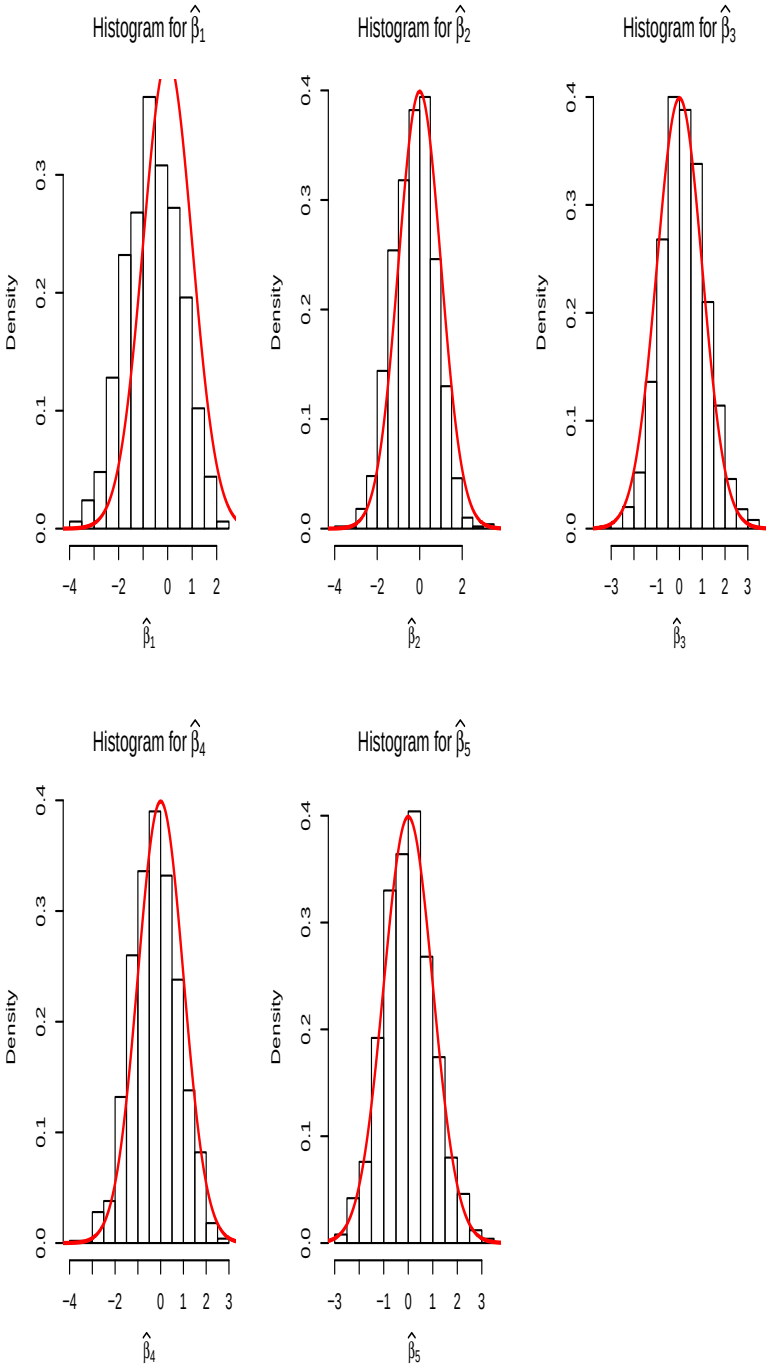

FIGURE 3.9: Histogram of the N normalized estimates $\widehat{\beta}^{(1)}_{n,\hat{\omega},j}, \ldots, \widehat{\beta}^{(N)}_{n,\hat{\omega},j}$, $j = 1, \ldots, 5$, $(n_1, n_2) = (110, 125)$ and censoring percentage = 10%. In red: density of the $\mathcal{N}(0, 1)$.

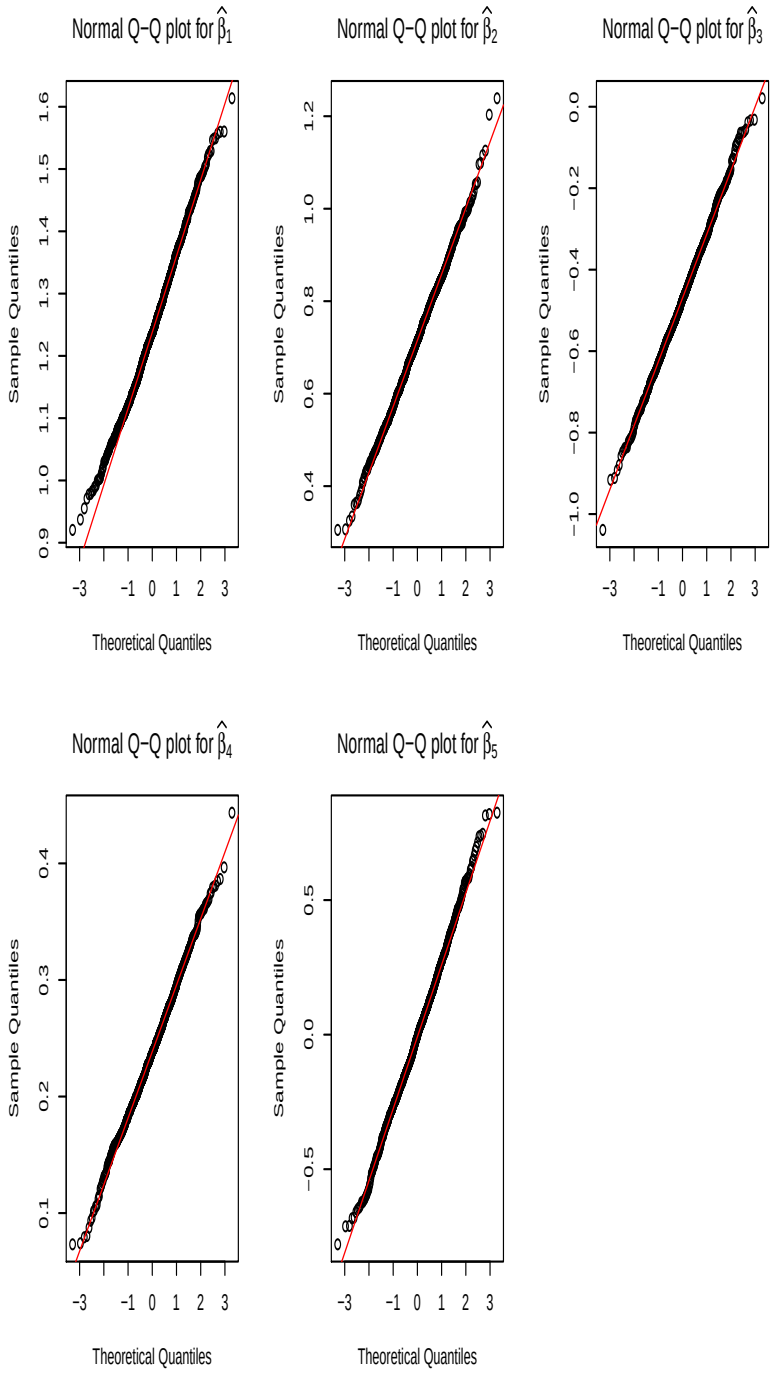

FIGURE 3.10: QQ-plots of the N estimates $\widehat{\beta}^{(1)}_{n,\hat{\omega},j}, \ldots, \widehat{\beta}^{(N)}_{n,\hat{\omega},j}$, $j = 1, \ldots, 5$, $(n_1, n_2) = (110, 125)$ and censoring percentage = 10%.

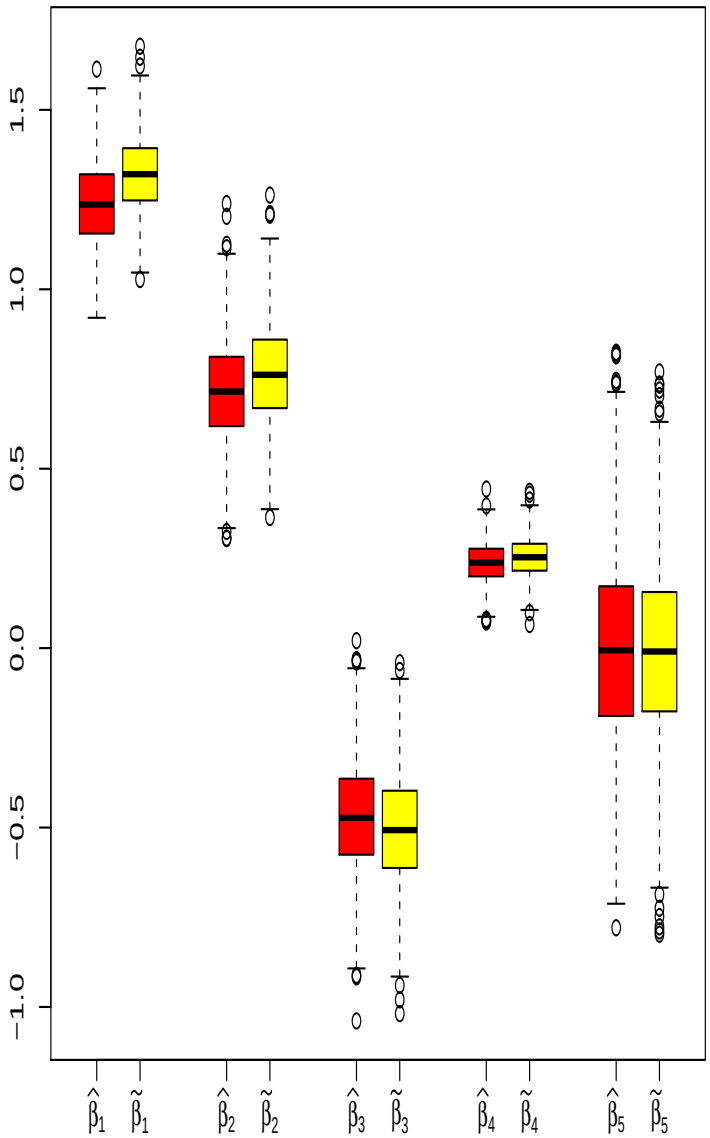

FIGURE 3.11: Boxplots-plots of the N **two-step estimates** $\widehat{\beta}^{(1)}_{n,\hat{\omega},j}, \ldots, \widehat{\beta}^{(N)}_{n,\hat{\omega},j}$ and **oracle estimates** $\widetilde{\beta}^{(1)}_{n,j}, \ldots, \widetilde{\beta}^{(N)}_{n,j}$, $j = 1, \ldots, 5$, $n = 100, 125$ and censoring percentage = 10%.

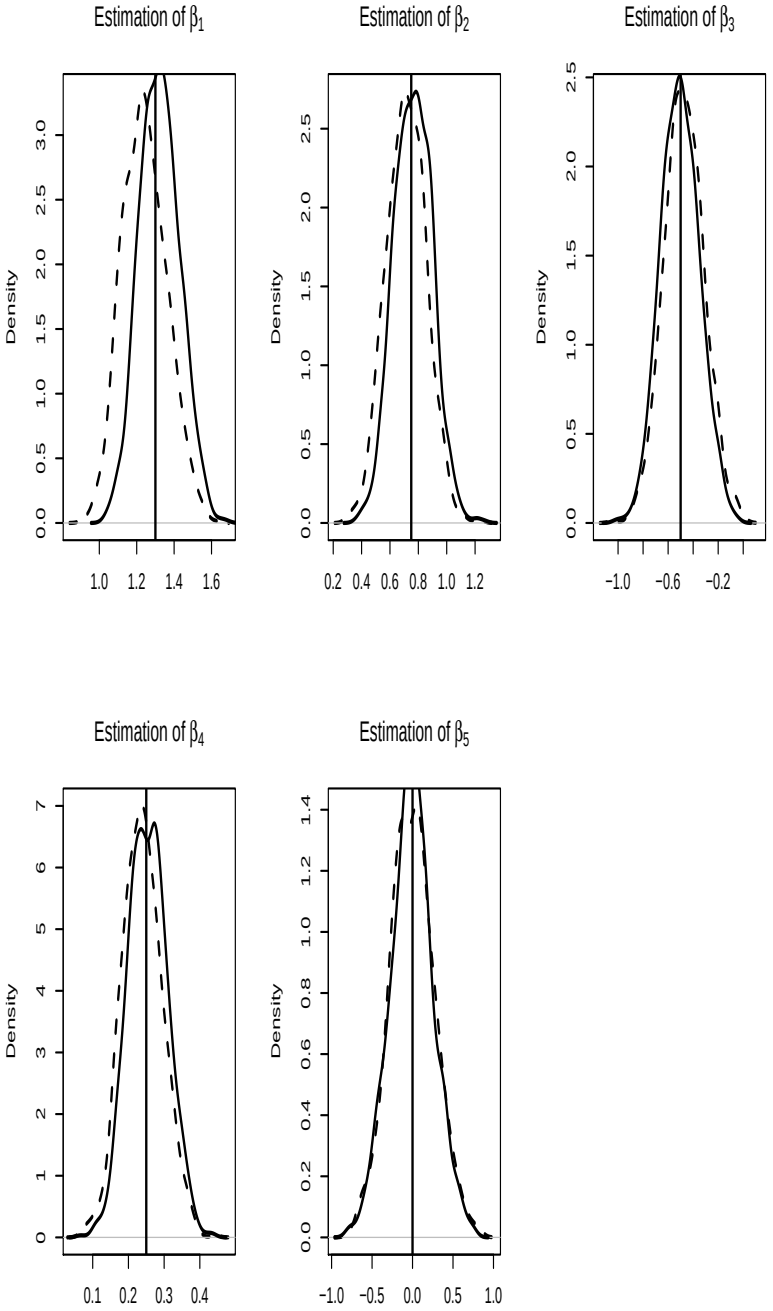

FIGURE 3.12: Density estimation from the N two-step estimates (dashed line) and oracle estimates (solid line), $(n_1, n_2) = (110, 125)$ and censoring percentage = 10%. The true value is indicated by a vertical line.

4

Nonparametric inference in stratified survival samples with missing data

Abstract

The stratified logrank test can be used to compare survival distributions of several groups of patients, while adjusting for the effect of some discrete variable that may be predictive of the survival outcome. In practice, it can happen that this discrete variable is missing for some patients. An inverse-probability-weighted version of the stratified logrank statistic is introduced to tackle this issue. Its asymptotic distribution is derived under the null hypothesis of equality of the survival distributions. A simulation study is conducted to assess behaviour of the proposed test statistic in finite samples. An analysis of a medical dataset illustrates the methodology.

Keywords:logrank test; asymptotic distribution; simulation; medical application.

Sommaire

Ce chapitre fait l'objet d'un article soumis pour publication à "Statistics and Probability Letters".

4.1 Introduction

THe logrank test is commonly used in medical research for comparing survival distributions across groups (such as treatment groups or groups of individuals exposed/non exposed to some risk factor: smoking, professional or environmental exposure...). In this setting, it is sometimes necessary to control for other potentially important factors that may affect the survival distribution. If these factors are discrete, or if they can be discretized into a workable number of levels, one can use a generalization of the logrank test, which is called the *stratified* logrank [34].

For example, consider a medical study involving n patients followed until death and assume that each patient belongs to one of two mutually exclusive groups. Suppose that we wish to compare the survival curves of these groups while adjusting for some discrete factor S with L modalities (also called strata, such as gender, disease stage...). If $\alpha_{1,\ell}(\cdot)$ (respectively $\alpha_{2,\ell}(\cdot)$) is the hazard function of death in the ℓ-th stratum ($\ell \in \{1, \ldots, L\}$) of the first (respectively second) group, the testing problem can be formulated as

$$H_0 : \alpha_{0,\ell}(\cdot) := \alpha_{1,\ell}(\cdot) = \alpha_{2,\ell}(\cdot) \text{ for every } \ell = 1, \ldots, L,$$

against

$$H_1 : \alpha_{1,\ell}(\cdot) \neq \alpha_{2,\ell}(\cdot) \text{ for some } \ell \in \{1, \ldots, L\},$$

where $\alpha_{0,\ell}(\cdot)$ denotes the common hazard function in the ℓ-th stratum of both groups, under H_0. Let $(T_1^0, \ldots, T_n^0)$ and $(C_1, \ldots, C_n)$ be respectively the survival and censoring times in the pooled groups. When censoring is present, one only observes pairs (T_i, δ_i), $i = 1, \ldots, n$, where $T_i = \min(T_i^0, C_i)$ and $\delta_i = 1_{\{T_i^0 \leq C_i\}}$ is the censoring indicator. Let G_i ($G_i \in \{1, 2\}$) and S_i ($S_i \in \{1, \ldots, L\}$) indicate respectively the group and stratum of the i-th individual. Finally, define the counting process $N_i(t) = \delta_i 1_{\{T_i \leq t\}}$ and the at-risk process $Y_i(t) = 1_{\{T_i \geq t\}}$. The stratified logrank statistic is obtained by taking the difference between the group-specific Nelson-Aalen estimators of the cumulative hazard functions in each stratum and the within-stratum Nelson-Aalen estimators calculated under H_0. Its expression is:

$$\mathbb{Z} = \sum_{i=1}^{n} \int_0^{\tau} \sum_{\ell=1}^{L} 1_{\{S_i = \ell\}} \left[1_{\{G_i = 1\}} - E_{\ell,n}(t) \right] dN_i(t), \tag{4.1}$$

where $\tau > 0$ is the end of follow-up time and

$$E_{\ell,n}(t) = \frac{\sum_{i=1}^{n} Y_i(t) 1_{\{S_i=\ell\}} 1_{\{G_i=1\}}}{\sum_{i=1}^{n} Y_i(t) 1_{\{S_i=\ell\}}}.$$

Let $\widehat{\sigma}^2$ be the estimated asymptotic variance of $\mathbb{Z}$. Then under H_0, $(\mathbb{Z}/\hat{\sigma})^2$ is asymptotically distributed as a χ^2 distribution with one degree of freedom [34].

In some applications, the factor S defining the strata may be missing for some individuals. For example, [2] describe a dataset containing time to esophageal cancer, baseline histology and several risk factors (such as smoking status: regular use for at least 6 months or not) for patients participating in a medical trial. A four-categories factor variable indicating the level of tissue zinc concentration is also recorded on each patient (zinc deficiency is a contributing factor to the development of esophageal squamous cell carcinoma in humans). One may use a stratified logrank test with four strata to assess influence of the smoking status on esophageal cancer incidence while adjusting for zinc exposure. However, measurement of zinc concentration requires an esophageal biopsy, which is a costly technique. Therefore, zinc level was measured for 32% of the patients only. This rules out direct application of formula (4.1). One simple solution to this issue is to discard patients with unobserved zinc level (in missing data problems, such a procedure is known as a complete-case analysis). This, however, may induce substantial loss of power, as will be illustrated in our simulation study, and calls for an alternative and more efficient method. This is the topic of the current paper.

Missing data problems in survival analysis have been widely investigated so far, but to the best of our knowledge, stratified logrank testing with missing stratum information was not yet considered thoroughly. One first attempt was provided by [15], who considered the special case where G_i is a randomized treatment group indicator and proposed a modified version of the stratified logrank statistic, based on the regression calibration idea. But the proposed methodology relies heavily on the assumption that G_i is independent of all other information. Moreover, the proposed test assumes that the probability distribution of the stratum variable is known to the investigator, which is rarely the case in practice. In this paper, we propose a new statistic for testing H_0 without making these strong assumptions. This statistic uses the inverse-probability-weighting-of-complete-cases principle. Basic idea is to correct for missing data by giving extra weight to subjects with fully observed data.

We first propose an inverse-probability-weighted version of Nelson-Aalen estimator of the cumulative hazard functions $\Lambda_{0,\ell}(\cdot) = \int_0^{\cdot} \alpha_{0,\ell}(u)\, du$ under H_0. From this, we construct an inverse-probability-weighted stratified logrank statistic appropriate for testing H_0 with missing stratum information. We investigate its properties, both

theoretically and numerically. The rest of the paper is organized as follows. In Section 4.2, we introduce our modified Nelson-Aalen estimator and stratified logrank statistic and we establish their aymptotic distribution. A simulation study and an application to a medical dataset are described in Section 4.3. A conclusion and some perspectives are given in Section 4.4. All proofs are postponed to an appendix.

4.2 Inverse-probability-weighted stratified logrank statistic

4.2.1 Setting and the proposed test statistic

Consider a sample of n independent individuals observed on the time interval $[0, \tau]$, with $\tau < \infty$. For each subject, we observe the random duration $T = \min(T^0, C)$ and the censoring indicator $\delta = 1_{\{T^0 \leq C\}}$. The censoring time C is assumed to be independent of T^0 and non informative. We let G denote the group indicator (to keep notations simple, we assume that $G \in \{1, 2\}$ but the proposed methodology can easily be generalized to the case of K groups, $K > 2$). G may be either fixed or random. Let S ($S \in \{1, \ldots, L\}$) be the stratum indicator, which may be unobserved for some individuals. Thus, we define the missingness indicator R which equals 1 if S is observed and 0 otherwise. We also introduce an auxiliary covariate Z which is not involved in the hazard function of T^0 but may be used to describe the missingness mechanism. The observed data consist of n independent vectors $(T_i, \delta_i, Z_i, G_i, R_i, R_i S_i)$, $i = 1, \ldots, n$. We assume that strata are missing at random (MAR), that is, the probability that S is missing given (T, δ, Z, G, S) depends only on the observed vector $\mathcal{O} = (T, \delta, Z, G)$. Thus, we have $\mathbb{P}(R = 1|\mathcal{O}, S) = \mathbb{P}(R = 1|\mathcal{O})$. MAR is a common assumption for statistical analysis with missing data [39]. Under this assumption, we propose an inverse-probability-weighted version of the stratified logrank statistic (4.1).

Inverse-probability-weighting-of-complete-cases was suggested by [27] and since then, the method has been adapted to various settings (see [56] for a detailed review of inverse-probability-weighting methods for missing data problems). In survival analysis, [50] and [67] propose inverse-probability-weighted estimation in Cox regression with missing covariates. [69] develop an adjusted Kaplan-Meier estimator to reduce confounding effects by using inverse-probability-of-treatment weighting. [60] propose an inverse-probability-weighted estimator of the survival function with left-truncated data and missing censoring indicators. [40] define inverse-probability-weighted estimators in the Cox model with missing failure indicators. [30] and [17] use a similar approach in competing risks models with missing cause of failure. [28]

propose inverse-probability-weighted estimation in the proportional hazards model with length-biased failure times and missing covariates. But to the best of our knowledge, our work constitutes the first attempt to adapt inverse-probability-weighting to stratified logrank testing with missing strata.

The probability $\mathbb{P}(R = 1|\mathcal{O})$ of observing a complete-case is usually unknown and has to be estimated. Here, we assume that this probability can be specified by a parametric model and we note $\Pi(\psi) = \mathbb{P}(R = 1|\mathcal{O})$ and $\Pi_i(\psi) = \mathbb{P}(R = 1|\mathcal{O}_i)$, where ψ is an unknown vector of regression parameters (for example, one may posit a logistic regression model: $\text{logit}(\Pi(\psi)) = \psi^\top \mathcal{O}$, or a probit model: $\Phi^{-1}(\Pi(\psi)) = \psi^\top \mathcal{O}$, where Φ is the distribution function of the standard normal distribution and $\top$ is the transpose operator). The parameter ψ can be estimated by maximizing the likelihood $L_n(\psi) = \prod_{i=1}^n \Pi_i(\psi)^{R_i}(1 - \Pi_i(\psi))^{1-R_i}$. Let $\widehat{\psi} = \arg\max_\psi L_n(\psi)$ denote the maximum likelihood estimator. If the model $\Pi(\psi)$ is correctly specified, $\widehat{\psi}$ is a consistent and asymptotically normal estimator of ψ [64].

In order to construct our test statistic, we first propose a modified version of Nelson-Aalen estimator of the within-stratum cumulative hazard functions $\Lambda_{0,\ell}(\cdot) = \int_0^\cdot \alpha_{0,\ell}(u)\, du$ ($\ell = 1, \ldots, L$) under H_0. Using the inverse-probability-weigthed idea, we propose to estimate $\Lambda_{0,\ell}(\cdot)$ ($\ell = 1, \ldots, L$) by:

$$\widehat{\Lambda}_\ell(t, \widehat{\psi}) = \frac{1}{n} \sum_{i=1}^n \int_0^t \frac{\frac{1_{\{S_i=\ell\}} R_i}{\Pi_i(\widehat{\psi})}}{n^{-1} \sum_{j=1}^n Y_j(u) \frac{1_{\{S_j=\ell\}} R_j}{\Pi_j(\widehat{\psi})}}\, dN_i(u),$$

which reduces to the usual Nelson-Aalen estimator $\sum_{i=1}^n \int_0^t \frac{1_{\{S_i=\ell\}}}{\sum_{j=1}^n Y_j(u) 1_{\{S_j=\ell\}}}\, dN_i(u)$ of the cumulative hazard function in the ℓ-th stratum, when all individual strata are observed. Group-specific estimators $\widehat{\Lambda}_{\ell,1}(t, \widehat{\psi})$ and $\widehat{\Lambda}_{\ell,2}(t, \widehat{\psi})$ can be constructed similarly, as:

$$\widehat{\Lambda}_{\ell,m}(t, \widehat{\psi}) = \frac{1}{n} \sum_{i=1}^n \int_0^t \frac{\frac{1_{\{S_i=\ell\}} R_i}{\Pi_i(\widehat{\psi})} 1_{\{G_i=m\}}}{n^{-1} \sum_{j=1}^n Y_j(u) \frac{1_{\{S_j=\ell\}} R_j}{\Pi_j(\widehat{\psi})} 1_{\{G_j=m\}}}\, dN_i(u), \quad m = 1, 2.$$

Taking the difference between $\widehat{\Lambda}_\ell(t, \widehat{\psi})$ and $\widehat{\Lambda}_{\ell,1}(t, \widehat{\psi})$ finally yields the following inverse-probability-weighted stratified logrank statistic:

$$\widetilde{\mathbb{Z}}(\widehat{\psi}) = \sum_{i=1}^n \int_0^\tau \sum_{\ell=1}^L \frac{1_{\{S_i=\ell\}} R_i}{\Pi_i(\widehat{\psi})} \left[1_{\{G_i=1\}} - \widetilde{E}_{\ell,n}(t, \widehat{\psi})\right] dN_i(t), \tag{4.2}$$

where $\widetilde{E}_{\ell,n}(t,\psi)$ is as:

$$\widetilde{E}_{\ell,n}(t,\psi) = \frac{\sum_{i=1}^{n} Y_i(t)\frac{1_{\{S_i=\ell\}}R_i}{\Pi_i(\psi)}1_{\{G_i=1\}}}{\sum_{i=1}^{n} Y_i(t)\frac{1_{\{S_i=\ell\}}R_i}{\Pi_i(\psi)}}.$$

In section 4.2.3, we establish asymptotic normality of $\widehat{\Lambda}_\ell(t,\widehat{\psi})$ and we derive the asymptotic distribution of (4.2) under H_0. Before that, we introduce some further notations.

4.2.2 Further notations

Let $\dot{\Pi}_i(\psi) = \partial\Pi_i(\psi)/\partial\psi$, $\ddot{\Pi}_i(\psi) = \partial\dot{\Pi}_i(\psi)/\partial\psi^\top$ and

$$\begin{aligned}
\widetilde{S}^{(0)}_{\ell,n}(t,\psi) &= \sum_{i=1}^{n} Y_i(t)\frac{1_{\{S_i=\ell\}}R_i}{\Pi_i(\psi)},\\
\widetilde{S}^{(1)}_{\ell,n}(t,\psi) &= \sum_{i=1}^{n} Y_i(t)\frac{1_{\{S_i=\ell\}}R_i}{\Pi_i(\psi)}1_{\{G_i=1\}},\\
\widetilde{s}^{(0)}_{\ell}(t,\psi) &= \mathbb{E}\left[Y_i(t)\frac{1_{\{S_i=\ell\}}R_i}{\Pi_i(\psi)}\right],\\
\widetilde{s}^{(1)}_{\ell}(t,\psi) &= \mathbb{E}\left[Y_i(t)\frac{1_{\{S_i=\ell\}}R_i}{\Pi_i(\psi)}1_{\{G_i=1\}}\right],\\
\widetilde{e}_{\ell}(t,\psi) &= \widetilde{s}^{(1)}_{\ell}(t,\psi)/\widetilde{s}^{(0)}_{\ell}(t,\psi),\\
m^{(0)}_{\ell}(t,\psi) &= \mathbb{E}\left[Y_i(t)\frac{1_{\{S_i=\ell\}}R_i}{(\Pi_i(\psi))^2}\dot{\Pi}_i(\psi)^\top\right],\\
m^{(1)}_{\ell}(t,\psi) &= \mathbb{E}\left[Y_i(t)\frac{1_{\{S_i=\ell\}}R_i}{(\Pi_i(\psi))^2}\dot{\Pi}_i(\psi)^\top 1_{\{G_i=1\}}\right].
\end{aligned}$$

Let $S_n(\psi)$ and $I(\psi)$ be respectively the score vector and information matrix for ψ under the parametric model $\Pi(\psi)$. Define

$$U_i(\psi) = \frac{(R_i - \Pi_i(\psi))\dot{\Pi}_i(\psi)}{\Pi_i(\psi)(1-\Pi_i(\psi))}.$$

Then we have

$$S_n(\psi) = \frac{\partial \log L_n(\psi)}{\partial\psi} = \sum_{i=1}^{n} U_i(\psi),$$

and

$$I(\psi) = \mathbb{E}\left[U_i(\psi)U_i(\psi)^\top - \frac{\ddot{\Pi}_i(\psi)(R_i-\Pi_i(\psi))}{\Pi_i(\psi)(1-\Pi_i(\psi))}\right]. \tag{4.3}$$

Calculation details for $S_n(\psi)$ and $I(\psi)$ are given in Appendix A. Recall that under H_0, $\alpha_{0,\ell}(\cdot)$ denotes the common hazard function of T^0 in the ℓ-th stratum of both groups. Then, define $M_i(t) = N_i(t) - \int_0^t Y_i(s) \sum_{\ell=1}^L \alpha_{0,\ell}(s) 1_{\{S_i=\ell\}}\, ds$. Finally, let

$$\begin{aligned}
\omega_{\ell,i}(\psi) &= \frac{1_{\{S_i=\ell\}} R_i}{\Pi_i(\psi)}, \ell = 1, \ldots, L, \\
\dot{\omega}_{\ell,i}(\psi) &= \frac{\partial \omega_{\ell,i}(\psi)}{\partial \psi}, \ell = 1, \ldots, L, \\
V(\psi) &= \mathbb{E}\left[-\int_0^\tau \sum_{\ell=1}^L \frac{1_{\{S_i=\ell\}} R_i \dot{\Pi}_i(\psi)^\top}{(\Pi_i(\psi))^2} \left(1_{\{G_i=1\}} - \widetilde{e}_\ell(t,\psi)\right) dN_i(t) \right. \\
&\qquad \left. + \int_0^\tau \sum_{\ell=1}^L \frac{1_{\{S_i=\ell\}} R_i}{\Pi_i(\psi)} \left(\frac{m_\ell^{(1)}(t,\psi)}{\widetilde{s}_\ell^{(0)}(t,\psi)} - \frac{\widetilde{s}_\ell^{(1)}(t,\psi) m_\ell^{(0)}(t,\psi)}{(\widetilde{s}_\ell^{(0)}(t,\psi))^2} \right) dN_i(t) \right], \\
\varphi_i &= \int_0^\tau \sum_{\ell=1}^L \frac{1_{\{S_i=\ell\}} R_i}{\Pi_i(\psi_0)} \left[1_{\{G_i=1\}} - \widetilde{e}_\ell(t,\psi_0)\right] dM_i(t) + V(\psi_0) I(\psi_0)^{-1} U_i(\psi_0),
\end{aligned} \tag{4.4}$$

and

$$W_\ell(t,\psi) = \mathbb{E}\left[\int_0^t \left\{ \frac{\dot{\omega}_{\ell,i}(\psi)^\top}{\widetilde{s}_\ell^{(0)}(u,\psi)} - \frac{\omega_{\ell,i}(\psi) \mathbb{E}[Y_i(u) \dot{\omega}_{\ell,i}(\psi)^\top]}{(\widetilde{s}_\ell^{(0)}(u,\psi))^2} \right\} dN_i(u) \right]. \tag{4.5}$$

4.2.3 Asymptotic results and decision rule

The following regularity conditions are needed to prove asymptotic normality of $\widehat{\Lambda}_\ell(t,\widehat{\psi})$ and to derive the null asymptotic distribution of $\widetilde{\mathbb{Z}}(\widehat{\psi})$:

1. $\mathbb{P}(Y(\tau) > 0) > 0$.
2. The true parameter ψ_0 in model $\Pi(\cdot)$ belongs to the interior of some compact set $\Psi \subset \mathbb{R}^p$ and Π is twice continuously differentiable with respect to ψ. There exists $\eta > 0$ such that $\eta < \Pi(\psi) < 1 - \eta$ for every ψ and $\mathcal{O}$.
3. For every $\ell \in \{1, \ldots, L\}$, $\int_0^\tau \alpha_{0,\ell}(t)\, dt < \infty$, $\inf_{t\in[0,\tau],\psi\in\Psi} \widetilde{s}_\ell^{(0)}(t,\psi) > 0$ and $\mathbb{P}(S = \ell) > 0$.
4. The auxiliary covariate Z is bounded.

We are now in position to state our first main result, which confirms that under H_0, $\widehat{\Lambda}_\ell(t,\widehat{\psi})$ is an appropriate estimator of $\Lambda_{0,\ell}(\cdot)$ when strata are missing at random. Consequently, one can also expect that $\widetilde{\mathbb{Z}}(\widehat{\psi})$ will be appropriate for testing H_0.

Theorem 4.1 *Assume that conditions 1-4 hold and that $\Pi(\psi)$ is correctly specified for $\mathbb{P}(R = 1|\mathcal{O})$. Let $t \in [0,\tau]$. Then under H_0, $n^{\frac{1}{2}}(\widehat{\Lambda}_\ell(t,\widehat{\psi}) - \Lambda_{0,\ell}(t))$ converges in distribution to a zero-mean Gaussian random variable with variance $\mathbb{E}[\xi_{\ell,i}(t,\psi_0)^2]$, where*

$$\xi_{\ell,i}(t,\psi_0) = W_\ell(t,\psi_0) I(\psi_0)^{-1} U_i(\psi_0) + \int_0^t \frac{\omega_{\ell,i}(\psi_0)}{\widetilde{s}_\ell^{(0)}(u,\psi_0)}\, dM_i(u).$$

Proof is given in Appendix B. Our second main result provides the asymptotic distribution of the proposed inverse-probability-weighted stratified logrank statistic under H_0. Proof is given in Appendix C.

Theorem 4.2 *Assume that conditions 1-4 hold and that* $\Pi(\psi)$ *is correctly specified for* $\mathbb{P}(R = 1|\mathcal{O})$. *Then under* H_0, $n^{-\frac{1}{2}}\widetilde{\mathbb{Z}}(\widehat{\psi})$ *converges in distribution to a zero-mean Gaussian random variable with variance* $\mathbb{E}[\varphi_i^2]$.

Remark 1. Asymptotics for usual unweighted Nelson-Aalen estimator and logrank statistic can be derived using martingale arguments [3]. Indeed, both quantities can be expressed as integrals of predictable processes with respect to the martingale $M_i(t)$. Such integrals are themselves martingales. This is not possible here since $\widehat{\Lambda}_\ell(t, \widehat{\psi})$ and $\widetilde{\mathbb{Z}}(\widehat{\psi})$ depend on $T_1, \ldots, T_n$ through $\widehat{\psi}$. Therefore, we take a different approach and express both $\widehat{\Lambda}_\ell(t, \widehat{\psi})$ and $\widetilde{\mathbb{Z}}(\widehat{\psi})$ as empirical means of independent and identically random variables.

Remark 2. Without missing data (that is, when $R_i = 1$ and $\Pi_i(\psi) = 1$ for every $i = 1, \ldots, n$), $\mathbb{E}[\xi_{\ell,i}(t, \psi_0)^2]$ and $\mathbb{E}[\varphi_i^2]$ reduce to the usual asymptotic variances of the Nelson-Aalen estimator and logrank statistic respectively.

The asymptotic variance $\mathbb{E}[\varphi_i^2]$ can be consistently estimated by $\widetilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \widehat{\varphi}_i^2$, where

$$\widehat{\varphi}_i = \int_0^\tau \sum_{\ell=1}^L \frac{1_{\{S_i=\ell\}} R_i}{\Pi_i(\widehat{\psi})} \left[1_{\{G_i=1\}} - \widetilde{E}_{\ell,n}(t, \widehat{\psi})\right] d\widehat{M}_i(t) + \widehat{V}(\widehat{\psi})\widehat{I}(\widehat{\psi})^{-1} U_i(\widehat{\psi}),$$

$$\widehat{M}_i(t) = N_i(t) - \sum_{\ell=1}^L \int_0^t Y_i(s) 1_{\{S_i=\ell\}}\, d\widehat{\Lambda}_\ell(s, \widehat{\psi}),$$

and $\widehat{I}$ and $\widehat{V}$ are the empirical counterparts of I and V, obtained by replacing all expectations in (4.3) and (4.4) by the corresponding sample means. Note that $\widehat{M}_i(t)$ can be calculated only for individuals with observed stratum. This is not a problem since in $\widehat{\varphi}_i$, we only need to calculate $\widehat{M}_i(t)$ for individuals with $R_i = 1$.

From Theorem 4.2, a decision rule for testing H_0 against H_1 can be constructed as follows: we reject H_0 at the asymptotic level α if $\mathcal{T} := n^{-1}(\widetilde{\mathbb{Z}}(\widehat{\psi})/\widetilde{\sigma})^2 \geq \chi^2_{1-\alpha}(1)$, where $\chi^2_{1-\alpha}(1)$ is the quantile of order $1-\alpha$ of the χ^2 distribution with one degree of freedom.

4.3 Numerical results

4.3.1 Simulation study: design and results

In this section, we conduct a simulation study to investigate finite sample performance of our test statistic under various conditions. This simulation study is car-

ried out using the statistical software R [51]. We consider 2 groups and 3 strata. In each group and stratum, failure times T^0 are simulated from a Weibull distribution $\mathcal{W}(\alpha, \lambda)$ with hazard function $\alpha(t) = \alpha\lambda t^{\alpha-1}$. Failure times in the ℓ-th stratum ($\ell = 1, 2, 3$) of groups 1 and 2 are simulated from $\mathcal{W}(\alpha_\ell, \lambda_\ell)$ and $\mathcal{W}(\alpha_\ell, \lambda_\ell \cdot r)$ respectively, where r denotes the hazard ratio of groups 1 and 2. We use $(\alpha_1, \lambda_1) = (2.1, 1)$, $(\alpha_2, \lambda_2) = (1.2, 0.75)$, $(\alpha_3, \lambda_3) = (1.8, 1.5)$ and consider three different values for the hazard ratio: (a) $r = 1$, (b) $r = 2.12$ and (c) $r = 2.72$. Case (a) corresponds to H_0 (no difference between groups within each stratum) while cases (b) and (c) provide increasing magnitude of difference between groups. Censoring times are simulated from an exponential distribution whose parameter is chosen to yield an average of 20% (first case) and 50% (second case) of censored individuals in the simulated samples. Let n_ℓ denote the sample size in stratum ℓ ($\ell = 1, 2, 3$) and $n = \sum_{\ell=1}^{3} n_\ell$ be the total sample size. We consider successively the following situations: $(n_1, n_2, n_3) = (40, 40, 40)$, $(n_1, n_2, n_3) = (70, 80, 60)$ and $(n_1, n_2, n_3) = (100, 100, 80)$. In a first scenario, approximately 30% (respectively 70%) of the individuals belong to group 1 (respectively group 2). In a second scenario, each group includes approximately half of the individuals. The binary indicator R is generated from a logistic regression model $\Pi(\psi) = \mathbb{P}(R = 1|\mathcal{O}) = \text{logit}^{-1}(\psi_1 + \psi_2 T + \psi_3 \delta + \psi_4 Z + \psi_5 G)$, where Z is an auxiliary covariate simulated from a uniform distribution on $[0, 1]$. For each simulation scenario, the parameter $\psi = (\psi_1, \psi_2, \psi_3, \psi_4, \psi_5)^\top$ is chosen to yield an average proportion of missing data equal to 70%. Design parameters of this simulation study are summarized in Table 4.1.

Parameter	Values	Description
(n_1, n_2, n_3)	$(40, 40, 40); (70, 80, 60); (100, 100, 80)$	Sample size
r	$1; 2.12; 2.72$	Groups hazard ratio
c	20%; 50%	Censoring percentage
$(p, 1-p)$	(0.3, 0.7); (0.5,0.5)	Proportions of individuals in groups 1 and 2

Tableau 4.1: Design parameters and values included in simulations.

For each configuration of the design parameters, we simulate 1000 samples. Based on them, we obtain the empirical size (case (a)) and power (cases (b) and (c)) of the proposed test at the significance level 0.05. For comparison, we report results of the stratified logrank test based on complete cases (that is, on individuals with observed stratum). In what follows, we denote this test by SLR_{CC}. We also calculate the stratified logrank test by using all stratum indicators, as if none were missing. In what follows, we denote this test by SLR_{FD} (for "full data"). Indeed, this test provides a natural benchmark for evaluating performance of our test. All calculations are carried out on a multicore processor server equipped with Ubuntu Linux

operating system. Parallel execution of our `R` code is provided by `R` package `doMC` (see `https://CRAN.R-project.org/package=doMC`). Table 4.2 summarizes the results.

From these results, it appears that powers of both complete-case and inverse-probability-weighted stratified logrank tests increase when sample size increases and censoring decreases. The empirical level of $\mathcal{T}$ tends to exceed (but only slightly) the nominal level. As expected, in cases (b) and (c), the inverse-probability-weighted stratified logrank has greater power than the complete-case test, for every scenario. Overall, the proposed test appears to improve on its complete-case aternative.

4.3.2 A real data example

In this section, we apply our test statistic to a medical dataset described by [2]. In these data, the survival time is time to esophageal cancer and we wish to investigate the effect of smoking habit on the risk of developing cancer, while adjusting for exposure to zinc (zinc deficiency is a contributing factor to the development of esophageal squamous cell carcinoma in humans). The dataset contains 166 male patients. 37 of them developed cancer, the others were censored at the end of the study period. 102 out of 166 smoked regularly for at least 6 months, the other 64 did not. Tissue zinc concentration is recorded as a four-categories factor variable, indicating the level of concentration.

Let $\alpha_{1,\ell}(\cdot)$ (respectively $\alpha_{2,\ell}(\cdot)$) be the hazard function of cancer incidence among smoker (respectively non-smoker) patients in the ℓ-th stratum, where indice ℓ refers to the level class of zinc concentration (zinc concentration increases with ℓ). Testing the effect of the smoking status while adjusting for zinc exposure can be expressed as the problem of testing $H_0 : \alpha_{1,\ell}(\cdot) = \alpha_{2,\ell}(\cdot)$ for $\ell = 1, 2, 3, 4$ against $H_1 : \alpha_{1,\ell}(\cdot) \neq \alpha_{2,\ell}(\cdot)$ for some $\ell \in \{1, 2, 3, 4\}$. The stratified logrank test provides an appropriate tool for answering this question. However, measurement of zinc concentration requires an esophageal biopsy, which is obtained through a costly measurement technique. Therefore, it was too difficult and expensive to measure the element concentration on every patient and it is only available for 53 subjects, leaving 113 patients with missing stratum information (that is, approximately 68% of the patients). Thus, we propose to test H_0 by using our inverse-probability-weighted stratified logrank statistic.

First, we fit a logistic regression model including all available variables to the missingness indicators R_i, $i = 1, \ldots, n$. This includes the survival time T, censoring status δ, group indicator G ($G_i = 0$ if patient i regularly smoked for at least 6 months and $G_i = 1$ otherwise), a binary indicator of the drinking status (any alcohol in the preceding 12 months: yes/no), age (a continuous variable) and dysplasia grade (mild,

moderate or severe). We use a stepwise procedure to select a smaller set of relevant predictors. The final set contains T, δ, age and dysplasia grade. Then, we fit a logistic regression including these four predictors only to the R_i, $i = 1, \ldots, n$, which yields the maximum likelihood estimate $\widehat{\psi}$, finally used to calculate the test statistic $\widetilde{\mathbb{Z}}(\widehat{\psi})$ with (4.2). The final value of the test statistic is $\mathcal{T} = 1.836$, with p-value equal to 0.175. Therefore, at the level 5%, we do reject the null hypothesis of within-strata equality of the distributions of cancer incidence in smokers and non-smokers groups. A complete-case stratified logrank test calculated on the 53 patients with observed zinc concentration yields the p-value 0.762. In this example, both tests give the same conclusion. However, the much smaller p-value obtained from our test may be the hint that smoking status is less non-significant than indicated by the complete-case logrank. Our approach uses all data information available from the 166 patients (including those with missing stratum), thus, we may expect that it provides a more realistic picture of the influence of risk factors on cancer incidence.

4.4 Conclusion

In this paper, we propose an inverse-probability-weighted stratified logrank test statistic for comparing survival groups with missing stratum information. We derive its asymptotic distribution under the null hypothesis of no difference between groups. This test is based on an inverse-probability-weighted version of Nelson-Aalen estimator of cumulative hazards. In our simulations, we found that the proposed test outperforms complete-case analysis. Now, several questions deserve attention.

First, we assume that the probability of observing a complete case can be modeled parametrically. If this model is incorrectly specified, it is known that inverse-probability-weighting methods can be biased. One solution is to estimate $\Pi(\cdot)$ nonparametrically (by using some kernel estimate for example). An alternative is to use augmented inverse-probability-weighting of complete cases, which is robust to misspecification of the missingness model see [65]. Both alternatives constitute stimulating topics for our future research. Then, in Theorem 4.1, we establish pointwise asymptotic normality of our modified Nelson-Aalen estimator. A uniform weak convergence result would allow us to construct simultaneous confidence bands for within-strata cumulative hazards. Tighness of the inverse-probability-weighted Nelson-Aalen estimator, viewed as a process indexed by t, is required. This is also a subject for our future work. Finally, in this paper, we assume that strata are missing-at-random (MAR). Assessing robustness of the proposed test to deviation to MAR constitutes an other interesting topic for future work.

Appendix A. Calculation details for $S_n(\psi)$ and $I(\psi)$.

The maximum likelihood estimator $\widehat{\psi}$ of ψ in model $\Pi(\psi) = \mathbb{P}(R = 1|\mathcal{O})$ can be obtained by solving the score equation $S_n(\psi) = \frac{\partial \log L_n(\psi)}{\partial \psi} = 0$, and

$$\begin{aligned}
S_n(\psi) &= \frac{\partial}{\partial \psi}\left[\sum_{i=1}^{n}\{R_i \log \Pi_i(\psi) + (1-R_i)\log(1-\Pi_i(\psi))\}\right], \\
&= \sum_{i=1}^{n}\left\{R_i \frac{\dot{\Pi}_i(\psi)}{\Pi_i(\psi)} - (1-R_i)\frac{\dot{\Pi}_i(\psi)}{1-\Pi_i(\psi)}\right\}, \\
&= \sum_{i=1}^{n}\frac{(R_i - \Pi_i(\psi))\dot{\Pi}_i(\psi)}{\Pi_i(\psi)(1-\Pi_i(\psi))}, \\
&= \sum_{i=1}^{n} U_i(\psi).
\end{aligned}$$

Moreover,

$$\begin{aligned}
\frac{\partial S_n(\psi)}{\partial \psi^\top} &= \frac{\partial}{\partial \psi^\top}\sum_{i=1}^{n} U_i(\psi), \\
&= \sum_{i=1}^{n}\frac{\left(\ddot{\Pi}_i(\psi)\,(R_i - \Pi_i(\psi)) - \dot{\Pi}_i(\psi)\dot{\Pi}_i(\psi)^\top\right)(\Pi_i(\psi)(1-\Pi_i(\psi)))}{(\Pi_i(\psi)(1-\Pi_i(\psi)))^2} \\
&\qquad - \frac{\dot{\Pi}_i(\psi)\,(R_i - \Pi_i(\psi))\left(\dot{\Pi}_i(\psi)^\top(1-\Pi_i(\psi)) - \Pi_i(\psi)\dot{\Pi}_i(\psi)^\top\right)}{(\Pi_i(\psi)(1-\Pi_i(\psi)))^2}, \\
&= \sum_{i=1}^{n}\left(\frac{\ddot{\Pi}_i(\psi)\,(R_i - \Pi_i(\psi))}{\Pi_i(\psi)(1-\Pi_i(\psi))} - \frac{\dot{\Pi}_i(\psi)\dot{\Pi}_i(\psi)^\top\,(R_i - \Pi_i(\psi))^2}{(\Pi_i(\psi)(1-\Pi_i(\psi)))^2}\right), \\
&= \sum_{i=1}^{n}\left(\frac{\ddot{\Pi}_i(\psi)\,(R_i - \Pi_i(\psi))}{\Pi_i(\psi)(1-\Pi_i(\psi))} - U_i(\psi)U_i(\psi)^\top\right).
\end{aligned}$$

Finally, we remark that as $n \to \infty$, $-\frac{1}{n}\frac{\partial S_n(\psi)}{\partial \psi^\top}$ converges in probability to

$$I(\psi) = \mathbb{E}\left[U(\psi)U(\psi)^\top - \frac{\ddot{\Pi}(\psi)(R - \Pi(\psi))}{\Pi(\psi)(1-\Pi(\psi))}\right].$$

Appendix B. Proof of Theorem 4.1.

We decompose $n^{\frac{1}{2}}(\widehat{\Lambda}_\ell(t,\widehat{\psi}) - \Lambda_{0,\ell}(t))$ as

$$n^{\frac{1}{2}}(\widehat{\Lambda}_\ell(t,\widehat{\psi}) - \Lambda_{0,\ell}(t)) = A_1 + A_2,$$

where $A_1 := n^{\frac{1}{2}}(\widehat{\Lambda}_\ell(t,\widehat{\psi}) - \widehat{\Lambda}_\ell(t,\psi_0))$ and $A_2 := n^{\frac{1}{2}}(\widehat{\Lambda}_\ell(t,\psi_0) - \Lambda_{0,\ell}(t))$. The term A_1 can be rewritten as:

$$A_1 = n^{-\frac{1}{2}} \sum_{i=1}^{n} \int_0^t \frac{[\omega_{\ell,i}(\widehat{\psi}) - \omega_{\ell,i}(\psi_0)]}{n^{-1}\widetilde{S}_{\ell,n}^{(0)}(u,\widehat{\psi})}\, dN_i(u)$$
$$+ n^{-\frac{1}{2}} \sum_{i=1}^{n} \int_0^t \omega_{\ell,i}(\psi_0) \left[\frac{1}{n^{-1}\widetilde{S}_{\ell,n}^{(0)}(u,\widehat{\psi})} - \frac{1}{n^{-1}\widetilde{S}_{\ell,n}^{(0)}(u,\psi_0)} \right] dN_i(u).$$

Now, under the assumption that $\widehat{\psi}$ is a consistent and asymptotically normal estimator of ψ_0, Taylor series expansions of $\omega_{\ell,i}(\widehat{\psi})$ and $1/(n^{-1}\widetilde{S}_{\ell,n}^{(0)}(u,\widehat{\psi}))$ around ψ_0 and some straightforward algebra yield:

$$\begin{aligned} A_1 &= n^{-1} \sum_{i=1}^{n} \int_0^t \left\{ \frac{\dot{\omega}_{\ell,i}(\psi_0)^\top}{n^{-1}\widetilde{S}_{\ell,n}^{(0)}(u,\widehat{\psi})} - \frac{\omega_{\ell,i}(\psi_0) n^{-1}\sum_{j=1}^{n} Y_j(u)\dot{\omega}_{\ell,j}(\psi_0)^\top}{(n^{-1}\widetilde{S}_{\ell,n}^{(0)}(u,\psi_0))^2} \right\} dN_i(u) \cdot n^{\frac{1}{2}}(\widehat{\psi} - \psi_0) \\ &\quad + o_{\mathbb{P}}(1), \\ &:= W_{\ell,n}(t,\widehat{\psi}) \cdot n^{\frac{1}{2}}(\widehat{\psi} - \psi_0) + o_{\mathbb{P}}(1). \end{aligned}$$

Consistency of $\widehat{\psi}$ and the law of large numbers imply that $W_{\ell,n}(t,\widehat{\psi})$ converges in probability to $W_\ell(t,\psi_0)$ given by (4.5). Therefore, $A_1 = W_\ell(t,\psi_0) \cdot n^{\frac{1}{2}}(\widehat{\psi} - \psi_0) + o_{\mathbb{P}}(1)$. Now, expanding $S_n(\widehat{\psi})$ in a neighbourhood of ψ_0 yields

$$\begin{aligned} n^{\frac{1}{2}}(\widehat{\psi} - \psi_0) &= I(\psi_0)^{-1} n^{-\frac{1}{2}} S_n(\psi_0) + o_{\mathbb{P}}(1), \\ &= I(\psi_0)^{-1} n^{-\frac{1}{2}} \sum_{i=1}^{n} U_i(\psi_0) + o_{\mathbb{P}}(1), \end{aligned}$$

and thus,

$$A_1 = n^{-\frac{1}{2}} \sum_{i=1}^{n} W_\ell(t,\psi_0) I(\psi_0)^{-1} U_i(\psi_0) + o_{\mathbb{P}}(1).$$

Next, we consider term A_2. Note first that

$$\begin{aligned}
\sum_{i=1}^{n}\int_0^t \frac{\omega_{\ell,i}(\psi_0)}{\widetilde{S}^{(0)}_{\ell,n}(u,\psi_0)}\,dM_i(u) &= n^{-1}\sum_{i=1}^{n}\int_0^t \frac{\omega_{\ell,i}(\psi_0)}{n^{-1}\widetilde{S}^{(0)}_{\ell,n}(u,\psi_0)}\left(dN_i(u)-Y_i(u)\sum_{\ell'=1}^{L}\alpha_{0,\ell'}(u)1_{\{S_i=\ell'\}}\,du\right),\\
&= \widehat{\Lambda}_\ell(t,\psi_0)-n^{-1}\sum_{i=1}^{n}\int_0^t \frac{\frac{1_{\{S_i=\ell\}}R_i}{\Pi_i(\psi_0)}}{n^{-1}\widetilde{S}^{(0)}_{\ell,n}(u,\psi_0)}Y_i(u)\sum_{\ell'=1}^{L}\alpha_{0,\ell'}(u)1_{\{S_i=\ell'\}}\,du,\\
&= \widehat{\Lambda}_\ell(t,\psi_0)-\int_0^t \frac{n^{-1}\sum_{i=1}^{n}Y_i(u)\frac{1_{\{S_i=\ell\}}R_i}{\Pi_i(\psi_0)}}{n^{-1}\widetilde{S}^{(0)}_{\ell,n}(u,\psi_0)}\alpha_{0,\ell}(u)\,du,\\
&= \widehat{\Lambda}_\ell(t,\psi_0)-\int_0^t \alpha_{0,\ell}(u)\,du,\\
&= \widehat{\Lambda}_\ell(t,\psi_0)-\Lambda_{0,\ell}(t).
\end{aligned}$$

Thus, we have:

$$\begin{aligned}
A_2 &= n^{\frac{1}{2}}(\widehat{\Lambda}_\ell(t,\psi_0)-\Lambda_{0,\ell}(t)),\\
&= n^{-\frac{1}{2}}\sum_{i=1}^{n}\int_0^t \frac{\omega_{\ell,i}(\psi_0)}{n^{-1}\widetilde{S}^{(0)}_{\ell,n}(u,\psi_0)}\,dM_i(u),\\
&= n^{-\frac{1}{2}}\sum_{i=1}^{n}\int_0^t \frac{\omega_{\ell,i}(\psi_0)}{\widetilde{s}^{(0)}_{\ell}(u,\psi_0)}\,dM_i(u)+o_{\mathbb{P}}(1),
\end{aligned}$$

and finally,

$$\begin{aligned}
n^{\frac{1}{2}}(\widehat{\Lambda}_\ell(t,\widehat{\psi})-\Lambda_{0,\ell}(t)) &= n^{-\frac{1}{2}}\sum_{i=1}^{n}\left(W_\ell(t,\psi_0)I(\psi_0)^{-1}U_i(\psi_0)+\int_0^t \frac{\omega_{\ell,i}(\psi_0)}{\widetilde{s}^{(0)}_{\ell}(u,\psi_0)}\,dM_i(u)\right)+o_{\mathbb{P}}(1),\\
&:= n^{-\frac{1}{2}}\sum_{i=1}^{n}\xi_{\ell,i}(t,\psi_0)+o_{\mathbb{P}}(1),
\end{aligned}$$

with $\mathbb{E}[\xi_{\ell,i}(t,\psi_0)]=0$. The result follows by applying the central limit theorem. □

Appendix C. Proof of Theorem 4.2.

We first prove some intermediate technical lemmas.

Lemma 4.1 *Under H_0, it holds that:*

$$\sum_{i=1}^{n}\int_0^\tau \sum_{\ell=1}^{L}\frac{1_{\{S_i=\ell\}}R_i}{\Pi_i(\psi_0)}\left[1_{\{G_i=1\}}-\widetilde{E}_{\ell,n}(t,\psi_0)\right]\left(Y_i(t)\sum_{\ell'=1}^{L}\alpha_{0,\ell'}(t)1_{\{S_i=\ell'\}}\,dt\right)=0.$$

Proof. We have:

$$\begin{aligned}
&\sum_{i=1}^{n}\int_0^{\tau}\sum_{\ell=1}^{L}\frac{1_{\{S_i=\ell\}}R_i}{\Pi_i(\psi_0)}\left[1_{\{G_i=1\}}-\widetilde{E}_{\ell,n}(t,\psi_0)\right]\left(Y_i(t)\sum_{\ell'=1}^{L}\alpha_{0,\ell'}(t)1_{\{S_i=\ell'\}}\,dt\right)\\
&\quad=\sum_{i=1}^{n}\int_0^{\tau}\sum_{\ell=1}^{L}\frac{1_{\{S_i=\ell\}}R_i}{\Pi_i(\psi_0)}\left[1_{\{G_i=1\}}-\widetilde{E}_{\ell,n}(t,\psi_0)\right]Y_i(t)\alpha_{0,\ell}(t)\,dt\\
&\quad=\int_0^{\tau}\sum_{\ell=1}^{L}\left(\sum_{i=1}^{n}\frac{1_{\{S_i=\ell\}}R_i}{\Pi_i(\psi_0)}1_{\{G_i=1\}}Y_i(t)-\sum_{i=1}^{n}\frac{1_{\{S_i=\ell\}}R_i}{\Pi_i(\psi_0)}Y_i(t)\frac{\widetilde{S}^{(1)}_{\ell,n}(t,\psi_0)}{\widetilde{S}^{(0)}_{\ell,n}(t,\psi_0)}\right)\alpha_{0,\ell}(t)\,dt\\
&\quad=\int_0^{\tau}\sum_{\ell=1}^{L}\left(\widetilde{S}^{(1)}_{\ell,n}(t,\psi_0)-\widetilde{S}^{(0)}_{\ell,n}(t,\psi_0)\frac{\widetilde{S}^{(1)}_{\ell,n}(t,\psi_0)}{\widetilde{S}^{(0)}_{\ell,n}(t,\psi_0)}\right)\alpha_{0,\ell}(t)\,dt\\
&\quad=0.
\end{aligned}$$

□

Lemma 4.2 *Assume that conditions 1-4 holds. Then*

$$\int_0^{\tau}\sum_{\ell=1}^{L}\left(\widetilde{e}_{\ell}(t,\psi_0)-\widetilde{E}_{\ell,n}(t,\psi_0)\right)d\left(n^{-\frac{1}{2}}\sum_{i=1}^{n}\frac{1_{\{S_i=\ell\}}R_i}{\Pi_i(\psi_0)}M_i(t)\right)\xrightarrow{\mathbb{P}}0\quad \textit{as } n\longrightarrow\infty.$$

Proof. Let $\ell\in\{1,\dots,L\}$. By Lemma 4.1 of [35], $\{Y(t):t\in[0,\tau]\}$ is Donsker. The classes $\{\psi\in\Psi\}$, $\{1_{\{S=\ell\}}\}$, $\{1_{\{G=1\}}\}$, $\{R\}$ and $\{\mathcal{O}\}$ are Donsker and so is the class $\{\psi^{\top}\mathcal{O}:\psi\in\Psi\}$ since products of bounded Donsker classes are Donsker. Now, the class $\{\Pi(\psi):\psi\in\Psi\}$ is Donsker since Π is Lipschitz continuous by condition 2 (Lipschitz continuous functions of Donsker classes are Donsker) and condition 2 implies that $\{1/\Pi(\psi):\psi\in\Psi\}$ is also Donsker. Finally, $\{Y(t)\frac{1_{\{S=\ell\}}R}{\Pi(\psi)}:t\in[0,\tau],\psi\in\Psi\}$ and $\{Y(t)\frac{1_{\{S=\ell\}}R}{\Pi(\psi)}1_{\{G=1\}}:t\in[0,\tau],\psi\in\Psi\}$ are Donsker since they are products of bounded Donsker classes. As Donsker classes, they are also Glivenko-Cantelli classes and since $\widetilde{s}^{(0)}_{\ell}(t,\psi)$ is bounded away from zero (by condition 3), we have that

$$\sup_{\psi\in\Psi,t\in[0,\tau]}\left|\widetilde{E}_{\ell,n}(t,\psi)-\widetilde{e}_{\ell}(t,\psi)\right|\xrightarrow{\mathbb{P}}0\quad \text{as } n\longrightarrow\infty.$$

It follows that $\sup_{t\in[0,\tau]}\left|\widetilde{E}_{\ell,n}(t,\psi_0)-\widetilde{e}_{\ell}(t,\psi_0)\right|\xrightarrow{\mathbb{P}}0$ as $n\longrightarrow\infty$.
Now, let $Q_n(t)=n^{-\frac{1}{2}}\sum_{i=1}^{n}\frac{1_{\{S_i=\ell\}}R_i}{\Pi_i(\psi_0)}M_i(t)$. By similar arguments as above, Q_n converges weakly to a mean zero process.
It follows from Lemma 4.2 of [35] that

$$\int_0^{\tau}\left(\widetilde{e}_{\ell}(t,\psi_0)-\widetilde{E}_{\ell,n}(t,\psi_0)\right)d\left(n^{-\frac{1}{2}}\sum_{i=1}^{n}\frac{1_{\{S_i=\ell\}}R_i}{\Pi_i(\psi_0)}M_i(t)\right)\xrightarrow{\mathbb{P}}0\quad \text{as } n\longrightarrow\infty,$$

which concludes the proof. □

The following lemma establishes a useful approximation of $n^{-\frac{1}{2}}\widetilde{\mathbb{Z}}(\widehat{\psi})$.

Lemma 4.3 *Assume that conditions 1-4 holds. Then, under H_0,*

$$n^{-\frac{1}{2}}\widetilde{\mathbb{Z}}(\widehat{\psi}) = n^{-\frac{1}{2}}\sum_{i=1}^{n}\int_0^{\tau}\sum_{\ell=1}^{L}\frac{1_{\{S_i=\ell\}}R_i}{\Pi_i(\psi_0)}\left[1_{\{G_i=1\}} - \widetilde{e}_\ell(t,\psi_0)\right]dM_i(t) + V(\psi_0)\sqrt{n}(\widehat{\psi}-\psi_0) + o_{\mathbb{P}}(1).$$

Proof. A Taylor expansion of $\widetilde{\mathbb{Z}}(\widehat{\psi})$ around ψ_0 yields

$$n^{-\frac{1}{2}}\widetilde{\mathbb{Z}}(\widehat{\psi}) = n^{-\frac{1}{2}}\widetilde{\mathbb{Z}}(\psi_0) + n^{-1}\left.\frac{\partial\widetilde{\mathbb{Z}}(\psi)}{\partial\psi^\top}\right|_{\psi=\psi_0}\sqrt{n}(\widehat{\psi}-\psi_0) + o_{\mathbb{P}}(1).$$

Then, some tedious albeit not difficult algebra yields $n^{-1}\partial\widetilde{\mathbb{Z}}(\psi)/\partial\psi^\top = V(\psi) + o_{\mathbb{P}}(1)$, where $V(\psi)$ is given by (4.4). It follows that

$$n^{-\frac{1}{2}}\widetilde{\mathbb{Z}}(\widehat{\psi}) = n^{-\frac{1}{2}}\widetilde{\mathbb{Z}}(\psi_0) + V(\psi_0)\sqrt{n}(\widehat{\psi}-\psi_0) + o_{\mathbb{P}}(1), \tag{4.6}$$

since under a correctly specified parametric model $\Pi(\psi)$, $\sqrt{n}(\widehat{\psi}-\psi_0) = O_{\mathbb{P}}(1)$. Now, by Lemma 4.1, we have that

$$\begin{aligned}
n^{-\frac{1}{2}}\widetilde{\mathbb{Z}}(\psi_0) &= n^{-\frac{1}{2}}\sum_{i=1}^{n}\int_0^{\tau}\sum_{\ell=1}^{L}\frac{1_{\{S_i=\ell\}}R_i}{\Pi_i(\psi_0)}\left[1_{\{G_i=1\}} - \widetilde{E}_{\ell,n}(t,\psi_0)\right]dN_i(t),\\
&= n^{-\frac{1}{2}}\sum_{i=1}^{n}\int_0^{\tau}\sum_{\ell=1}^{L}\frac{1_{\{S_i=\ell\}}R_i}{\Pi_i(\psi_0)}\left[1_{\{G_i=1\}} - \widetilde{E}_{\ell,n}(t,\psi_0)\right]\left(dN_i(t) - Y_i(t)\sum_{\ell'=1}^{L}\alpha_{0,\ell'}(t)1_{\{S_i=\ell'\}}\,dt\right),\\
&= n^{-\frac{1}{2}}\sum_{i=1}^{n}\int_0^{\tau}\sum_{\ell=1}^{L}\frac{1_{\{S_i=\ell\}}R_i}{\Pi_i(\psi_0)}\left[1_{\{G_i=1\}} - \widetilde{E}_{\ell,n}(t,\psi_0)\right]dM_i(t).
\end{aligned}$$

Then,

$$n^{-\frac{1}{2}}\widetilde{\mathbb{Z}}(\psi_0) = n^{-\frac{1}{2}}\sum_{i=1}^{n}\int_0^{\tau}\sum_{\ell=1}^{L}\frac{1_{\{S_i=\ell\}}R_i}{\Pi_i(\psi_0)}\left[1_{\{G_i=1\}} - \widetilde{e}_\ell(t,\psi_0) + \widetilde{e}_\ell(t,\psi_0) - \widetilde{E}_{\ell,n}(t,\psi_0)\right]dM_i(t),$$

and by Lemma 4.2,

$$n^{-\frac{1}{2}}\widetilde{\mathbb{Z}}(\psi_0) = n^{-\frac{1}{2}}\sum_{i=1}^{n}\int_0^{\tau}\sum_{\ell=1}^{L}\frac{1_{\{S_i=\ell\}}R_i}{\Pi_i(\psi_0)}\left[1_{\{G_i=1\}} - \widetilde{e}_\ell(t,\psi_0)\right]dM_i(t) + o_{\mathbb{P}}(1). \tag{4.7}$$

Finally, it follows from (4.6) and (4.7) that

$$n^{-\frac{1}{2}}\widetilde{\mathbb{Z}}(\widehat{\psi}) = n^{-\frac{1}{2}}\sum_{i=1}^{n}\int_0^{\tau}\sum_{\ell=1}^{L}\frac{1_{\{S_i=\ell\}}R_i}{\Pi_i(\psi_0)}\left[1_{\{G_i=1\}} - \widetilde{e}_\ell(t,\psi_0)\right]dM_i(t) + V(\psi_0)\sqrt{n}(\widehat{\psi}-\psi_0) + o_{\mathbb{P}}(1).$$

□

We now turn to proof of Theorem 4.2. A Taylor expansion of $S_n(\widehat{\psi})$ around ψ_0 yields

$$0 = S_n(\widehat{\psi}) = S_n(\psi_0) + \left.\frac{\partial S_n(\psi)}{\partial \psi^\top}\right|_{\psi=\psi_0} (\widehat{\psi} - \psi_0) + o_{\mathbb{P}}(\|\widehat{\psi} - \psi_0\|).$$

Thus, under a correctly specified model $\Pi(\psi)$, we have:

$$\widehat{\psi} - \psi_0 = \frac{1}{n}\sum_{i=1}^{n} I(\psi_0)^{-1} U_i(\psi_0) + o_{\mathbb{P}}(n^{-\frac{1}{2}}). \tag{4.8}$$

By Lemma 4.3 and (4.8), we obtain:

$$\begin{aligned} n^{-\frac{1}{2}}\widetilde{\mathbb{Z}}(\widehat{\psi}) &= n^{-\frac{1}{2}}\sum_{i=1}^{n}\left(\int_0^\tau \sum_{\ell=1}^{L} \frac{1_{\{S_i=\ell\}}R_i}{\Pi_i(\psi_0)}\left[1_{\{G_i=1\}} - \widetilde{e}_\ell(t,\psi_0)\right] dM_i(t) + V(\psi_0)I(\psi_0)^{-1}U_i(\psi_0)\right) \\ &\quad + o_{\mathbb{P}}(1), \\ &= n^{-\frac{1}{2}}\sum_{i=1}^{n}\varphi_i + o_{\mathbb{P}}(1), \end{aligned}$$

with $\mathbb{E}[\varphi_i] = 0$. Hence, by the central limit theorem, $n^{-\frac{1}{2}}\widetilde{\mathbb{Z}}(\widehat{\psi})$ converges in distribution to a zero-mean Gaussian random variable with variance $\mathbb{E}[\varphi_i^2]$. □

				SLR_{FD}	SLR_{CC}	$\mathcal{T}$
(n_1, n_2, n_3)	c	case	p			
$(40, 40, 40)$	20%	(a)	0.3	0.055	0.051	0.060
			0.5	0.056	0.059	0.070
		(b)	0.3	0.893	0.273	0.338
			0.5	0.931	0.257	0.326
		(c)	0.3	0.988	0.392	0.468
			0.5	0.995	0.412	0.485
	50%	(a)	0.3	0.048	0.061	0.072
			0.5	0.056	0.042	0.070
		(b)	0.3	0.709	0.172	0.196
			0.5	0.761	0.173	0.250
		(c)	0.3	0.915	0.323	0.327
			0.5	0.947	0.306	0.382
$(70, 80, 60)$	20%	(a)	0.3	0.054	0.059	0.070
			0.5	0.050	0.058	0.050
		(b)	0.3	0.989	0.427	0.450
			0.5	0.997	0.438	0.480
		(c)	0.3	1	0.659	0.684
			0.5	1	0.686	0.711
	50%	(a)	0.3	0.034	0.049	0.054
			0.5	0.048	0.052	0.061
		(b)	0.3	0.929	0.345	0.359
			0.5	0.952	0.329	0.382
		(c)	0.3	0.996	0.537	0.543
			0.5	0.997	0.548	0.582
$(100, 100, 80)$	20%	(a)	0.3	0.049	0.070	0.075
			0.5	0.052	0.044	0.072
		(b)	0.3	0.997	0.572	0.610
			0.5	1	0.583	0.634
		(c)	0.3	1	0.808	0.819
			0.5	1	0.819	0.827
	50%	(a)	0.3	0.047	0.037	0.059
			0.5	0.061	0.055	0.058
		(b)	0.3	0.974	0.410	0.438
			0.5	0.988	0.441	0.474
		(c)	0.3	0.999	0.667	0.689
			0.5	1	0.683	0.698

Tableau 4.2: Empirical size (case (a)) and power (cases (b) and (c)) for SLR_{FD}, SLR_{CC} and the proposed test $\mathcal{T}$ (all results are based on 1000 replications).

Conclusion et perspectives

Dans ce travail, nous nous intéressons à l'inférence statistique dans deux modèles semi-paramétrique et non-paramétrique stratifiés de durées de vie censurées. Nous proposons tout d'abord une statistique de test d'ajustement pour le modèle de régression stratifié à risques proportionnels. Nous établissons sa distribution asymptotique sous l'hypothèse nulle d'un ajustement correct du modèle aux données. Nous étudions les propriétés numériques de ce test (niveau, puissance sous différentes alternatives) au moyen de simulations. Nous proposons ensuite une procédure permettant de stratifier le modèle à risques proportionnels suivant un seuil inconnu d'une variable de stratification. Cette procédure repose sur l'utilisation du test d'ajutement proposé précédemment. Une étude de simulation exhaustive est conduite pour évaluer les performances de cette procédure.

Notre test d'ajustement est réalisable lorsque le nombre de covariables est modéré, donc il est envisageable d'explorer certaines directions pour étendre ce nombre et résoudre les problèmes de calcul.

Dans une seconde partie de notre travail, nous nous intéressons à l'application du test du logrank stratifié dans un contexte de données manquantes (nous considérons la situation où les strates ne peuvent êtres observées chez tous les individus de l'échantillon). Nous construisons une version pondérée du logrank stratifié adaptée à ce problème. Nous en établissons la loi limite sous l'hypothèse nulle d'égalité des fonctions de risque dans les différents groupes. Les propriété de cette nouvelle statistique de test sont évaluées au moyen de simulations. Le test est ensuite appliqué à un jeu de données médicales.

Il serait intéressant d'étendre ce travail dans plusieures directions. Tout d'abord, nous pourrions nous intéresser à la sensibilité du test proposé à une mauvaise spécification des pondérations individuelles. Puis nous pourrions envisager des modélisations alternatives de ces pondérations, faisant intervenir des modèles plus flexibles et plus robustes: modèles logistiques semi ou non paramétriques. Nous pourrions également nous intéresser à la sensibilité du test proposé à un écart à l'hypothèse de données manquantes au hasard. L'ensemble de ces pistes de recherche nécessite la mise en oeuvre d'études de simulations poussées, qui sont le sujet de nos travaux futurs.

A

Programmes des simulations

Code de calcul du test d'ajustement au modèle de Cox stratifié lorsque sous l'alternative, les risques instantanés dans les différentes strates contiennent des prédicteurs linéaires différents

```
    library(MASS)
library(survival)

require(doMC, quietly = TRUE)
registerDoMC()
nw=getDoParWorkers()
nbrerep=1e5 # nombre d'échantillons simulés

    ###### INITIALISATIONS DIVERSES
Q=400 # nombre de z_i
J=3 # nombre de strates
n1=150 # effectifs des strates
n2=175
n3=120

nj=c(n1,n2,n3)
n=sum(nj)
m=0.525 # .237 assure 10% de censure, .525 assure 20% de censure, 1.325 assure 40%
de censure

a1=2.1 #alpha1
a2=1.2 #alpha2
a3=1.8 #alpha3
```

```
l1=1 # lambda ds strate 1
l2=0.75 # lambda ds strate 2
l3=1.5 # lambda ds strate 3
b11=0.2 # paramètre de régression beta11
b12=0.7 # paramètre de régression beta12
b21=1 # paramètre de régression beta21
b22=1 # paramètre de régression beta22
b31=0.2 # paramètre de régression beta31
b32=0.2 # paramètre de régression beta32

    ###### FIN INITIALISATIONS
res=foreach(u=1:nbrerep, .combine=rbind) %dopar%
N=1
essai1=matrix(rep(0,N*Q),ncol=Q)
essai2=matrix(rep(0,N*Q),ncol=Q)
essai3=matrix(rep(0,N*Q),ncol=Q)
S1=rep(0,N)
S2=rep(0,N)
S3=rep(0,N)
Z1=rnorm(n1)
W1=runif(n1,min=1,max=3)
X1=(-log(runif(n1))/l1*exp(-b11*Z1-b12*W1))^(1/a1) #durées de vie de S1
Z2=rnorm(n2) #covariables de strate S2
W2=runif(n2,min=1,max=3) #covariables
X2=(-log(runif(n2))/l2*exp(-b21*Z2-b22*W2))^(1/a2) #durées de vie de S2
Z3=rnorm(n3) #covariables
W3=runif(n3,min=1,max=3) #covariables
X3=(-log(runif(n3))/l3*exp(-b31*Z3-b32*W3))^(1/a3) #durées de vie de S3
X=c(X1,X2,X3)
Z=c(Z1,Z2,Z3)
W=c(W1,W2,W3)
S=rep(1:J,nj) #indicatrice d'appartenance aux strates
C=rexp(n,m) # censure
T=pmin(X,C) #durées observée
delta=as.integer(X==T)
u=seq(-1.6,1.6,length=Q)
v=seq(1.2,2.8,length=Q)
Q1=matrix(rep(1,Q),ncol=Q)
```

```
Q2=matrix(rep(1,Q),ncol=Q)
Q3=matrix(rep(1,Q),ncol=Q)

fit=coxph(Surv(T,delta)Z̃+W+strata(S)) # ajustement modele Cox stratifié

mr=resid(fit) # récupération résidus de martingale

for(l in 1:Q){
Q1[l]=1/sqrt(n)*sum(mr*(Z<=u[l])*(W<=v[l])*(S==1))
Q2[l]=1/sqrt(n)*sum(mr*(Z<=u[l])*(W<=v[l])*(S==2))
Q3[l]=1/sqrt(n)*sum(mr*(Z<=u[l])*(W<=v[l])*(S==3))
}

essai1=Q1
essai2=Q2
essai3=Q3
resultat=c(essai1,essai2,essai3)
essai1=res[,1:Q]
essai2=res[,(Q+1):(2*Q)]
essai3=res[,(2*Q+1):(3*Q)]

    # une simulation pour injection dans bootstrap
N=1
S1=rep(0,N)
S2=rep(0,N)
S3=rep(0,N)
Z1=rnorm(n1)
W1=runif(n1,min=1,max=3)
X1=(-log(runif(n1))/l1*exp(-b11*Z1-b12*W1))̂(1/a1)
Z2=rnorm(n2)
W2=runif(n2,min=1,max=3)
X2=(-log(runif(n2))/l2*exp(-b21*Z2-b22*W2))̂(1/a2)
Z3=rnorm(n3)
W3=runif(n3,min=1,max=3)
X3=(-log(runif(n3))/l3*exp(-b31*Z3-b32*W3))̂(1/a3)
X=c(X1,X2,X3)
Z=c(Z1,Z2,Z3)
W=c(W1,W2,W3)
```

```
S=rep(1:J,nj)
C=rexp(n,m)
T=pmin(X,C)
delta=as.integer(X==T) # crée les indicatrices delta
data=cbind(T,Z,W,S,delta)

    # replications bootstrap
B=2000
matricebootQ1=matrix(rep(1,B*Q),ncol=B)
matricebootQ2=matrix(rep(1,B*Q),ncol=B)
matricebootQ3=matrix(rep(1,B*Q),ncol=B)
for (b in 1:B)
print(b)
Q1b=matrix(rep(0,Q),ncol=1)
Q2b=matrix(rep(0,Q),ncol=1)
Q3b=matrix(rep(0,Q),ncol=1)
echboot=data[sample(1:nrow(data),nrow(data),replace=TRUE),]
fitboot=coxph(Surv(echboot[,1],echboot[,5]) echboot[,2]+echboot[,3]+strata(echboot[,4]))
mr=resid(fitboot)
u=seq(-1.6,1.6,length=Q)
v=seq(1.2,2.8,length=Q)
for(l in 1:Q)
Q1b[l]=1/sqrt(n)*sum(mr*(echboot[,2]<=u[l])*(echboot[,3]<=v[l])*(echboot[,4]==1))
Q2b[l]=1/sqrt(n)*sum(mr*(echboot[,2]<=u[l])*(echboot[,3]<=v[l])*(echboot[,4]==2))
Q3b[l]=1/sqrt(n)*sum(mr*(echboot[,2]<=u[l])*(echboot[,3]<=v[l])*(echboot[,4]==3))

################### fin boucle sur l
matricebootQ1[,b]=Q1b
matricebootQ2[,b]=Q2b
matricebootQ3[,b]=Q3b
matricebootQ123=rbind(matricebootQ1,matricebootQ2,matricebootQ3)
vbQ123=var(t(matricebootQ123)) # matrice S

    ############ fin boucle bootstrap
vbQ123temp=vbQ123
vbQ123=matrix(rep(0,(J*Q)2̂),ncol=J*Q)
vbQ123[1:Q,1:Q]=vbQ123temp[1:Q,1:Q]
vbQ123[(Q+1):(2*Q),(Q+1):(2*Q)]=vbQ123temp[(Q+1):(2*Q),(Q+1):(2*Q)]
```

```
vbQ123[(2*Q+1):(3*Q),(2*Q+1):(3*Q)]=vbQ123temp[(2*Q+1):(3*Q),(2*Q+1):(3*Q)]
stat.test=apply(cbind(apply(abs(essai1),1,max),apply(abs(essai2),1,max),apply(abs(essai3),1,max)),
# test max-max
loi.theo.boot=abs(mvrnorm(n=5000,rep(0,J*Q),vbQ123))
comp1=apply(loi.theo.boot[,1:Q],1,max)
comp2=apply(loi.theo.boot[,(Q+1):(2*Q)],1,max)
comp3=apply(loi.theo.boot[,(2*Q+1):(3*Q)],1,max)
stat.theo.boot=pmax(comp1,comp2,comp3) # loi "theorique bootstrappé" et sa densité
estimée pour le test max-max
rm(list=(ls()[ls()!="stat.test"&ls()!="stat.theo.boot"]))
print(mean(stat.test>=quantile(stat.theo.boot,probs=.95)))
save(list = ls(all=TRUE), file = "sortieDiffPL20-150-175-120.RData");
```

Code de calcul du test du logrank stratifié avec strates manquantes

```
#############fonction réalisant le calcul de la statistique de test
stat.test=function(R,S,T,delta,G,Z)
W=cbind(rep(1,n),T,delta,G,Z)
m.logit=glm(R T+delta+G+Z,family=binomial(link=logit))
p.hat=predict(m.logit,newdata=data.frame(W[,-1]),type='response') # récupération des
probabilités estimées P(R=1 | W)

############ calcul de la statistique de test
M1=matrix(as.integer(t(replicate(n,T))>=replicate(n,T)),ncol=n)
M2=matrix(t(replicate(n,as.integer(G==1))),ncol=n)
l=1
M3l1=matrix(t(replicate(n,as.integer(S==l))),ncol=n)
l=2
M3l2=matrix(t(replicate(n,as.integer(S==l))),ncol=n)
l=3
M3l3=matrix(t(replicate(n,as.integer(S==l))),ncol=n)
M4=matrix(t(replicate(n,R)),ncol=n)
M5=matrix(t(replicate(n,as.vector(p.hat))),ncol=n)
M6l1=M1*M2*M3l1*M4/M5
M6l2=M1*M2*M3l2*M4/M5
M6l3=M1*M2*M3l3*M4/M5
# # récupération des S̃^(1)_{l,n}(T_i, ψ̂) pour i=1 à n et l=1,2,3
# S1l1=apply(M6l1,1,sum)
S1l2=apply(M6l2,1,sum)
S1l3=apply(M6l3,1,sum)
# # # M6l1sansG=M1*M3l1*M4/M5
M6l2sansG=M1*M3l2*M4/M5
M6l3sansG=M1*M3l3*M4/M5
#
# récupération des S̃^(0)_{l,n}(T_i, ψ̂) pour i=1 à n et l=1,2,3
#
S0l1=apply(M6l1sansG,1,sum)
S0l2=apply(M6l2sansG,1,sum)
S0l3=apply(M6l3sansG,1,sum)
#
# récupération des Ẽ^(0)_{l,n}(T_i, ψ̂) pour i=1 à n et l=1,2,3
```

```
#
El1=S1l1/S0l1
El2=S1l2/S0l2
El3=S1l3/S0l3
indNaN=which(is.na(El1)|is.na(El2)|is.na(El3))
Z1=sum(as.integer(S[-indNaN]==1)*R[-indNaN]/p.hat[-indNaN]*(as.integer(G[-indNaN]==1)-
El1[-indNaN])*delta[-indNaN])
Z2=sum(as.integer(S[-indNaN]==2)*R[-indNaN]/p.hat[-indNaN]*(as.integer(G[-indNaN]==1)-
El2[-indNaN])*delta[-indNaN])
Z3=sum(as.integer(S[-indNaN]==3)*R[-indNaN]/p.hat[-indNaN]*(as.integer(G[-indNaN]==1)-
El3[-indNaN])*delta[-indNaN])
statistique.test=(Z1+Z2+Z3)/sqrt(n-length(indNaN))
```

Détermination du seuil de stratification dans le modèle de Cox stratifié

```
nbrerep=1e3 # nombre d'échantillons simulés

    ############ INITIALISATIONS DIVERSES

Q=600 # nombre de z_i
J=2 # nombre de strates
n1=100 # effectifs des strates 200, 225, 190
n2=125
nj=c(n1,n2)
n=sum(nj)
m=1 # m=.36 assure 10% de censure / 1 assure 20% de censure / 2.15 assure 30% de
censure / 3.9 assure 40% de censure
ws=1.5 # threshold for stratifying
a1=1
a2=1
l1=3
l2=1
b=c(1.3,.75,-0.5,0.25,0) # paramètre de régression beta1

    ############ FIN INITIALISATIONS

res=foreach(u=1:nbrerep, .combine=rbind) %dopar%
print(u)
# début simulation des données
C1=rnorm(n)
C2=runif(n,min=1,max=3)
C3=rbinom(n,1,.4)
C4=rnorm(n,1,1.5)
C5=runif(n,0,1)
Cov=rbind(C1,C2,C3,C4,C5)
W=runif(n,0,4)
X1=(-log(runif(n))/l1*exp(-t(b)%*%Cov))^(1/a1)
X2=(-log(runif(n))/l2*exp(-t(b)%*%Cov))^(1/a2)
X=X1*(W<=ws)+X2*(W>ws)
C=rexp(n,m)
```

```
T=pmin(X,C)
delta=as.integer(X==T)
A=t(rbind(T,delta,C1,C2,C3,C4,C5,W))
# fin simulation des données

w=seq(0,4,length=Q)
v=seq(-3,8,length=Q)
Q1=matrix(rep(1,(Q*Q)),ncol=Q)
Q2=Q1
for (j in 1:Q){
fit=coxph(Surv(A[,1],A[,2]) A[,3]+A[,4]+A[,5]+A[,6]+A[,7]+strata(W<=w[j]))
mr=resid(fit) # récupération résidus de martingale
#
for(l in 1:Q){
Q1[l,j]=1/sqrt(n)*sum(mr*(t(fit$coeff)%*%Cov<=v[l])*(W<=w[j]))
Q2[l,j]=1/sqrt(n)*sum(mr*(t(fit$coeff)%*%Cov<=v[l])*(W>w[j]))
}
#
}
est.ws=w[which.min(apply(pmax(abs(Q1),abs(Q2)),2,max))]
# estimation du modèle de Cox avec threshold estimé
fit2=coxph(Surv(A[,1],A[,2]) A[,3]+A[,4]+A[,5]+A[,6]+A[,7]+strata(W<=est.ws))
# oracle: estimation du modèle de Cox avec vrai threshold
fitoracle=coxph(Surv(A[,1],A[,2]) A[,3]+A[,4]+A[,5]+A[,6]+A[,7]+strata(W<=ws))
#
# estimation bootstrap de la variance des estimateurs
#
B=1000
coeff=matrix(rep(0,B*dim(Cov)[1]),ncol=B)
for(b in 1:B)
Aboot=A[sample(1:n,n,replace=TRUE),]
fitboot=coxph(Surv(Aboot[,1],Aboot[,2]) Aboot[,3]+Aboot[,4]+Aboot[,5]+Aboot[,6]+Aboot[,7]+strata(A
coeff[,b]=fitboot$coeff
sdestboot=apply(coeff,1,sd)
#
resultat=rbind(c(est.ws,fit2$coeff,sdestboot,fitoracle$coeff,sqrt(diag(fitoracle$var))))
save(list = ls(all=TRUE), file = "stratification20pc2.RData");
```

B

Résultats des Simulations $(n_1, n_2) = (200, 225)$

		proposed method					oracle				
c		β_1	β_2	β_3	β_4	β_5	β_1	β_2	β_3	β_4	β_5
10%	RMSE	0.0130	0.0119	0.0140	0.0016	0.0362	0.0064	0.0088	0.0118	0.0013	0.0355
	bias	-0.0640	-0.0407	0.0263	-0.0125	0.0036	0.0151	0.0068	-0.0051	0.0027	0.0027
	SE	0.0763	0.0988	0.1133	0.0378	0.1885	0.0760	0.0962	0.1098	0.0369	0.1823
	SD	0.0945	0.1012	0.1154	0.0381	0.1902	0.0783	0.0937	0.1084	0.0362	0.1886
	CP	0.7850	0.9140	0.9400	0.9460	0.9460	0.9490	0.9490	0.9450	0.9590	0.9480
	ℓ	0.2992	0.3872	0.4441	0.1483	0.7388	0.2981	0.3773	0.4303	0.1445	0.7147
20%	RMSE	0.0125	0.0128	0.0156	0.0018	0.0403	0.0063	0.0101	0.0145	0.0016	0.0378
	bias	-0.0612	-0.0354	0.0222	-0.0109	0.0066	0.0131	0.0076	-0.0063	0.0032	0.0059
	SE	0.0798	0.1046	0.1200	0.0400	0.1991	0.0791	0.1015	0.1163	0.0388	0.1926
	SD	0.0933	0.1074	0.1230	0.0410	0.2008	0.0786	0.1001	0.1204	0.0395	0.1945
	CP	0.8250	0.9280	0.9420	0.9360	0.9400	0.9530	0.9620	0.9560	0.9460	0.9480
	ℓ	0.3129	0.4100	0.4704	0.1569	0.7805	0.3099	0.3980	0.4558	0.1522	0.7550
40%	RMSE	0.0123	0.0167	0.0203	0.0022	0.0567	0.0084	0.0144	0.0187	0.0020	0.0521
	bias	-0.0481	-0.0305	0.0204	-0.0074	-0.0049	0.0215	0.0110	-0.0086	0.0062	-0.0019
	SE	0.0902	0.1211	0.1402	0.0464	0.2323	0.0884	0.1167	0.1353	0.0447	0.2236
	SD	0.1000	0.1256	0.1411	0.0461	0.2381	0.0893	0.1196	0.1364	0.0443	0.2284
	CP	0.8770	0.9260	0.9480	0.9440	0.9460	0.9460	0.9440	0.9400	0.9460	0.9460
	ℓ	0.3534	0.4745	0.5496	0.1817	0.9107	0.3465	0.4576	0.5304	0.1752	0.8766

Table 2. Simulation results with $(n_1, n_2) = (200, 225)$. RMSE: empirical root mean square error. SD: empirical standard deviation. CP: empirical coverage probability of 95%-level confidence intervals. ℓ: average length of 95%-level confidence intervals.

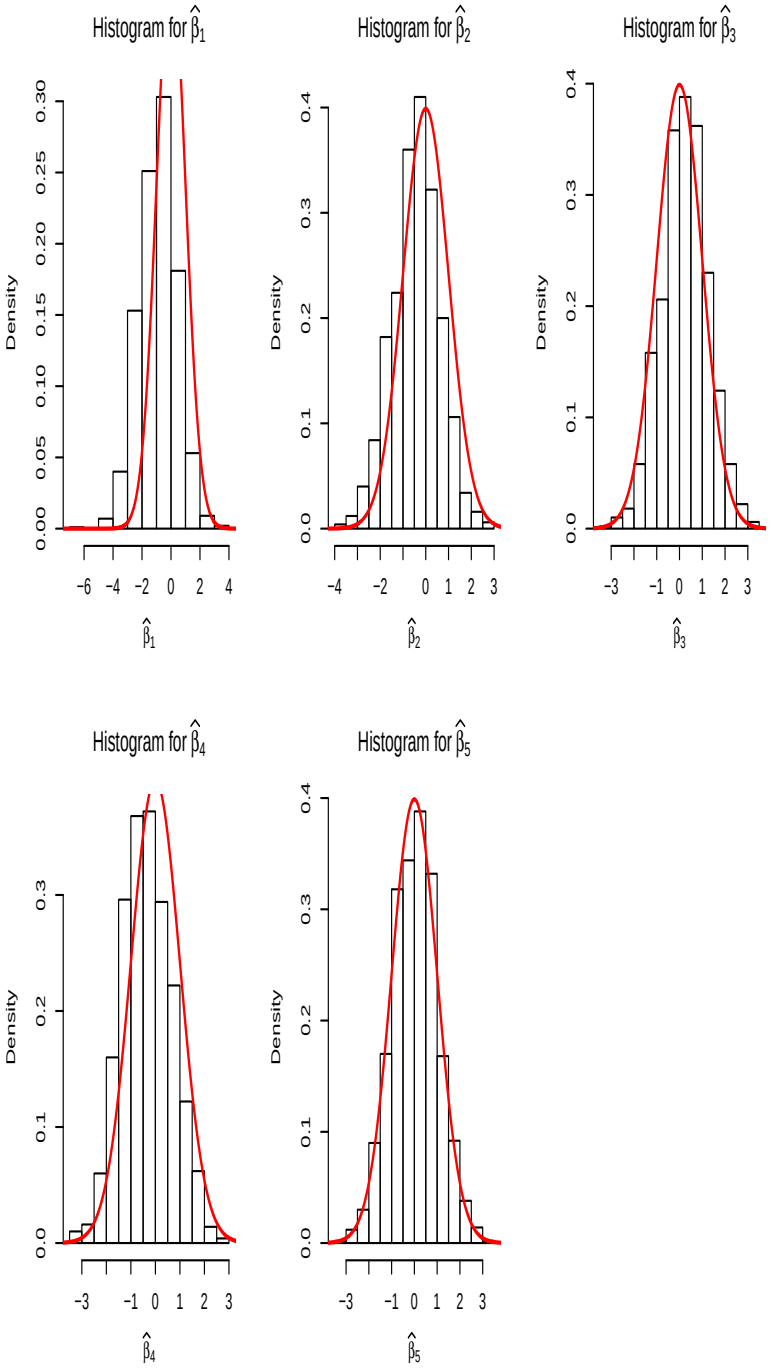

FIGURE B.1: Histogram of the N normalized estimates $\widehat{\beta}^{(1)}_{n,\hat{\omega},j}, \ldots, \widehat{\beta}^{(N)}_{n,\hat{\omega},j}$, $j = 1, \ldots, 5$, $(n_1, n_2) = (200, 225)$ and censoring percentage = 10% . In red: density of the $\mathcal{N}(0, 1)$.

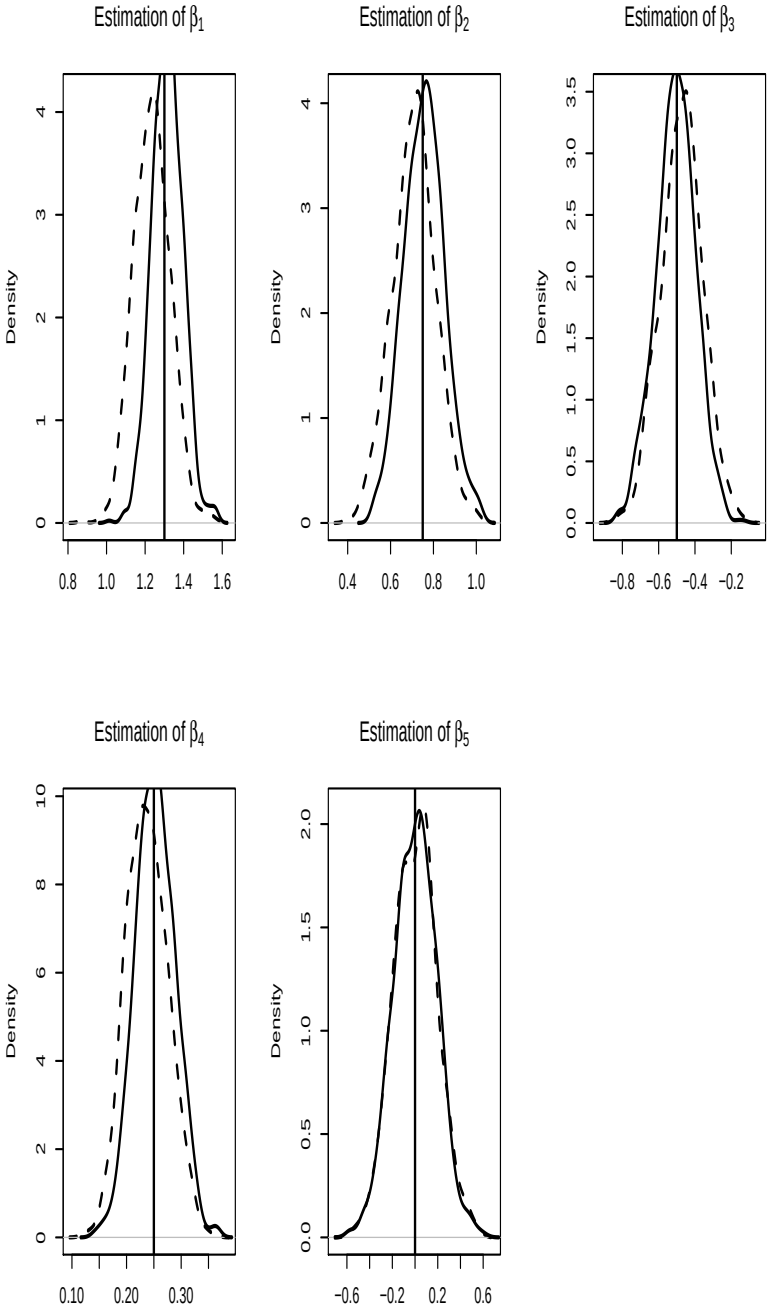

FIGURE B.2: Density estimation from the N two-step estimates (dashed line) and oracle estimates (solid line), $(n_1, n_2) = (200, 225)$ and censoring percentage = 10% . The true value is indicated by a vertical line.

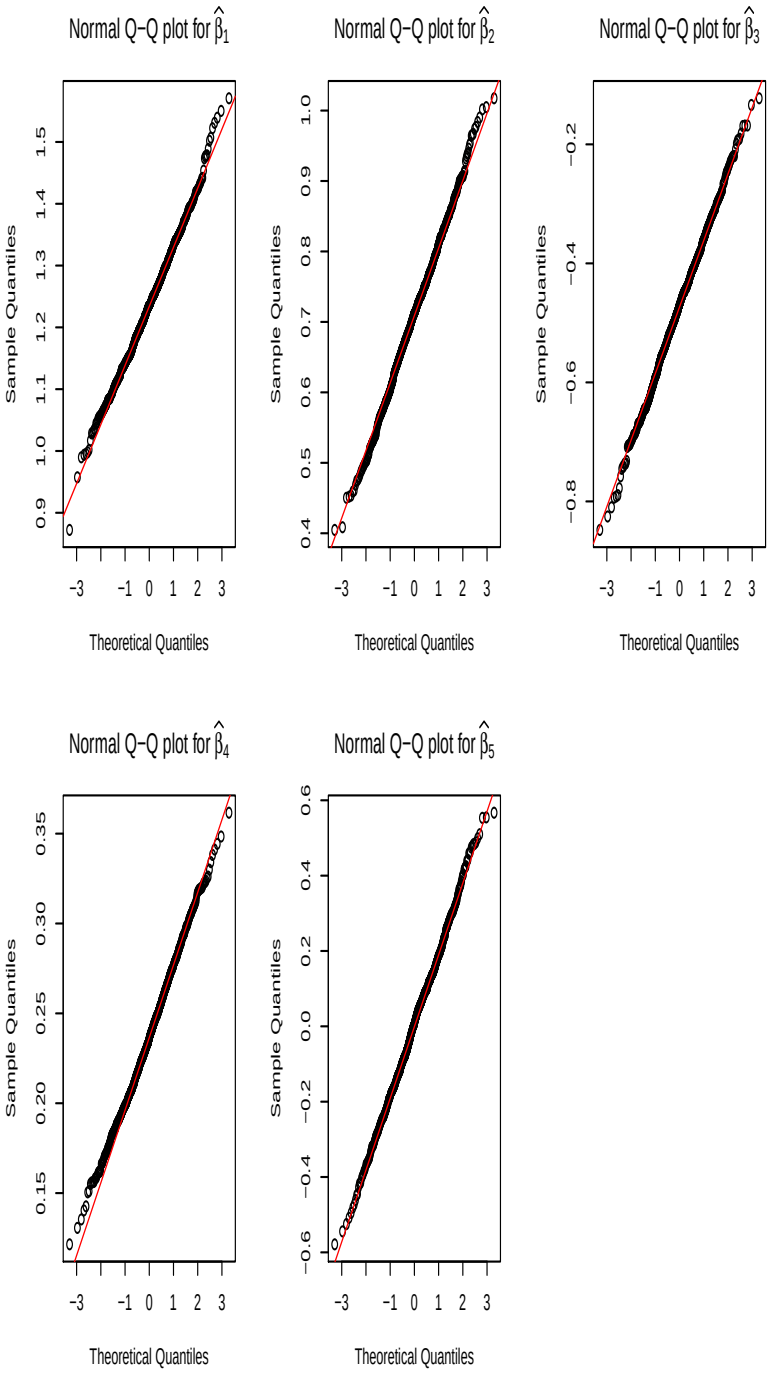

FIGURE B.3: QQ-plots of the N estimates $\widehat{\beta}^{(1)}_{n,\hat{\omega},j}, \ldots, \widehat{\beta}^{(N)}_{n,\hat{\omega},j}$, $j = 1, \ldots, 5$, $(n_1, n_2) = (200, 225)$ and censoring percentage = 10%.

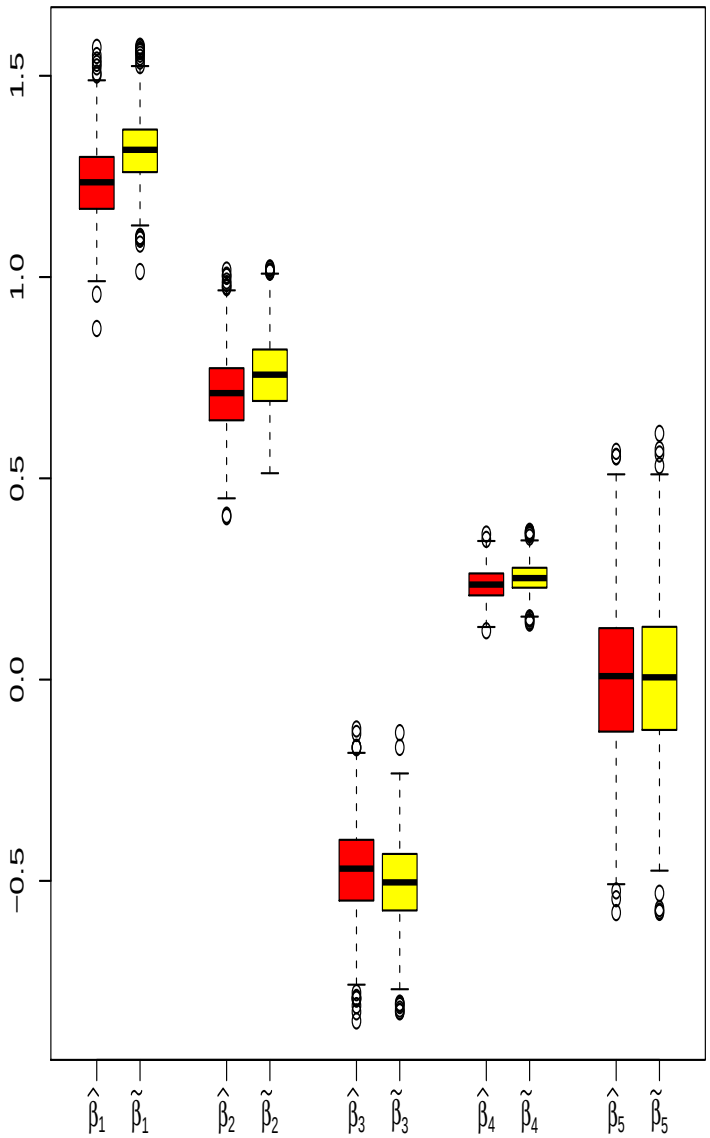

FIGURE B.4: Boxplots-plots of the N **two-step estimates** $\widehat{\beta}^{(1)}_{n,\hat{\omega},j}, \ldots, \widehat{\beta}^{(N)}_{n,\hat{\omega},j}$ and **oracle estimates** $\widetilde{\beta}^{(1)}_{n,j}, \ldots, \widetilde{\beta}^{(N)}_{n,j}$, $j = 1, \ldots, 5$, $(n_1, n_2) = (200, 225)$ and censoring percentage = 10%

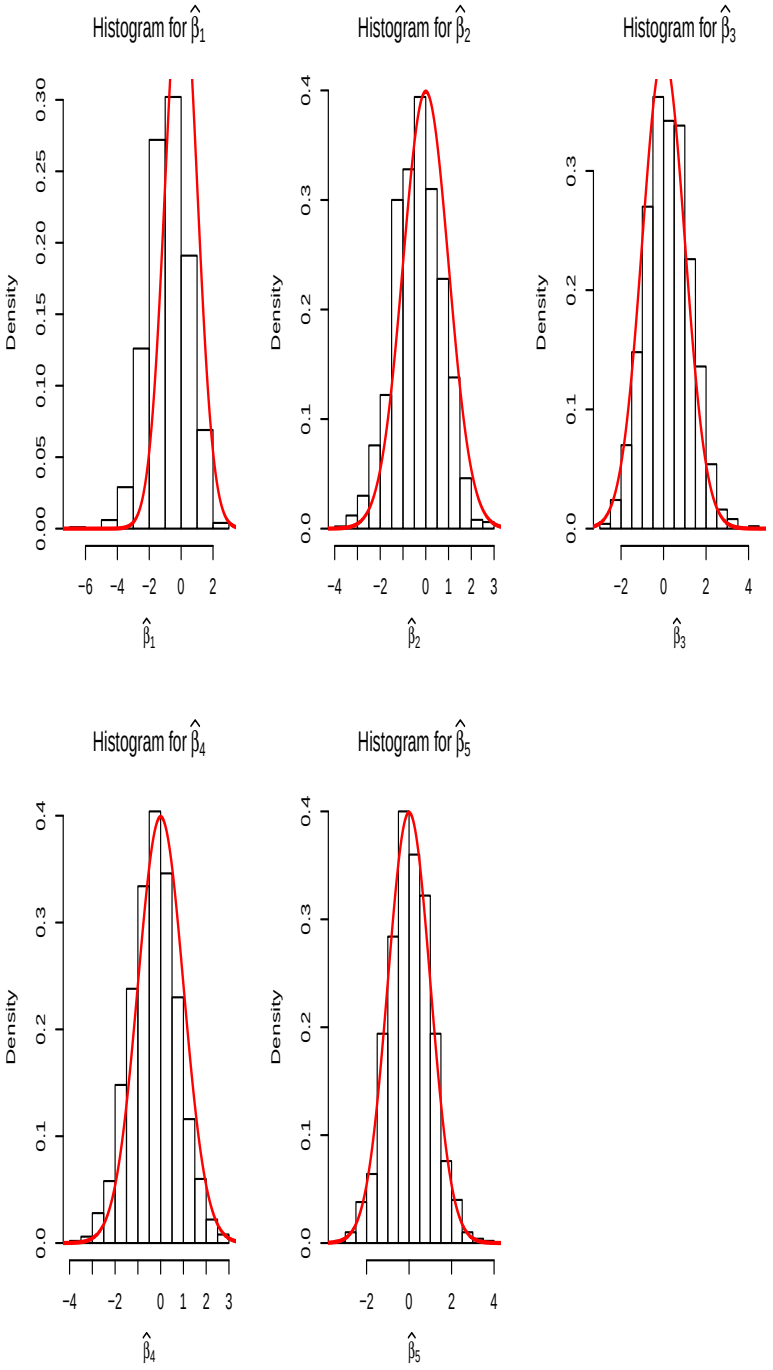

FIGURE B.5: Histogram of the N normalized estimates $\widehat{\beta}^{(1)}_{n,\hat{\omega},j}, \ldots, \widehat{\beta}^{(N)}_{n,\hat{\omega},j}$, $j = 1, \ldots, 5$, $(n_1, n_2) = (200, 225)$ and censoring percentage = 20% . In red: density of the $\mathcal{N}(0, 1)$.

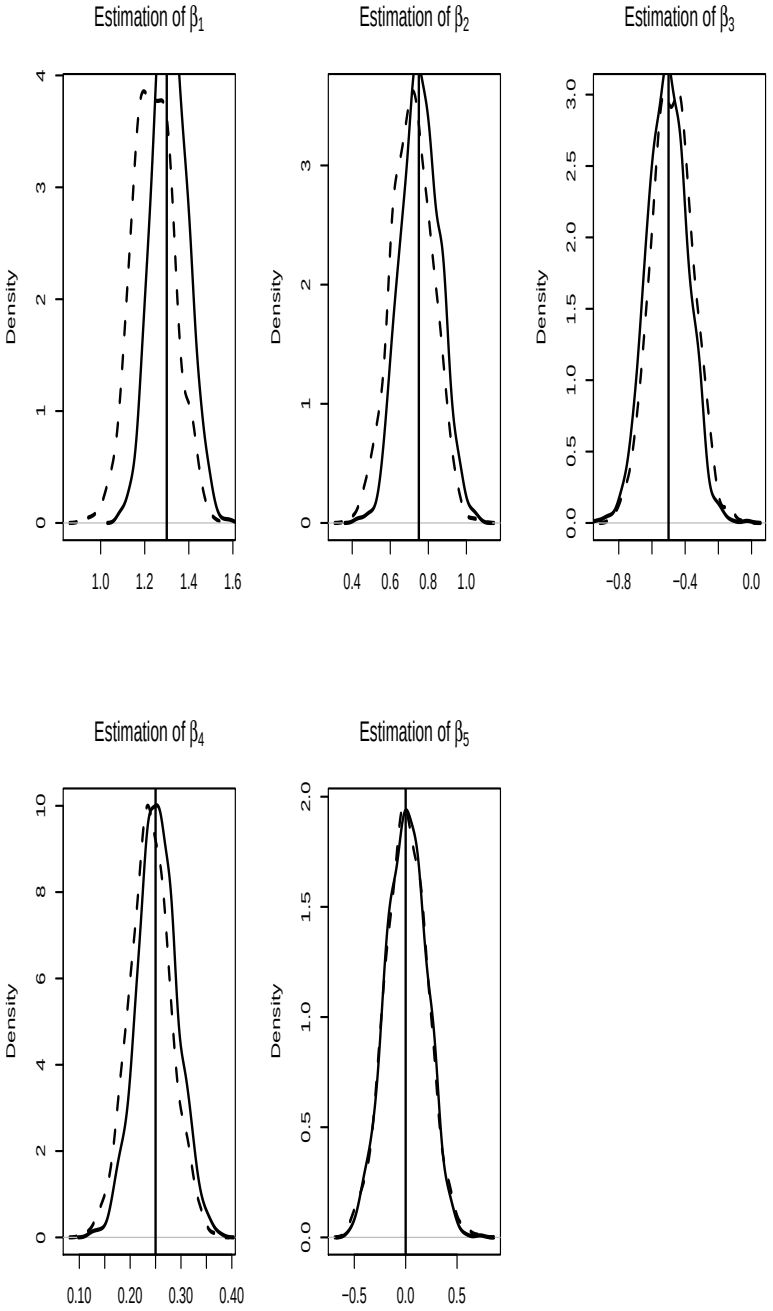

FIGURE B.6: Density estimation from the N two-step estimates (dashed line) and oracle estimates (solid line), $(n_1, n_2) = (200, 225)$ and censoring percentage = 20% . The true value is indicated by a vertical line.

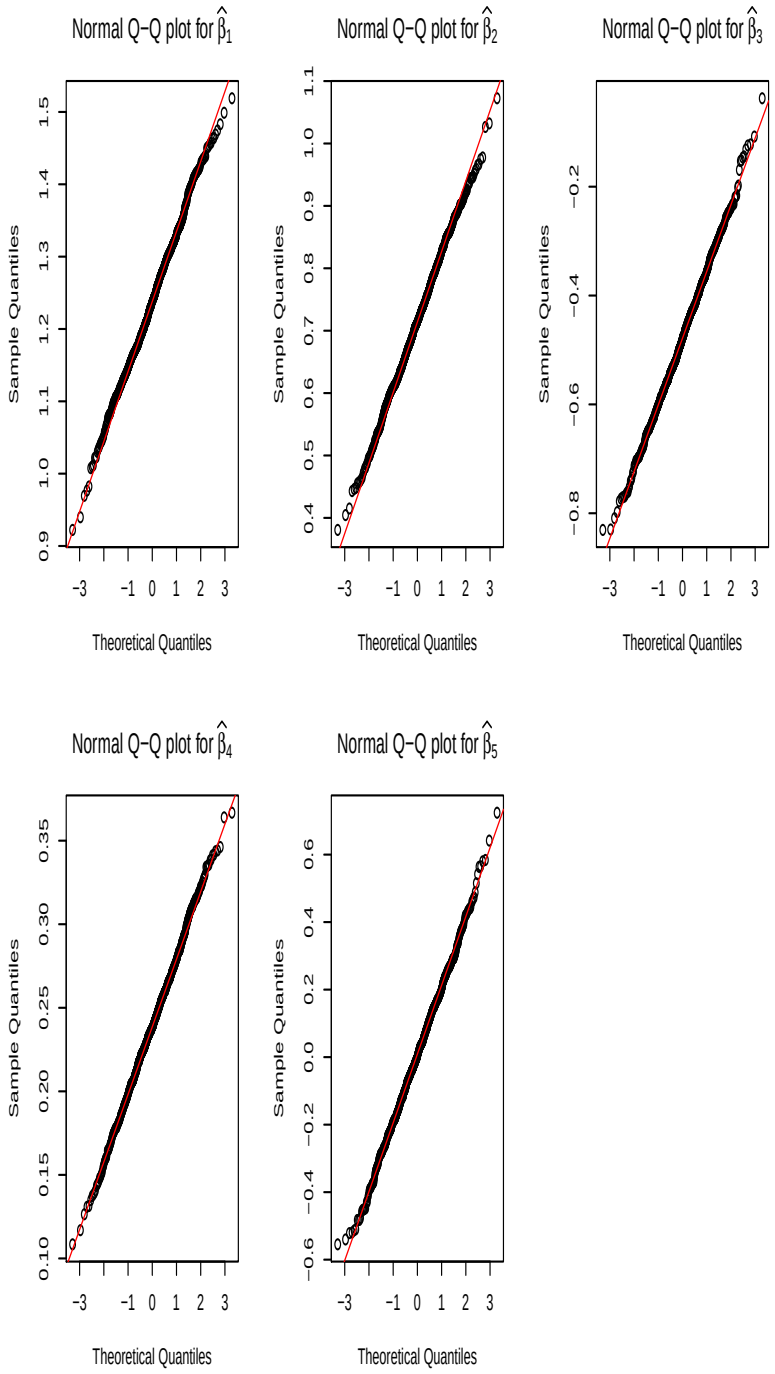

FIGURE B.7: QQ-plots of the N estimates $\widehat{\beta}^{(1)}_{n,\hat{\omega},j}, \ldots, \widehat{\beta}^{(N)}_{n,\hat{\omega},j}$, $j = 1, \ldots, 5$, $(n_1, n_2) = (200, 225)$ and censoring percentage = 20%.

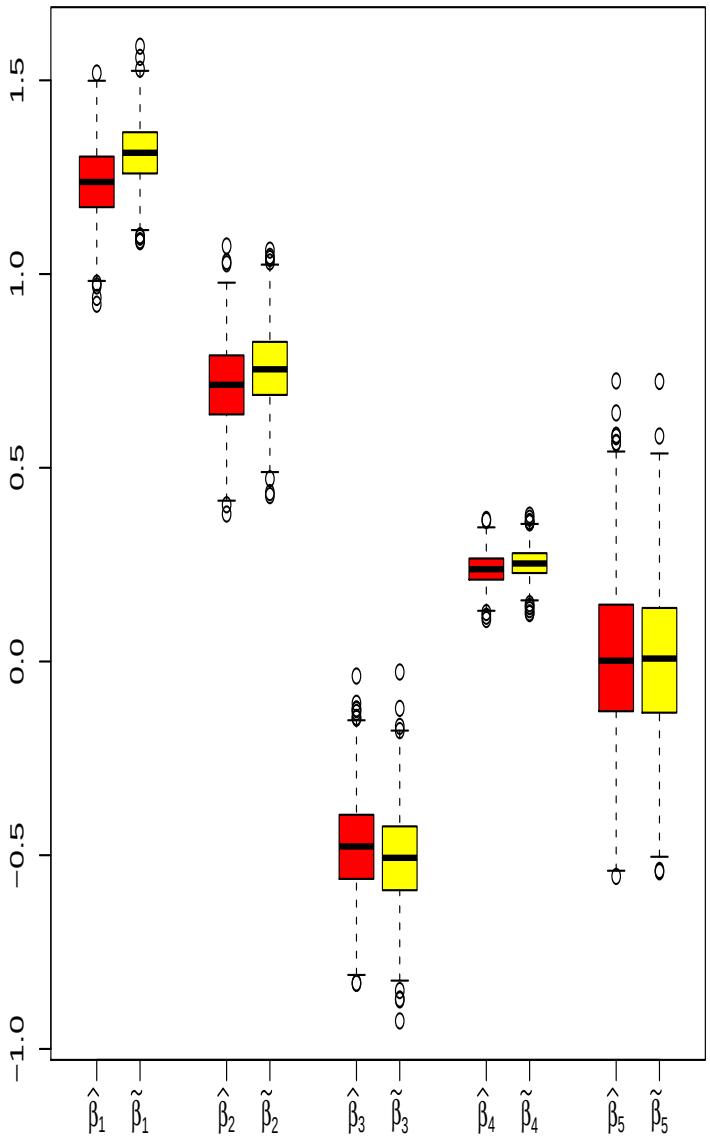

FIGURE B.8: Boxplots-plots of the N **two-step estimates** $\widehat{\beta}^{(1)}_{n,\hat{\omega},j}, \ldots, \widehat{\beta}^{(N)}_{n,\hat{\omega},j}$ and **oracle estimates** $\widetilde{\beta}^{(1)}_{n,j}, \ldots, \widetilde{\beta}^{(N)}_{n,j}$, $j = 1, \ldots, 5$, $(n_1, n_2) = (200, 225)$ and censoring percentage = 20%

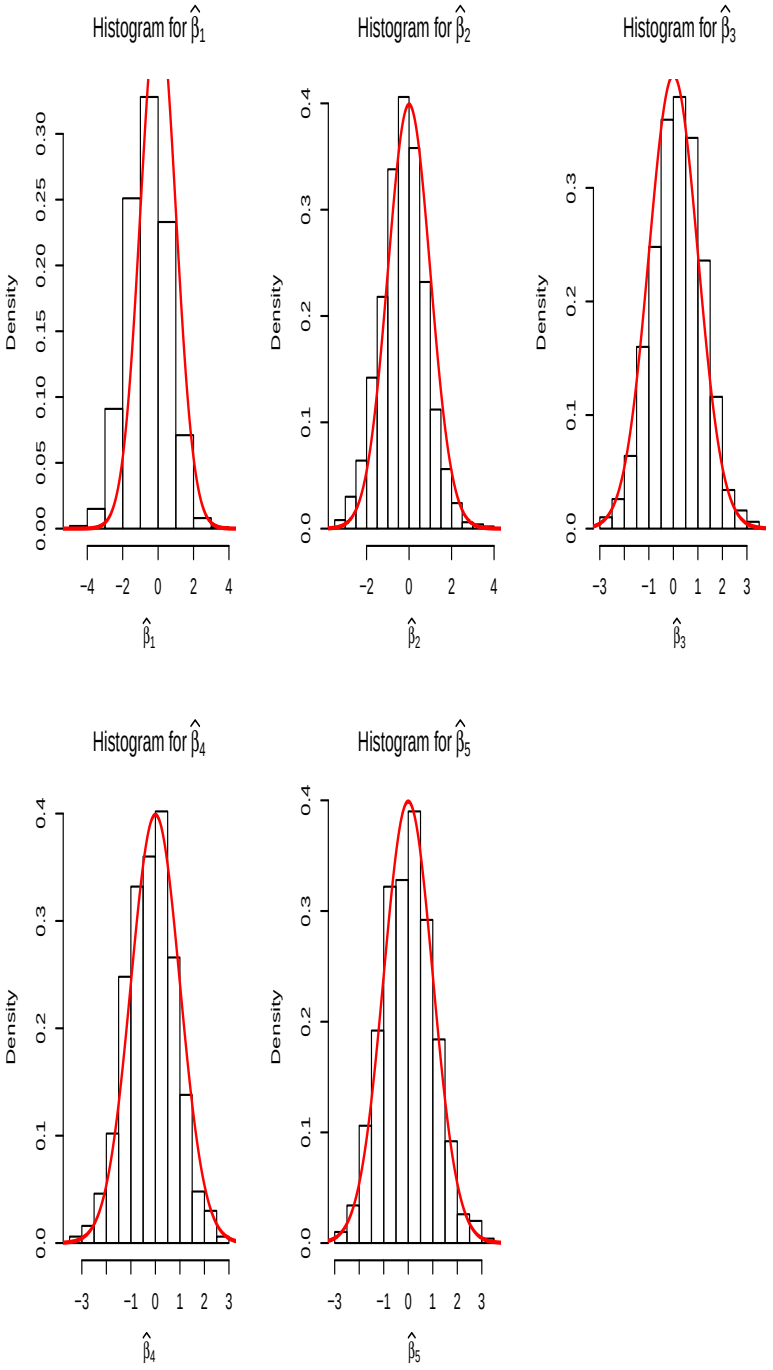

FIGURE B.9: Histogram of the N normalized estimates $\widehat{\beta}^{(1)}_{n,\hat{\omega},j}, \ldots, \widehat{\beta}^{(N)}_{n,\hat{\omega},j}$, $j = 1, \ldots, 5$, $(n_1, n_2) = (200, 225)$ and censoring percentage = 40% . In red: density of the $\mathcal{N}(0, 1)$.

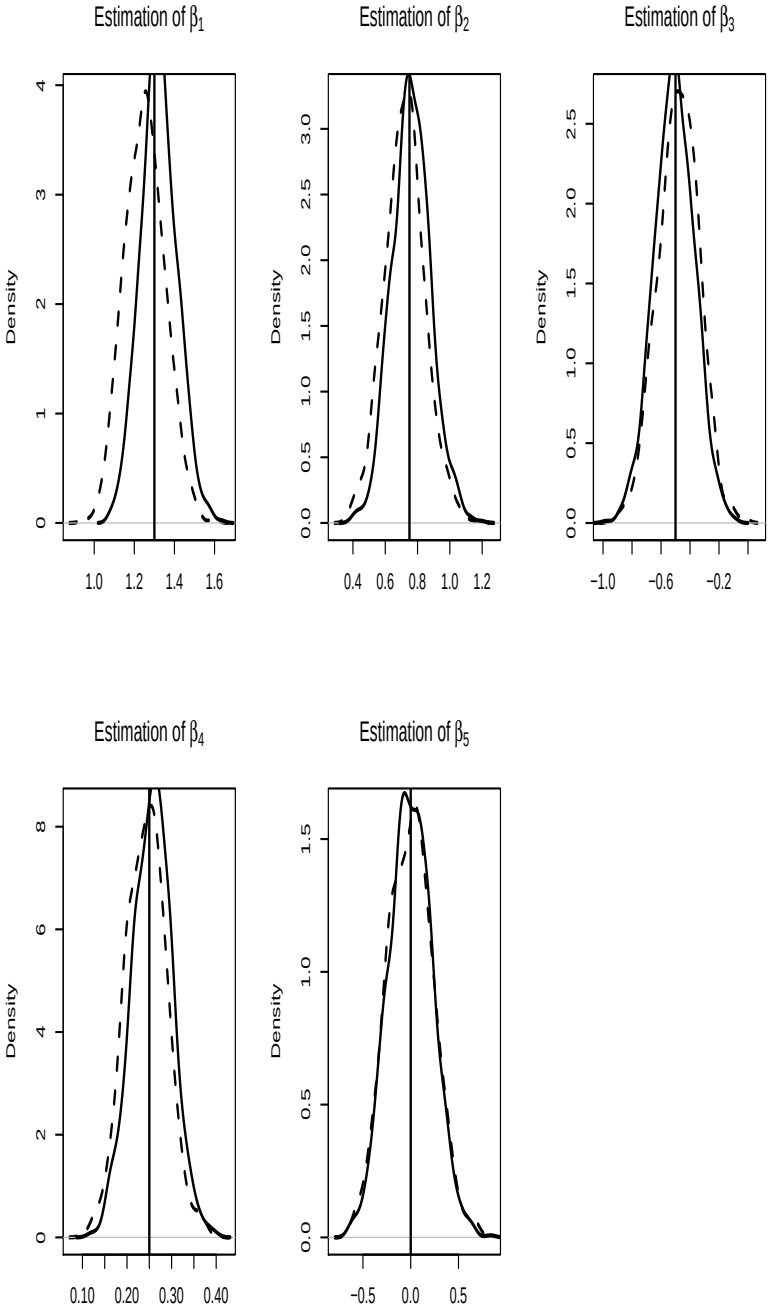

FIGURE B.10: Density estimation from the N two-step estimates (dashed line) and oracle estimates (solid line), $(n_1, n_2) = (200, 225)$ and censoring percentage = 40% . The true value is indicated by a vertical line.

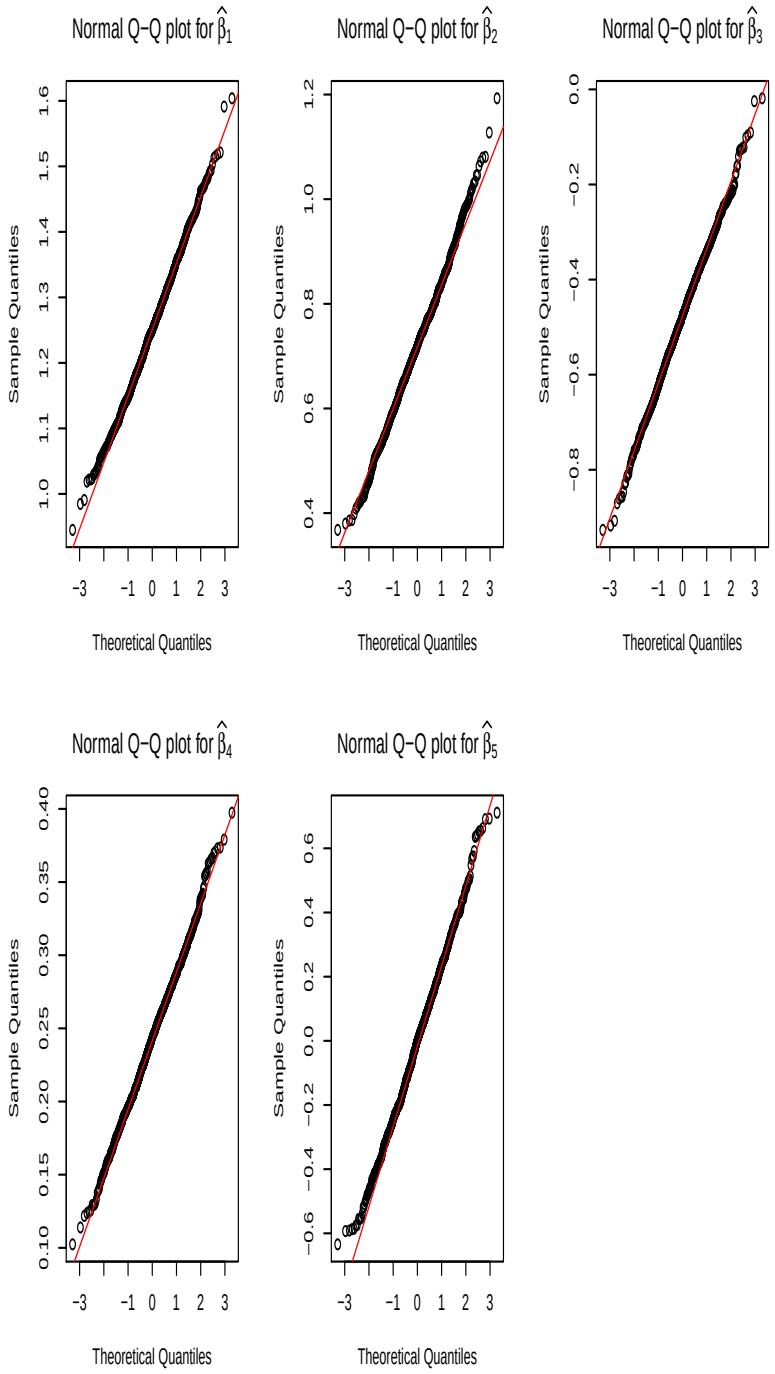

FIGURE B.11: QQ-plots of the N estimates $\widehat{\beta}^{(1)}_{n,\hat{\omega},j}, \ldots, \widehat{\beta}^{(N)}_{n,\hat{\omega},j}$, $j = 1, \ldots, 5$, $(n_1, n_2) = (200, 225)$ and censoring percentage = 40%.

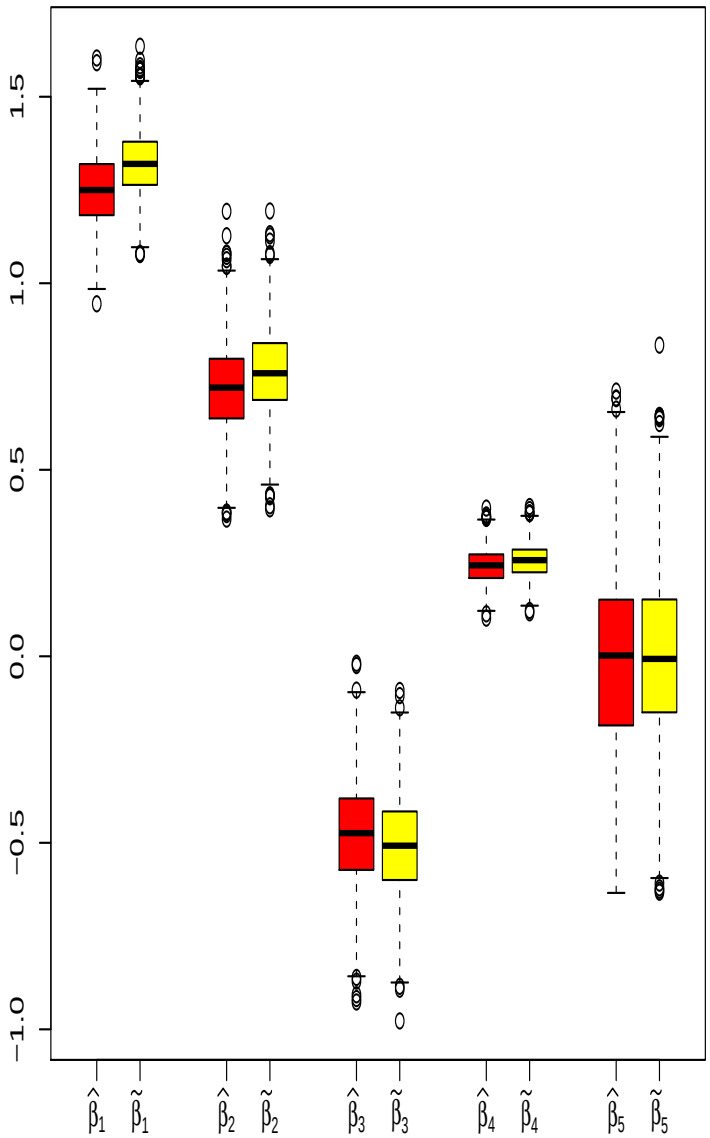

FIGURE B.12: Boxplots-plots of the N **two-step estimates** $\widehat{\beta}^{(1)}_{n,\hat{\omega},j}, \ldots, \widehat{\beta}^{(N)}_{n,\hat{\omega},j}$ and **oracle estimates** $\widetilde{\beta}^{(1)}_{n,j}, \ldots, \widetilde{\beta}^{(N)}_{n,j}$, $j = 1, \ldots, 5$, $(n_1, n_2) = (200, 225)$ and censoring percentage = 40%

C

Résultats des simulations $(n_1, n_2) = (70, 85)$

		proposed method					oracle				
c		β_1	β_2	β_3	β_4	β_5	β_1	β_2	β_3	β_4	β_5
10%	RMSE	0.0231	0.0301	0.0377	0.0050	0.1117	0.0201	0.0303	0.0362	0.0044	0.1077
	bias	-0.0436	-0.0151	0.0188	-0.0075	0.0091	0.0360	0.0311	-0.0145	0.0071	0.0109
	SE	0.1409	0.1820	0.2071	0.0696	0.3450	0.1321	0.1673	0.1901	0.0640	0.3170
	SD	0.1458	0.1729	0.1933	0.0701	0.3342	0.1371	0.1715	0.1898	0.0660	0.3282
	CP	0.9130	0.9630	0.9630	0.9480	0.9550	0.9460	0.9390	0.9540	0.9440	0.9440
	ℓ	0.5522	0.7134	0.8117	0.2727	1.3523	0.5178	0.6557	0.7450	0.2509	1.2427
20%	RMSE	0.0255	0.0389	0.0463	0.0050	0.1410	0.0234	0.0366	0.0438	0.0050	0.1289
	bias	-0.0345	-0.0247	0.0177	-0.0076	-0.0010	0.0443	0.0187	-0.0157	0.0076	0.0041
	SE	0.1495	0.1925	0.2196	0.0741	0.3674	0.1386	0.1762	0.2015	0.0677	0.3349
	SD	0.1560	0.1959	0.2145	0.0705	0.3757	0.1466	0.1904	0.2087	0.0704	0.3592
	CP	0.9350	0.9400	0.9570	0.9590	0.9480	0.9360	0.9260	0.9460	0.9400	0.9330
	ℓ	0.5859	0.7545	0.8608	0.2905	1.4403	0.5434	0.6908	0.7901	0.2653	1.3127
40%	RMSE	0.0297	0.0468	0.0601	0.0072	0.1609	0.0296	0.0431	0.0581	0.0070	0.1560
	bias	-0.0261	-0.0220	0.0096	-0.0041	-0.0031	0.0436	0.0201	-0.0147	0.0071	-0.0051
	SE	0.1701	0.2267	0.2605	0.0869	0.4330	0.1542	0.2038	0.2344	0.0776	0.3882
	SD	0.1704	0.2154	0.2450	0.0846	0.4014	0.1667	0.2066	0.2407	0.0833	0.3951
	CP	0.9340	0.9590	0.9680	0.9590	0.9660	0.9400	0.9500	0.9540	0.9300	0.9380
	ℓ	0.6670	0.8888	1.0213	0.3407	1.6975	0.6043	0.7988	0.9187	0.3042	1.5218

Table 3. Simulation results with $(n_1, n_2) = (70, 85)$. RMSE: empirical root mean square error. SD: empirical standard deviation. CP: empirical coverage probability of 95%-level confidence intervals. ℓ: average length of 95%-level confidence intervals.

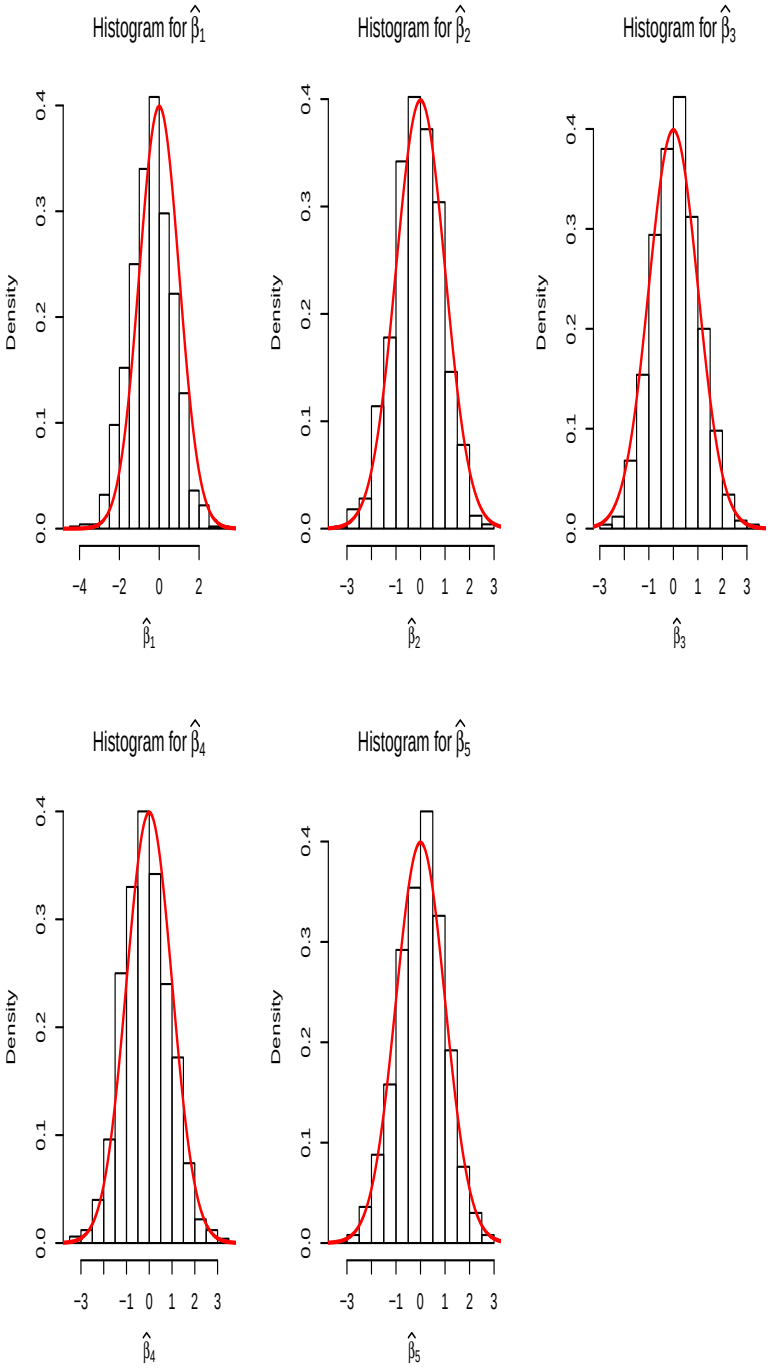

FIGURE C.1: Histogram of the N normalized estimates $\widehat{\beta}^{(1)}_{n,\hat{\omega},j}, \ldots, \widehat{\beta}^{(N)}_{n,\hat{\omega},j}$, $j = 1, \ldots, 5$, $(n_1, n_2) = (70, 85)$ and censoring percentage = 10% . In red: density of the $\mathcal{N}(0, 1)$.

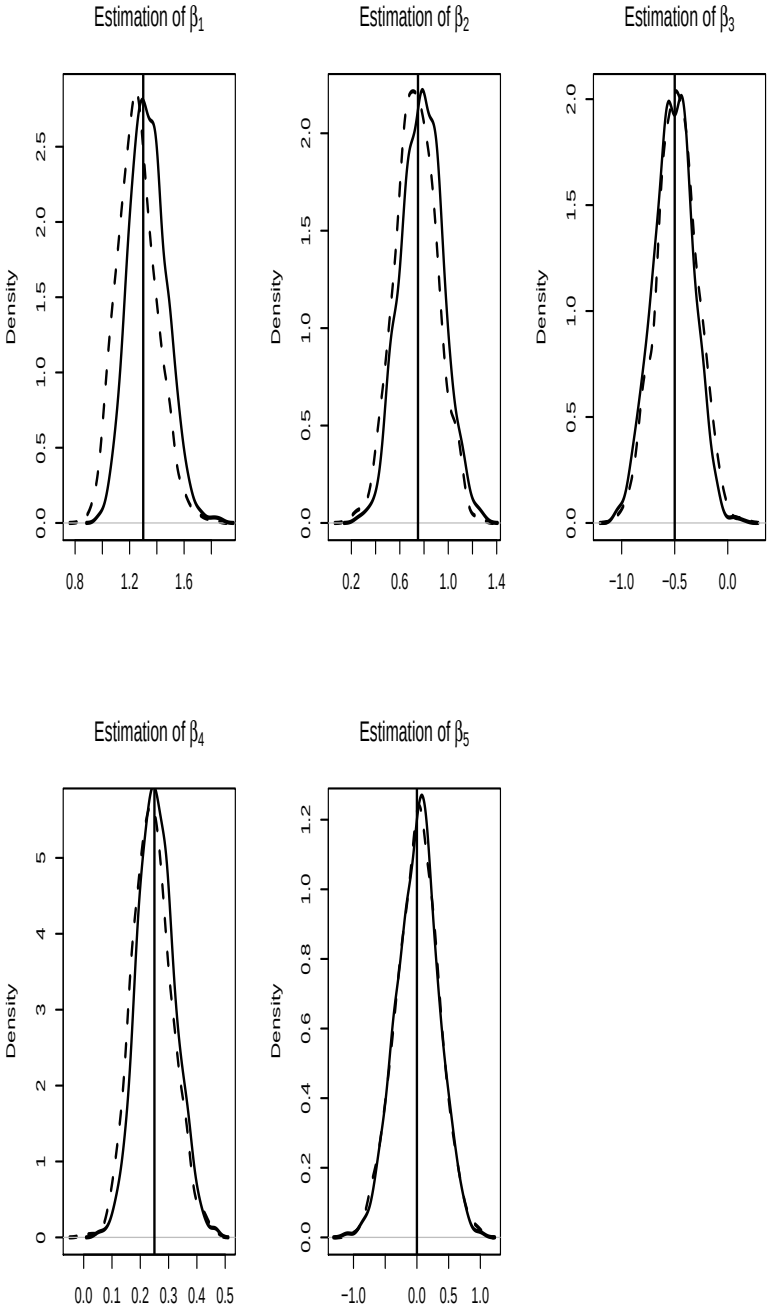

FIGURE C.2: Density estimation from the N two-step estimates (dashed line) and oracle estimates (solid line), $(n_1, n_2) = (70, 85)$ and censoring percentage = 10% . The true value is indicated by a vertical line.

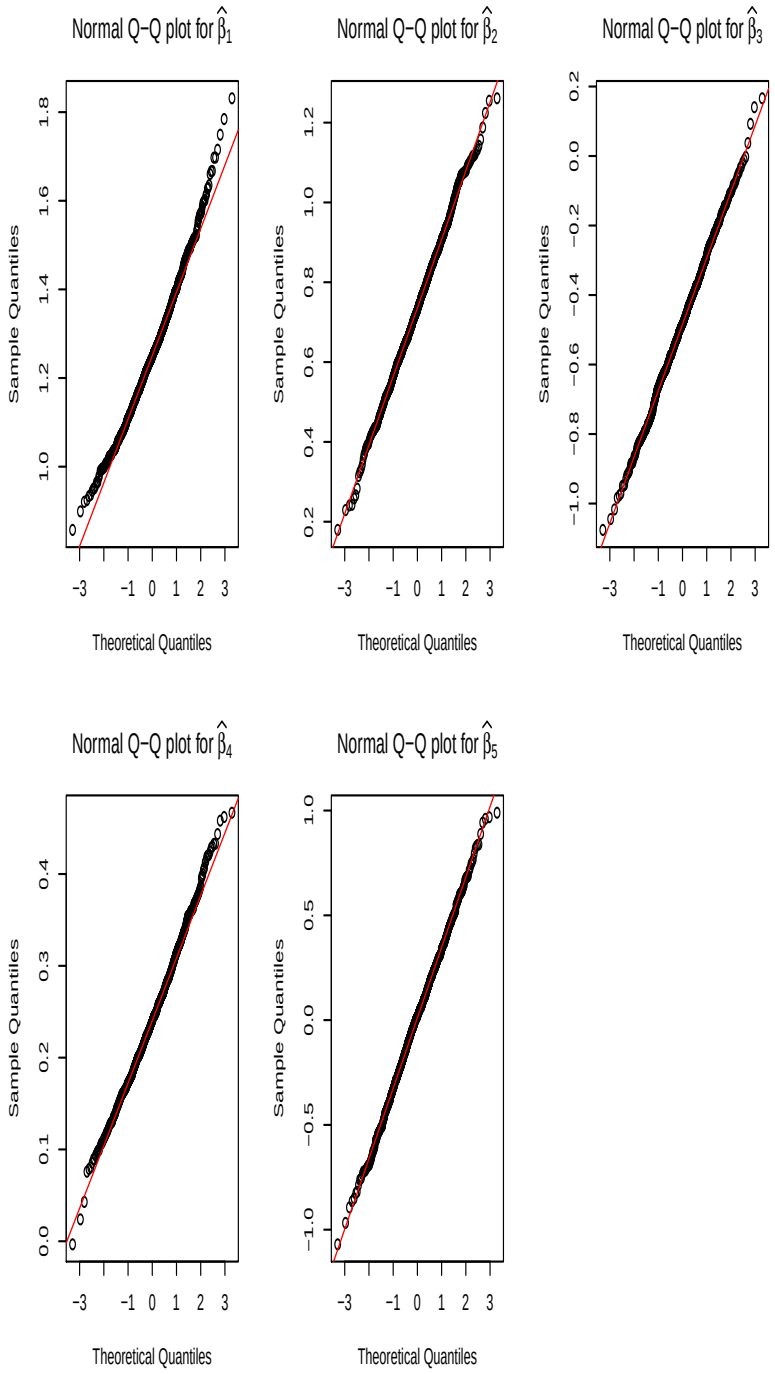

FIGURE C.3: QQ-plots of the N estimates $\widehat{\beta}^{(1)}_{n,\hat{\omega},j}, \ldots, \widehat{\beta}^{(N)}_{n,\hat{\omega},j}$, $j = 1, \ldots, 5$, $(n_1, n_2) = (70, 85)$ and censoring percentage = 10%.

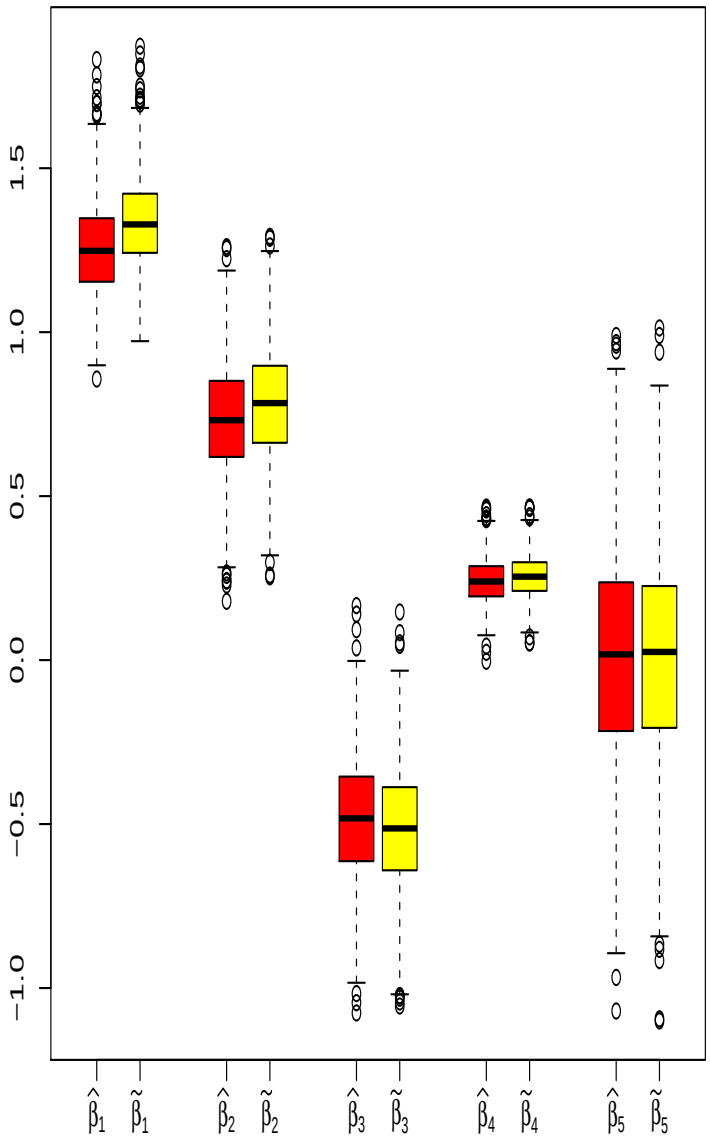

FIGURE C.4: Boxplots-plots of the N **two-step estimates** $\widehat{\beta}^{(1)}_{n,\hat{\omega},j}, \ldots, \widehat{\beta}^{(N)}_{n,\hat{\omega},j}$ and **oracle estimates** $\widetilde{\beta}^{(1)}_{n,j}, \ldots, \widetilde{\beta}^{(N)}_{n,j}$, $j = 1, \ldots, 5$, $(n_1, n_2) = (70, 85)$ and censoring percentage = 10%

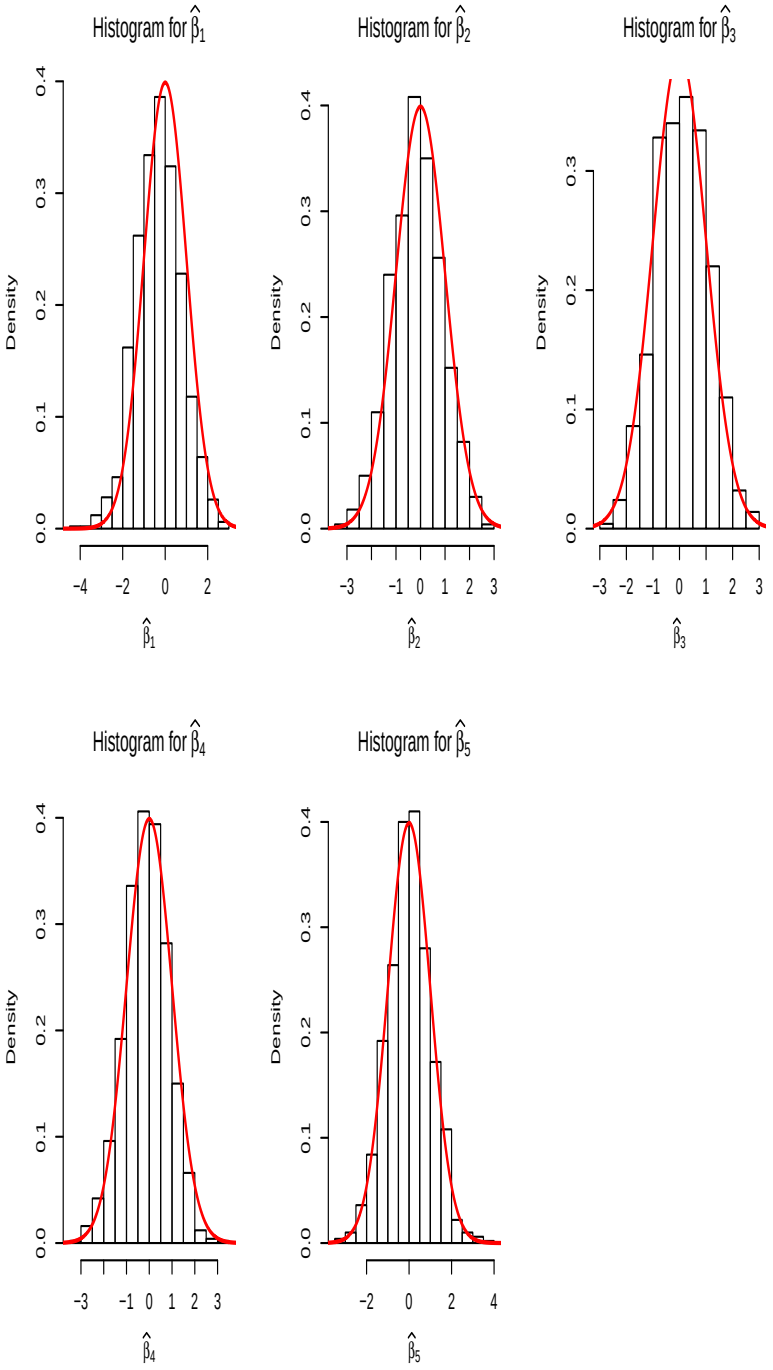

FIGURE C.5: Histogram of the N normalized estimates $\widehat{\beta}^{(1)}_{n,\hat{\omega},j}, \ldots, \widehat{\beta}^{(N)}_{n,\hat{\omega},j}$, $j = 1, \ldots, 5$, $(n_1, n_2) = (70, 85)$ and censoring percentage = 20% . In red: density of the $\mathcal{N}(0, 1)$.

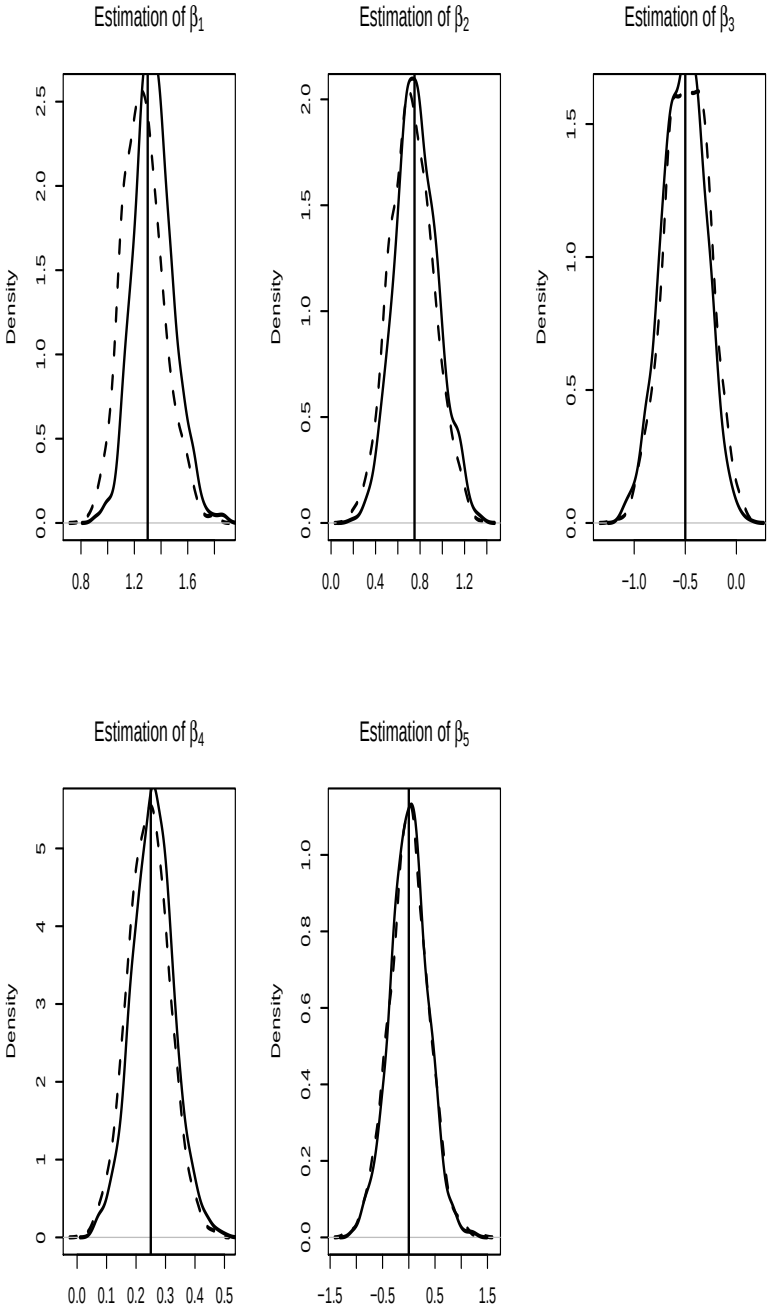

FIGURE C.6: Density estimation from the N two-step estimates (dashed line) and oracle estimates (solid line), $(n_1, n_2) = (70, 85)$ and censoring percentage = 20% . The true value is indicated by a vertical line.

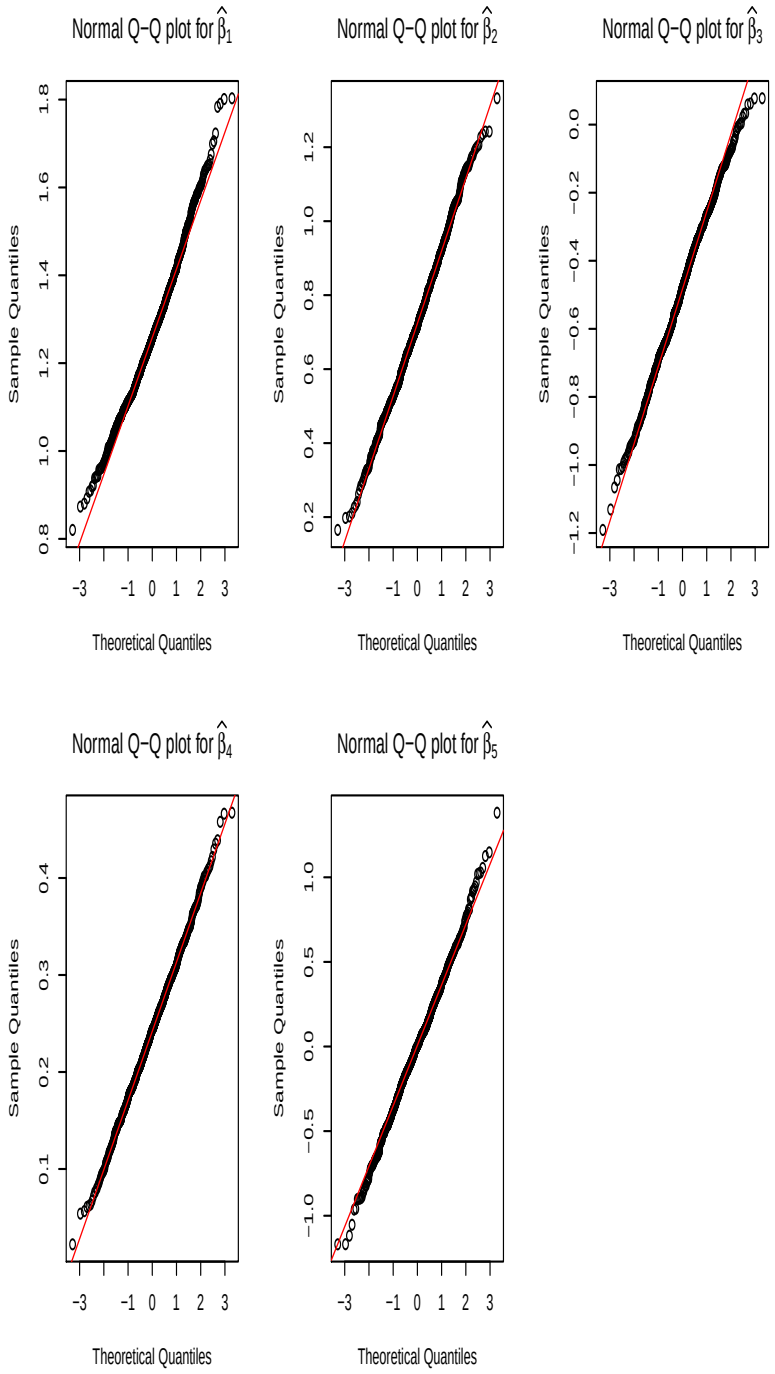

FIGURE C.7: QQ-plots of the N estimates $\widehat{\beta}^{(1)}_{n,\hat{\omega},j}, \ldots, \widehat{\beta}^{(N)}_{n,\hat{\omega},j}$, $j = 1, \ldots, 5$, $(n_1, n_2) = (70, 85)$ and censoring percentage = 20%.

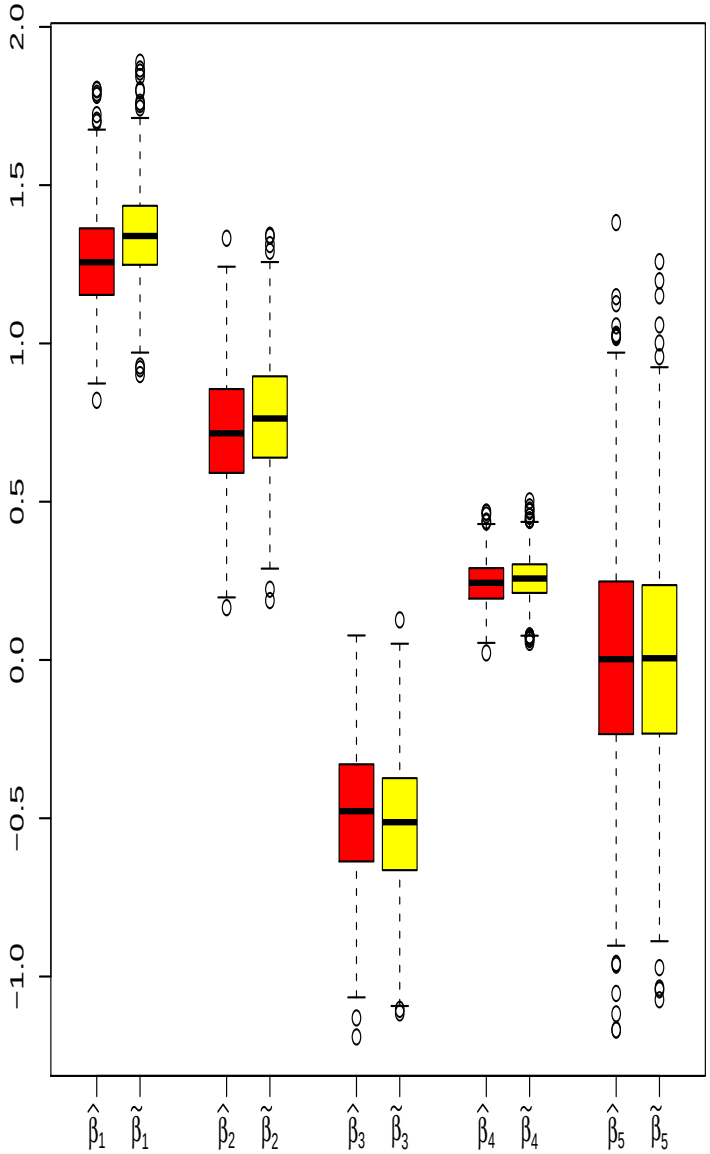

FIGURE C.8: Boxplots-plots of the N **two-step estimates** $\widehat{\beta}^{(1)}_{n,\hat{\omega},j}, \ldots, \widehat{\beta}^{(N)}_{n,\hat{\omega},j}$ and **oracle estimates** $\widetilde{\beta}^{(1)}_{n,j}, \ldots, \widetilde{\beta}^{(N)}_{n,j}$, $j = 1, \ldots, 5$, $(n_1, n_2) = (70, 85)$ and censoring percentage = 20%

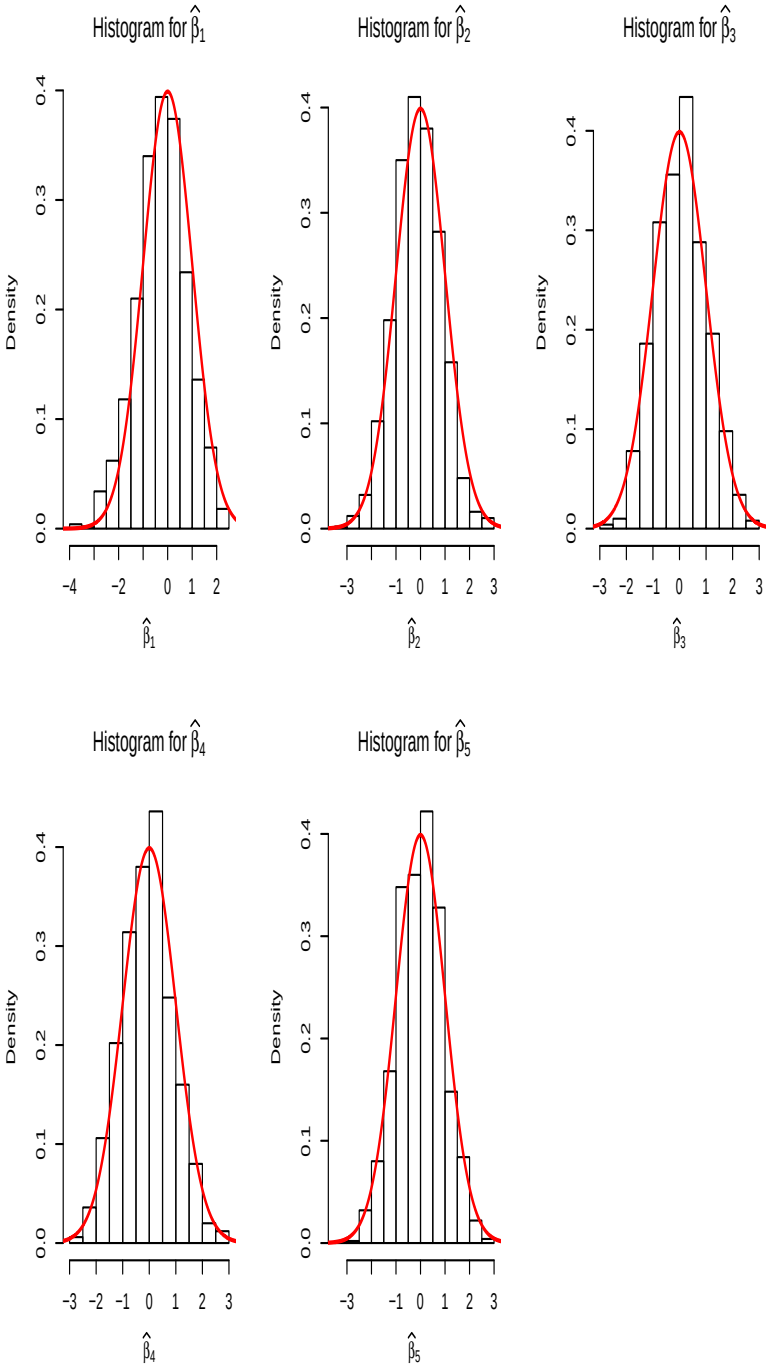

FIGURE C.9: Histogram of the N normalized estimates $\widehat{\beta}^{(1)}_{n,\hat{\omega},j}, \ldots, \widehat{\beta}^{(N)}_{n,\hat{\omega},j}$, $j = 1, \ldots, 5$, $(n_1, n_2) = (70, 85)$ and censoring percentage = 40% . In red: density of the $\mathcal{N}(0, 1)$.

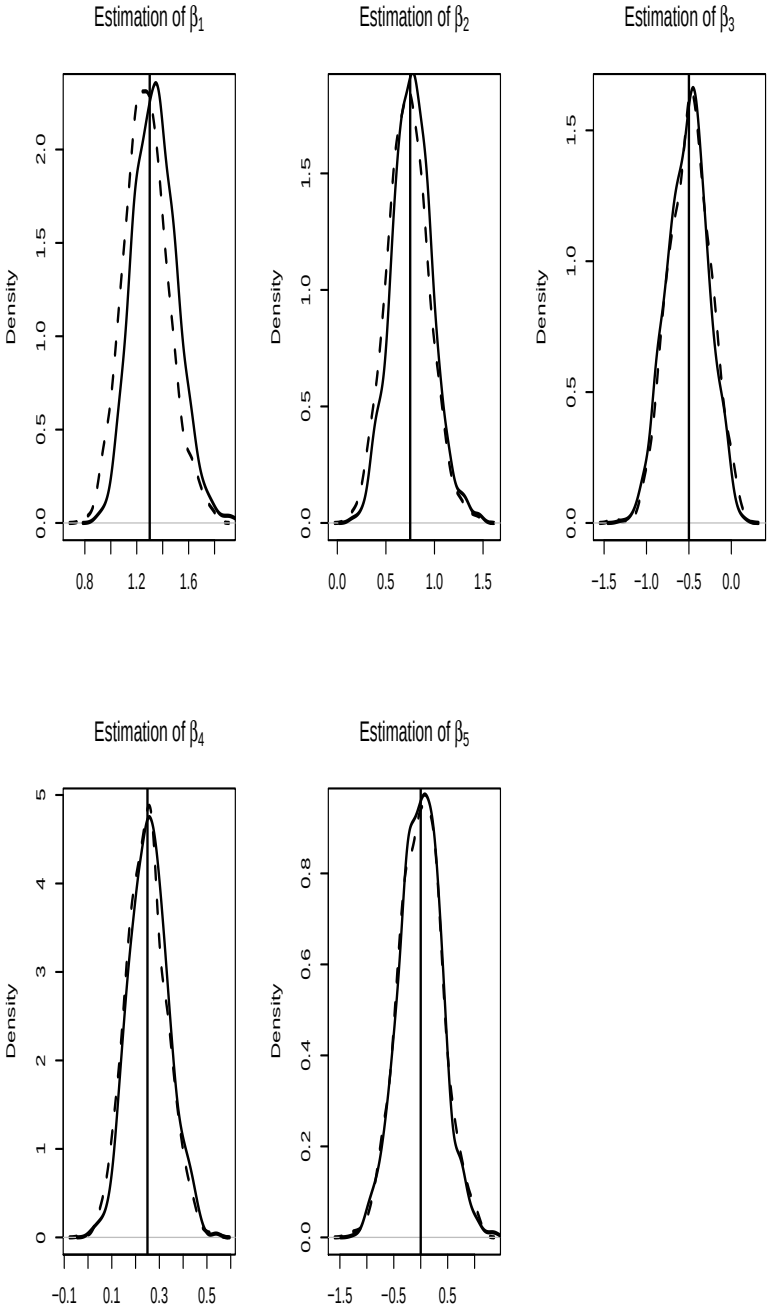

FIGURE C.10: Density estimation from the N two-step estimates (dashed line) and oracle estimates (solid line), $(n_1, n_2) = (70, 85)$ and censoring percentage = 40% . The true value is indicated by a vertical line.

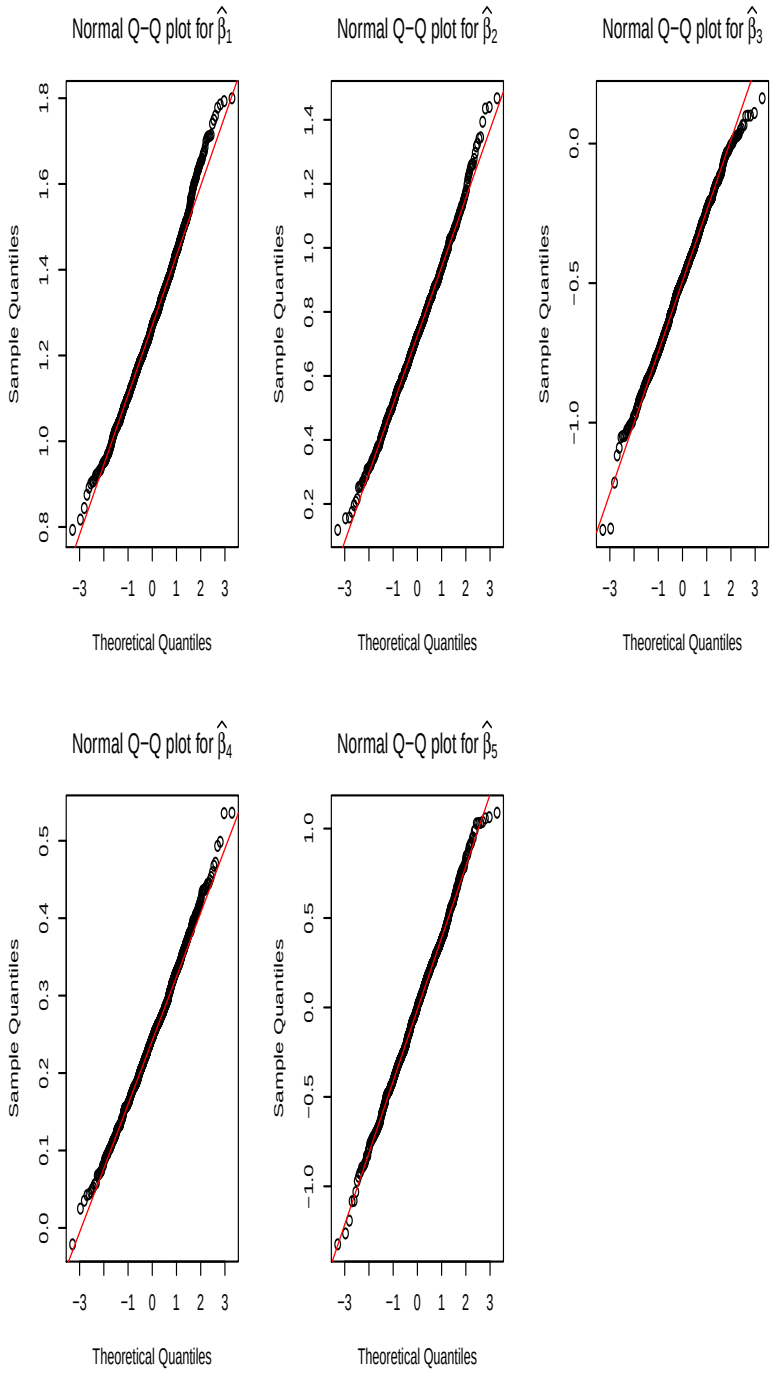

FIGURE C.11: QQ-plots of the N estimates $\widehat{\beta}^{(1)}_{n,\hat{\omega},j}, \ldots, \widehat{\beta}^{(N)}_{n,\hat{\omega},j}$, $j = 1, \ldots, 5$, $(n_1, n_2) = (70, 85)$ and censoring percentage = 40%.

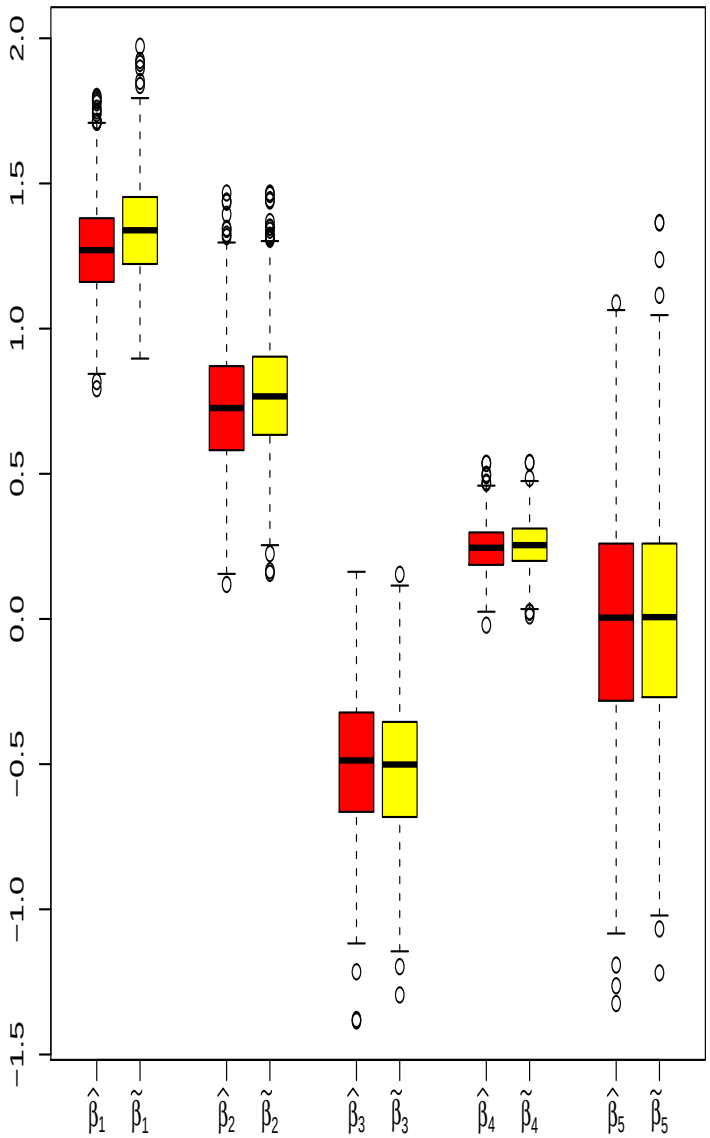

FIGURE C.12: Boxplots-plots of the N **two-step estimates** $\widehat{\beta}^{(1)}_{n,\hat{\omega},j}, \ldots, \widehat{\beta}^{(N)}_{n,\hat{\omega},j}$ and **oracle estimates** $\widetilde{\beta}^{(1)}_{n,j}, \ldots, \widetilde{\beta}^{(N)}_{n,j}$, $j = 1, \ldots, 5$, $(n_1, n_2) = (70, 85)$ and censoring percentage = 40%

Bibliographie

[1] O. Aalen, Ø. Borgan, and H. K. Gjessing. *Suvival and Event History Analysis. A Process Point of View*. Statistics for Biology and Health. Springer, New York, 2008. (Cité en pages 16 et 17.)

[2] Lai B. Qiao Y.-L. Vogt S. Luo X.-M. Taylor P.R. Dong Z.-W. Mark S.D. Dawsey S.M. Abnet, C.C. Zinc concentration in esophageal biopsy specimens measured by x-ray fluorescence and esophageal cancer risk. *Journal of the National Cancer Institute*, (97(4)):301–306, 2005. (Cité en pages 83 et 90.)

[3] P. K. Andersen, Ø. Borgan, R. D. Gill, and N. Keiding. *Statistical models based on counting processes*. Springer Series in Statistics. Springer-Verlag, New York, 1993. (Cité en pages 17, 46, 52 et 88.)

[4] P. K. Anderson and R. D. Gill. Cox's regression model for counting processes: A large sample study. *The Annals of Statistics,*. (Cité en pages 29, 32 et 33.)

[5] V. Bagdonavičius and M. S. Nikulin. *Accelerated life models: modeling and statistical analysis*. Boca Raton, FL: CRC Press, 2002. (Cité en pages 10, 27, 32, 45, 46, 47 et 48.)

[6] C. A. J. Ritov Y. Bikel, P. J. Klaassem and J. A. Weller. *Efficient and adaptive estimation for semiparametric models*. John Hopkins Series in the Mathematical Sciences. Baltimore, MD: John Hopkins University Press, 1993. (Cité en page 13.)

[7] N. Breslow. Contribution à la discussion sur l'article de d.r. cox, regression models and life-tables. *J. Roy. Statist. Soc. Ser. B*, 34:216–217, 1972. (Cité en page 31.)

[8] N. Breslow. Covariance analysis of censored survival data. *Biometrics*, 30:89–99, 1974. (Cité en page 31.)

[9] H. Y Chen and R. J. A. Little. Proportional hazards regression with missing covariates. *J. Amer. Statist. Assoc.*, 94:896–908, 1999. (Cité en page 35.)

[10] The Fibrinogen Studies Collaboration. Measures to assess the prognostic ability of the stratified Cox proportional hazards model. *Stat. Med.*, 28(3):389–411, 2009. (Cité en page 46.)

[11] D. R. Cox. Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B*, 34:187–220, 1972. With discussion. (Cité en pages 24, 27 et 45.)

[12] D. R. Cox. Partial likelihood. *Biometrika*, 62(2):269–276, 1975. (Cité en pages 27 et 45.)

[13] N. M. Dempster, A. P. Laird and D. B. Rubin. Maximum likelihood from incomplete data via the em alogorithm. with discussion. *J. Roy. Statist. Soc. Ser. B*, 39:1–38, 1977. (Cité en page 41.)

[14] A. Detais and J.-F Dupuy. Maximum likelihood estimation in a partially observed stratified regression model with censored data. *Inst. Statist. Math*, 2008. (Cité en pages xix, 36, 40, 41 et 42.)

[15] J.-F. Dupuy and E. Leconte. Stratified logrank test of no randomized treatment effect with missing stratum information. *J. Statist. Plann. Inference*, (37(17)):2760–2776, 2008. (Cité en page 83.)

[16] J.-F. Dupuy and E. Leconte. A study of regression calibration in a partially observed stratified Cox model. *J. Statist. Plann. Inference*, 139(2):317–328, 2009. (Cité en pages xix, 36, 39, 40 et 46.)

[17] Dahl C. M. Effraimidis, G. Nonparametric estimation of cumulative incidence functions for competing risks data with missing cause of failure. *Statistics & Probability Letters*, 89:1–7, 2014. (Cité en page 84.)

[18] T. R. Fleming and D. P. Harrington. *Counting Processes and Survival Analysis*. Wiley Series in Probability and Statistics. Wiley, 1991. (Cité en pages 13, 24, 27, 55 et 57.)

[19] A. Gandy and U. Jensen. On goodness-of-fit tests for Aalen's additive risk model. *Scand. J. Statist.*, 32(3):425–445, 2005. (Cité en pages xviii et 48.)

[20] A. Gandy and U. Jensen. Model checks for Cox-type regression models based on optimally weighted martingale residuals. *Lifetime Data Anal.*, 15(4):534–557, 2009. (Cité en pages 46, 48 et 64.)

[21] M. J. et Robins J. M. Gill, R. D. Van der Laan. Coarsening at random, in 1st seattle symposium in biostatistics: survival analysis. *Springer*, pages 255–294, 1997. (Cité en page 9.)

[22] M. Gorfine, L. Hsu, and R. L. Prentice. Nonparametric correction for covariate measurement error in a stratified Cox model. *Biostatistics*, 5(1):75–87, 2004. (Cité en page 46.)

[23] M. Haenlein. Social interactions in customer churn decisions: The impact of relationship directionality. *Int. J. Research Marketing*, 30(3):236–248, 2013. (Cité en pages xviii et 46.)

[24] T. E. Hanson, A. Jara, and L. Zhao. A Bayesian semiparametric temporally-stratified proportional hazards model with spatial frailties. *Bayesian Anal.*, 7(1):147–188, 2012. (Cité en page 46.)

[25] H. Hardin, J. W. Schmiedich and R. J. Carroll. The regression calibration method for fitting generalized linear models with additive measurement error. *The stata. Journal*, pages 1–11, 2003. (Cité en page 36.)

[26] H. Heinzl, J. Stare, and M. Mittlböck. A measure of dependence for the stratified Cox proportional hazards regression model. *Biom. J.*, 44(6):671–682, 2002. (Cité en page 46.)

[27] Thompson D. J. Horvitz, D. G. A generalization of sampling without replacement from a finite universe. 47:663–685, 1952. (Cité en page 84.)

[28] Chen X. Sun J. Hu, N. A generalization of sampling without replacement from a finite universe. 42(2):438–452, 2015. (Cité en page 84.)

[29] S. Y. H. Huang. Regression calibration using response variables in linear models. 15:685–696, 2005. (Cité en page 36.)

[30] Lee J. Sun Y. Hyun, S. Proportional hazards model for competing risks data with missing cause of failure. (7):1767–1779, 2012. (Cité en page 84.)

[31] J. D. Kalbfleisch and R. L. Prentice. *The statistical analysis of failure time data*. Wiley Series in Probability and Mathematical Statistics. Jhon Wiley and Sons. New York-Chichester-Brisbane, 1980. (Cité en pages 10 et 27.)

[32] N. M. Kiefer and C. E. Larson. Counting processes for retail default modeling. Technical Report CREATES Research Paper 2015-17, Department of Economics and Business Economics, Aarhus University, Denmark, 2015. (Cité en pages xviii et 46.)

[33] J. Kim. Confidence intervals for the difference of median survival times using the stratified Cox proportional hazards model. *Biom. J.*, 43, 2001. (Cité en page 46.)

[34] Moeschberger M.L. Klein, J.P. Survival analysis: Techniques for censored and truncated data. 1997. (Cité en pages 33, 82 et 83.)

[35] M. R. Kosorok. Introduction to empirical processes and semiparametric inference. 2008. (Cité en page 95.)

[36] J. F. Lawless. Statistical models and methods for lifetime data. *Wiley Series in Probability and Statistics. Wiley, Hoboken, second edition*, 2003. (Cité en page 10.)

[37] D. Y. Lin, L. J. Wei, and Z. Ying. Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika*, 80(3):557–572, 1993. (Cité en pages xviii, 46, 48 et 64.)

[38] R. J. Little and B. D. Rubin. Statistical analysis with missing data. *John Wiley, New-york*, 1987. (Cité en page 9.)

[39] R. J. Little and B. D. Rubin. Statistical analysis with missing data. *Wiley-Interscience*, 2002. (Cité en page 84.)

[40] Wang Q. Liu, C. Semiparametric estimation for regression coefficients in the cox model with failure indicators missing at random. *Statistica Sinica*, 20(3):1125–1142, 2010. (Cité en page 84.)

[41] T. Martinussen. Cox regression with incomplete covariate measurements using the em-algorithm. *Scand. J. Statist*, 26:479–491, 1999. (Cité en page 35.)

[42] T. Martinussen and T. H. Scheike. *Dynamic regression models for survival data*. Statistics for Biology and Health. Springer, New York, 2006. (Cité en pages 24, 45 et 46.)

[43] L. Marzec and P. Marzec. Generalized martingale-residual processes for goodness-of-fit inference in Cox's type regression models. *Ann. Statist.*, 25(2):683–714, 1997. (Cité en pages xviii, 46, 48 et 64.)

[44] L. Natarajan and J. O'Quigley. Predictive capability of stratified proportional hazards models. *J. Appl. Stat.*, 29(8):1153–1163, 2002. (Cité en page 46.)

[45] A. K. Örtqvist, C. Lundholm, H. Kieler, J. F. Ludvigsson, T. Fall, W. Ye, and C. Almqvist. Antibiotics in fetal and early life and subsequent childhood asthma: nationwide population based study with sibling analysis. *BMJ*, 349:doi: 10.1136/bmj.g6979, 2014. (Cité en page 46.)

[46] M. C. Paik. Multiple imputation for the cox proportional hazards model with missing covariates. *Lifetime Data Anal*, 3:289–298, 1997. (Cité en page 35.)

[47] M. C. Paik and W. Y. Tsai. On using the proportional hazards model with missing covariates. *Biometrika*, 84:579–593, 1997. (Cité en page 35.)

[48] O. Pons. Estimation in the cox model with missing covariate data. *J. Nonparametr. Stat.*, 14:223–247, 2002. (Cité en page 35.)

[49] R. L. Prentice. Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*, 96:331–342, 1982. (Cité en page 36.)

[50] Robins J. Lipsitz S. Harrington D. Pugh, M. Inference in the cox proportional hazards model with missing covariate data. *Technical report. Harvard School of Public Health, Department of Biostatistics*, 51:269–286, 1980. (Cité en pages xix et 84.)

[51] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. (Cité en pages 52 et 89.)

[52] R. Robolledo. Central limit theorems for local martingales. *Z. Wahrsch, Verw. Geb*, 51:269–286, 1980. (Cité en page 17.)

[53] Spiegelman D. Willett W. C. Rosner, B. Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. *Amer. J. Epidemiol*, 132:734–745, 1990. (Cité en page 36.)

[54] Willett W. C. Spiegelman D. Rosner, B. Correction of logistic relative risk estimates for non-random measurement error. *Statist. Med*, 8:1051–1069, 1989. (Cité en page 36.)

[55] F. Scavonetto, T. Y. Yeoh, E. C. Umbreit, T. N. Weingarten, M. T. Gettman, I. Frank, S. A. Boorjian, R. J. Karnes, D. R. Schroeder, L. J. Rangel, A. C. Hanson, R. E. Hofer, D. I. Sessler, and J. Sprung. Association between neuraxial analgesia, cancer progression, and mortality after radical prostatectomy: a large, retrospective matched cohort study. *Br. J. Anaesth.*, 113:doi: 10.1093/bja/aet467, 2014. (Cité en page 46.)

[56] White I.R. Seaman, S.R. Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*, 22(3):278–295, 2005. (Cité en page 84.)

[57] Carroll R. Kipnis V Spiegelman, D. Efficient regression calibration for logistic regression in main study/internal validation study designs with an imperfect reference instrument. *Statist. Med*, 10:139–160, 2001. (Cité en page 36.)

[58] W. Stute and L.-X. Zhu. Model checks for generalized linear models. *Scand. J. Statist.*, 29(3):535–545, 2002. (Cité en pages 48 et 64.)

[59] J. Q. Su and L. J. Wei. A lack-of-fit test for the mean function in a generalized linear model. *J. Amer. Statist. Assoc.*, 86(414):420–426, 1991. (Cité en pages 48 et 64.)

[60] Bandyopadhyay D. Subramanian, S. Semiparametric left truncation and right censorship models with missing censoring indicators. 78(16):2572–2577, 2008. (Cité en page 84.)

[61] T. M. Therneau and P. M. Grambsch. *Modeling suvival data: extending the Cox model*. Statistic and Biology and Health. Springer-Verlag, New York, 2000. (Cité en page 24.)

[62] A. A. Tsiatis. A large sample study of cox's regression model. *The Annals of Statistics*, 9:93–108, 1981. (Cité en page 29.)

[63] A. A. Tsiatis. Semi parametric theory and missing data. *Springer*, 2006. (Cité en pages 9 et 36.)

[64] A.W. van der Vaart. *Asymptotic Statistics*. Statistic and Biology and Health. Cambridge University Press, 1998. (Cité en page 85.)

[65] A.W. van der Vaart. *Semiparametric Theory and Missing Data*. New York: Springer, 2006. (Cité en page 91.)

[66] P. J. M. Verweij, H. C. van Houwelingen, and T. Stijnen. A goodness-of-fit test for Cox's proportional hazards model based on martingale residuals. *Biometrics*, 54(4):1517–1526, 1998. (Cité en pages xviii, 46 et 48.)

[67] Chen H.Y. Wang, C.Y. Augmented inverse probability weighted estimator for cox missing covariate regression. *Biometrics*, 57:414–419, 2001. (Cité en page 84.)

[68] Milton D.K. Eisen E.A. et Spiegelman D. Weller, E.A. Regression calibration for logistic regression with multiple surrogates for one exposure. *Journal of Statistical Planning and Inference*, 137:449–461, 2007. (Cité en page 36.)

[69] Liu C. Xie, J. Adjusted kaplan-meier estimator and log-rank test with inverse probability of treatment weighting for survival data. *Statistics in Medicine*, 24(20):3089–3110, 2005. (Cité en page 84.)

Résumé:

DAns ce travail, nous nous intéressons à l'inférence statistique dans deux modèles semi-paramétrique et non-paramétrique stratifiés de durées de vie censurées. Nous proposons tout d'abord une statistique de test d'ajustement pour le modèle de régression stratifié à risques proportionnels. Nous établissons sa distribution asymptotique sous l'hypothèse nulle d'un ajustement correct du modèle aux données. Nous étudions les propriétés numériques de ce test (niveau, puissance sous différentes alternatives) au moyen de simulations. Nous proposons ensuite une procédure permettant de stratifier le modèle à risques proportionnels suivant un seuil inconnu d'une variable de stratification. Cette procédure repose sur l'utilisation du test d'ajutement proposé précédemment. Une étude de simulation exhaustive est conduite pour évaluer les performances de cette procédure.

Dans une seconde partie de notre travail, nous nous intéressons à l'application du test du logrank stratifié dans un contexte de données manquantes (nous considérons la situation où les strates ne peuvent être observées chez tous les individus de l'échantillon). Nous construisons une version pondérée du logrank stratifié adaptée à ce problème. Nous en établissons la loi limite sous l'hypothèse nulle d'égalité des fonctions de risque dans les différents groupes. Les propriété de cette nouvelle statistique de test sont évaluées au moyen de simulations. Le test est ensuite appliqué à un jeu de données médicales.

Mots clés: Analyse de survie, données manquantes, modèle de Cox stratifié, résultats asymptotiques, simulations, test d'ajustement, test du logrank stratifié.

Titre: Inférence statistique dans des modèles de mélange à risques proportionnels.

Spécialité: Mathématiques appliquées.

Lieux: Thèse préparée à la Faculté des Sciences de Monastir (Tunisie) et à l'Institut National des Sciences Appliquées de Rennes (France).

Abstract:

IN this work, we are interested in the statistical inference in two semi-parametric and non-parametric stratified models for censored data. We first propose a goodness-of-fit test statistic for the stratified proportional hazards regression model. We establish its asymptotic distribution under the null hypothesis of a correct fit of the model. We investigate the numerical properties of this test (level, power under different alternatives) by means of simulations. Then, we propose a procedure allowing to stratify the proportional hazards model according to an unknown threshold in a stratification variable. This procedure is based on the goodness-of-fit test proposed earlier. An exhaustive simulation study is conducted to evaluate the performance of this procedure. In a second part of our work, we consider the stratified logrank test in a context of missing data (we consider the situation where strata can not be observed on all sample individuals). We construct a weighted version of the stratified logrank, adapted to this problem. We establish its asymptotic distribution under the null hypothesis of equality of the hazards functions in the different groups. The properties of this new test statistic are assessed using simulations. Finally, the test is applied to a medical dataset.

Keywords: Survival analysis, missing data, stratified Cox model, asymptotic results, simulation, goodness-of-fit test, stratified logrank test.

Title: Statistical inference in mixture of proportional hazards models.

Speciality: Mathematic and Applications

Thesis prepared at the Monastir University (Tunisia) and the National Institute of Applied Sciences of Rennes (France).

INSA de RENNES
Service des Formations

AVIS DU JURY SUR LA REPRODUCTION DE LA THESE SOUTENUE

Titre de la thèse:
Inférence statistique dans des modèles de mélange à risques proportionnels

Nom Prénom de l'auteur : BEN ELOUEFI RIM

Membres du jury :
- Madame DABO Sophie
- Monsieur ZARATI Saïd
- Monsieur KOKONENDJI Célestin
- Monsieur DUPUY Jean-François
- Madame CHAGNEAU Pierrette
- Madame HEUTTE Natacha

Président du jury : Monsieur KOKONENDJI Célestin

Date de la soutenance : 05 Septembre 2017

Reproduction de la these soutenue

☒ Thèse pouvant être reproduite en l'état
☐ Thèse pouvant être reproduite après corrections suggérées

Fait à Rennes, le 05 Septembre 2017

Signature du président de jury

Célestin KOKONENDJI

Le Directeur

M'hamed DRISSI

Résumé

Dans ce travail, nous nous intéressons à l'inférence statistique dans deux modèles semi-paramétrique et non-paramétrique stratifiés de durées de vie censurées. Nous proposons tout d'abord une statistique de test d'ajustement pour le modèle de régression stratifié à risques proportionnels. Nous établissons sa distribution asymptotique sous l'hypothèse nulle d'un ajustement correct du modèle aux données. Nous étudions les propriétés numériques de ce test (niveau, puissance sous différentes alternatives) au moyen de simulations. Nous proposons ensuite une procédure permettant de stratifier le modèle à risques proportionnels suivant un seuil inconnu d'une variable de stratification. Cette procédure repose sur l'utilisation du test d'ajustement proposé précédemment. Une étude de simulation exhaustive est conduite pour évaluer les performances de cette procédure.

Dans une seconde partie de notre travail, nous nous intéressons à l'application du test du logrank stratifié dans un contexte de données manquantes (nous considérons la situation où les strates ne peuvent être observées chez tous les individus de l'échantillon). Nous construisons une version pondérée du logrank stratifié adaptée à ce problème. Nous en établissons la loi limite sous l'hypothèse nulle d'égalité des fonctions de risque dans les différents groupes. Les propriétés de cette nouvelle statistique de test sont évaluées au moyen de simulations. Le test est ensuite appliqué à un jeu de données médicales..

Abstract

In this work, we are interested in the statistical inference in two semi-parametric and non-parametric stratified models for censored data. We first propose a goodness-of-fit test statistic for the stratified proportional hazards regression model. We establish its asymptotic distribution under the null hypothesis of a correct fit of the model. We investigate the numerical properties of this test (level, power under different alternatives) by means of simulations. Then, we propose a procedure allowing stratifying the proportional hazards model according to an unknown threshold in a stratification variable. This procedure is based on the goodness-of-fit test proposed earlier. An exhaustive simulation study is conducted to evaluate the performance of this procedure.

In a second part of our work, we consider the stratified logrank test in a context of missing data (we consider the situation where strata cannot be observed on all sample individuals). We construct a weighted version of the stratified logrank, adapted to this problem. We establish its asymptotic distribution under the null hypothesis of equality of the hazards functions in the different groups. The properties of this new test statistic are assessed using simulations. Finally the test is applied to a medical dataset..

N° d'ordre : 17ISAR 23 / D17 - 23

Institut National des Sciences Appliquées de Rennes
20, Avenue des Buttes de Coëmes • CS 70839 • F-35708 Rennes Cedex 7
Tel : 02 23 23 82 00 - Fax : 02 23 23 83 96