Développement d'un indice de séparabilité adapté aux données de génomique en analyse de survie

Sigrid Laure Rouam

Année 2011 N°

Thèse

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITE PARIS SUD

Spécialité : Santé Publique
Option : Biostatistiques

Présentée et soutenue publiquement par

M^elle^ Sigrid ROUAM

le 30 mars 2011

Développement d'un indice de séparabilité adapté aux données de génomique en analyse de survie

Directeur de thèse : Monsieur le Docteur Philippe BROËT
Co-directeur : Monsieur le Docteur Thierry MOREAU

Membres du Jury :

M. Jean-Christophe THALABARD (PU-PH)	Président
M. Jean-Louis GOLMARD (MCU-PH)	Rapporteur
M. Jean-Pierre DAURÈS (PU-PH)	Rapporteur
M. Khê HOANG XUAN (PU-PH)	Examinateur
M. Philippe BROËT (MCU-PH)	Directeur de thèse
M. Thierry MOREAU (DR)	Co-directeur

Thèse préparée dans les laboratoires suivants :

Département de Méthodologie Biostatistique de la génomique en épidémiologie clinique
Hôpital Paul Brousse
16 av. Paul Vaillant Couturier
94807 Villejuif cedex
France
http ://ifr69.vjf.inserm.fr/je2492/index.html

Équipe Biostatistiques
INSERM UMRS 1018
Hôpital Paul Brousse
16 av. Paul Vaillant Couturier
94807 Villejuif cedex
France
http ://www.cesp.idf.inserm.fr/page.asp ?page=1098

Genome Institute of Singapore
60 Biopolis Street, Genome
Singapore 138672
Singapore
http ://www.gis.a-star.edu.sg/internet/site/

Remerciements

J'aimerais ici remercier toutes les personnes qui m'ont soutenues au cours de ces trois années et demi de thèse, en France et à Singapour.

Tout d'abord, je tiens à remercier Philippe Broët, mon directeur de thèse, pour ses conseils, sa patience et son aide considérable, pour avoir toujours pris le temps de m'éclairer sur les points obscurs et pour m'avoir donner l'opportunité de réaliser une grande partie de mon travail à Singapour.

J'aimerais également exprimer ma gratitude à Thierry Moreau, qui m'a également beaucoup apporté d'un point de vue scientifique, qui m'a soutenue et encouragée, et m'a suivie durant ces trois années (pas toujours évident avec la distance).

Je remercie sincèrement les membres de mon jury de thèse. Merci à Monsieur Jean Christophe Thalabard de m'avoir fait l'honneur d'être président, ainsi qu'à Messieurs Jean-Louis Golmard et Jean-Pierre Daurès, qui ont bien voulu rapporter cette thèse. Merci Monsieur Khê Hoang Xuan d'avoir accepté de faire partie de mon jury.

J'aimerais remercier le Ministère de l'Enseignement Supérieur et de la Recherche , ainsi que le Genome Institute of Singapore pour leur financement.

Mes remerciements s'adressent à toutes les personnes que j'ai rencontrées au cours de mon parcours et qui ont contribué à rendre ces trois années agréables et enrichissantes, aussi bien à l'INSERM, à l'Université Paris Sud qu' au Genome Institute of Singapore.

Un grand merci à mes amis qui m'ont soutenu et aidé à persévérer dans mon travail.

Je tiens à remercier ma famille pour leur amour et l'intérêt porté à mon travail : mon père, ma sœur et mon grand-père.

Finally, I would like to express my gratitude to Lawrence, who has been very comprehensive, supportive and always available.

Résumé

Dans le domaine de l'oncogénomique, l'un des axes actuels de recherche est l'identification de nouveaux marqueurs génétiques permettant entre autres de construire des règles prédictives visant à classer les patients selon le risque d'apparition d'un événement d'intérêt (décès ou récidive tumorale). En présence de telles données de haute dimension, une première étape de sélection parmi l'ensemble des variables candidates est généralement employée afin d'identifier les marqueurs ayant un intérêt explicatif jugé suffisant. Une question récurrente pour les biologistes est le choix de la règle de sélection. Dans le cadre de l'analyse de survie, les approches classiques consistent à ranger les marqueurs génétiques à partir du risque relatif ou de quantités issues de test statistiques (p-value, q-value). Cependant, ces méthodes ne sont pas adaptées à la combinaison de résultats provenant d'études hétérogènes dont les tailles d'échantillons sont très différentes.

Utiliser un indice tenant compte à la fois de l'importance de l'effet pronostique et ne dépendant que faiblement de la taille de l'échantillon permet de répondre à cette problématique. Dans ce travail, nous proposons un nouvel indice de capacité de prédiction afin de sélectionner des marqueurs génomiques ayant un impact pronostique sur le délai de survenue d'un évènement. Cet indice étend la notion de pseudo-R^2 dans le cadre de l'analyse de survie. Il présente également une interprétation originale et intuitive en terme de « séparabilité ». L'indice est tout d'abord construit dans le cadre du modèle de Cox, puis il est étendu à d'autres modèles plus complexes à risques non-proportionnels. Des simulations montrent que l'indice est peu affectée par la taille de l'échantillon et la censure. Il présente de plus une meilleure séparabilité que les indices classiques de la littérature. L'intérêt de l'indice est illustré sur deux exemples. Le premier consiste à identifier des marqueurs génomiques communs à différents types de cancers. Le deuxième, dans le cadre d'une étude sur le cancer broncho-pulmonaire, montre l'intérêt de l'indice pour sélectionner des facteurs génomiques entraînant un croisement des fonctions de risques instantanés pouvant être expliqué par un effet « modulateur » entre les marqueurs. En conclusion, l'indice proposé est un outil prometteur pouvant aider les chercheurs à identifier des listes de gènes méritant des études plus approfondies.

Mots clés : Analyse de survie, Génomique, Oncologie, Pseudo-R^2

Abstract : Development of a separability index for genomic data in survival analysis

In oncogenomics research, one of the main objectives is to identify new genomic markers so as to construct predictive rules in order to classify patients according to time-to-event outcomes (death or tumor relapse). Most of the studies dealing with such high throughput data usually rely on a selection process in order to identify, among the candidates, the markers having a prognostic impact. A common problem among biologists is the choice of the selection rule. In survival analysis, classical procedures consist in ranking genetic markers according to either the estimated hazards ratio or quantities derived from a test statistic (p-value, q-value). However, these methods are not suitable for gene selection across multiple genomic datasets with different sample sizes.

Using an index taking into account the magnitude of the prognostic impact of factors without being highly dependent on the sample size allows to address this issue. In this work, we propose a novel index of predictive ability for selecting genomic markers having a potential impact on time-to-event outcomes. This index extends the notion of "pseudo-R^2" in the framework of survival analysis. It possesses an original and straightforward interpretation in terms of "separability". The index is first derived in the framework of the Cox model and then extended to more complex non-proportional hazards models. Simulations show that our index is not substantially affected by the sample size of the study and the censoring. They also show that its separability performance is higher than indices from the literature. The interest of the index is illustrated in two examples. The first one aims at identifying genomic markers with common effects across different cancer types. The second shows, in the framework of a lung cancer study, the interest of the index for selecting genomic factor with crossing hazards functions, which could be explained by some "modulating" effects between markers. The proposed index is a promising tool, which can help researchers to select a list of features of interest for further biological investigations.

Key words : Survival Analysis, Genomics, Oncology, Pseudo-R^2

Liste des travaux relatifs à la thèse

Publications

(1) **S. Rouam, T. Moreau and P. Broët**. Identifying common prognostic factors in genomic cancer studies : A novel index for censored outcomes. *BMC Bioinformatics*, 11(1) :150, 2010.
(2) **S. Rouam, T. Moreau and P. Broët**. A pseudo-R^2 measure for selecting genomic markers with crossing hazard functions *BMC Medical Research Methodology*, 11(1) :28, 2011.
(3) **S. Rouam, T. Moreau and P. Broët.** . A note on crossing hazard functions in survival models. En préparation.

Posters

(1) **S. Rouam**. Identifying common prognostic factors in genomic cancer studies : A novel discrimination index for survival data. *Singapore Symposium on Computational Biology*, 8 septembre 2009, A*Star, Singapore.

Table des matières

Chapitre 1

INTRODUCTION

Le contexte de la génomique à haut débit

L'apparition, à la fin du siècle dernier des biotechnologies de génomique dites « à haut débit », représente une nouvelle source d'information pour l'étude des pathologies humaines. On distingue très schématiquement la **génomique structurale**, qui s'intéresse à la structure du génome, et la **génomique fonctionnelle** dont l'objectif est de déterminer la fonction des gènes. La génomique structurale peut se définir comme la connaissance complète des génomes, tant en ce qui concerne le nombre et l'organisation spatiale sur les chromosomes des gènes qui les constituent, que leur séquence chimique et les produits cellulaires qui résultent de leur fonctionnement. La génomique structurale englobe les techniques de cartographie, de séquençage et d'annotation du génome ainsi que la détermination de la structure tridimensionnelle des protéines. La génomique fonctionnelle s'intéresse, quant-à-elle, à la connaissance des mécanismes régulateurs des gènes et l'étude de leur fonctionnement intégré dans la cellule et l'organisme.

Les outils technologiques, qui sont actuellement plus communément utilisés en génomique fonctionnelle, sont les **« puces à ADN »** (Acide DésoxyriboNucléique), de part leur efficacité et la diversité de leurs champs d'application. L'apparition des premières « puces à ADN » remonte à une quinzaine d'années, la première puce ayant été commercialisée en 1994 par la société Affymetrix (*www.affymetrix.com*). D'une manière générale, une puce comporte plusieurs dizaines de milliers d'unités d'hybridation ("spots" en anglais), chacune de ces sondes étant constituée d'un court fragment d'oligonucléotides (par dépôt ou par synthèse *in situ*) correspondant à des séquences données (cible). Les sondes sont déposées/fixées/synthétisées sur un support solide selon une disposition ordonnée. Le fonctionnement des puces repose sur le principe d'hybridation entre la sonde et la cible, qui est, le plus souvent, marquée par une molécule fluorescente, permettant de détecter et de quantifier l'ensemble des cibles présentes (ADN ou ARN, Acide Ribonucléique) en une seule expérience.

On distingue différents types de puces en fonction du support, de la densité des puits, de la nature des sondes, de la méthode d'hybridation (simple ou compétitive).

Une des premières applications du principe des puces, et encore largement dominante à l'heure actuelle, est l'**analyse du transcriptome**. L'objectif est la détection de la présence, dans une cellule ou un organisme, des ARN messagers (ARNm). Outre l'analyse de l'expression des gènes, les puces à ADN sont également utilisées dans d'autres domaines, dont en particulier le génotypage avec l'identification de polymorphismes génétiques ponctuels (ou **SNP**, Single Nucleotide Polymorphism, (Hacia *et al.*, 1998)) ; la recherche de **variants de nombre**, comme la détection des variations du nombre de copies de l'ADN (amplifications et délétions de régions chromosomiques constitutionnelles ou tumorales). Plus récemment, de nouvelles applications voient le jour, comme par exemple les **ChIP-on-Chip** (Chromatin-ImmunoPrecipitation on Chip), qui combinent les principes d'immuno-précipitation de la chromatine et des puces à ADN et visent à étudier l'interaction entre ADN et protéines, typiquement les facteurs de transcription.

L'oncogénomique

En cancérologie, les puces à ADN sont actuellement de plus en plus largement utilisées. Les études en oncogénomique portent sur le génome tumoral et/ou le génome constitutionnel et recherchent les **facteurs de susceptibilité ou d'évolution** de la maladie. Ainsi, les études d'association pan-génomiques (GWAS, Genome Wide Assocation Studies) centrées sur l'ADN constitutionnel visent à identifier des génotypes associés à l'augmentation ou la diminution du risque d'apparition de certains cancers. L'utilisation des puces à ADN centrées sur l'ADN tumoral permet également l'identification de profils d'aberrations chromosomiques et de modifications transcriptionnelles. La technologie de type CGH-array (ou hybridation génomique compétitive) permet d'identifier des variations du nombre de copies de l'ADN tumoral (i.e. des délétions ou amplifications). Ces altérations génomiques peuvent être associées à des modifications d'expressions de gènes « clés » de la cellule (amplification d'oncogènes, délétion de gènes suppresseurs de tumeurs).

L'objectif de l'oncogénomique est d'apporter de nouveaux éléments favorisant une meilleure compréhension de la biologie de la progression tumorale, conduisant potentiellement au développement de nouvelles stratégies diagnostiques, pronostiques et prédictives (réponse des patients à la thérapie). Le transfert des outils de la génomique en clinique représente l'un des défis de la médecine de demain. Lors des cinq dernières années, de très nombreux marqueurs génomiques ont été proposés en cancérologie et une minorité d'entre eux a déjà été implémenté en clinique (e.g. Oncotype DX®, MammaPrint®). Il est hautement probable que ce type d'approche sera l'un des enjeux des prochaines années.

Malgré un accroissement des connaissances dans le domaine de la biologie du cancer, de nombreux mécanismes restent encore inconnus. L'un des axes de recherche actuel en cancérologie génomique est **l'identification de nouveaux marqueurs moléculaires** afin de construire des règles prédictives visant à classer les patients selon le risque d'apparition d'un événement d'intérêt (décès ou récidive tumorale). En présence de telles données de haute dimension, une première

étape de **sélection** parmi l'ensemble des variables candidates est généralement employée afin d'identifier les marqueurs liés à un critère de jugement principal. En oncologie, ce critère est souvent le délai d'apparition d'un événement.

Les données censurées

En cancérologie, l'**analyse de survie** est fréquemment utilisée pour relier le délai d'apparition d'un évènement dans la population étudiée à des variables explicatives (e.g. les biomarqueurs génétiques). L'analyse de survie est apparue au XVII$^{\text{ème}}$ siècle dans le domaine de la démographie et des sciences actuarielles. Son utilisation dans d'autres domaines tels que la physique, l'industrie, les sciences médicales n'est apparue qu'au XX$^{\text{ème}}$ siècle. Comme son nom l'indique, l'analyse de survie vise, à l'origine, à étudier la survie d'un ensemble de patients, c'est-à-dire le taux de mortalité. A l'heure actuelle, l'analyse de survie a une définition plus large et désigne l'**analyse du temps d'apparition** de tout type d'événement. L'analyse des données de survie a deux principales particularités : la première est de ne concerner que des variables aléatoires **positives**, la deuxième est la présence de données incomplètes car possiblement **censurées**. Des méthodes spécifiques ont donc été développées pour analyser ce type de données.

Les avancées majeures dans ce domaine ont vu le jour à partir des années cinquante. En 1951, **Weibull** conçoit un modèle paramétrique dans le domaine de la fiabilité (Weibull, 1951). A cet effet, il propose une nouvelle distribution de probabilité qui sera par la suite fréquemment utilisée en analyse de la survie : la «loi de Weibull ». En 1958, **Kaplan et Meier** présentent d'importants résultats concernant l'estimation non-paramétrique de la fonction de survie (Kaplan et Meier, 1958). En 1972, **Cox** introduit un modèle statistique semi-paramétrique permettant de prendre en compte, dans la modélisation de la fonction de risque, des variables explicatives (Cox, 1972). Il définit également la notion de vraisemblance partielle. Cette approche a fait l'objet de développements méthodologiques majeurs et a servi de cadre théorique pour le développement de nombreux autres modèles. En outre, de nouveaux développements dans le cadre de la théorie des martingales et des processus de comptage ont fait l'objet de travaux lors des deux dernières décennies (Fleming et Harrington, 2005).

En oncogénomique, l'analyse de survie permet d'étudier la **relation entre le risque d'apparition d'un événement et des modifications génomiques**. Les méthodes classiquement utilisées pour sélectionner les biomarqueurs liées à ce risque consistent à ordonner les gènes en fonction d'une mesure basée sur le risque relatif et/ou sur le degré de signification (p-value associée à un test statistique), et à choisir un seuil permettant de déterminer le sous-ensemble de gènes d'intérêt. Une question récurrente pour les biologistes est **la méthode de sélection**. Dans le cadre de la comparaison de deux groupes, on voit très fréquemment l'utilisation de règles heuristiques combinant effet biologique minimum (log supérieur à 2) et significativité (p-value inférieure à 0.001). Dans le cas des données censurées, l'utilisation du risque relatif estimé comme mesure d'effet ne tient pas compte de la variabilité et le degré de signification dépend fortement

de la taille de l'échantillon pénalisant une analyse combinant des études de tailles différentes. Utiliser une mesure tenant compte à la fois de la **variabilité des données et ne dépendant que faiblement de la taille de l'échantillon**, présente un intérêt majeur, notamment pour comparer, voire combiner, les résultats issus d'études hétérogènes de tailles différentes produites par des groupes distincts.

Indices de capacité de prédiction

Les mesures de **capacité de prédiction** permettent de répondre à cette problématique. Elles visent à déterminer la capacité d'une ou plusieurs variables explicatives à prédire la variable réponse et permettent ainsi d'évaluer la contribution de variables pronostiques au modèle. Dans le **modèle linéaire**, la mesure de capacité de prédiction la plus utilisée est le **coefficient de détermination** ou $\mathbf{R^2}$. Ce dernier est défini comme le pourcentage de variation expliquée par le modèle. Il permet à la fois de mesurer la qualité d'ajustement du modèle et également de quantifier la force de la relation entre la ou les variables explicatives et la variable à expliquer. Comme le souligne Magee (1990), il peut être interprété de différentes façons : comme un pourcentage de variance expliqué, mais également comme le carré du coefficient de corrélation, comme une fonction de la vraisemblance (et du score). Dans le modèle linéaire, toutes ces quantités sont liées. Dans des **modèles plus complexes**, comme le modèle logistique et l'analyse de survie, ce n'est pas nécessairement le cas et la transposition des différentes interprétations de la notion de R^2 n'est pas aisément réalisable. Ainsi il n'y a pas de consensus sur la façon de calculer l'équivalent du R^2, ou **pseudo-$\mathbf{R^2}$**, dans le cadre de la survie et de nombreuses mesures ont été proposées.

Dans ce travail, nous proposons une **nouvelle mesure de capacité de prédiction** afin de sélectionner des marqueurs génomiques ayant un impact pronostique sur le délai de survenue d'un évènement. En pratique, l'indice proposé présente une interprétation originale et intuitive en terme de capacité d'un marqueur génomique à séparer les patients en fonction de leur temps de survie et de leurs mesures d'expression génétique. L'indice est compris entre 0 et 1 et sa valeur augmente lorsque l'effet du gène augmente.

Pour écrire notre pseudo-R^2, plusieurs modèles sont considérés. Dans un premier temps, notre indice est construit à partir du **modèle de Cox à risques proportionnels**. Comme pour toute modélisation, bien que l'hypothèse des risques proportionnels soit une simplification de la réalité, ce modèle est largement employé car il permet de résumer l'effet moyen d'une variable sur la fonction de risque de base, et une inférence simple peut être obtenue à partir de la vraisemblance partielle. Dans ce travail, nous considérons également le **modèle à odds proportionnels**. Il constitue une alternative au modèle de Cox avec variables dépendantes du temps dans des situations où l'effet de la covariable diminue au cours du temps. Enfin, nous nous intéressons à un **modèle particulier** visant à décrire l'effet de **facteurs entraînant un croisement des fonctions de risques instantanés** et ne pouvant être modélisé par les deux modèles évoqués précédemment.

Plan de la thèse

Dans le chapitre 2, nous rappelons la définition du R^2 dans le cadre du modèle linéaire simple et généralisé. Dans le chapitre 3, nous présentons les principales mesures de capacité de prédiction proposées en analyse de survie. Le chapitre 4 est composé de deux parties. La première vise à décrire les différents modèles étudiés, et la deuxième est consacrée à la présentation de l'indice sous ces modèles et à l'étude de ses propriétés. Le chapitre 5 expose les schémas et résultats de simulations visant à étudier le comportement de l'indice sous les différents modèles de survie. Le chapitre 6 décrit deux exemples d'application de l'indice pour la sélection de variables génétiques. Le premier s'inscrit dans le cadre du modèle de Cox et permet d'identifier des facteurs génétiques communs à divers types de cancers solides. Le deuxième montre l'intérêt de l'indice pour sélectionner des gènes aux risques qui se croisent. Le chapitre 7 discute la méthode présentée et conclut sur le travail effectué.

Chapitre 2

RAPPELS SUR LE R^2 DANS LE MODÈLE LINÉAIRE GÉNÉRALISÉ

Contenu

Dans ce chapitre, la notion de coefficient de détermination (R^2) est introduite dans le cadre le plus simple, la régression linéaire. Le R^2 estime la fraction de la variance (dispersion) qui est expliquée par une ou plusieurs variables explicatives $\mathbf{Z}$ dans un modèle de régression linéaire.

La définition du R^2 dans le modèle linéaire est tout d'abord rappelée ; puis, les généralisations du R^2 pour le modèle logistique sont présentées.

2.1 R^2 dans le modèle de régression linéaire

Dans un premier temps, le coefficient de détermination dans le cadre du modèle linéaire simple sans faire d'hypothèse particulière sur la distribution des termes d'erreurs est présenté. Dans un second temps, les extensions de la définition du R^2 dans le cas d'erreurs de type gaussiennes sont exposées.

2.1.1 Cas de la régression linéaire sans hypothèse gaussienne

a. Rappel de la définition du modèle de régression linéaire

Définition 2.1 *Le **modèle de régression linéaire** (multiple) standard est défini par la relation suivante :*

$$Y = \mathbf{Z}\beta + \epsilon \tag{2.1}$$

où Y est le vecteur $(n \times 1)$ de variables à expliquer ou variables réponses, $\mathbf{Z}$ la matrice $(n \times (p+1))$ de variables explicatives ou régresseurs, β le vecteur $((p+1) \times 1)$ des paramètres de la régression et ϵ le vecteur $(n \times 1)$ d'erreurs. Leurs expressions peuvent être présentées sous forme matricielle comme suit :

$$\underset{n\times 1}{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \underset{n\times(p+1)}{\mathbf{Z}} = \begin{pmatrix} 1 & z_{11} & \cdots & z_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & z_{n1} & \cdots & z_{np} \end{pmatrix} \quad \underset{(p+1)\times 1}{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \quad \underset{n\times 1}{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

On suppose que le terme d'erreur ϵ a une moyenne nulle et une matrice de variance-covariance $\sigma^2\mathbf{I}_n$ ($\mathbf{I}_n$ est la matrice identité de dimension $n \times n$).

Dans un premier temps, on ne fait aucune hypothèse sur la distribution du terme d'erreur ϵ.

b. Définition du coefficient de détermination

On note $\bar{y} = \sum_{i=1}^{n} \frac{y_i}{n}$ la moyenne empirique des y_i , $\bar{Y}$ le n-vecteur $(\bar{y}, \cdots, \bar{y})^T$ et $\widehat{Y}$ le n-vecteur des valeurs prédites de Y.

Définition 2.2 *Le* $\boldsymbol{R^2}$ *ou* ***coefficient de détermination*** *est défini par la relation suivante :*

$$\boxed{R^2 = 1 - \frac{SCE}{SCT} = \frac{SCM}{SCT}}$$

$$\begin{aligned} \text{avec } SCE &= \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \left\|Y - \widehat{Y}\right\|^2 \\ SCM &= \sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2 = \left\|\widehat{Y} - \bar{Y}\right\|^2 \\ SCT &= \sum_{i=1}^{n}(y_i - \bar{y})^2 = \left\|Y - \bar{Y}\right\|^2 \end{aligned} \tag{2.2}$$

SCE est la somme des carrés des erreurs, SCT la somme des carrés totaux et SCM la somme des carrés du modèle.

c. Représentation géométrique

La figure 2.1 donne la **représentation géométrique** des différentes sommes de carrés dans le modèle linéaire.

On note $M(Z)$ le sous-espace de $\mathbb{R}^n$ engendré par les $p+1$ vecteurs colonne de $\mathbf{Z}$, souvent appelé espace image ou espace des solutions. Soit $M^{\perp}(Z)$ l'espace des résidus orthogonal à $M(Z)$. D'après le théorème de Pyhtagore, on a :

$$||Y - \bar{Y}||^2 = ||\widehat{Y} - \bar{Y}||^2 + ||\widehat{\epsilon}||^2 \quad \Longleftrightarrow \quad SCT = SCM + SCE$$

La variabilité totale ($||Y - \bar{Y}||^2$) est égale a la somme de la variabilité expliquée par le modèle ($||\widehat{Y} - \bar{Y}||^2$) et de la variabilité résiduelle ($||\widehat{\epsilon}||^2$).
Si la constante ne fait pas partie du modèle, i.e. $\mathbf{Z}$ est de dimension $(n \times p)$, le théorème de Pythagore devient :

$$||Y||^2 = ||\widehat{Y}||^2 + ||\widehat{\epsilon}||^2 = ||\mathbf{Z}\widehat{\beta}||^2 + ||Y - \mathbf{Z}\widehat{\beta}||^2$$

Dans un modèle sans ordonnée à l'origine, il faut donc prendre $SCM = ||\widehat{Y}||^2$ et $SCT = ||Y||^2$ pour que la décomposition en sommes de carrés ($SCT = SCM + SCE$) soit vérifiée.

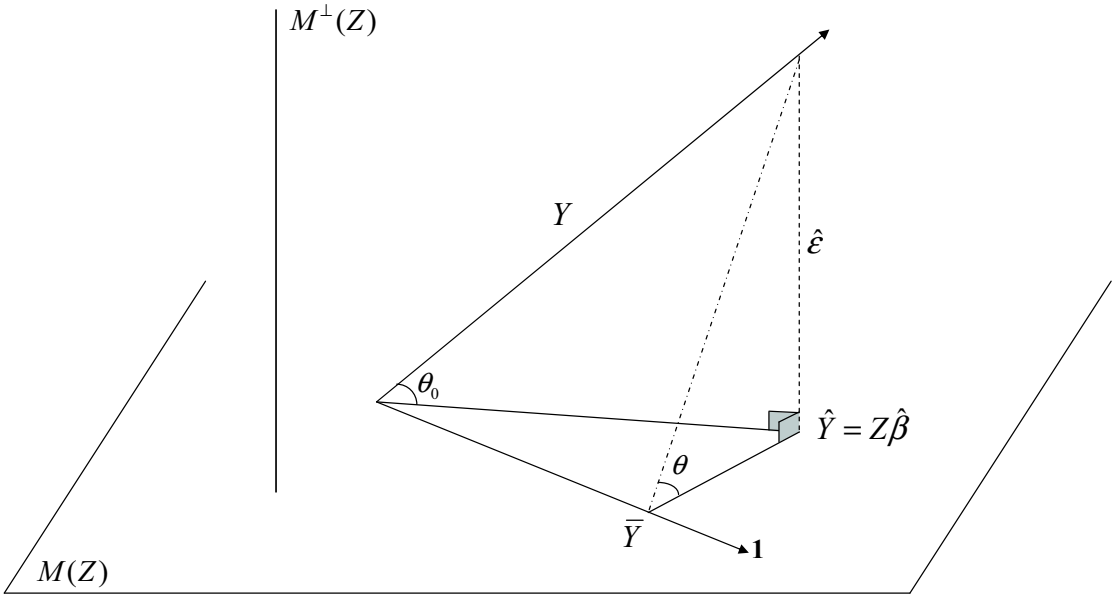
Dans ce cadre, le coefficient de détermination s'interprète comme le **cosinus carré de l'angle θ entre Y et $\widehat{Y}$** pris en $\bar{Y}$. En l'absence de constante, il est égal au cosinus carré de l'angle θ_0 entre Y et $\widehat{Y}$ pris à l'origine.

d. Caractéristiques du $\mathbf{R^2}$

Le coefficient de détermination est compris entre 0 et 1.

La valeur du R^2 augmente lorsque le nombre p de prédicteurs augmente. Son utilisation pour comparer des modèles avec un nombre de variables différentes n'est donc pas approprié, comme le souligne Healy (1984).

FIGURE 2.1 – Représentation des sommes de carrés dans le modèle linéaire



Pour remédier à ce problème, un **coefficient de détermination ajusté** a été proposé. Dans le cadre du modèle linéaire, le R^2 ajusté est une modification du R^2 qui tient compte du nombre de variables explicatives. Contrairement au R^2 non ajusté, le R^2 ajusté augmente uniquement si la nouvelle variable améliore la prédiction du modèle.

Le R^2 ajusté est défini de la manière suivante :

$$\boxed{\mathrm{R}^2_{\mathrm{adj}} = 1 - \frac{SCE/(n-p-1)}{SCT/(n-1)} = 1 - (1-\mathrm{R}^2)\frac{n-1}{n-p-1}} \tag{2.3}$$

Le R^2 ajusté peut être négatif et est toujours inférieur au R^2. Son interprétation est donc différente de celle du R^2 non ajusté.

e. Interprétations du coefficient de détermination

Le coefficient de détermination peut être interprété de différentes manières.

• La plus simple et plus communément utilisée consiste à considérer le R^2 comme une mesure de **pourcentage de variation expliquée** par la ou les covariables incluses dans le

modèle. En effet, le R^2 peut s'écrire sous la forme :

$$\boxed{R^2 = \frac{\mathbb{V}[\mathbb{E}(Y|Z)]}{\mathbb{E}[\mathbb{V}(Y|Z)] + \mathbb{V}[\mathbb{E}(Y|Z)]} = \frac{\mathbb{V}(Y) - \mathbb{E}[\mathbb{V}(Y|Z)]}{\mathbb{V}(Y)} = 1 - \frac{\mathbb{E}[\mathbb{V}(Y|Z)]}{\mathbb{V}(Y)}} \tag{2.4}$$

L'interprétation en terme de pourcentage de variation expliquée ne peut se faire que conditionnellement à un modèle. Sous cette forme, le coefficient de détermination permet de décrire la réduction de la variance de Y en passant de la distribution marginale à la distribution conditionnelle sachant Z. Il constitue alors un moyen de quantifier l'amélioration de la prédiction par l'ajout d'une covariable dans un modèle donné par rapport au modèle nul, i.e. sans covariables.

• Une autre interprétation est également parfois utilisée et correspond à l'utilisation du coefficient de détermination comme une mesure d'**ajustement du modèle**, ce qu'en anglais on désigne par «goodness of fit ». Le R^2 permet alors de quantifier l'adéquation du modèle aux données.

Dans ce cadre, le R^2 peut être relié à la notion de **perte** introduite par Korn et Simon (1991) et définie ci-après.
Soit $L(y, \widetilde{y})$ la perte encourue en faisant la prédiction $\widetilde{y}$ de la vraie valeur observée y de la variable aléatoire Y. Par exemple, la fonction de perte la plus couramment utilisée est l'erreur au carré $||y - \widetilde{y}||^2$.
En présence d'une covariable Z, la perte attendue est $\int L(y, \widetilde{y}) dF(y|z)$, où F est la fonction de répartition de Y conditionnellement à Z. Le risque $R(z)$ est défini comme la perte minimale atteinte pour $\widetilde{y} = \widetilde{y}(z)$, et vaut $R(z) = \min_{\widetilde{y}} \int L(y, \widetilde{y}) dF(y|z)$.
En l'absence de covariables, la perte attendue est $\int L(y, \widetilde{y}) dF_0(y)$, où F_0 peut être décomposée comme un mélange de ses composantes, telle que $F_0(y) = \frac{1}{n} \sum_{i=1}^{n} F(y|z_i)$. Le minimum est atteint pour $\widetilde{y} = \widetilde{y}_0$. Le risque vaut alors $R_0 = \min_{\widetilde{y}} \int L(y, \widetilde{y}) dF_0(y)$.
Le coefficient de détermination peut alors s'exprimer comme

$$\boxed{R^2 = \frac{\displaystyle\sum_{i=1}^{n} L(y_i, \widehat{y}_0) - \sum_{i=1}^{n} L(y_i, \widehat{y}(z_i))}{\displaystyle\sum_{i=1}^{n} L(y_i, \widehat{y}_0)}} \tag{2.5}$$

où $\widehat{y}(z_i)$ et $\widehat{y}_0$ sont les prédicteurs respectifs de $\widetilde{y}(z_i)$ et de $\widetilde{y}_0$, et en prenant $L(y, \widetilde{y}) = ||y - \widetilde{y}||^2$. Sous cette forme, le R^2 permet de quantifier la perte relative encourue en faisant la prédiction basée sur l'utilisation des covariables dans le modèle, par rapport à la prédiction ne faisant pas intervenir les covariables dans le modèle.

Il existe, dans la littérature, une certaine confusion entre les deux interprétations, en terme de pourcentage de variance expliquée et d'ajustement au modèle, car, en réalité, la notion de coefficient de détermination englobe les deux concepts dans le cadre du modèle linéaire (voir Korn et Simon, 1991).

2.1.2 Cas du modèle de régression linéaire gaussien

Ce paragraphe montre qu'en faisant l'hypothèse de distribution gaussienne sur les termes d'erreur, le **R^2 peut s'écrire sous différentes formes**. Il peut être relié à d'autres quantités, comme le coefficient de corrélation ainsi que de la divergence de Kullback-Leibler. Magee (1990) montre également qu'il peut être relié aux statistiques de Fisher, de Wald, du log-rapport de vraisemblance et du score pour tester l'hypothèse nulle $\mathcal{H}_0 : \{\beta_1 = \cdots = \beta_p = 0\}$ (i.e. tous les paramètres de la régression sont nuls à l'exception de l'ordonnée à l'origine).

a. Rappels dans le cadre du modèle linéaire avec erreurs gaussiennes

Dans ce paragraphe, on considère le modèle linéaire défini p. 22. De plus, on suppose que les termes d'erreur ϵ_i sont iid (indépendants et identiquement distribués), de distribution normale de moyenne nulle et de matrice de variance-covariance $\sigma^2\mathbf{I}_n$, soit $\epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I}_n)$.

Dans ce cas, la densité de probabilité des y_i, $i = 1, \cdots, n$ est

$$f(y_i, \beta, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(y_i - \beta_0 - \sum_{j=1}^{p} z_{ij}\beta_j)^2}{2\sigma^2}\right\}$$

La **vraisemblance** des observations est

$$\mathcal{L}(y, \beta, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2} \|Y - \mathbf{Z}\beta\|^2\right\}$$

où $\|Y - \mathbf{Z}\beta\|^2 = (Y - \mathbf{Z}\beta)^T(Y - \mathbf{Z}\beta)$.

La **log-vraisemblance** s'écrit alors

$$\log \mathcal{L}(y, \beta, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \|Y - \mathbf{Z}\beta\|^2$$

Le vecteur des **dérivées**, par rapport aux $\beta_j, j = 1, \cdots, p$, **de la log-vraisemblance** se déduit de ce qui précède :

$$U(\beta) = \nabla \log \mathcal{L}(y, \beta, \sigma^2) = \left(\frac{\partial \log \mathcal{L}(y, \beta, \sigma^2)}{\partial \beta_1}, \cdots, \frac{\partial \log \mathcal{L}(y, \beta, \sigma^2)}{\partial \beta_p}\right)^T = \frac{1}{\sigma^2}\mathbf{Z}^T(Y - \mathbf{Z}\beta)$$

ainsi que

$$\frac{\partial \log \mathcal{L}(y, \beta, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \|Y - \mathbf{Z}\beta\|^2$$

Les **estimateurs des paramètres** de la régression $\beta_j, j = 1, \cdots, p$ et de la variance σ^2 sont alors

$$\widehat{\beta} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T Y$$
$$\widehat{\sigma}^2 = \frac{||Y - \mathbf{Z}\widehat{\beta}||^2}{n} = \frac{||Y - \widehat{Y}||^2}{n} = \frac{||\widehat{\epsilon}||^2}{n} = \frac{SCE}{n}$$

La dérivée seconde de la log-vraisemblance permet d'obtenir l'**information de Fisher** ainsi que son inverse, la **variance des** β :

$$I(\beta) = -\mathbb{E}\left[\frac{\partial^2 \log \mathcal{L}(y, \beta, \sigma^2)}{\partial \beta^2}\right] = \mathbb{V}(\beta)^{-1}$$

b. Lien entre le R^2 et le coefficient de corrélation multiple

Dans le cadre de la régression linéaire avec erreurs gaussiennes, le coefficient de détermination est relié au **coefficient de corrélation**. Dans le cas d'une seule variable explicative et avec une estimation des β par la méthode des moindres carrés, le R^2 est exactement égal au carré du coefficient de corrélation de Pearson entre la variable réponse et le régresseur, qui s'écrit de la manière suivante :

$$r = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(z_i - \bar{z}) \sum_{i=1}^{n}(y_i - \bar{y})}}$$

Dans le cadre multivarié, on peut définir un coefficient de corrélation multiple. Pour ce faire, il est plus pratique, dans le modèle (2.1), de séparer l'effet moyen des autres variables de la matrice $\mathbf{Z}$:

$$Y = \beta_0 \mathbf{1} + \mathbf{Z}\beta + \epsilon$$

où $\mathbf{1}$ est la matrice colonne unité $(p \times 1)$, $\mathbf{Z}$ est une matrice $(n \times p)$ et β le vecteur colonne de dimension $(p \times 1)$ des paramètres.
Sans perte de généralité, on peut supposer que les colonnes de $\mathbf{Z}$ sont centrées de moyenne nulle. Le modèle s'écrit alors (Mardia *et al.*, 1979)

$$Y - \bar{Y} = \mathbf{Z}\beta + \omega$$

avec $\omega = \epsilon - \epsilon^T \mathbf{1}$ qui dénote le vecteur centré des erreurs.
Soit

$$S = \begin{pmatrix} s_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} = \frac{1}{n} \begin{pmatrix} Y^T - \bar{Y}^T \\ \mathbf{Z}^T \end{pmatrix} (Y - \bar{Y}; \mathbf{Z})$$

la matrice de covariance de Y et $\mathbf{Z}$. Le coefficient de corrélation multiple est défini par

$$r = \left(\frac{S_{12} S_{22}^{-1} S_{21}}{s_{11}} \right)^{1/2}$$

Le coefficient de détermination est alors égal au carré du coefficient de corrélation multiple.

Dans les cas univarié et multivarié, on a donc la relation suivante :

$$\boxed{R^2 = r^2} \tag{2.6}$$

Preuve. Dans le cas multivarié, les composantes du coefficient de corrélation s'écrivent :

$$s_{11} = ||Y - \bar{Y}||^2 \quad \text{et} \quad S_{12} S_{22}^{-1} S_{21} = (Y - \bar{Y})^T \mathbf{Z} (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T (Y - \bar{Y})$$

D'autre part, on a

$$\begin{aligned} SCE &= ||Y - \bar{Y} - \widehat{Y}||^2 \\ &= ||Y - \bar{Y} - \mathbf{Z}\widehat{\beta}||^2 \\ &= (Y - \bar{Y})^T(Y - \bar{Y}) - (Y - \bar{Y})^T\mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T(Y - \bar{Y}) \\ &= ||Y - \bar{Y}||^2 \left(1 - \frac{(Y - \bar{Y})^T\mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T(Y - \bar{Y})}{||Y - \bar{Y}||^2}\right) \\ &= ||Y - \bar{Y}||^2 \left(1 - r^2\right) \\ &= SCT(1 - r^2) \end{aligned}$$

D'où

$$r^2 = 1 - \frac{SCE}{SCT} = R^2$$

Le cas univaré est un cas particulier du cas multivarié. La relation est donc également vérifiée. □

c. Lien avec la statistique de Fisher

La statistique de Fisher F visant à tester l'hypothèse nulle $\mathcal{H}_0$ s'écrit comme suit

$$F = \frac{SCM/(p)}{SCE/(n - p - 1)}$$

Elle suit un loi de Fisher à p et $n - p - 1$ degrés de liberté. On a alors

$$F = \frac{R^2/p}{(1 - R^2)/(n - p - 1)}$$

Par conséquent, le R^2 peut s'écrire comme une **fonction de la statistique de Fisher** :

$$\boxed{R^2 = \frac{nF}{n - p - 1 + nF}} \tag{2.7}$$

d. Lien avec la statistique de Wald

La statistique F peut être reliée à la statistique de Wald W pour tester $\mathcal{H}_0$:

$$W = \frac{n(SCM)}{SCE} = \frac{npF}{n - p - 1}$$

et donc la relation entre le R^2 et **la statistique de Wald** est

$$\boxed{R^2 = \frac{W}{W + n}} \tag{2.8}$$

Preuve. La statistique de Wald testant $\mathcal{H}_0$ est définie par la relation suivante :

$$W = (\widehat{\beta} - \beta^*)^T\widehat{\mathbb{V}}(\widehat{\beta})^{-1}(\widehat{\beta} - \beta^*) \qquad \text{avec} \quad \underset{(p+1)\times 1}{\beta^*} = \begin{pmatrix} \bar{y} \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

La variance des coefficients $\widehat{\beta}$ s'écrit

$$\mathbb{V}(\widehat{\beta}) = \sigma^2(\mathbf{Z}^T\mathbf{Z})^{-1}$$

Un estimateur est donné par

$$\widehat{\mathbb{V}}(\widehat{\beta}) = \widehat{\sigma}^2(\mathbf{Z}^T\mathbf{Z})^{-1} \quad \text{avec} \quad \widehat{\sigma}^2 = \frac{1}{n}||Y - \widehat{Y}||^2 = \frac{1}{n}SCE$$

La statistique de Wald devient

$$\begin{aligned} W &= \frac{1}{\widehat{\sigma}^2}(\widehat{\beta} - \beta^*)^T(\mathbf{Z}^T\mathbf{Z})(\widehat{\beta} - \beta^*) \\ &= \frac{1}{\widehat{\sigma}^2}\left[\widehat{\beta}^T\mathbf{Z}^T\mathbf{Z}\widehat{\beta} - \widehat{\beta}^T(\mathbf{Z}^T\mathbf{Z})\beta^* - \beta^*(\mathbf{Z}^T\mathbf{Z})\widehat{\beta} + \beta^{*T}(\mathbf{Z}^T\mathbf{Z})\beta^*\right] \\ &= \frac{1}{\widehat{\sigma}^2}\left[\widehat{\beta}^T\mathbf{Z}^T\mathbf{Z}\widehat{\beta} - Y^T\mathbf{Z}\beta^* - \beta^{*T}\mathbf{Z}^TY - \beta^{*T}(\mathbf{Z}^T\mathbf{Z})\beta^*\right] \end{aligned}$$

En développant, on montre que

$$Y^T\mathbf{Z}\beta^* = \beta^{*T}\mathbf{Z}^TY = \beta^{*T}(\mathbf{Z}^T\mathbf{Z})\beta^* = n\bar{y}^2$$

On en déduit donc que

$$W = \frac{||\mathbf{Z}\widehat{\beta} - \bar{Y}||^2}{\widehat{\sigma}^2} = n\frac{SCM}{SCE} = n\frac{R^2}{1 - R^2}$$

Par conséquent,

$$R^2 = \frac{W}{W + n}$$

□

e. Lien avec la statistique du rapport de vraisemblance

Dans le modèle (2.1), la relation entre la définition initiale du R^2 et la statistique du rapport de vraisemblance s'écrit :

$$\boxed{R^2 = 1 - \exp\left\{-\frac{LR}{n}\right\}} \tag{2.9}$$

Preuve. Nous avons vu précédemment que la vraisemblance du modèle linéaire s'écrit

$$\mathcal{L}(y, \beta, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}||Y - \mathbf{Z}\beta||^2\right\}$$

Cette vraisemblance est maximale lorsque $\beta = \widehat{\beta}$ est l'estimateur du maximum de vraisemblance ou des moindres carrés et que $\sigma^2 = \widehat{\sigma}^2_U = ||Y - \mathbf{Z}\widehat{\beta}||^2/n = SCE/n$. Donc

$$\begin{aligned} \sup_{\beta,\sigma^2} \mathcal{L}(y, \beta, \sigma^2) &= \left(\frac{n}{2\pi||Y - \mathbf{Z}\widehat{\beta}||^2}\right)^{n/2} e^{-n/2} \\ &= \left(\frac{n}{2\pi SCE}\right)^{n/2} e^{-n/2} \\ &= \mathcal{L}(y, \widehat{\beta}, \widehat{\sigma}^2_U) = \mathcal{L}_1 \end{aligned}$$

Sous l'hypothèse $\mathcal{H}_0$, nous obtenons de manière équivalente :

$$\begin{aligned}\sup_{\beta,\sigma^2} \mathcal{L}_R(y,\beta,\sigma^2) &= \left(\frac{n}{2\pi SCE_R}\right)^{n/2} e^{-n/2} \\ &= \mathcal{L}_R(y,\widehat{\beta}_0,\widehat{\sigma}^2_R) = \mathcal{L}_0\end{aligned}$$

où SCE_R correspond à la somme des carrés résiduels sous $\mathcal{H}_0$, c'est-à-dire $SCE_R = ||y - \widehat{\beta}_0||^2 = ||Y - \bar{Y}||^2 = SCT$ et $\widehat{\sigma}^2_R = SCE_R/n$. La statistique du rapport de vraisemblance qui teste $\mathcal{H}_0$ est défini par

$$LR = 2\log\left(\frac{\mathcal{L}_1}{\mathcal{L}_0}\right)$$

Donc

$$LR = 2\log\left(\frac{\mathcal{L}_1}{\mathcal{L}_0}\right) = n\log\left(\frac{SCT}{SCE}\right)$$

D'où on déduit

$$R^2 = 1 - \exp\left\{-\frac{LR}{n}\right\}$$

□

f. Lien avec la statistique du score

Le lien entre le R^2 et **la statistique du score** est exposé ci-dessous. La statistique du score peut s'écrire comme suit

$$LM = \frac{nSCM}{SCT} = \frac{W}{1+W/n}$$

soit

$$\boxed{R^2 = \frac{LM}{n}} \tag{2.10}$$

Preuve. La statistique du score visant à tester que $\beta = \beta^*$ est définie par

$$LM = \widehat{U}(\beta^*)^T \widehat{\mathbf{I}}(\beta^*)^{-1} \widehat{U}(\beta^*)$$

avec

$$\widehat{U}(\beta^*) = \frac{1}{\widehat{\sigma}^{*2}}\mathbf{Z}^T(Y - \mathbf{Z}\beta^*) \quad \text{et} \quad \widehat{I}(\beta^*)^{-1} = \widehat{\sigma}^{*2}(\mathbf{Z}^T\mathbf{Z})^{-1}$$

On a alors

$$\begin{aligned}LM &= \frac{1}{\widehat{\sigma}^{*2}}\left[(Y^T\mathbf{Z} - \beta^{*T}\mathbf{Z}^T\mathbf{Z})(\mathbf{Z}^T\mathbf{Z})^{-1}(\mathbf{Z}^TY - \mathbf{Z}^T\mathbf{Z}\beta^*)\right] \\ &= \frac{1}{\widehat{\sigma}^{*2}}\left[Y^T\mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^TY - Y^T\mathbf{Z}\beta^* - \beta^{*T}\mathbf{Z}^TY + \beta^{*T}(\mathbf{Z}^T\mathbf{Z})\beta^*\right]\end{aligned}$$

Sous l'hypothèse nulle, le vecteur des paramètres se réduit à

$$\beta^{*T} = \widehat{\beta}^T_{\mathcal{H}_0} = (\widehat{\beta}_0, 0, \cdots, 0) = (\bar{y}, 0, \cdots, 0)$$

On montre facilement que

$$Y^T\mathbf{Z}\beta^* = \beta^{*T}\mathbf{Z}^TY = \beta^{*T}(\mathbf{Z}^T\mathbf{Z})\beta^* = n\bar{y}^2$$

De plus, la variance σ^{*2} est estimée par

$$\widehat{\sigma}^{*2} = \frac{1}{n}||Y - \mathbf{Z}\widehat{\beta}_{\mathcal{H}_0}||^2 = \frac{1}{n}||Y - \bar{Y}||^2 = \frac{1}{n}SCT$$

La statistique du score est donc

$$LM = n\frac{||\widehat{Y} - \bar{Y}||^2}{||Y - \bar{Y}||^2} = n\frac{SCM}{SCT}$$

Par conséquent, on a $R^2 = \dfrac{LM}{n}$. □

Sous cette formulation, le R^2 possède la propriété surprenante de pouvoir être interprétée comme une mesure d'ajustement du modèle qui ne nécessite pas d'en estimer les paramètres.

g. Lien avec l'information et la divergence de Kullback-Leibler

Dans le cas gaussien, le R^2 peut être relié à deux mesures proposées par Kullback-Leibler, l'information et la divergence, définies ci-après.

Définition 2.3 *L'**information de Kullback et Leibler (1951)** permettant de mesurer l'écart entre deux fonctions de densité f et g est définie par*

$$I_{KL}(f,g) = \int f(z)\log\frac{f(z)}{g(z)}dz$$

Sous forme d'espérance, cette information peut s'écrire

$$I_{KL}(f,g) = \mathbb{E}_{P_f}\{\log(f(z))\} - \mathbb{E}_{P_f}\{\log(g(z))\}$$

où P_f désigne la mesure de probabilité sous-jacente à la fonction de densité f.

L'intégrale ci-dessus n'est pas toujours définie. Une condition nécessaire pour que l'intégrale converge est que P_f , la mesure de probabilité sous-jacente à la fonction de densité f, est absolument continue par rapport à P_g la mesure de probabilité induite par g.

L'information de Kullback-Leibler permet de quantifier la proximité de deux lois f et g.

Appliquée dans le cadre de la régression linéaire de $\mathbf{Z}$ sur Y, l'information de Kullback-Leibler s'écrit (Linde et Tutz, 2008) :

$$I_{KL}(\mathbf{Z},Y) = \int\int f_{Z,Y}(z,y)\log\left\{\frac{f_{Z,Y}(z,y)}{f_Z(z)f_Y(y)}\right\}dzdy$$

où $f_{Z,Y}$, f_Y et f_Z sont respectivement les densités jointe, marginale de Y et marginale de Z.

Dans le cas où les distributions marginales de Y et Z sont normales et leur distribution jointe est normale bivariée, le **coefficient de détermination peut s'écrire en fonction de l'information de Kullback-Leibler** par la relation suivante :

$$\boxed{R^2 = 1 - \exp\{-2I_{KL}(Z,Y)\}} \tag{2.11}$$

Preuve. Dans le cas gaussien avec une seule covariable Z (p=1), les densités marginales de Y et Z s'écrivent respectivement

$$f_Y(y) = \frac{1}{\sigma_Y\sqrt{2\pi}} \exp\left\{-\frac{(y-\mu_Y)^2}{2\sigma_Y^2}\right\}$$

et

$$f_Z(z) = \frac{1}{\sigma_Z\sqrt{2\pi}} \exp\left\{-\frac{(z-\mu_Z)^2}{2\sigma_Z^2}\right\}$$

La densité jointe de Y et Z vaut

$$f_{Z,Y}(z,y) = \frac{1}{2\pi\sigma_Z\sigma_Y\sqrt{1-r^2}} \exp\left\{-\frac{1}{2(1-r^2)}\left(\frac{(z-\mu_Z)^2}{\sigma_Z^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - \frac{2r(z-\mu_Z)(y-\mu_Y)}{\sigma_Z\sigma_Y}\right)\right\}$$

où r est le coefficient de corrélation entre Z et Y.
D'autre part, l'information de Kullback-Leibler peut être réarrangée comme suit

$$\begin{aligned} I_{KL}(Z,Y) &= \int\int f_{Z,Y}(z,y)\log\{f_{Z,Y}(z,y)\}dzdy - \int\int f_{Z,Y}(z,y)\log\{f_Z(z)\}dzdy \\ &\quad - \int\int f_{Z,Y}(z,y)\log\{f_Y(y)\}dzdy \\ &= \int\int f_{Z,Y}(z,y)\log\{f_{Z,Y}(z,y)\}dzdy - \int f_Z(z)\log\{f_Z(z)\}dz \\ &\quad - \int f_Y(y)\log\{f_Y(y)\}dy \end{aligned}$$

De plus, on a

$$\begin{aligned} \int f_Z(z)\log\{f_Z(z)\}dz &= -\frac{1}{2}\int \log\left\{2\pi\sigma_Z^2\right\} f_Z(z)dz - \int \frac{1}{\sigma_Z\sqrt{2\pi}}\frac{(z-\mu_Z)^2}{2\sigma_Z^2}\exp\left\{-\frac{(z-\mu_Z)^2}{2\sigma_Z^2}\right\} \\ &= -\frac{1}{2}\log\left\{2\pi\sigma_Z^2\right\} - \mathbb{E}\left[\frac{(z-\mu_Z)^2}{2\sigma_Z^2}\right] \\ &= -\frac{1}{2}\log\left\{2\pi\sigma_Z^2\right\} - \frac{\sigma_Z^2}{2\sigma_Z^2} \\ &= -\frac{1}{2}\left[\log\left\{2\pi\sigma_Z^2\right\} + 1\right] \end{aligned}$$

De même façon, on a

$$\int f_Y(y)\log\{f_Y(y)\}dy = -\frac{1}{2}\left[\log\left\{2\pi\sigma_Y^2\right\} + 1\right]$$

et

$$\begin{aligned} \int\int f_{Z,Y}(z,y)\log\{f_{Z,Y}(z,y)\}dzdy &= -\frac{1}{2}\log\left\{(2\pi)^2\sigma_Z^2\sigma_Y^2(1-r^2)\right\} \\ &\quad -\mathbb{E}\left[\frac{1}{2(1-r^2)}\left(\frac{(z-\mu_Z)^2}{\sigma_Z^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - \frac{2r(z-\mu_Z)(y-\mu_Y)}{\sigma_Z\sigma_Y}\right)\right] \\ &= -\frac{1}{2}\left[\log\left\{(2\pi)^2\sigma_Z^2\sigma_Y^2(1-r^2)\right\} + 1\right] \end{aligned}$$

On en déduit que

$$I_{KL}(Z,Y) = -\frac{1}{2}\log(1-r^2) = -\frac{1}{2}\log(1-R^2)$$

et donc

$$R^2 = 1 - \exp\{-2I_{KL}(Z,Y)\}$$

□

L'information de Kullback-Leibler n'est pas une distance au sens mathématique, car l'inégalité triangulaire et la propriété de symétrie ne sont pas respectées.

Pour résoudre le problème de non-symétrie de cette information, Kullback et Leibler (1951) ont proposé une autre mesure que l'on désigne par les termes «divergence de Kullback-Leibler » définie ci-après.

Définition 2.4 *La **divergence de Kullback et Leibler (1951)** entre deux densités de probabilité f et g est définie par*

$$J_{KL}(f,g) = I_{KL}(f,g) + I_{KL}(g,f) = \int (f(z) - g(z)) log \frac{f(z)}{g(z)} dz$$

Dans le cadre de la régression linéaire de $\mathbf{Z}$ sur Y, la divergence de Kullback-Leibler s'écrit

$$\begin{aligned} J_{KL}(\mathbf{Z},Y) &= \int\int (f_{Z,Y}(z,y) - f_Z(z)f_Y(y)) \log\left\{\frac{f_{Z,Y}(z,y)}{f_Z(z)f_Y(y)}\right\} dzdy \\ &= \mathbb{E}_Z \int (f(y|\mathbf{Z}) - f_Y(y)) log \frac{f(y|\mathbf{Z})}{f_Y(y)} dy \end{aligned}$$

où $f(y|\mathbf{Z})$ dénote la densité de probabilité de y conditionnellement à $\mathbf{Z}$. $J_{KL}(\mathbf{Z},Y)$ mesure la déviation entre la densité conditionnelle de $\mathbf{Z}$ et la densité marginale de Y. Elle décrit donc le pouvoir discriminant de $\mathbf{Z}$ dans le modèle de régression. Une valeur de 0 indique que Y et $\mathbf{Z}$ sont indépendants, tandis que des valeurs élevées reflètent la variabilité de $f(y|\mathbf{Z})$ en fonction de $\mathbf{Z}$.

Dans le cadre de la régression linéaire gaussienne, Linde et Tutz (2008) montrent que le **coefficient de détermination peut être relie à la divergence de Kullback-Leibler** par la relation suivante :

$$\boxed{R^2 = \frac{J_{KL}(\mathbf{Z},Y)}{1 + J_{KL}(\mathbf{Z},Y)}} \tag{2.12}$$

h. Résumé

Le tableau 2.1 résume les différentes formulations possibles du coefficient de détermination dans le cadre de la régression linéaire.

TABLEAU 2.1 – Tableau récapitulatif donnant la relation entre le R^2 et plusieurs quantités statistiques

Quantité	Formule
Erreurs quelconques	
Somme de carrés	$R^2 = 1 - \frac{SCE}{SCT} = \frac{SCM}{SCT}$
Pourcentage de variance expliquée	$R^2 = \frac{\mathbb{V}[\mathbb{E}(Y\|Z)]}{\mathbb{V}(Y)} = 1 - \frac{\mathbb{E}[\mathbb{V}(Y\|Z)]}{\mathbb{V}(Y)}$
Fonction de perte	$R^2 = \frac{\sum_{i=1}^{n} L(y_i, \widehat{y}_0) - \sum_{i=1}^{n} L(y_i, \widehat{y}(z_i))}{\sum_{i=1}^{n} L(y_i, \widehat{y}_0)}$ avec $L(y, \widetilde{y}) = \|\|y - \widetilde{y}\|\|^2$
Erreurs gaussiennes	
Coefficient de corrélation	$R^2 = r^2$
Statistique de Fisher	$R^2 = \frac{nF}{n - p - 1 + nF}$
Statistique de Wald	$R^2 = \frac{W}{W + n}$
Statistique du rapport de vraisemblance	$R^2 = 1 - \exp\left\{-\frac{LR}{n}\right\}$
Statistique du score	$R^2 = \frac{LM}{n}$
Information de Kullback-Leibler	$R^2 = 1 - \exp\{-2I_{KL}(\mathbf{Z}, Y)\}$
Divergence de Kullack-Leibler	$R^2 = \frac{J_{KL}(\mathbf{Z}, Y)}{1 + J_{KL}(\mathbf{Z}, Y)}$

2.2 Pseudo-R^2 et régression logistique

Les propositions de généralisation du coefficient de détermination au modèle linéaire généralisé sont nombreuses et proviennent des différentes écritures dans le modèle linéaire. Dans la suite, nous nous intéressons plus particulièrement à la régression logistique, qui est largement utilisée en épidémiologie et permet de mieux comprendre les outils utilisés en analyse de survie. La définition de ce modèle est rappelée ci-dessous.

Définition 2.5 *Soit Y une variable binaire et Z est une variable explicative, le* ***modèle de régression logistique*** *suppose que Y sachant $(Z = z_i)$ suit une loi binomiale de paramètres (n_i, p_i), soit*

$$(Y|Z = z_i) \sim \mathcal{B}(n_i, p_i), \text{ où } i = 1, \cdots, n$$

avec

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \sum_{j=1}^{p} z_{ij}\beta_p$$

et donc

$$p_i = \frac{\exp\left\{\beta_0 + \sum_j z_{ij}\beta_j\right\}}{1 + \exp\left\{\beta_0 + \sum_j z_{ij}\beta_j\right\}}$$

Dans le cadre du modèle de régression logistique, de nombreux pseudo-R^2 ont été proposés : on distingue les indices issus de la définition originelle du R^2, de la statistique du rapport de vraisemblance, du coefficient de corrélation ou encore de l'information et de la divergence de Kullback-Leibler. A notre connaissance, aucune généralisation n'a été proposée en rapport avec les statistiques de Fisher, de Wald et du score.

2.2.1 Généralisations issues de la définition originelle du R^2

Dans le modèle linéaire, le coefficient de détermination R^2 s'interprète en terme de pourcentage de variance expliquée par le modèle d'une part, et de mesure d'adéquation du modèle aux données, d'autre part (voir équations 2.4 et 2.5). La généralisation de ces interprétations dans le cadre de la régression logistique a donné naissance à deux grandes familles d'indices : la première est basée sur la notion de **proportion de variation expliquée** ; la deuxième sur l'utilisation de **fonctions de perte**.

• Tout d'abord, Mittlböck et Schemper (1996) utilisent la notion de « **proportion de variation expliquée** » (PEV ; en anglais : « proportion of explained variation ») sous la forme

$$PEV = \frac{\sum_{i=1}^{n} D(y_i) - \sum_{i=1}^{n} D(y_i|z_i)}{\sum_{i=1}^{n} D(y_i)}$$

où $D(y_i)$ et $D(y_i|z_i)$ sont des mesures de dispersion des y_i autour d'un paramètre de centralité calculé soit à partir de la distribution marginale de Y, soit à partir de sa distribution conditionnelle au vecteur des covariables pour la i$^{\text{ème}}$ observation. Autrement dit, $D(y_i|z_i)$ et $D(y_i)$ sont des mesures de dispersion conditionnelle et non conditionnelle de Y. L'indice PEV s'interprète en termes de **pourcentage de variation expliquée entre le modèle nul** (ne prenant pas en compte les covariables) **et le modèle alternatif** (prenant en compte les covariables). Il permet de quantifier l'influence d'une ou plusieurs variables pronostiques sur Y.

• Une autre façon de généraliser le coefficient de détermination est basée sur l'utilisation de **fonction de pertes**, comme détaillé précédemment (p. 25). Le pseudo-R^2 s'exprime alors comme

$$KS = \frac{\sum_{i=1}^{n} L(y_i, \widehat{y}_0) - \sum_{i=1}^{n} L(y_i, \widehat{y}(z_i))}{\sum_{i=1}^{n} L(y_i, \widehat{y}_0)}$$

Sous cette forme, le R^2 permet de quantifier la perte relative encourue en faisant la prédiction basée sur l'utilisation des covariables dans le modèle par rapport à la prédiction ne faisant par intervenir les covariables.

Les indices présentés dans la suite peuvent être interprétées sous l'un des deux angles définis précédemment, i.e. sous l'angle de la « proportion de variation expliquée » ou sous celui des fonctions de perte. Dans le cas de la régression logistique, les deux interprétations du coefficient de détermination, PEV et KS, donnent naissance à des indices identiques, car la définition du modèle nul est la même : $\widehat{y}_0 = \frac{1}{n}\sum_i y_i = \bar{p}$.

a. Indices basés sur la somme des carrés

La distance initialement proposée dans le cadre de la régression linéaire est, comme nous l'avons vu, la **somme des carrés**.

Dans le cadre de la proportion de variation expliquée, Efron (1978) et Mittlböck et Schemper (1996), proposent, entre autres, l'indice suivant :

$$R^2_{SC} = \frac{\sum_{i=1}^{n} D(y_i) - \sum_{i=1}^{n} D(y_i|z_i)}{\sum_{i=1}^{n} D(y_i)} = \frac{\sum_{i=1}^{n}(y_i - \bar{p})^2 - \sum_{i=1}^{n}(y_i - \widehat{p}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{p})^2} = \frac{2\sum_{i}^{n} y_i\widehat{p}_i - \sum_{i}^{n}\widehat{p}_i^2 - n\bar{p}}{n\bar{p}(1-\bar{p})}$$

avec $\bar{p} = \dfrac{\sum_i y_i}{n}$ et $\widehat{p}_i = \dfrac{\exp(\widehat{\beta} z_i)}{1 + \exp(\widehat{\beta} z_i)}$.

L'indice de Korn et Simon (1991) est égal à

$$KS = \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_0)^2 - \sum_{i=1}^{n}(y_i - \widehat{y}(z_i))^2}{\sum_{i=1}^{n}(y_i - \widehat{y}_0)^2} = \frac{\frac{2}{n}\sum_{i=1}^{n} y_i\widehat{y}(z_i) - \frac{1}{n}\sum_{i=1}^{n}\widehat{y}(z_i)^2 - \widehat{y}_0^2}{\widehat{y}_0(1-\widehat{y}_0)}$$

avec $\widehat{y}_0 = \frac{1}{n}\sum_{i=1}^{n} y_i$ et $\widehat{y}(z_i) = \frac{\exp(\widehat{\beta}z_i)}{1+\exp(\widehat{\beta}z_i)}$.

Il coïncide avec celui de Mittlböck et Schemper (1996), car $\widehat{y}_0 = \bar{p}$ et $\widehat{y}(z_i) = \widehat{p}_i$.

L'interprétation du R^2 basé sur la somme des carrés est la même que dans le modèle linéaire.

b. Indices basés sur l'entropie

La notion d'entropie a initialement été introduite par Shannon (1948) dans le cadre de la théorie d'information pour décrire la quantité d'information contenue ou délivrée par une source.

Définition 2.6 *Soit une variable aléatoire discrète, $U = \{u_1, \cdots, u_K\}$. L'**entropie** H de U est définie comme (Haberman, 1982) :*

$$H(U) = -\sum_{l=1}^{K} \Pr(U = u_l)\log\{\Pr(U = u_l)\}$$

La définition de l'entropie peut s'étendre au cas où U est une variable continue de densité de probabilité f

$$H_{P_f}(u) = \mathbb{E}_{P_f}\{\log f(u)\} = -\int f(u)\log f(u)du$$

où P_f désigne la mesure de probabilité sous-jacente à la fonction de densité f.

Interprétation. Dans le cas discret, l'entropie est comprise entre 0 et $\log(K)$. Elle est nulle si $\exists i|P_i = 1$. On montre qu'elle est maximale pour une distribution uniforme pour K fixé. Elle augmente avec le nombre de valeurs possibles de la valeur discrète, K. La quantité H croit proportionnellement avec l'incertitude de l'information qui manque. L'entropie peut donc s'interpréter comme une mesure d'incertitude associée à la variable étudiée.

Dans le cadre de la régression logistique ($K = 2$), les deux indices basés sur l'entropie proposés par Korn et Simon (1991), d'une part, et Efron (1978); Mittlböck et Schemper (1996), d'autre part, sont les mêmes.

Pour Efron (1978) et Mittlböck et Schemper (1996), dans le cadre de la proportion de variation expliquée, le pseudo-R^2 déduit de l'entropie s'écrit à partir de :

$$\begin{aligned} D(Y) &= \sum_{i=1}^{n} -\left[y_i\log(\bar{p}) + (1-y_i)\log(1-\bar{p})\right] \\ &= -n\left[\bar{p}\log(\bar{p}) + (1-\bar{p})\log(1-\bar{p})\right] \end{aligned}$$

et de

$$D(Y|Z) = \sum_{i=1}^{n} - [y_i \log(\widehat{p}_i) + (1 - y_i) \log(1 - \widehat{p}_i)]$$

d'où

$$PEV_H = \frac{n [\bar{p} \log(\bar{p}) + (1 - \bar{p}) \log(1 - \bar{p})] - \sum_{i=1}^{n} [y_i \log(\widehat{p}_i) + (1 - y_i) \log(1 - \widehat{p}_i)]}{n [\bar{p} \log(\bar{p}) + (1 - \bar{p}) \log(1 - \bar{p})]}$$

Pour Korn et Simon (1991), dans le cadre des fonctions de perte, le pseudo-R^2 dérivé de l'entropie s'écrit avec :

$$L(y_i, \widehat{y}_0) = y_i \log(\widehat{y}_0) + (1 - y_i) \log(1 - \widehat{y}_0)$$

et

$$L(y_i, \widehat{y}(z_i)) = y_i \log(\widehat{y}(z_i)) + (1 - y_i) \log(1 - \widehat{y}(z_i))$$

et donc

$$KS_H = \frac{\sum_{i=1}^{n} [y_i \log(\widehat{y}_0) + (1 - y_i) \log(1 - \widehat{y}_0)] - \sum_{i=1}^{n} [y_i \log(\widehat{y}(z_i)) + (1 - y_i) \log(1 - \widehat{y}(z_i))]}{\sum_{i=1}^{n} [y_i \log(\widehat{y}_0) + (1 - y_i) \log(1 - \widehat{y}_0)]}$$

c. Indices basés sur la concentration

Le coefficient de concentration a été introduit par Gini (1912). Il existe de nombreuses façons de le présenter et la définition d'Haberman (1982) a été considérée ici.

Définition 2.7 *Soit une variable aléatoire discrète, $U = \{u_1, \cdots, u_K\}$. La* ***concentration*** *est définie, selon Haberman (1982), par*

$$C(U) = 1 - \sum_{l=1}^{K} \Pr(U = u_i)^2$$

Interprétation. Ce coefficient mesure la concentration. Il peut également être interprété sous l'angle de la dispersion, une valeur égale à 0 correspondant à une absence de disparité et une valeur de 1 à une disparité maximale.

Dans le cas de la régression logistique, la concentration de Gini permet de construire un indice de proportion de variation expliquée à partir des quantités suivantes (voir Mittlböck et Schemper, 1996)

$$D(y_i) = 1 - \bar{p}^2 - (1 - \bar{p})^2 = 2\bar{p}(1 - \bar{p})$$
$$D(y_i|z_i) = 1 - \widehat{p}_i^2 - (1 - \widehat{p}_i)^2 = 2\widehat{p}_i(1 - \widehat{p}_i)$$

Le pseudo-R^2 basé sur la concentration de Gini se simplifie comme suit

$$PEV_G = \frac{\sum_{i=1}^{n} \widehat{p}_i^2 - n\bar{p}}{n\bar{p}(1-\bar{p})}$$

L'utilisation de la concentration de Gini dans le contexte des fonctions de perte n'a pas été explorée.

d. Indices d'erreur de classement

L'**erreur de classement** est définie dans l'article d'Efron (1977) comme une mesure de dispersion qui vaut 0 si le prédicteur de la variable réponse est inférieur à 1/2, 1 s'il est supérieur à 1/2 et 0.5 s'il est égal à 1/2.

Pour la régression logistique, les termes $D(y_i)$ et $D(y_i|z_i)$ permettant de construire l'indice de proportion de variation expliquée sont définies de la manière suivante (Mittlböck et Schemper, 1996) :

$$D(y_i) = \begin{cases} 1 & \text{si } |y_i - \bar{p}| > 0.5 \\ 0.5 & \text{si } |y_i - \bar{p}| = 0.5 \\ 0 & \text{si } |y_i - \bar{p}| < 0.5 \end{cases} \quad \text{et } D(y_i|z_i) = \begin{cases} 1 & \text{si } |y_i - \widehat{p}_i| > 0.5 \\ 0.5 & \text{si } |y_i - \widehat{p}_i| = 0.5 \\ 0 & \text{si } |y_i - \widehat{p}_i| < 0.5 \end{cases}$$

d'où

$$PEV_E = \frac{\sum_{i=1}^{n} D(y_i) - \sum_{i=1}^{n} D(y_i|z_i)}{\sum_{i=1}^{n} D(y_i)}$$

Korn et Simon (1991) n'ont pas proposé d'interprétation de l'erreur de classement en terme de fonction de perte.

Remarque:
Un autre indice de proportion de variation expliquée peut être construite à partir de l'information de Kullback-Leibler :

$$R_K^2 = 1 - \frac{I_{KL}(y, \widehat{y})}{I_{KL}(y, \widehat{y}_0)}$$

où $\widehat{y}$ et $\widehat{y}_0$ sont les estimateurs de la moyenne de y en présence ou en l'absence de covariables.
Dans le cas gaussien, on retrouve l'indice basé sur les sommes de carrés (voir p. 36).
Dans le cas d'un modèle logistique, on obtient l'indice basé sur l'entropie (p. 37).

2.2.2 Généralisations issues du coefficient de corrélation

Dans le modèle linéaire, nous avons vu que le coefficient de détermination peut s'écrire comme le **carré du coefficient de corrélation** (équation 2.6, p. 27). La transposition de cette définition dans le cadre de la régression logistique a donné naissance à plusieurs indices basés sur les coefficients de Pearson, Spearman, Kendall, Somers et Goddman et Kruskal.

Ces coefficients décrivent la relation entre la variable observée et les covariables prédites, en utilisant soit la valeur des covariables, soit leur rang. Ils sont compris entre -1 et 1 et leur carré peut donc être utilisé pour mesurer la capacité de prédiction du modèle.

Tout d'abord, Mittlböck et Schemper (1996) utilisent le carré du coefficient de Pearson r^2 comme coefficient de détermination, avec

$$r = \frac{\sum_{i=1}^{n} y_i \widehat{p}_i - n\bar{p}^2}{\sqrt{n\bar{p}(1-\bar{p})\sum_{i=1}^{n}(\widehat{p}_i - \bar{p})^2}}$$

D'autres coefficients (au carré) ont également été considérés par les auteurs. Il s'agit tout d'abord du coefficient de corrélation de Spearman. Ce dernier s'exprime comme le coefficient de corrélation de Pearson en remplaçant les valeurs des variables par leurs rangs (voir Snedecor et Cochran, 1989; Conover et Iman, 1981) :

$$r_s = \frac{\sum_{i=1}^{n}(R(y_i) - \bar{R})(R(\widehat{p}_i) - \bar{R})}{\sqrt{\sum_{i=1}^{n}(R(y_i) - \bar{R})^2 \sum_{i=1}^{n}(R(\widehat{p}_i) - \bar{R})^2}}$$

où $R(z)$ représente le rang de z et $\bar{R} = (n+1)/2$ est le « rang moyen ».

Ils ont également suggéré le coefficient de Kendall (Kendall et Gibbons, 1990), qui permet de tester l'indépendance entre 2 variables aléatoires :

$$\tau_a = \frac{\sum_{i<j} \text{sign}(y_j - y_i)\text{sign}(\widehat{p}_j - \widehat{p}_i)}{n(n-1)/2}$$

avec

$$\text{sign}(z) = \begin{cases} 1 & \text{si } z > 0 \\ 0 & \text{si } z = 0 \\ -1 & \text{si } z < 0 \end{cases}$$

ou bien alternativement

$$\tau_b = \frac{\sum_{i<j} \text{sign}(y_j - y_i)\text{sign}(\widehat{p}_j - \widehat{p}_i)}{\sqrt{\sum_{i<j} \text{sign}^2(y_j - y_i) \sum_{i<j} \text{sign}^2(\widehat{p}_j - \widehat{p}_i)}}$$

Le coefficient de Somers (1962) est ensuite présenté

$$D = \frac{\sum_{i<j} \text{sign}(y_j - y_i)\text{sign}(\widehat{p}_j - \widehat{p}_i)}{\sum_{i<j} \text{sign}^2(y_j - y_i)}$$

Enfin, la possibilité d'utiliser le coefficient de Goodman et Kruskal (1954) est évoquée :

$$\gamma = \frac{\sum_{i<j} \text{sign}(y_j - y_i)\text{sign}(\widehat{p}_j - \widehat{p}_i)}{\sum_{i<j} \text{sign}^2(y_j - y_i) \sum_{i<j} \text{sign}^2(\widehat{p}_j - \widehat{p}_i)}$$

2.2.3 Généralisations issues de la statistique du rapport de vraisemblance

Dans le modèle linéaire, le coefficient de détermination peut être relié à la **statistique du rapport de vraisemblance** (équation 2.9, p. 29). Allison (1995); Maddala (1983); Magee (1990) ont utilisé cette relation pour généraliser le R^2 au modèle logistique, en calculant la vraisemblance correspondante. Ce pseudo-R^2 est noté R^2_{LR}.

Cependant, comme R^2_{LR} ne peut pas atteindre la valeur 1, Nagelkerke (1991) a suggéré l'utilisation d'une version modifiée de cet indice pour remédier à cet inconvénient :

$$R^2_N = \frac{R^2_{LR}}{R^2_{max}}$$
$$\text{avec} \quad R^2_{max} = 1 - \exp\left(\frac{2}{n}\log \mathcal{L}_0\right)$$

$\mathcal{L}_0$ étant la vraisemblance du modèle nul.

2.2.4 Généralisations issues de l'information et de la divergence de Kullback-Leibler

Dans le modèle linéaire, le coefficient de détermination peut être relié à l'**information de Kullback-Leibler** (équation 2.11, p. 32). Linde et Tutz (2008) utilisent la même relation dans le cadre du modèle logistique :

$$R^2_I = 1 - \exp\{-2I_{KL}(Z, Y)\}$$

Cet indice est en fait équivalent à celui reposant sur la statistique du rapport de vraisemblance (voir p. 41).

Dans le cadre de la régression linéaire, le R^2 peut également être relié à la **divergence de Kullback-Leibler** (équation 2.12, p. 33).

Dans le cadre de la régression logistique, Linde et Tutz (2008) ont proposé une généralisation du coefficient de détermination dans deux cas particulier. Dans un premier temps, lorsque la variance σ^2_Z des covariables est petite, le pseudo-R^2 peut être approximé par la relation suivante

$$R^2_J = \frac{\beta^2\sigma^2_Z}{\bar{y}(1-\bar{y}) + \beta^2\sigma^2_Z}$$

Linde et Tutz (2008) présentent également une généralisation du coefficient de détermination, lorsque Z est une variable gaussienne. Ils supposent que $\mathbb{E}(Z|Y=1) = \mu_1$ et $\mathbb{E}(Z|Y=0) = \mu_0$ et que $P(Y=1) = p_1$ et $p(Y=0) = p_0$. Pour simplifier, ils posent $E(Z) = 0$, ce qui implique que $\mu_0 = -\mu_1 p_1/p_0$.
Sous ces hypothèses, le pseudo-R^2 est donné par la relation suivante :

$$R^2_J = \frac{p_1\mu_1^2}{p_0\sigma^2 + p_1\mu^2} = \frac{p_1 p_0 \beta^2}{1/\sigma^2 + p_1 p_0 \beta^2}$$

2.3 Conclusion

Dans le cas du **modèle de régression linéaire gaussien**, toutes les **valeurs du R^2** présentées précédemment, basées sur les notions de proportion de variance expliquée et de fonctions de pertes, sur les statistiques de Fisher, de Wald, du log-rapport de vraisemblance et du score, sur le coefficient de corrélation et les information et divergence de Kullback-Leibler, **sont égales**. Dans le modèle de régression logistique, cette égalité n'est pas nécessairement maintenue et les très nombreux indices proposés conduisent à des résultats difficilement comparables.

Dans ce chapitre, nous avons uniquement considéré les généralisations du R^2 dans le cadre de la régression logistique. Cependant, il existe d'autres indices adaptés à d'autres modèles linéaires généralisés. Ainsi Cameron et Windmeijer (1996) ont proposé des indices pour le modèle de Poisson. Par ailleurs, Agresti (1986) s'est intéressé à la généralisation des indices basés sur l'entropie et la concentration dans le cas de données catégorielles multivariées.

Chapitre 3

REVUE DE LA LITTÉRATURE : INDICES DE CAPACITÉ DE PRÉDICTION EN ANALYSE DE SURVIE

Contenu

Comme nous l'avons évoqué précédemment (chapitre 2), dans le modèle linéaire, le coefficient de détermination possède différentes interprétations et peut être relié à différentes statistiques. Dans le cas de ce modèle particulier, les différentes écritures du R^2 sont égales entre elles. Pour des modèles plus généraux (modèle logistique ou de survie), ce n'est pas nécessairement le cas, et, par conséquent, plusieurs pseudo-R^2 peuvent être proposés à partir des différentes interprétations du R^2 dans le modèle linéaire.

Cette section est une revue des **mesures de capacité de prédiction** rencontrées dans la littérature en **analyse de survie**.

Dans un premier temps, quelques définitions et notations sont rappelées dans le cadre de l'analyse de survie, puis, les différents indices de type **pseudo-R^2** proposés dans la littérature sont présentés.

3.1 Définitions et notations en analyse de survie

Cette section vise à donner les définitions des principaux outils utilisés en analyse de survie.

3.1.1 Modélisation de la survie en l'absence de covariables

a. Fonctions de survie et de risque

Une étude de survie est une étude ayant pour but d'étudier le **délai de survenue d'un évènement**, qui peut être la récidive tumorale ou métastasique, la réponse à un traitement, le décès. Le vocabulaire utilisé fait souvent référence au décès bien que d'autres événements puissent être concernés.

On note X la variable aléatoire représentant le temps de survie réel d'un sujet, i.e. le délai entre le début du suivi et la date de survenue de l'évènement étudié, éventuellement non observé si la fin de la durée de surveillance du sujet lui est antérieure (voir p. 45, paragraphe b.).

Définition 3.1 *La **fonction de survie** d'une variable aléatoire X est définie par*

$$S(x) = 1 - F(x) = \Pr(X > x)$$

où $F(x)$ est la fonction de répartition de la variable aléatoire X.

Définition 3.2 *La **fonction de risque instantané** est la fonction*

$$\begin{aligned}\lambda(x) &= \lim_{\Delta x \to 0} \frac{1}{\Delta x} \Pr\{x \leq X < x + \Delta x | X \geq x\} \\ &= -\left[\frac{d}{dx} S(x)\right] / S(x) \\ &= \frac{f(x)}{S(x)}\end{aligned}$$

où f est la densité de probabilité de X.

La quantité $\lambda(x)dx$ est la probabilité que l'événement se produise entre x et $x+dx$ sachant qu'il ne s'est pas produit auparavant.

Définition 3.3 *La **fonction de risque cumulé** est donnée par*

$$A(x) = \int_0^x \lambda(s)ds$$

Des définitions précédentes, il s'en suit que

$$S(x) = \exp\left(-\int_0^x \lambda(s)ds\right) = \exp\left(-A(x)\right)$$

b. Censure

Définition 3.4 *La **variable de censure** C est définie par la possible non-observation de l'événement. Si l'on observe C, et non X, et que l'on sait que $X > C$ (respectivement $X < C$, $C1 < X < C2$), on dit qu'il y a censure à droite (respectivement censure à gauche, censure par intervalle).*

En pratique, en cas de censure à droite, le temps de survie observé T, i.e. le délai entre le début du suivi et la date de survenue de l'événement étudié ou la censure lorsque ce dernier n'est pas observé, est égal au minimum entre le temps de survie réel et le temps de censure : $T = \min(X, C)$.

Parmi les grands types de censure à droite, on distingue la **censure fixe** et la **censure aléatoire**. La censure fixe a lieu lorsque l'étude s'arrête après une durée de suivi fixée pour chaque sujet. Dans ce cas, tous les individus, pour lesquels l'événement d'intérêt n'a pas eu lieu au cours de l'étude, auront tous le même temps de censure. Dans le cas le plus simple de censure aléatoire, l'étude s'arrête après une date donnée et le temps de censure attaché à chaque sujet est égal à la durée écoulée entre son entrée dans l'étude et la date de « point ».

En général, le mécanisme de censure C est supposé indépendant de l'évènement étudié X : la censure est dite « **non-informative** ». L'hypothèse est ainsi faite que la raison du départ des patients de l'étude est indépendante du risque d'apparition de l'évènement, et que le risque de survenue de l'évènement des sujets censurés et identique à celui des patients encore présents dans l'étude. Si la censure est **informative**, i.e. que leur survenue n'est pas liée au hasard, l'inférence de modèles standards, ne tenant pas compte de ce type de données, peut mener à des conclusions biaisées (Kalbfleisch et Prentice, 2002).
Dans la suite de ce travail, les méthodes présentées reposent sur l'hypothèse de censure non-informative.

c. Processus de comptage

Définition 3.5 *Pour chaque sujet $i = 1, \cdots, n$, une* ***observation*** *consiste en (T_i, δ_i), où*

$$\begin{cases} T_i = \min(X_i, C_i) \\ \delta_i = \mathbf{1}_{\{X_i \leq C_i\}} \end{cases},$$

X_i désignant le temps de survenue de l'événement et C_i la variable de censure. δ_i est l'indicatrice de survenue de l'évènement, souvent appelée indicatrice de décès.

Définition 3.6 *Soit le* ***processus de comptage*** *suivant*

$$N_i(t) = \mathbf{1}\{T_i \geq t; \delta_i = 1\}$$

indiquant le nombre d'événements observés dans l'intervalle de temps $(0, t]$ pour l'individu i. Le processus N_i est croissant et augmente par pas de taille $+1$, avec $N_i(0) = 0$.

On note $dN_i(t) = N_i(t^- + dt) - N_i(t^-)$ le nombre d'événements observés dans l'intervalle $[t, t + dt)$.

Définition 3.7 *Soit le* ***processus à risque*** *défini par*

$$Y_i(t) = \mathbf{1}\{T_i \geq t\}$$

Il vaut 1 lorsque l'individu i est à risque juste avant le temps t, et 0 sinon.

On note

$$\bar{N}(t) = \sum_{i=1}^{n} N_i(t) \quad ; \quad \bar{N}(\infty) = k$$

et

$$\bar{Y}(t) = \sum_{i=1}^{n} Y_i(t), 0 < t < \infty$$

Clairement, $\bar{N}(t)$ est le nombre total d'événements observés dans l'intervalle $(0, t]$, k est le nombre total de décès observés et $\bar{Y}(t)$ est le nombre total d'individus à risque au temps t.

De ce qui précède, nous pouvons donc écrire

$$\Pr\{dN_i(t) = 1 | \mathcal{F}_{t^-}\} = \lambda_i(t) Y_i(t) dt$$

$\mathcal{F}_{t^-}$ est l'historique du processus ou **filtration** définie par $\mathcal{F}_{t^-} = \{N_i(u), Y_i(u^+)) \, ; i = 1, \cdots, n; 0 \leq u < t\}$. Dans le cas où p covariables dépendantes du temps $Z_i(t) = \{Z_{1i}(t), \cdots, Z_{pi}(t)\}$ sont observées pour chaque sujet i, $\mathcal{F}_{t^-}$ est donnée par $\mathcal{F}_{t^-} = \{N_i(u), Y_i(u^+)), Z_i(u^+) \, ; i = 1, \cdots, n; 0 \leq u < t\}$.

Définition 3.8 *Les processus*

$$\lambda_i(t)Y_i(t)$$

et

$$\Lambda_i(t) = \int_0^t \lambda_i(s)Y_i(s)ds$$

sont désignés respectivement sous les termes de ***processus d'intensité*** *et* ***processus d'intensité cumulée*** *de N_i.*

Il est possible d'exprimer $\lambda_i(t)$ en fonction du processus de comptage $N_i(t)$:

$$\lambda_i(t) = \lim_{dt\to 0} \frac{1}{dt} \Pr\{N_i(t+dt) - N_i(t) = 1 | \mathcal{F}_{t^-}\}$$

Proposition 3.1 *Pour un individu i donné, le processus suivant*

$$M_i(t) = N_i(t) - \int_0^t \lambda_i(s)Y_i(s)ds$$

est une ***martingale*** *si et seulement si*

$$\lambda(x) = \lim_{\Delta x\to 0} \frac{1}{\Delta x} \Pr\{x \le X < x + \Delta x | X \ge x, C \ge x\} \text{ lorsque } \Pr\{T > x\} > 0 \tag{3.1}$$

De plus, elle est de moyenne nulle (Fleming et Harrington, 2005). De manière équivalente, on a

$$dM_i(t) = dN_i(t) - d\Lambda_i(t) = dN_i(t) - \lambda_i(t)Y_i(t)dt$$

Le processus d'intensité cumulé, $\Lambda_i(t)$, est également appelé **compensateur** du processus $N_i(t)$.

Dans l'équation (3.1), le terme de droite correspond à la fonction e risque instantanné de X en présence de censure, qui diffère de la définition de fonction de risque donnée à la page 44, qui est la fonction de risque de X. Cette dernière est appelée **fonction de risque nette**, alors que la fonction de la condition (3.1) est la **fonction de risque brute**. La proposition ci-dessus stipule donc que M est une martingale si et seulement si le risque net et le risque brut sont égaux.

Ainsi, la condition (3.1) peut être interprétée comme

$$\Pr\{x \le X < x + dx | X \ge x\} = \Pr\{x \le X < x + dx | X \ge x, C \ge x\}$$

Elle est alors légèrement plus faible que la **condition d'indépendance** entre X et C.

Enfin, notons que la **condition de non-information** suppose que la distribution de la variable de censure ne dépend pas des mêmes paramètres que la survie.

d. Estimation des fonctions du risque cumulé et de survie

Définition 3.9 *L'**estimateur de Nelson-Aalen** de la fonction du risque cumulé est défini par (Nelson, 1972; Aalen, 1978)*

$$\widehat{A}(t) = \int_0^t \frac{\mathbf{1}_{\{\bar{Y}(s)>0\}}}{\bar{Y}(s)} d\bar{N}(s) \tag{3.2}$$

Définition 3.10 *L'**estimateur de Kaplan-Meier** de la fonction de survie est défini par (Kaplan et Meier, 1958)*

$$\widehat{S}_{KM}(t) = \prod_{s \leq t} \left\{ 1 - \frac{\Delta \bar{N}(s)}{\bar{Y}(s)} \right\} \tag{3.3}$$

3.1.2 Modèle de Cox : Rappels et Notations

Le modèle de Cox (1972) est le plus couramment utilisé en pratique pour l'analyse de données censurées à droite. Il permet de modéliser l'effet d'une ou plusieurs covariables sur la probabilité d'apparition d'un évènement. La plupart des indices de capacité de prédiction proposés en analyse de survie reposent sur ce modèle. Nous y ferons référence tout au long de la thèse, et c'est pour cette raison qu'il est introduit dans cette section, avec la méthode d'estimation correspondante.

a. Définition

Définition 3.11 *Le **modèle de Cox** est défini par la fonction de risque suivante*

$$\lambda\left(t|Z_i\right) = \lambda_0\left(t\right) \exp\left\{\beta' Z_i\right\} \quad i = 1, \cdots, n \tag{3.4}$$

où $\lambda_0\left(t\right)$ est une fonction de risque de base fixée, non précisée, $Z_i' = (Z_{1i}, \cdots, Z_{pi})$ est le vecteur de dimension p des covariables du sujet i, et β est un vecteur $p \times 1$ de paramètres à estimer.

Le modèle peut également s'écrire à l'aide de la **fonction de survie**, dont l'expression à un temps t donné et pour un sujet i est

$$S(t|Z_i) = \exp\left\{ -\int_0^t \lambda_0(s) \exp\{\beta' Z_i\} ds \right\} = S_0(t)^{\exp(\beta' Z_i)} \tag{3.5}$$

où $S_0(t)$ désigne la fonction de survie en l'absence de covariables.

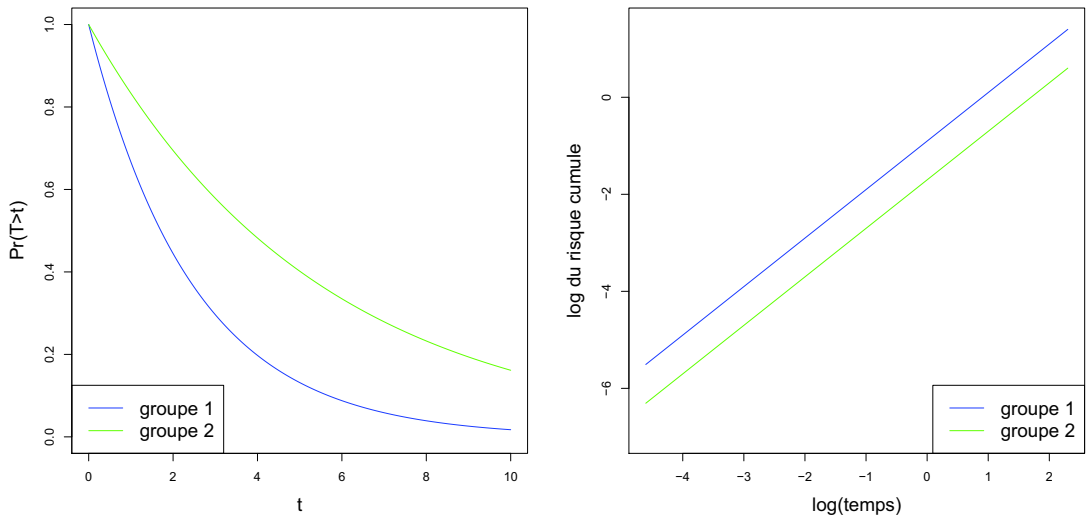
Lorsque les covariables ne dépendent pas du temps, le modèle obtenu est dit à **risques proportionnels**. Dans ce cas, le rapport des risques instantanés, $\dfrac{\lambda(t|Z_i)}{\lambda(t|Z_j)} = \exp\{\beta'(Z_i - Z_j)\}, \forall (i,j) \in \{1, \cdots, n\}$, ne dépend pas du temps.

De plus, il s'agit d'un **modèle semi-paramétrique**, car il comprend une partie paramétrique $\exp(\beta' Z_i)$ qui modélise la relation entre le temps de survenue de l'événement et les covariables, et une partie non-paramétrique, la fonction de risque de base $\lambda_0(t)$ qui est inconnue.

Le modèle de Cox est également **log-linéaire** puisque le log-risque cumulé est une fonction linéaire des covariables : $\log(A(t|Z_i)) = \beta' Z_i + \log(A_0(t))$, où $A_0(t) = \int_0^t \lambda_0(s)ds$.

La figure 3.1 (a) montre un exemple théorique de courbes de survie pour un modèle à risques proportionnels tracées à partir de la loi exponentielle. Les deux courbes représentent la survie associée à deux groupes d'individus définis par les deux niveaux d'une covariable binaire de risques instantanés respectifs égaux à 1.5 (groupe 1) et 1.2 (groupe 2). La figure 3.1 (b), qui représente la courbe du log-risque cumulé pour les deux groupes d'individus, permet de vérifier visuellement l'hypothèse des risques proportionnels (courbes parallèles).

FIGURE 3.1 – Courbes (a) de survie et (b) du log-risque cumulé théoriques pour deux groupes d'individus définis par les deux valeurs d'une covariable binaire, dans le cadre du modèle de Cox.



(a) (b)

b. Estimation

Pour estimer les paramètres du modèle de Cox, on définit, en utilisant les notations relatives aux processus de comptage, la **vraisemblance partielle** de la façon suivante

$$\mathcal{L}(\beta) = \prod_{t \leq \tau} \prod_{i=1}^{n} \left\{ \frac{Y_i(t) \exp(\beta' Z_i)}{S^{(0)}(\beta; t)} \right\}^{\Delta N_i(t)}$$

avec

$$S^{(0)}(\beta; t) = \sum_{i=1}^{n} Y_i(t) \exp(\beta' Z_i)$$

et τ désignant la plus grand temps de décès observé.
On en déduit

$$\log \mathcal{L}(\beta) = \sum_{i=1}^{n} \int_0^{\tau} \left[\beta' Z_i - \log S^{(0)}(\beta; s)\right] dN_i(s)$$

Les termes ne dépendant pas de β n'apparaissent pas dans la vraisemblance partielle.

D'une façon plus générale, la **log-vraisemblance partielle** du modèle de Cox peut s'écrire à un temps t fixé :

$$\log\{\mathcal{L}(\beta; t)\} = \sum_{i=1}^{n} \int_0^{t} \left[\beta' Z_i - \log S^{(0)}(\beta; s)\right] dN_i(s) \tag{3.6}$$

Le vecteur **score**, déduit de la vraisemblance partielle, s'écrit

$$U(\beta; t) = \frac{\partial \log\{\mathcal{L}(\beta; t)\}}{\partial \beta} = \sum_{i=1}^{n} \int_0^{t} [Z_i - E(\beta; s)]\, dN_i(s)$$

où

$$E(\beta; t) = \frac{S^{(1)}(\beta; t)}{S^{(0)}(\beta; t)}$$

et

$$S^{(1)}(\beta; t) = \sum_{i=1}^{n} Y_i(t) Z_i \exp(\beta' Z_i)$$

Les estimateurs $\widehat{\beta}$ du maximum de vraisemblance des paramètres du modèle sont calculés à partir du vecteur score par la relation $U(\widehat{\beta}, \tau) = 0$. Ces estimateurs permettent de donner une estimation de la fonction de survie comme suit

$$\widehat{S}(t|\mathbf{Z}) = \widehat{S}_0(t)^{\exp(\widehat{\beta}'\mathbf{Z})} \tag{3.7}$$

avec $\widehat{S}_0(t) = \exp\{-\widehat{A}_0(t)\}$, où $\widehat{A}_0(t)$ est l'estimateur de Nelson-Aalen du risque cumulé (équation (3.2), p. 48)

Afin de tester l'hypothèse nulle $\mathcal{H}_0 : \{\beta = 0\}$, plusieurs tests peuvent être utilisés. Dans la suite, il sera essentiellement question de la statistique du score qui est basé sur le vecteur score calculé en $\beta = 0$:

$$U(0) = \sum_{i=1}^{n} \int_0^{\tau} [Z_i - E(0; s)]\, dN_i(s) = \sum_{i=1}^{n} \int_0^{\tau} \left[Z_i - \frac{\sum_{l=1}^{n} Y_l(s) Z_l}{\bar{Y}(s)}\right] dN_i(s) \tag{3.8}$$

3.2 Présentation des indices

La transposition de la notion de R^2 à l'analyse de survie a donné naissance à une multitude d'indices visant à généraliser les indices existant dans le modèle linéaire. Nous avons regroupé ces indices en quatre catégories :

- les **indices fondés sur la somme des écarts** ;
- les **indices dérivés du rapport de vraisemblance** ;
- les **indices basés sur la notion de corrélation** ;
- les **indices de concordance**.

Les indices de somme des écarts sont une généralisation de la définition de base du R^2 dans le modèle linéaire. Les indices dérivés du rapport de vraisemblance sont une simple transposition des indices du modèle linéaire correspondants, en utilisant la vraisemblance partielle du modèle de Cox. Les indices basés sur la notion de corrélation décrivent la relation entre la variable aléatoire de survie et la ou les variable(s) explicative(s). Enfin, les indices de concordance sont spécifiques à la survie et quantifient le degré d'adéquation entre l'ordre des temps de décès et les covariables explicatives.

La plupart des indices présentés ci-après sont définis dans la cadre du modèle de Cox à risques proportionnels (sauf indication contraire).

Notations

Soit un sujet $i, i = 1, \cdots, n$. On note

- $X = (X_1, \cdots, X_n)$ le vecteur aléatoire des temps de survenue de l'évènement étudié (ou temps de survie) ;
- $t_1, \cdots, t_n$ les valeurs prises par les variables aléatoires $(T_1, \cdots, T_n)$ du temps de surveillance (ou suivi) et $t_{(1)}, \cdots, t_{(n)}$ les valeurs des temps de suivi ordonnées ;
- $t^*_{(1)}, \cdots, t^*_{(k)}$ les temps de décès (non censurés) ordonnés et k le nombre total d'individus non censurés (en l'absence d'ex-æquo) ;
- $\mathbf{Z} = \{Z_1, \cdots, Z_n\}^T$ la matrice $(n \times p)$ des covariables avec $Z_i = \{Z_{1i}, \cdots, Z_{pi}\}^T$;
- δ_i l'indicatrice de décès en t_i, valant 0 en cas de censure, et $\delta_{(i)}$ l'indicatrice de décès en $t_{(i)}$;
- $S(t)$ la fonction de survie marginale et $\widehat{S}_{KM}(t)$ son estimateur ;
- $S(t|Z_i)$ la fonction de survie conditionnelle aux covariables et $\widehat{S}(t|Z_i)$ son estimateur.

Les expressions des estimateurs de la fonction de survie marginale, $\widehat{S}_{KM}(t)$, et de $\widehat{S}(t|Z)$, sont données par les formules (3.3) et (3.7) pages 48 et 50.

3.2.1 Les indices fondés sur la somme des écarts

Dans le cadre de la survie, les indices fondés sur la somme des écarts ne portent pas directement sur les temps de survie, mais se basent sur le processus à risque $Y(t)$ (ou sur des processus dérivés) qui vaut 1 si l'individu est toujours vivant en t avec la probabilité $S(t)$ et 0 sinon avec la probabilité $(1 - S(t))$. Ces indices mesurent ainsi un **écart** entre le processus à risque observé (ou dérivé) et l'estimation $\widehat{S}$ de son espérance avec ou sans covariables.

a. Les indices basés sur une dispersion

Comme dans le cadre de la régression logistique (voir p. 35), Schemper (1990) a proposé une généralisation de la notion de « **proportion de variation expliquée** » dans le cadre de la survie :

$$PEV = \frac{D(T) - D(T|Z)}{D(T)} = \frac{\sum_{i=1}^{n} D(t_i) - \sum_{i=1}^{n} D(t_i|Z_i)}{\sum_{i=1}^{n} D(t_i)} \tag{3.9}$$

Dans cette expression, $D(t_i)$ et $D(t_i|Z_i)$ désignent des mesures de **dispersion** non-conditionnelle et conditionnelle à la valeur des covariables, qui dépendent du temps au travers des fonctions de survie. Les indices PEV présentés ci-après sont tous compris entre 0 et 1 et diffèrent par leurs définitions de $D\left(t_i\right)$ et $D\left(t_i \mid Z_i\right)$.

• Schemper (1990, 1994) a introduit deux indices reposant sur la dispersion entre les fonctions de survie estimées, conditionnelles ou non aux covariables, et un processus $YL_i(t)$ analogue du processus à risque $Y_i(t)$ et défini par

$$YL_i(t_{(j)}) = \begin{cases} 0 & \text{si } T_i < t_{(j)} \quad \text{(individu déjà décédé ou censuré)} \\ 1/2 & \text{si } T_i = t_{(j)} \quad \text{(individu décédé ou censuré en } t_{(j)}\text{)} \\ 1 & \text{si } T_i > t_{(j)} \quad \text{(individu à risque)} \end{cases}$$

Dans un premier temps, l'auteur a proposé l'indice V_2 où les mesures de dispersion sont données par :

$$D(t_i) = \left(\frac{1}{k_i} \sum_{j=1}^{k_i} \left| YL_i(t^*_{(j)}) - \widehat{S}_{KM}(t^*_{(j)}) \right| \right)^2$$

$$D(t_i|Z_i) = \left(\frac{1}{k_i} \sum_{j=1}^{k_i} \left| YL_i(t^*_{(j)}) - \widehat{S}(t^*_{(j)}|Z_i) \right| \right)^2$$

où k_i a deux définitions différentes selon que le sujet i est décédé ou censuré. Si le sujet i est décédé, k_i est le nombre total de décès non censurés (i.e. $k_i = k$), les auteurs utilisent dans ce cas l'information de l'ensemble des individus décédés. Lorsque l'individu i est censuré, k_i est le nombre de décès avant le temps considéré ($k_i = \sum_{l|t_{(l)} \leq t_i} \delta_l$), les auteurs utilisent l'information disponible sur tous les individus décédés avant le temps de censure.
Enfin, la fonction $\widehat{S}(t|z)$ est l'estimateur de la fonction de survie en présence de covariables (équation (3.7) p. 50), et $\widehat{S}_{KM}(t)$ est l'estimateur de la fonction de survie marginale (équation (3.3) p. 48).

Finalement, l'indice est donné par

$$V_2 = \frac{\displaystyle\sum_{i=1}^{n}\left(\frac{1}{k_i}\sum_{j=1}^{k_i}\left|YL_i(t^*_{(j)}) - \widehat{S}_{KM}(t^*_{(j)})\right|\right)^2 - \sum_{i=1}^{n}\left(\frac{1}{k_i}\sum_{j=1}^{k_i}\left|YL_i(t^*_{(j)}) - \widehat{S}(t^*_{(j)}|Z_i)\right|\right)^2}{\displaystyle\sum_{i=1}^{n}\left(\frac{1}{k_i}\sum_{j=1}^{k_i}\left|YL_i(t^*_{(j)}) - \widehat{S}_{KM}(t^*_{(j)})\right|\right)^2}$$

L'indice V_2 peut être considéré comme une généralisation des indices basés sur la somme des écarts carrés.

Une alternative à l'indice V_2 est le coefficient V_1, qui est basé sur la même somme des écarts mais sans le carré (Schemper, 1990, 1994), avec :

$$D(t_i) = \frac{1}{k_i}\sum_{j=1}^{k_i}\left|YL_i(t^*_{(j)}) - \widehat{S}_{KM}(t^*_{(j)})\right|$$
$$D(t_i|Z_i) = \frac{1}{k_i}\sum_{j=1}^{k_i}\left|YL_i(t^*_{(j)}) - \widehat{S}(t^*_{(j)}|Z_i)\right|$$

et donc

$$V_1 = \frac{\displaystyle\sum_{i=1}^{n}\frac{1}{k_i}\sum_{j=1}^{k_i}\left|YL_i(t^*_{(j)}) - \widehat{S}_{KM}(t^*_{(j)})\right| - \sum_{i=1}^{n}\frac{1}{k_i}\sum_{j=1}^{k_i}\left|YL_i(t^*_{(j)}) - \widehat{S}(t^*_{(j)}|Z_i)\right|}{\displaystyle\sum_{i=1}^{n}\frac{1}{k_i}\sum_{j=1}^{k_i}\left|YL_i(t^*_{(j)}) - \widehat{S}_{KM}(t^*_{(j)})\right|}$$

Ces deux indices quantifient le **degré avec lequel la prédiction du processus de survie est améliorée** en remplaçant l'estimateur de la survie marginale par l'estimateur de la survie conditionnelle du processus. Ils permettent de valider ou d'invalider le modèle. Une proportion de variation expliquée faible peut indiquer que la prédiction n'est pas la meilleure possible et que d'autres facteurs pronostiques doivent être envisagés.

• Schemper et Henderson (2000) proposent deux indices, V et V_w, présentant plusieurs avantages par rapport aux indices V_2 et V_1. Ils améliorent la prise en compte de la censure en distinguant les individus à risque, décédés et censurés et sont plus robustes aux mauvaises spécifications du modèle, comme l'ont montré des simulations. Les indices V et V_w sont basées sur des mesures de **déviation absolue moyenne** (voir définition ci-après), et s'écrivent, après estimation, sous la forme de somme des écarts (i.e. comme dans la formule (3.9), p. 52).

Définition 3.12 *La **déviation absolue moyenne** d'une population (« mean absolute deviation » en anglais) est définie par (Read, 2006)*

$$d = \mathbb{E}\left(|U - \mathbb{E}(U)|\right)$$

Son estimation à partir d'un échantillon $u_1, \cdots, u_n$ *s'écrit*

$$d = \frac{1}{n}\sum_{i=1}^{n} |u_i - \bar{u}| \quad avec \quad \bar{u} = \frac{1}{n}\sum_{i=1}^{n} u_i$$

Par exemple, si U suit une loi binomiale de paramètres (n, p), la déviation absolue moyenne vaut $d = |1 - \mathbb{E}(U)| \Pr\{U = 1\} + |0 - \mathbb{E}(U)| \Pr\{U = 0\} = 2p(1 - p)$.

- *Cadre théorique*

Schemper et Henderson considèrent le processus à risque $Y(t)$ de déviation absolue moyenne égale à $2S(t)(1 - S(t))$, puisque $0 \leq S(t) \leq 1$. A partir de cette dernière quantité, ils définissent, d'un point de vue théorique, d'une part, la notion de capacité de prédiction marginale (en l'absence de covariables) :

$$D(\tau) = \frac{2\displaystyle\int_0^\tau S(t)(1 - S(t))f(t)dt}{\displaystyle\int_0^\tau f(t)dt}$$

et, d'autre part, la notion de capacité de prédiction conditionnelle (avec les covariables) :

$$D_Z(\tau) = \frac{2\displaystyle\int_0^\tau \mathbb{E}_Z\left[S(t|Z)(1 - S(t|Z))\right] f(t)dt}{\displaystyle\int_0^\tau f(t)dt}$$

où $f(t)$ est la densité marginale de T et $\mathbb{E}_Z$ est l'espérance selon la loi des covariables.

Afin de quantifier le gain de la capacité de prédiction dû aux covariables, ils ont proposé l'indice théorique suivant :

$$V(\tau) = \frac{D(\tau) - D_Z(\tau)}{D(\tau)}$$

qui s'interprète en termes de **proportion de variation expliquée** par les covariables.
Un indice théorique alternatif est donné par la formule :

$$V_w(\tau) = \left\{\int_0^\tau f(t)dt\right\}^{-1} \times \int_0^\tau \frac{S(t)(1 - S(t)) - \mathbb{E}_Z\left[S(t|Z)(1 - S(t|Z))\right]}{S(t)(1 - S(t))} f(t)dt$$

- *Estimation*

Schemper et Henderson ont proposé des estimateurs de $D(\tau)$, $D_Z(\tau)$, $V(\tau)$ et $V_w(\tau)$ en présence de censure. Pour ce faire, ils définissent la capacité de prédiction comme la différence entre un « processus de survie observé », dont ils donnent une définition presque identique à celle du processus à risque $Y(t)$, et les fonctions de survie prédites avec ou sans covariables, $\widehat{S}(t|Z_i)$ et $\widehat{S}_{KM}(t)$ entre les temps 0 et τ. Le « processus de survie observé » prend la valeur 1 si l'individu i est vivant, à risque au temps t, la valeur 0 si l'individu i décède ou est déjà décédé au temps t et n'est pas défini si l'individu est censuré avant le temps t.

Il en résulte que la **déviation absolue moyenne** entre les processus de survie observé et prédit en l'absence de covariables, $\widehat{B}(t_{(j)})$, s'écrit

$$\widehat{B}\left(t_{(j)}\right) = \frac{1}{n}\sum_{i=1}^{n}\left\{\mathbf{1}_{\{T_i>t_{(j)}\}}\left|1-\widehat{S}_{KM}(t_{(j)})\right| + \delta_i\mathbf{1}_{\{T_i\leq t(j)\}}\left|0-\widehat{S}_{KM}(t_{(j)})\right|\right.$$
$$\left.+(1-\delta_i)\mathbf{1}_{\{T_i\leq t(j)\}}\left[\left|1-\widehat{S}_{KM}(t_{(j)})\right|\frac{\widehat{S}_{KM}(t_{(j)})}{\widehat{S}_{KM}(t_i)} + \left|0-\widehat{S}_{KM}(t_{(j)})\right|\left(1-\frac{\widehat{S}_{KM}(t_{(j)})}{\widehat{S}_{KM}(t_i)}\right)\right]\right\}$$

Dans cette expression, le premier terme s'applique aux individus vivants, à risque en $t_{(j)}$, le deuxième s'applique aux individus décédés avant ou en $t_{(j)}$ et le troisième aux individus censurés avant ou en $t_{(j)}$. Concernant le troisième groupe d'individus, l'hypothèse est faite que leur risque de décès ainsi que leur probabilité de survie est identique à ceux des individus dont le statut est connu en $t_{(j)}$. Dû au fait que leur processus de survie n'est pas défini, ils sont considérés, dans le dernier terme, comme vivants avec la probabilité (conditionnelle au fait qu'ils soient vivants en t_i) $\widehat{S}_{KM}(t_{(j)})/\widehat{S}_{KM}(t_i)$ ou comme décédés avec la probabilité $1-\widehat{S}_{KM}(t_{(j)})/\widehat{S}_{KM}(t_i)$.
De plus, $\widehat{B}\left(t_{(j)}|Z\right)$, la déviation moyenne absolue entre les processus de survie observé et prédit en présence de covariables, est obtenue en remplaçant, dans l'expression des $\widehat{B}(t_{(j)})$ ci-dessus, l'estimateur de Kaplan-Meier $\widehat{S}_{KM}(t)$ (équation (3.3) p. 48) de la fonction de survie par l'estimateur de Cox $\widehat{S}(t|Z)$ de la fonction de survie (équation (3.7) p. 50).

Les estimateurs de V et V_w sont alors

$$\widehat{V} = \frac{w^{-1}\sum_{j=1}^{n}\widehat{G}(t_{(j)})^{-1}\delta_{(j)}\widehat{B}(t_{(j)}) - w^{-1}\sum_{j=1}^{n}\widehat{G}(t_{(j)})^{-1}\delta_{(j)}\widehat{B}(t_{(j)}|Z)}{w^{-1}\sum_{j=1}^{n}\widehat{G}(t_{(j)})^{-1}\delta_{(j)}\widehat{B}(t_{(j)})}$$

et

$$\widehat{V}_w = w^{-1}\sum_{j=1}^{n}\widehat{G}(t_{(j)})^{-1}\delta_{(j)}\left[\frac{\widehat{B}(t_{(j)})-\widehat{B}(t_{(j)}|Z)}{\widehat{B}(t_{(j)})}\right]$$

avec

$$w = \sum_{j}\widehat{G}(t_{(j)})^{-1}\delta_{(j)}$$

Dans ces expressions, $\widehat{G}(t)$ est une pondération égale à la fonction de survie de la variable de censure, qui se calcule comme $\widehat{S}_{KM}(t)$ mais en utilisant l'indicatrice de censure $(1-\delta)$ à la place de l'indicatrice de décès δ (Schemper et Smith, 1996). Elle reflète l'amélioration de la qualité du suivi lorsque le temps de suivi augmente et permet de compenser la perte d'information liée à la censure.
Les auteurs montrent que $\widehat{V}$ et $\widehat{V}_w$ sont des estimateurs convergents de V et V_w.
Ils ont une préférence pour $\widehat{V}$ par rapport à $\widehat{V}_w$, car la liaison entre $\widehat{V}$ et la notion de proportion de variation expliquée, comme définie en (3.9) p. 52, est plus directe qu'avec $\widehat{V}_w$. Les deux indices $\widehat{V}$ et $\widehat{V}_w$ sont présentés comme permettant de valider le modèle. En efftet, lorsque le pourcentage

de variation expliquée est faible, la valeur prédictive des covariables, même lorsqu'elles sont hautement significatives, est alors perçue comme faible, évitant ainsi la sur-interprétation des résultats.

• Enfin, Schemper (2003) propose l'indice $\widehat{V}_s$ suivant :

$$\widetilde{V}_s = \frac{w^{-1}\sum_{j=1}^{n}\widehat{G}(t_{(j)})^{-1}\delta_{(j)}\widetilde{B}(t_{(j)}) - w^{-1}\sum_{j=1}^{n}\widehat{G}(t_{(j)})^{-1}\delta_{(j)}\widetilde{B}(t_{(j)}|Z)}{w^{-1}\sum_{j=1}^{n}\widehat{G}(t_{(j)})^{-1}\delta_{(j)}\widetilde{B}(t_{(j)})}$$

mais avec

$$\widetilde{B}(t_{(j)}) = 2\widehat{S}_{KM}(t_{(j)})\left(1-\widehat{S}_{KM}(t_{(j)})\right)$$

et

$$\widetilde{B}(t_{(j)}|Z) = \frac{2}{n}\sum_{i=1}^{n}\widehat{S}(t_{(j)}|z_i)\left(1-\widehat{S}(t_{(j)}|Z_i)\right)$$

L'indice $\widetilde{V}_s$ est une version simplifiée de V, pour lequel Schemper utilise la déviation absolue moyenne calculée de façon globale, sans séparer les individus décédés, à risque et censurés en chaque temps. L'indice $\widetilde{V}_s$ quantifie la **gain relatif de capacité de prédiction** entre le modèle tenant compte des covariables par rapport au modèle nul. Comme pour les indices précédents, l'indice $\widetilde{V}_s$ permet de valider le modèle et de vérifier si des covariables à effets importants améliore la capacité de prédiction du moèle.

b. Les indices basés sur une fonction de perte

Comme dans le cadre de la régression logistique (cf. p. 36), Korn et Simon (1990) ont proposé des indices de variation expliquée, adaptés à l'analyse de survie, basés sur des **fonctions de perte $\boldsymbol{L(y,\tilde{y})}$** entre la valeur d'une variable y et sa prédiction $\tilde{y}$. Les auteurs n'utilisent pas directement les temps de survie pour définir leurs fonctions de perte, mais se basent sur le processus à risque $Y(t)$ qui vaut 1 si l'individu est toujours vivant en t avec la probabilité $S(t)$ et 0 sinon avec la probabilité $(1-S(t))$. Soit $L(y(t),\widetilde{y}(t))$ la perte entre la valeur observée $y(t)$ de la variable $Y(t)$ et sa prédiction $\widetilde{y}(t)$. La fonction de risque est définie comme la perte minimale par

$$R_L(t) = \min_{\widetilde{y}(t)}\int L(y(t),\widetilde{y}(t))dF_Y(y(t)) = \min_{\widetilde{y}(t)}\mathbb{E}\left[L(y(t),\widetilde{y}(t))\right]$$

où F_Y est la fonction de répartition de la variable de Bernoulli $Y(t)$. La quantité $\widetilde{y}(t)$ qui minimise cette fonction est appelée « prédicteur optimal ». Le risque s'exprime alors pour un temps donné. La généralisation est obtenue en intégrant le risque R_L sur le temps.

Notons que Korn et Simon utilisent la définition de la fonction de risque la plus fréquemment utilisée. Il existe d'autres définitions dans le cadre des statistiques bayésiennes (voir Berger, 1985).

Dans le modèle tenant compte des covariables, le risque R_L est calculé à partir du processus $Y_i(t), i = 1, \cdots, n$ d'espérance $S(t|z_i)$. Dans le modèle nul, i.e. sans covariables, R_{L0} est calculé pour le processus $Y_i(t)$ d'espérance $\bar{S}(t) = (1/n) \sum_{i=1}^{n} S(t|z_i), \forall i = 1, \cdots, n$.

Le tableau 3.1 donne un aperçu des différentes fonctions de perte proposées par les auteurs, ainsi que les prédicteurs optimaux et les risques à un temps donné et intégré associés.

Par exemple, pour l'erreur au carré, la fonction de perte vaut $[y(t) - \widetilde{y}(t)]^2$. Son espérance est minimale pour $\widetilde{y}(t) = \mathbb{E}[Y(t)]$, soit $S(t)$. Le risque associé vaut alors $\int [y(t) - \mathbb{E}[Y(t)]]^2 \, dF(y(t)) = \mathbb{E}\left[(Y(t) - \mathbb{E}(Y(t)))^2\right] = \mathbb{V}[Y(t)]$, c'est-à-dire $S(t)(1-S(t))$. Pour obtenir le risque intégré, il suffit ensuite d'écrire l'intégrale du risque au temps t sur toute la période de suivi $(0, \tau)$. En pratique, le risque intégré se calcule en faisant la somme discrète sur les temps de décès ordonnés.

L'erreur de prédiction au carré est facile à interpréter et recommandée par Korn et Simon en pratique. L'entropie est utile pour donner plus de poids aux erreurs de prédiction relatives aux premiers temps de survie.
La **variation expliquée par le modèle** de survie avec covariables est définie comme la réduction du risque obtenu avec le modèle de prédiction par rapport au modèle nul :

$$KS = \frac{\int R_{L0}(t)dt - \frac{1}{n}\sum_{i=1}^{n} \int R_L(t|Z_i)dt}{\int R_{L0}(t)dt}$$

L'indice KS est compris entre 0 et 1. Dans cette expression, $\int R_{L0}(t)dt$ correspond à la somme de ce que Schemper appelle dispersion non-conditionnelle $D(T)$ et $\sum_{i=1}^{n} \int \frac{R_L(t|Z_i)}{n} dt$ est la moyenne des dispersions conditionnelles $D(T|Z)$ (voir équation (3.9) p. 52). L'indice KS permet de valider le modèle de survie et d'éviter une possible sur-estimation de sa capacité de prédiction, dans le cas de covariables hautement significatives mais à faible pouvoir prédictif.
Dans le cas de l'erreur au carré, l'**indice de variation expliquée** s'écrit

$$KS = \frac{\int \bar{S}(t)(1 - \bar{S}(t))dt - \frac{1}{n}\sum_{i=1}^{n} \int S(t|Z_i)(1 - S(t|Z_i))dt}{\int \bar{S}(t)(1 - \bar{S}(t))dt}$$

dont l'estimation est obtenue, comme mentionné plus haut, en calculant les intégrales comme des sommes discrètes sur les temps de décès ordonnés.

c. Les indices basés sur le score de Brier

Graf *et al.* (1999) ont proposé plusieurs indices de proportion de variation expliquée basés le **score de Brier**. Ces indices peuvent être interprétés d'une façon semblable à ceux proposés par Korn et Simon, en considérant le score de Brier comme une fonction de risque R_L particulière. Le score de Brier a initialement été développé en météorologie pour évaluer les erreurs commises

TABLEAU 3.1 – Exemples de fonctions de perte entre $y(t)$ et $\widetilde{y}(t)$, avec les prédicteurs optimaux et risques associés

Nom	$L(y(t), \widetilde{y}(t))$	Prédicteur optimal	Risque au temps t $R_L(t)$	Risque intégré R_L
1. Erreur au carré	$[y(t) - \widetilde{y}(t)]^2$	$S(t)$	$S(t)(1 - S(t))$	$\int S(t)(1 - S(t))dt$
2. Entropie	$-2\Big\{y(t) \log\{\widetilde{y}(t)\} + [1 - y(t)] \log[1 - \widetilde{y}(t)]\Big\}$	$S(t)$	$-2\Big\{S(t) \log S(t) + [1 - S(t)] \log[1 - S(t)]\Big\}$	$\int \Big\{S(t) \log S(t) + (1 - S(t)) \log(1 - S(t))\Big\}dt$
3. Erreur de prédiction	$\begin{cases} 1 & \text{si } \lvert y(t) - \widetilde{y}(t)\rvert > 1/2 \\ 1/2 & \text{si } \widetilde{y}(t) = 1/2 \\ 0 & \text{si } \lvert y(t) - \widetilde{y}(t)\rvert < 1/2 \end{cases}$	$\begin{cases} 1 & \text{si } S(t) > 1/2 \\ 1/2 & \text{si } S(t) = 1/2 \\ 0 & \text{si } S(t) < 1/2 \end{cases}$	$\min(S(t), 1 - S(t))$	$\int \min(S(t), 1 - S(t))\, dt$

par des prévisions météorologiques probabilistes (Brier, 1950; Winkler, 1967; Winkler et Murphy, 1968).

• Dans son article, Brier (1950) considère un évènement répété n fois et pouvant prendre r valeurs possibles. Par exemple, il s'intéresse à la présence ou absence de pluie ($r = 2$) pour $n = 10$ prévisions différentes. Pour chaque répétition $i, i = 1, \cdots, n$, la probabilité que l'évènement appartienne à la catégorie $j, j = 1, \cdots, r$ est noté f_{ij}, avec $\sum_j f_{ij} = 1, i = 1, \cdots, n$. De plus, Brier considère la quantité E_{ij} qui vaut 1 si l'évènement associé à la répétition i appartient bien à la catégorie j et 0 sinon. En reprenant l'exemple de la pluie avec $j = 1$ correspondant à la présence de pluie et $j = 2$ à l'absence de pluie, E_{i1} vaut 1 s'il a effectivement plu au moment correspondant à la répétition i et E_{i2} vaut 1 s'il n'a pas plu au moment de la répétition i.

Le **score défini par Brier** est le suivant :

$$B = \frac{1}{n} \sum_{j=1}^{r} \sum_{i=1}^{n} (E_{ij} - f_{ij})^2$$

Ici, la probabilité f_{ij} est calculée pour un modèle donné. Le score de Brier dépend donc du modèle utilisé pour le calcul des f_{ij}. Sous un modèle donné, il permet de quantifier le **degré de confiance** d'une prévision.

• Dans leur article, Graf *et al.* (1999) construisent des indices visant à comparer le score de Brier calculé dans le modèle sans covariable au score de Brier calculé dans le modèle avec covariables. Ces indices sont construits soit pour un temps t fixé soit globalement en intégrant les scores sur le temps. Ils sont compris entre 0 et 1. Pour définir les scores de Brier, les auteurs utilisent le même processus à risque observé que celui défini par Schemper pour l'estimation de V et V_w, qui vaut 1 si l'individu est vivant, à risque au temps t, 0 si l'individu est décédé avant ou au temps t, et qui n'est pas défini si l'individu est censuré avant ou en t.
Les auteurs s'intéressent d'abord au cas non-censuré, puis au cas censuré.

- *Cas non censuré.*

Dans le cadre de la survie, en l'absence de censure, le score de Brier, à un temps t fixé, calculé dans le modèle nul, est donné par la formule suivante :

$$BS_0(t) = \frac{1}{n} \sum_{i=1}^{n} \left(\mathbf{1}\{T_i > t\} - \widehat{S}_{KM}(t) \right)^2$$

où $\widehat{S}_{KM}(t)$ est l'estimateur de Kaplan-Meier de la fonction de survie (cf. équation (3.3) p. 48). Le score de Brier s'interprète comme le carré de l'**erreur de prédiction** entre les survies individuelles observées et leur prédiction estimée sur tout l'échantillon sous le modèle nul. Le score de Brier peut également être interprété comme une **fonction de risque** R_L particulière, pour reprendre la terminologie de Korn et Simon.

Sous le modèle tenant compte des covariables, le score de Brier est donné par

$$BS(t) = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{1}\{T_i > t\} - \widehat{S}(t|Z_i)\right)^2$$

Un indice de **proportion de variation expliquée** peut alors être défini, **à un temps** t fixé, par :

$$R^2_G(t) = \frac{BS_0(t) - BS(t)}{BS_0(t)}$$

Une version intégrée du score de Brier peut également être calculée. Sous le modèle nul, il s'écrit

$$IBS_0 = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left(\mathbf{1}(T_i > u) - \widehat{S}_{KM}(u)\right)^2 dW(u)$$

où W est une pondération qui tient compte la perte graduelle d'information au cours du temps liée à la baisse du nombre de patients de l'ensemble à risque (dû aux décès et aux censures). Dans leur article, Graf *et al.* proposent $W(u) = u/\tau$ ou encore $W(u) = [1 - \widehat{S}(u)]/[1 - \widehat{S}(\tau)]$. En pratique, l'intégrale est remplacée par la somme discrète sur tous les temps de décès ordonnés.

Sous le modèle avec covariables, la version intégrée du score de Brier IBS s'obtient manière similaire :

$$IBS = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left(\mathbf{1}(T_i > u) - \widehat{S}(u|Z_i)\right)^2 dW(u)$$

Un indice de **proportion de variation expliquée global** peut alors être défini :

$$R^2_G = \frac{IBS_0 - IBS}{IBS_0}$$

- *Cas censuré.*

Le score de Brier peut être adapté au cas de données censurées. Sous le modèle nul, à un temps t donné, il s'écrit comme suit :

$$\begin{aligned} BS^c_0(t) &= \frac{1}{n}\sum_{i=1}^{n}\Bigg[\left(0 - \widehat{S}_{KM}(t)\right)^2 \delta_i \mathbf{1}\{T_i \le t\}\left(1/\widehat{G}(T_i)\right) \\ &\quad + \left(1 - \widehat{S}_{KM}(t)\right)^2 \mathbf{1}\{T_i > t\}\left(1/\widehat{G}(t)\right)\Bigg] \end{aligned}$$

Ici, $\widehat{G}(t)$ est, comme pour les indices $\widehat{V}$ et $\widehat{V}_w$ de Schemper, une pondération définie par l'estimation de Kaplan-Meier de la fonction de survie de la censure calculée comme $\widehat{S}_{KM}(t)$ mais à partir de $(T, 1-\delta)$ et non pas (T, δ) (Schemper et Smith, 1996).
Dans l'expression ci-dessus, le premier terme correspond aux individus décédés avant t, dont le processus à risque $Y_i(t)$ est nul. Leur contribution au score de Brier est donc égale à $(0 - \widehat{S}_{KM}(t))$. Le deuxième terme correspond aux individus à risque en t pour lesquels $Y_i(t)$ vaut 1. Leur contribution au score de Brier vaut donc $(1 - \widehat{S}_{KM}(t))$. Enfin, pour les individus dont la censure a lieu

avant le temps t, la valeur de $Y_i(t)$ est inconnue et leur contribution au score de Brier ne peut pas être calculée. Ils n'apparaissent donc pas dans l'expression de $BS_0^c(t)$.
Par ailleurs, pour compenser la perte d'information due à la censure, les deux premiers termes de la somme sont respectivement pondérés par $1/\widehat{G}(T_i)$ et $1/\widehat{G}(t)$.
En remplaçant l'estimateur de Kaplan-Meier de la fonction de survie, $\widehat{S}_{KM}(t)$ (équation (3.3) p. 48), par l'estimateur calculé grâce au modèle de Cox, $\widehat{S}(t|Z)$ (équation 3.7 p. 50), on peut définir $BS^c(t)$.

On en déduit l'**indice de proportion de variation expliquée** suivant, calculé à un temps t fixé :

$$R^{2\,c}_G(t) = \frac{BS_0^c(t) - BS^c(t)}{BS_0^c(t)}$$

Comme précédemment, on peut intégrer $BS_0^c(t)$ (resp. $BS^c(t)$) sur le temps conduisant à une version intégrée du score de Brier IBS_0^c (resp. IBS^c). En pratique, l'intégrale est remplacée par la somme discrète sur tous les temps de décès ordonnés.
Il s'ensuit un **indice global** égal à la différence relative entre IBS_0^c et IBS^c :

$$R^{2\,c}_G = \frac{IBS_0^c - IBS^c}{IBS_0^c}$$

Ici encore, on peut faire le parallèle avec la terminologie de Schemper (équation (3.9) p. 52) en considérant IBS^c et IBS_0^c respectivement comme la somme des dispersions conditionnelles et non-conditionnelles. Différentes fonctions de risque peuvent être utilisées dans le calcul du score de Brier, conduisant à autant d'indices de variation expliquée. Le score de Brier est utilisé pour comparer l'effet prognostique de différentes covariables. Dans leur article, Graf *et al.* l'utilisent pour comparer des règles de classification pronostique basées sur des données cliniques dans le cadre d'une étude de caner du sein. Enfin, Le score de Brier dans sa version non intégrée peut également être utilisé, pour construire un indice de proportion de variance expliquée à un temps donné.

3.2.2 Les indices dérivés de la vraisemblance

Dans le modèle linéaire (p. 29 et 41), le R^2 peut s'écrire à partir de la vraisemblance : $R^2_{\text{LR}} = 1 - \exp\{-\text{LR}/n\}$, où LR est la **statistique du log-rapport de vraisemblance**. En analyse de survie, la formule a d'abord été reprise en utilisant la vraisemblance partielle du modèle, puis des modifications ont été apportées pour améliorer les propriétés de l'indice.

a. Les indices issus de la statistique du log-rapport de vraisemblance

L'ensemble des indices présentés dans cette section **compare la vraisemblance du modèle nul** (ne prenant pas en compte les covariables) **et du modèle alternatif** (prenant en compte toutes les covariables). Ils permettent de tester l'intérêt des covariables et de comparer

l'effet de différentes covariables sur la survie. Ils sont compris entre 0 et 1. Pour ces raisons, il sont interprétés en terme de pourcentage de variation expliquée.

• Plusieurs auteurs (Allison, 1995; Maddala, 1983; Magee, 1990) ont repris l'indice R^2_{LR} calculé avec la vraisemblance partielle du modèle de Cox :

$$\rho_n^2 = 1 - \left(\frac{\mathcal{L}(0)}{\mathcal{L}(\widehat{\beta})}\right)^{2/n} = 1 - \exp\left(-\frac{2}{n}\left[\log \mathcal{L}(\widehat{\beta}) - \log \mathcal{L}(0)\right]\right)$$

On rappelle que dans le cadre du modèle de Cox à risques proportionnels, la log-vraisemblance partielle s'écrit comme suit

$$\log\{\mathcal{L}(\beta)\} = \sum_{i=1}^{n} \delta_i \left[\beta' Z_i - \log\left\{\sum_{i=1}^{n} Y_i(t_i)\exp(\beta' Z_i)\right\}\right]$$

• Cependant, l'indice ρ_n^2 possède un majorant strictement inférieur à 1, qui vaut $R^2_{max} = 1 - \exp\left(\frac{2}{n} \times \log \mathcal{L}(0)\right)$ (Nagelkerke, 1991) . En effet, la vraisemblance partielle étant un produit de probabilités, son maximum est atteint pour $\mathcal{L}(\widehat{\beta}) = 1$.

Pour cette raison, Nagelkerke (1991) a proposé la **modification** suivante :

$$R_N^2 = \frac{\rho_n^2}{R^2_{max}}$$

• Une **version modifiée** de l'indice d'Allison a également été proposée par O'Quigley *et al.* (2005) :

$$\rho_k^2 = 1 - \exp\left(-\frac{2 \times \left[\log \mathcal{L}(\widehat{\beta}) - \log \mathcal{L}(0)\right]}{k}\right)$$

D'après les simulations d'Oquigley *et al.*, il apparaît en effet que ρ_k^2 est moins sensible à la censure que ρ_n^2.

L'interprétation de ces indices en terme de pourcentage de variation expliquée est peut-être un peu abusive, car ils reposent sur une transformation exponentielle de la différence entre la vraisemblance du modèle nul et du modèle alternatif.

b. Les indices de pourcentage « d'aléatoire » expliqué par le modèle

O'Quigley *et al.* ont introduit la notion de « **caractère aléatoire expliqué par le modèle** » (en anglais : « explained randomness »). Les indices d' O'Quigley *et al.* sont basés sur un **indice de gain d'information** entre le modèle nul (en l'absence de covariables) et le modèle alternatif (tenant compte des covariables) et sont compris entre 0 et 1. Ils sont utilisés pour évaluer la valeur prédictive d'un covariable sur la survie. Ils correspondent à la généralisation de l'expression (2.11) de la p. 32 dérivée de **l'information de Kullback-Leibler**. Les auteurs ont considéré le cas d'une covariable.

Comme nous l'avons évoqué dans le chapitre précédent, l'information de Kullback-Leibler n'est pas symétrique. Dans le cadre de la survie, elle peut s'écrire de deux façons différentes en faisant intervenir soit la distribution de la variable de survie X conditionnelle à la covariable Z, soit la distribution de la covariable Z conditionnelle à la survie X. Ainsi, deux indices distincts ρ^2_{KO} et ρ^2_{XOQ} ont été proposés (voir Xu et O'Quigley, 1999; O'Quigley *et al.*, 2005). Le premier indice repose sur la distribution de X conditionnelle à Z ; le second repose sur la distribution de Z conditionnelle à X.
Dans les deux cas, la forme générale de l'indice est :

$$\rho^2 = 1 - \exp(-\Gamma(\beta)) = 1 - \exp\left\{2I_{KL}(\beta;0)\right\}$$

où $I_{KL}(\beta;0)$ est l'information de Kullback-Leibler entre le modèle tenant compte de la covariable et le modèle nul.

• Le premier indice ρ^2_{KO} repose sur l'information de Kullback-Leibler entre le modèle alternatif et le modèle nul, calculée à partir de la **distribution de X conditionnelle à Z** :

$$I^{(1)}_{KL}(\beta;0) = \int_{\mathcal{Z}}\int_{\mathcal{X}} f(x|\beta;z)\log\left\{\frac{f(x|\beta;z)}{f(x|0;z)}\right\}dxdF_Z(Z)$$

Dans cette expression $\mathcal{Z}$ et $\mathcal{X}$ désignent les domaines de définition des variables Z et X, et F_Z est la fonction de répartition de Z.

Les auteurs présentent l'information de Kullback-Leibler de façon un peu différente, en l'écrivant comme la différence $I^{(1)}(\beta) - I^{(1)}(0)$ où $I^{(1)}$ est la quantité suivante :

$$I^{(1)}(\theta) = E\{\log f(X|Z;\theta)\} = \int_{\mathcal{Z}}\int_{\mathcal{X}} \log\left\{f(x|z;\theta)\right\} f(x|z;\beta)dxdF_Z(z)$$

Cette écriture permet de décomposer l'indice ρ^2_{KO} sous la forme $\rho^2_{KO} = \dfrac{\exp\{-2I^{(1)}(0)\} - \exp\{-2I^{(1)}(\beta)\}}{\exp\{-2I^{(1)}(0)\}}$.

L'information de Kullback-Leibler $I^{(1)}_{KL}$ est estimée par

$$\widehat{I}^{(1)}_{KL}(\widehat{\beta};0) = \frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{X}} f(x|\widehat{\beta};z)\log\left\{\frac{f(x|\widehat{\beta};z)}{f(x|0;z)}\right\}dx$$

Dans cette expression, la distribution marginale de Z a été remplacée par son estimateur empirique usuel. Pour pouvoir calculer l'information de Kullback-Leibler, il faut se placer dans un cadre paramétrique et supposer une distribution particulière de la variable X. Le cas particulier où X suit une distribution de Weibull a été étudié par les auteurs et a permis de donner une estimation de ρ^2_{KO} (Kent et O'Quigley, 1988).

• Pour le deuxième indice ρ^2_{XOQ}, l'information de Kullback-Leibler entre le modèle alternatif et le modèle nul est calculée différemment, et repose non pas sur la densité f de X conditionnelle à Z, mais sur **celle g de Z conditionnelle à X** :

$$I^{(2)}_{KL}(\beta;0) = \int_{\mathcal{X}}\int_{\mathcal{Z}} g(z|\beta;x)\log\left\{\frac{g(z|\beta;x)}{g(z|0;x)}\right\}dzdF(x)$$

F étant la fonction de répartition de X.
Comme précédemment, l'information de Kullback-Leibler peut être décomposée comme la différence $I^{(2)}(\beta) - I^{(2)}(0)$ avec

$$I^{(2)}(\theta) = E\{\log g(Z|X;\theta)\} = \int_{\mathcal{X}} \int_{\mathcal{Z}} \log\{g(z|x;\theta)\}\, g(z|x;\beta) dz dF(x)$$

L'indice ρ^2_{XOQ} peut alors s'écrire comme la différence relative $\rho^2_{XOQ} = \dfrac{\exp\{-2I^{(2)}(0)\} - \exp\{-2I^{(2)}(\beta)\}}{\exp\{-2I^{(2)}(0)\}}$.

L'estimation de l'information de Kullback-Leibler $I^{(2)}_{KL}$ est obtenue en remplaçant $dF(x)$ par l'estimation du « saut » de l'estimateur de Kaplan-Meier en chaque temps de suivi observé. De plus, la densité $g(z|x;\beta)$ est définie sous le modèle de Cox. Elle peut donc être estimée à partir de l'expression de la vraisemblance partielle sous ce modèle par la probabilité $\pi_j(t;\widehat{\beta})$ que l'individu j, caractérisé par la valeur de sa covariable Z_j, décède au temps t étant donné tous les individus à risque au temps t. Son expression est la suivante :

$$\pi_j(t;\beta) = Y_j(t) \exp\left[\beta' Z_j(t)\right] / \sum_{l=1}^{n} Y_l(t) \exp\left[\beta' Z_l(t)\right]$$

Finalement, Γ est estimé par

$$\widehat{\Gamma}(\widehat{\beta}) = 2 \sum_{i=1}^{n} \left(\widehat{F}(t_i+) - \widehat{F}(t_i)\right) \sum_{j=1}^{n} \pi_j(t_i;\widehat{\beta}) \log\left(\frac{\pi_j(t_i;\widehat{\beta})}{\pi_j(t_i;0)}\right)$$

L'expression ci-dessus est divisée par $\sum_{i=1}^{n}\left(\widehat{F}(t_i+) - \widehat{F}(t_i)\right)$. Cette dernière quantité peut en effet être inférieure à 1 lorsque le dernier temps de suivi est censuré. Diviser par cette quantité permet de ramener la fonction de répartition entre 0 et 1.

• Les auteurs proposent pour ces deux indices une interprétation en terme de **pourcentage d'aléatoire expliqué**. En effet, dans le premier cas, l'indice peut s'écrire sous la forme $\rho^2 = \{D(X) - D(X|Z)\}/D(X)$ avec $D(X) = \exp\{-2I^{(1)}(0)\}$ et $D(X|Z) = \exp\{-2I^{(1)}(\beta)\}$, $I^{(1)}$ estimé à partir de la distribution de X sachant Z, et s'interprète comme un indice de pourcentage d'aléatoire de X sachant les covariables Z. Dans le second cas, l'indice est présenté sous la forme $\rho^2 = D(Z) - D(Z|X)/D(Z)$ avec $D(Z) = \exp\{-2I^{(2)}(0)\}$ et $D(Z|X) = \exp\{-2I^{(2)}(\beta)\}$, $I^{(2)}$ estimé à partir de la distribution de Z sachant X, et possède une interprétation en pourcentage d'aléatoire de Z sachant X.
La notion de « pourcentage de variation expliquée par le modèle » définie par Schemper (3.9, p. 52) n'est cependant pas exactement transposable pour ces indices puisque l'on ne peut pas les écrire sous la forme $\dfrac{\sum_i D(t_i) - \sum_i D(t_i|Z_i)}{\sum_i D(t_i)}$.

3.2.3 Les indices basés sur la notion de corrélation

Comme dans le cadre du modèle logistique (p. 40), plusieurs **coefficients de corrélation** ont été proposés pour l'analyse de survie. Leur **carré** peut être utilisé pour définir des indices

de type pseudo-R^2 compris entre 0 et 1. Les coefficients présentés dans la suite ne s'appliquent que dans le cas d'une seule covariable. Ils permettent d'évaluer l'intérêt prédictif des covariables sur la survie. Seul l'un d'entre eux (coefficient de Kendall) est directement utilisable en présence de données censurées. Pour d'autres, une procédure permettant de compléter les échantillons censurés en échantillons non censurés a été proposée (voir paragraphe e.).

a. Coefficient de corrélation de Kendall

Le coefficient de Kendall permet de mesurer la concordance entre l'ordre des temps de survie et ceux de la covariable Z.

- *Cas non censuré.*

Dans le cas de données de survie non censurées, le coefficient de Kendall (Kendall et Gibbons, 1990) entre un temps de survie X et une unique covariable Z s'écrit :

$$\tau(X,Z) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}b_{ij}}{(\sum_{i,j} a_{ij}^2 \sum_{i,j} b_{ij}^2)^{1/2}}$$

avec

$$a_{ij} = \begin{cases} 1 & \text{si } Z_i > Z_j \\ 0 & \text{si } Z_i = Z_j \\ \text{-1} & \text{si } Z_i < Z_j \end{cases}$$

et

$$b_{ij} = \begin{cases} 1 & \text{si } X_i > X_j \\ 0 & \text{si } X_i = X_j \\ \text{-1} & \text{si } X_i < X_j \end{cases}$$

- *Cas censuré.*

Dans le cas censuré, Brown *et al.* (1974) ont proposé deux possibilités pour généraliser le coefficient de Kendall. Dans les deux cas, les a_{ij} sont définis comme ci-dessus. La différence réside dans le calcul des b_{ij}.

• Dans un premier temps, Brown *et al.* proposent d'évaluer les b_{ij} de la façon suivante :

$$b_{ij} = \begin{cases} 1 & \text{si } X_i > X_j \\ 0 & \text{si } X_i = X_j \text{ ou si incertain} \\ \text{-1} & \text{si } X_i < X_j \end{cases}$$

le cas « incertain » correspond au cas où l'on ne peut pas être sûr de l'ordre d'apparition de l'événement entre les individus i et j du fait de la censure. Le tableau 3.2 donne la valeur de b_{ij} en fonction de la valeur du couple (δ_i, δ_j) et de l'ordre de T_i et T_j. Les cas « incertains » apparaissent en bleu.

TABLEAU 3.2 – Valeur des b_{ij} en fonction de la censure et de l'ordre des temps de survie

(δ_i, δ_j)	$T_i > T_j$	$T_i = T_j$	$T_i < T_j$
$(1,1)$	1	0	-1
$(0,1)$	1	1	0
$(1,0)$	0	-1	-1
$(0,0)$	0	0	0

• La deuxième possibilité pour le calcul des b_{ij}, basée sur les estimateurs de Kaplan-Meier $\widehat{S}_{KM}$ de la survie (équation (3.3) p. 48) calculés en T_i et T_j, est résumée dans le tableau 3.3. Ces nouveaux b_{ij} permettent d'éviter d'avoir des cas incertains et ne se calculent non pas sur les X mais sur les T. Sauf dans le cas « $T_i = T_j$ », les b_{ij} sont égaux à $2\Pr\left\{X_i > X_j | \delta_i, \delta_j, T_i, T_j, \widehat{S}_{KM}\right\}$, en supposant $\widehat{S}_{KM}$ connue. Par exemple, dans le cas $T_i < T_j$ et $(\delta_i, \delta_j) = (1,0)$, la probabilité $\Pr\left\{X_i > X_j | \delta_i, \delta_j, T_i, T_j, \widehat{S}_{KM}\right\}$ est prise égale à la probabilité conditionnelle que $X_j > T_j$ est inférieur à $X_i = T_i$ (voir aussi Efron, 1967).

TABLEAU 3.3 – Valeur des b_{ij} basées sur l'estimateur de Kaplan Meier

(δ_i, δ_j)	$T_i > T_j$	$T_i = T_j$	$T_i < T_j$
$(1,1)$	1	0	-1
$(0,1)$	1	1	$2\dfrac{\widehat{S}_{KM}(T_j)}{\widehat{S}_{KM}(T_i)} - 1$
$(1,0)$	$1 - 2\dfrac{\widehat{S}_{KM}(T_i)}{\widehat{S}_{KM}(T_j)}$	-1	-1
$(0,0)$	$1 - \dfrac{\widehat{S}_{KM}(T_i)}{\widehat{S}_{KM}(T_j)}$	$1 - \dfrac{\widehat{S}_{KM}(T_i)}{\widehat{S}_{KM}(T_j)}$	$\dfrac{\widehat{S}_{KM}(T_j)}{\widehat{S}_{KM}(T_i)} - 1$

b. Coefficient de Somers

En l'absence de censure, le coefficient de Somers (1962) est le suivant :

$$D = \frac{\sum_{i<j} \text{sign}(T_i - T_j)\text{sign}(Z_i - Z_j)}{\sum_{i<j} \text{sign}^2(T_i - T_j)}$$

Ce coefficient vaut 1 lorsque les T_i et les Z_i sont rangés dans le même ordre et -1 lorsqu'ils sont rangés en ordre contraire.

c. Coefficient de corrélation de Spearman

Ce coefficient s'applique en l'absence de censure.

Dans le cas de la survie, on a :

$$r_s(T,Z) = \frac{\sum_{i=1}^{n} R(T_i)R(Z_i) - n\left(\frac{n+1}{2}\right)^2}{\sqrt{\sum_{i=1}^{n}\left(R(T_i)^2 - n\left(\frac{n+1}{2}\right)^2\right)\sum_{i=1}^{n}\left(R(Z_i)^2 - n\left(\frac{n+1}{2}\right)^2\right)}}$$

où $R(T)$ désigne le rang de T.

d. Coefficient de Pearson transformé

Un dernier coefficient permet de quantifier la corrélation entre la survie et les facteurs pronostiques dans le cas non-censuré (Schemper et Stare, 1996) :

$$r_{pr}(T,Z) = \frac{\sum_{i=1}^{n}\left(R(T_i) - \bar{R}\right)\left(Z_i - \bar{Z}\right)}{\sqrt{\sum_{i=1}^{n}\left(R(T_i) - \bar{R}\right)^2\sum_{i=1}^{n}\left(Z_i - \bar{Z}\right)^2}} \quad \text{avec} \quad \bar{R} = \frac{n+1}{2}$$

Cet indice est le coefficient de corrélation de Pearson entre Z et l'ordre des T. Il est cohérent avec le caractère semi-paramétrique du modèle de Cox, qui est invariant par transformation monotone de T, mais pas de Z. En revanche, aucune généralisation au cas censuré n'a été proposée.

e. Estimation de plusieurs coefficients de corrélation en présence de censure

Schemper et Kaider (1997) ont proposé un algorithme pour pouvoir utiliser les coefficients de Spearman r_s et Pearson transformé r_{pr} sur des données censurées. Le coefficient de Kendall τ, qui, quant-à-lui, peut être utilisé en présence de censure, a cependant été inclus dans ce travail par les auteurs. Schemper et Kaider proposent donc de **reconstruire un jeu de données non censurées** à partir de données de survie censurées. Pour ce faire, ils pratiquent des imputations multiples reposant sur des régressions linéaires (Rubin, 1991, 1987). La procédure de reconstruction de données est répétée plusieurs fois et la moyenne des coefficients de corrélation des jeux de données reconstruits est calculée.
Notons enfin que ce travail aurait également pu être appliqué au coefficient de Somers, bien que les auteurs n'y fassent pas référence.

3.2.4 Les indices de concordance

La concordance est un concept utilisé en analyse discriminante et qui peut être appliqué en analyse de survie. Elle quantifie le **degré d'accord entre les rangs des temps de décès et ceux des covariables explicatives** et permet de mesurer la **capacité de prédiction** du modèle. Elle est utilisée pour déterminer l'intérêt des covariables sur la survie. L'indice de

concordance le plus répandu est l'aire sous la courbe ROC (AUC), utilisée dans le cas de données binaires non censurées. Des indices plus complexes adaptés à des données censurées ont ensuite été proposés. Tous ces indices sont compris entre 0 et 1.

a. Aire sous la courbe ROC : définition et estimation dans le cas particulier d'un critère de décès binaire.

- *Définition*

Le cadre est celui d'une étude où chaque sujet est suivi pendant une durée fixée à l'issue de laquelle il peut être soit décédé, soit vivant.

Soit $\phi(\mathbf{Z})$ un prédicteur linéaire du décès calculé à partir des covariables de chaque sujet, et dont il s'agit d'évaluer la capacité prédictive. Il peut s'agir par exemple de la fonction de risque linéaire $\beta'\mathbf{Z}$ estimée par un modèle logistique. On suppose que plus la valeur de $\phi(\mathbf{Z})$ est élevée, plus la probabilité de survenue du décès est grande. La règle de classification évaluée repose sur le choix d'un seuil v tel qu'un sujet soit classé comme « décédé » si $\phi(\mathbf{Z}) > v$, et comme « non décédé » si $\phi(\mathbf{Z}) \leq v$. A chaque valeur v choisie pour seuil correspond une **sensibilité** Se et une **spécificité** Sp, qui sont les proportions de sujets correctement classés dans le groupe des décédés et dans celui des non décédés, respectivement. On peut écrire :

$$\begin{aligned} \text{Se} &= \Pr\{\phi(\mathbf{Z}) > v | \delta = 1\} \\ \text{Sp} &= \Pr\{\phi(\mathbf{Z}) < v | \delta = 0\} \end{aligned}$$

où δ est la variable indicatrice du décès.

Par définition, la **courbe ROC** représente la sensibilité Se en fonction de (1 - Sp) lorsque le seuil v parcourt l'ensemble de ses valeurs possibles.

Soit un couple (i, j) de sujets, tels que le sujet i est tiré au hasard dans la population des individus décédés et le sujet j est tiré au hasard dans la population des individus vivants. L'**aire sous la courbe ROC** est donnée par :

$$AUC_{ij} = \Pr\{\phi(Z_i) > \phi(Z_j) | \delta_i = 1; \delta_j = 0\}$$

Cette expression suppose que la distribution de $\phi(Z_i)$ conditionnelle à $\delta_i = 1$ (respectivement $\delta_i = 0$) est identique pour tous les sujets décédés (respectivement vivants).

Preuve. Soit $F^i(v)$ la probabilité que, pour un seuil v donné, le sujet i classé comme décédé appartienne au groupe des individus décédés, i.e. $F^i(v) = \Pr\{\phi(Z_i) < v | \delta_i = 1)\}$. Soit $G^j(v)$ la probabilité que, pour un seuil v donné, le sujet j classé comme décédé appartienne au groupe des individus vivants, i.e. $G^i(v) = \Pr\{\phi(Z_i) < v | \delta_i = 0)\}$. L'aire sous la courbe ROC, qui représente la fonction de survie de $\phi(Z)$ dans la population des vivants en fonction de la fonction de survie de $\phi(Z)$ parmi les décédés, s'écrit, pour un couple (i, j), comme suit :

$$AUC_{ij} = \int_{-\infty}^{+\infty} F^i(u) dG^j(u)$$

Par ailleurs, on a

$$\begin{aligned}\Pr\{\phi(Z_i) > \phi(Z_j)|\delta_i = 1; \delta_j = 0\} &= \int_{-\infty}^{+\infty}\int_{-\infty}^{u} dF_i(y)dG_j(u) \\ &= \int_{-\infty}^{+\infty} F_i(u)dG_j(u)\end{aligned}$$

Par conséquent,

$$AUC_{ij} = \Pr\{\phi(Z_i) > \phi(Z_j)|\delta_i = 1; \delta_j = 0\}$$

□

L'aire sous la courbe ROC mesure la concordance entre la covariable et l'indicatrice de décès. $AUC = 1$ correspond à une discrimination maximale ; $AUC = 0.5$ correspond à l'absence de discrimination.

- *Estimation*

Une estimation de l'aire sous la courbe ROC est donnée ci-après pour un échantillon aléatoire constitué de $n + m$ sujets, dont n sont décédés (indexés par $i = 1, \cdots, n$) et m sont non-décédés (indexés par $j = n + 1, \cdots, n + m$). Soit $\phi(Z_1), \cdots, \phi(Z_{n+m})$ les valeurs du prédicteur linéaire de $Z_1, \cdots, Z_{n+m}$. Pour $i = 1, \cdots, n$ et $j = n + 1, \cdots, n + m$, on définit la fonction indicatrice suivante

$$conc_{ij} = \mathbf{1}\{\phi(Z_i) > \phi(Z_j)\} ,$$

qui vaut 1 si $\{i, j\}$ sont « concordants », i.e. si l'ordre des valeurs du prédicteur linéaire est en accord avec le statut décédé/non-décédé des patients, et 0 sinon.
L'aire sous la courbe ROC est estimée par le **rapport du nombre de paires « concordantes » sur le nombre de paires « comparables »** (ici il s'agit de toutes les paires possibles dont le nombre est de $n \times m$) dans l'échantillon :

$$\widehat{AUC} = \frac{\sum_{i=1}^{n}\sum_{j=n+1}^{m} conc_{ij}}{n \cdot m}$$

Cet estimateur est directement relié à la statistique de Mann et Whitney (1947) pour comparer la distribution de $\phi(\mathbf{Z})$ entre le groupe des décédés et celui des vivants.

Notons enfin que des extensions de la courbe ROC à des variables dépendantes du temps ont été proposées (voir Heagerty *et al.*, 2000), ainsi que leur lien avec l'indice de Harrell (décrit dans ce qui suit). La notion de courbe ROC a également été étendue dans le cadre de variables multi-niveaux, (par exemple Ma et Huang, 2007).

b. Indice de Harrell en présence de censure

- *Définition*

L'**indice C de Harrell *et al.* (1982)** est un indice de **concordance** reposant sur une **analogie avec l'aire sous la courbe ROC** dans le cas de données censurées. L'indice de Harrell estime la probabilité que, sur deux patient choisis au hasard, le patient avec le prédicteur linéaire $\phi(\mathbf{Z})$ le plus grand vivra plus longtemps que le patient avec le prédicteur le plus petit. Pour un couple (i, j) d'individus, cette probabilité peut s'écrire

$$C_{ij} = \Pr\{\phi(Z_i) > \phi(Z_j) | T_i < T_j; \delta_i = 1\}$$

où δ_i est l'indicatrice d'apparition de l'événement au temps t_i.
Lorsque cette probabilité vaut 0.5, le prédicteur linéaire ne permet pas de prédire quel patient vivra plus longtemps. Des valeurs proches de 0 ou 1 permettent de prédire quel patient aura les meilleures chances de survie.
L'indice C de Harell s'écrit comme une combinaison des C_{ij} pour tous les couples (i, j) et son estimation est donnée plus loin (p. 71).

- *Autre formulation de l'indice de Harrell*

Heagerty et Zheng (2005) proposent de **reformuler l'indice de Harrell** en intégrant un indice analogue à l'aire sous la courbe ROC défini en un temps donné. Dans une première étape, les auteurs définissent les notions de spécificité et sensibilité pour un temps $t_{(l)}$ fixé. A partir de ces quantités, ils construisent une aire sous la courbe ROC au temps $t_{(l)}$, qui, moyenné sur le temps, permet alors de retrouver l'indice C de Harrell.

Soit $\delta(t_{(l)})$ le statut au temps $t_{(l)}$, défini uniquement pour les individus à risque en $t_{(l)}$, qui vaut 1 si le sujet est décédé en $t_{(l)}$ et 0 si le sujet est vivant après $t_{(l)}$.
Pour définir les notions de spécifité et sensibilité à un temps fixé, les auteurs utilisent une règle de classification similaire à celle présentée pour l'AUC. Pour un seuil v donné, un sujet est classé comme « décédé en $t_{(l)}$ » si $\phi(\mathbf{Z}) \leq v$, et comme « non-décédé jusqu'en $t_{(l)}$ » si $\phi(\mathbf{Z}) > v$. Les **sensibilité et spécificité au temps** $t_{(l)}$ sont respectivement définies comme suit :

$$\mathrm{Se}(t_{(l)}) = \Pr\left\{\phi(\mathbf{Z}) > v | \delta(t_{(l)}) = 1\right\}$$

$$\mathrm{Sp}(t_{(l)}) = \Pr\left\{\phi(\mathbf{Z}) \leq v | \delta(t_{(l)}) = 0\right\}$$

Un indice $AUC^*(t_{(l)})$ est ensuite construit pour un temps $t_{(l)}$ donné. Pour un couple de sujets (i,j), où le sujet i appartient au groupe des individus décédés au temps $t_{(l)}$ et le sujet j appartient au groupe des individus à risque non-décédés au temps $t_{(l)}$, l'**aire sous la courbe ROC au temps** $t_{(l)}$ est définie par la probabilité suivante :

$$AUC^*_{ij}(t_{(l)}) = \Pr\left\{\phi(Z_i) < \phi(Z_j) |\ \delta_i(t_{(l)}) = 1; \delta_j(t_{(l)}) = 0\right\}$$

Une moyenne pondérée des $AUC^*_{ij}(t_{(l)})$ permet d'écrire l'indice suivant :

$$AUC^*_{ij} = \frac{\sum_{l=1}^{n} AUC^*_{ij}(t_{(l)}) \cdot \omega_{ij}(t_{(l)})}{\sum_{l=1}^{n} \omega_{ij}(t_{(l)})}$$

où

$$\omega_{ij}(t_{(l)}) = \Pr\left\{\delta_i(t_{(l)}) = 1; \delta_j(t_{(l)}) = 0\right\}$$

On montre que cette écriture est équivalente à l'indice de Harrell C_{ij} (Antolini *et al.*, 2005).

- *Estimation*

Pour un échantillon de n sujets, l'indice C est estimé par

$$\widehat{C} = \frac{\sum_{i=1}^{n} \sum_{j=1;i\neq j}^{n} \text{conc}_{ij}}{\sum_{i=1}^{n} \sum_{j=1;i\neq j}^{n} \text{comp}_{ij}}$$

où

$$\text{comp}_{ij} = \mathbf{1}\left\{T_i < T_j; \delta_i = 1\right\} + \mathbf{1}\left\{T_i = T_j; \delta_i = 1; \delta_j = 0\right\}$$

et

$$\text{conc}_{ij} = \mathbf{1}\left\{\phi(Z_i) > \phi(Z_j)\right\} \cdot \text{comp}_{ij}$$

La quantité $comp_{ij}$ vaut 1 si le couple (i, j) est comparable, i.e. lorsque l'ordre des temps de décès entre les individus i et j peut être déterminé de façon certaine, et 0 dans le cas où l'ordre n'est pas connu du fait de la censure. La quantité $conc_{ij}$ vaut 1 si le couple est concordant, i.e. lorsque l'ordre des valeurs du prédicteur linéaire $\phi(\mathbf{Z})$ est en accord avec l'ordre des temps de décès.

Pencina et D'Agostino (2004) ont proposé de construire un intervalle de confiance pour l'indice de Harrell en se basant sur la normalité asymptotique de son estimateur.

c. Indice d'Antolini pour des données censurées dépendantes du temps

- *Définition*

Dans leur article, Antolini *et al.* (2005) proposent un **indice qui étend la notion de concordance dans le cas de variables dépendantes du temps**, i.e. $\mathbf{Z} = \mathbf{Z}(\mathbf{t})$ (dans la suite la notation $\mathbf{t}$ correspond aux temps où les covariables varient).

La démarche utilisée dans ce travail est très proche de celle introduite par Heagerty et Zheng (2005). La différence réside dans la définition de la règle de classification, qui est basée,

non pas sur le prédicteur linéaire $\phi(\mathbf{Z})$, mais sur la fonction de survie prédite. Ainsi, pour un seuil v donné, un sujet est classé comme « décédé en $t_{(l)}$ » si $S(t_{(l)}|\mathbf{Z}(\mathbf{t})) \leq v$, et comme « non-décédé jusqu'en $t_{(l)}$ » si $S(t_{(l)}|\mathbf{Z}(\mathbf{t})) > v$.

Par analogie avec le travail de Heagerty et Zheng (2005), un indice $C_{ij}^{td}(t_{(l)})$ est défini pour un temps $t_{(l)}$ donné et un couple (i, j) de sujets :

$$C_{ij}^{td}(t_{(l)}) = \Pr\left\{S(t_{(l)}|Z_i(\mathbf{t})) < S(t_{(l)}|Z_j(\mathbf{t})) \mid \delta_i(t_{(l)}) = 1; \delta_j(t_{(l)}) = 0\right\}$$

Il représente la capacité de discrimination de la fonction de survie $S(t_l|\mathbf{Z}(t))$.

Finalement, l'indice d'Antolini moyenné sur le temps s'écrit, pour un couple (i,j) de sujets, de la manière suivante :

$$C_{ij}^{td} = \Pr\left\{S(T_i|Z_i(\mathbf{t})) < S(T_i|Z_j(\mathbf{t})) \mid T_i < T_j; \delta_i = 1\right\}$$

Notons que dans le cas d'un modèle à risques proportionnels, i.e. où les covariables ne dépendent pas du temps, l'ordre des fonctions de survie est identique à l'ordre des valeurs du prédicteur linéaire. L'indice d'Antolini est alors égal à l'indice de Harrell.

- *Estimation*

Pour un échantillon de n sujets, l'indice d'Antolini est estimé par

$$\widehat{C}^{td} = \frac{\sum_{i=1}^{n}\sum_{j=1;j\neq i}^{n} conc_{ij}^{td}}{\sum_{i=1}^{n}\sum_{j=1;j\neq i}^{n} comp_{ij}}$$

avec comp_{ij} défini comme pour l'indice de Harell :

$$\text{comp}_{ij} = \mathbf{1}\left\{T_i < T_j; \delta_i = 1\right\} + \mathbf{1}\left\{T_i = T_j; \delta_i = 1; \delta_j = 0\right\}$$

et

$$conc_{ij}^{td} = \mathbf{1}\left\{\widehat{S}(t_i|Z_i(\mathbf{t})) < \widehat{S}(t_j|Z_j(\mathbf{t}))\right\} \cdot comp_{ij}$$

Les auteurs construisent ensuite des intervalles de confiance de $\widehat{C}^{td}$ basés sur des méthodes « jacknife ».

3.3 Comparaison des indices

Qu'est ce qu'un « bon » indice de capacité de prédiction ? Pour tenter de répondre à cette question, plusieurs auteurs (O'Quigley *et al.*, 2005; Schemper et Stare, 1996; Xu et O'Quigley, 1999) ont proposé des **critères** afin d'évaluer les indices de capacité de prédiction en analyse de survie. Nous avons retenu les suivants :

(1) L'indice doit être compris entre 0 et 1.
(2) Lorsque les coefficients de régression sont égaux à 0, l'indice doit être égal à 0.
(3) L'indice doit être une fonction croissante de la valeur absolue des coefficients de régression.
(4) L'indice doit atteindre ses bornes, sans posséder de majorant strictement inférieur à 1 ou de minorant strictement supérieur à 0.
(5) La valeur de l'indice ne doit pas être modifiée en présence de censure.
(6) La valeur de l'indice ne doit pas être affectée par la taille de l'échantillon.
(7) L'indice ne doit pas être affecté par une transformation monotone sur l'échelle des temps.
(8) L'indice doit avoir une interprétation intuitive.
(9) L'indice doit être facilement transposable à des modèles de survie à risques non-proportionnels.

Dans le tableau qui suit, nous résumons les propriétés des différents indices. Certains de ces résultats ont été établis théoriquement, et d'autres ont été obtenus par simulations sans pouvoir faire de démonstration formelle. Les résultats des simulations proviennent soit des publications directement, soit de simulations complémentaires permettant de compléter les informations manquantes.

TABLEAU 3.4 – Résumé des propriétés des différents indices de capacité de prédiction

Mesures	Auteurs	Propriétés								
		$\rho \in [0,1]$	Si $\beta = 0$, $\rho = 0$	$\rho \nearrow$ quand $\lvert\beta\rvert \nearrow$	Atteints ses bornes	Affecté par la censure	Affecté par la taille	Affecté par une transf. des T_i	Interprétation intuitive	Transposition modèle NPH
KS	Korn, Simon	oui	oui	oui	oui	oui	*non*	oui	oui	*non*
$R^{2\,c}_G$	Graf et *al.*	oui	oui	oui	oui	*non*	*non*	*non*	oui	*non*
V_2	Schemper	oui	oui	oui	oui	oui	*non*	*non*	oui	*non*
V_1	Schemper	oui	oui	oui	oui	oui	*non*	*non*	oui	*non*
$\widehat{V}$	Schemper	oui	oui	oui	oui	oui	*non*	*non*	oui	*non*
$\widehat{V}_\omega$	Schemper	oui	oui	oui	oui	oui	*non*	*non*	*non*	*non*
$\tilde{V}_s$	Schemper	oui	oui	oui	oui	oui	*non*	*non*	*non*	*non*
ρ^2_n	Allison	oui	oui	oui	*non*	oui	*non*	*non*	*non*	*non*
R^2_N	Nagelkerke	oui	oui	oui	oui	oui	*non*	*non*	*non*	*non*
ρ^2_k	O'Quigley et *al.*	oui	oui	oui	oui	*non*	*non*	*non*	*non*	*non*
ρ^2_{KO}	Kent,O'Quigley	oui	oui	oui	oui	*non*	*non*	*non*	*non*	*non*
ρ^2_{XOQ}	Xu, O'Quigley	oui	oui	oui	oui	*non*	*non*	*non*	*non*	*non*
r^2_s	Spearman	oui	oui	oui	oui	NA	*non*	*non*	oui	*non*
τ^2	Kendall	oui	oui	oui	oui	oui	*non*	*non*	oui	*non*
D^2	Somers	oui	oui	oui	oui	oui	*non*	*non*	oui	*non*
r^2_{pr}	Pearson	oui	oui	oui	oui	oui	*non*	*non*	oui	*non*
$\widehat{\text{AUC}}$	Hanley	oui	*non*	*non*	oui	NA	*non*	*non*	oui	*non*
$\widehat{\text{C}}$	Harell	oui	oui	oui	oui	oui	*non*	*non*	oui	*non*
$\widehat{\text{C}}^{td}$	Antolini et *al.*	oui	oui	oui	oui	oui	*non*	*non*	oui	oui

En bleu : propriétés issues de la définition des indices ; en vert : propriétés résultant des simulations des articles ; en orange : propriétés résultant de simulations complémentaires à celles des différents articles

Les critères (1), (6) et (7) sont identiques pour tous les indices : ils sont compris entre 0 et 1 et ne sont pas affectés par une transformation sur l'échelle des temps. De plus, d'après les simulations, ils ne sont pas affectés par la taille de l'échantillon.

Le coefficient AUC est le seul à ne pas remplir les conditions (2) et (3). En effet, lorsque $\beta = 0$, AUC= 0.5 et lorsque $|\beta|$ augmente, AUC peut soit augmenter vers 1, soit diminuer vers 0 en fonction du signe de β.

L'indice d'Allison, quant-à-lui est le seul à ne pas atteindre ses bornes, car il possède un majorant strictement inférieur à 1, qui dépend notamment de la distribution des covariables.

Seuls quelques indices ne sont pas ou peu affectés par la censure (KS, $\widehat{V}$, $\bar{V}_s$, ρ_k^2, ρ_{KO}^2, ρ_{XOQ}^2).

D'un point de vue interprétation, certains indices sont basés sur des transformations exponentielles plus ou moins simples du rapport de vraisemblance (ρ_n^2, R_N^2, ρ_k^2, ρ_{KO}^2, ρ_{XOQ}^2) , d'autres, dans le but d'améliorer leur comportement vis-à-vis de la censure, reposent sur des pondérations complexes et sont difficilement interprétables ($\widehat{V}_\omega$, $\bar{V}_s$).

Remarquons également que, en refaisant quelques simulations, nous n'avons pas trouvé les mêmes résultats concernant le comportement vis-à-vis de la censure de certains indices. En particulier, les indices KS, $\widehat{V}$ et $\tilde{V}_s$ semblent affectés par la censure puisqu'ils diminuent fortement lorsque le pourcentage de censure augmente. De plus, l'indice de Graf *et al.* ne semble pas être affecté par la censure.

Enfin, très peu d'indices ont une distribution connue. L'indice d'Allison peut être relié à un chi-deux puisqu'il est basé sur une transformation de la statistique du rapport de vraisemblance. L'indice AUC est, quant-à-lui, estimé par la statistique de Wilcoxon. Enfin, le score de Brier peut être relié à une distribution normale (voir Lai *et al.*, 2010)

3.4 Conclusion

Ce chapitre donne un aperçu des indices de **capacité de prédiction** proposés dans la littérature dans le cadre de l'**analyse de survie**. Un certain nombre d'entre eux sont des généralisations d'indices proposés dans le cadre des modèles de régression linéaire ou logistique. Certains indices sont des extensions de la notion de pseudo-R^2 dans le cas de données censurées. Ils peuvent alors parfois s'écrire sous la forme de somme de dispersions conditionnelles et non-conditionnelles, ce que Schemper appelle « proportion de variation expliquée ». D'autres tentent de généraliser la notion d'aire sous la courbe ROC à l'analyse de survie.

Sans consensus sur la façon de définir la notion de capacité de prédiction en analyse de survie, un **grand nombre d'indices** ont été proposés. Cependant, aucun d'entre elles ne généralise la relation entre le coefficient de détermination et la **statistique du score** dans le cas linéaire.

Dans le chapitre suivant, un nouvel indice de capacité de prédiction, ou pseudo-R^2, basé sur la statistique du score est présenté.

Chapitre 4

MATÉRIELS ET MÉTHODES : PRÉSENTATION DE L'INDICE

Contenu

Dans ce chapitre, un nouvel indice de capacité de prédiction est proposé. Cet indice est en relation avec la statistique du score robuste pour tester l'hypothèse nulle $\mathcal{H}_0$ d'absence d'effet. Ce pseudo-R^2 est calculé pour trois grand types de modèles : le modèle de Cox à risques proportionnels, le modèle à odds proportionnels et un modèle à risques non-proportionnels conduisant à un croisement des fonctions de risque instantané. Dans un premier temps, la définition de ces deux derniers modèles est rappelée (le modèle de Cox ayant déjà été introduit dans le chapitre 3). Dans un second temps, le nouvel indice de capacité de prédiction est présenté (paragraphe 4.2.1), ainsi que ses principales caractéristiques (paragraphe 4.2.2). Le cas multivarié est ensuite étudié plus en détails dans le paragraphe 4.2.3, ainsi que la prise en compte des ex-æquo dans le paragraphe 4.2.4.

4.1 Deux modèles de survie à risques non-proportionnels

Les deux modèles présentés dans cette section permettent d'analyser des données pour lesquelles le rapport des risques instantanés n'est pas constant au cours du temps.

Les mêmes notations que celles introduites dans le chapitre 3 sont reprises (voir paragraphe 3.1.2 p. 48).

4.1.1 Un modèle à risques non-proportionnels dont les risques convergent : le modèle à odds proportionnels

Le **modèle à odds proportionnels** a initialement été introduit par Cox (1972) puis repris par Bennett (1983) pour analyser des données telles que les taux de mortalité de différents groupes de patients convergent dans le temps. Ce modèle permet de prendre en compte la variation réelle de l'effet au cours du temps. Il peut également être interprété en termes de fragilité (voir paragraphe c.).

a. Définition et Propriétés

Le **modèle semi paramétrique à odds proportionnels** peut se définir par la relation suivante

$$S(t|\mathbf{Z}) = \frac{1}{1 + e^{\beta' \mathbf{Z}} H_0(t)} \quad \text{avec} \quad S(t|0) = S_0(t) = \frac{1}{1 + H_0(t)} \tag{4.1}$$

où $S_0(t)$ désigne la fonction de survie en l'absence de covariables, et $H_0(t)$ est une fonction croissante du temps, telle que $H_0(0) = 0$. Le modèle peut également s'écrire sous la forme

$$\frac{S(t|\mathbf{Z})}{1 - S(t|\mathbf{Z})} = e^{-\beta' \mathbf{Z}} \frac{S_0(t)}{1 - S_0(t)} \tag{4.2}$$

Pour un individu $i, i = 1, \cdots, n$ le rapport

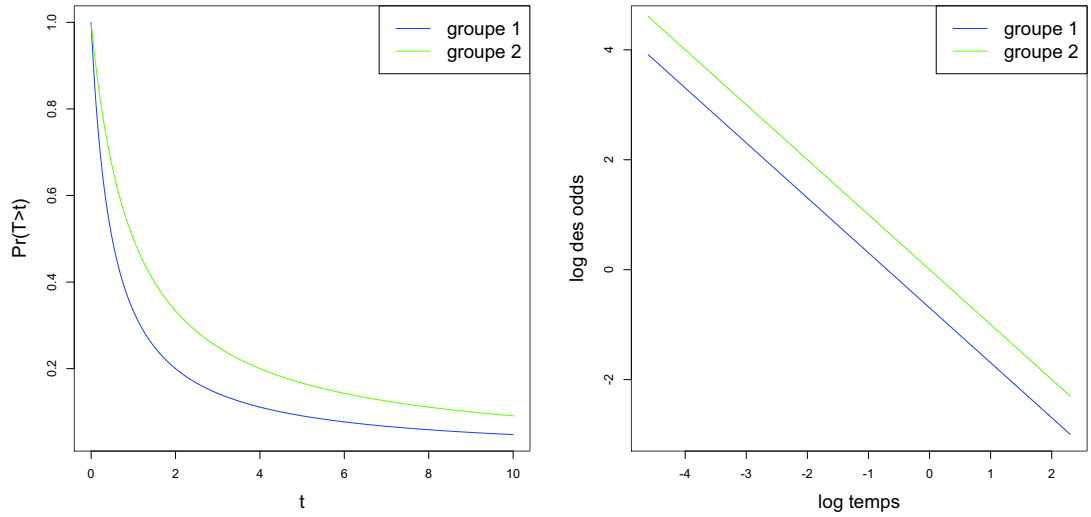
$$\text{OR} = \frac{S(t|Z_i)/(1 - S(t|Z_i))}{S_0(t)/(1 - S_0(t))}$$

est l'**odds ratio** entre la survie de l'un individu i et la survie d'un individu dont les variables explicatives sont nulles.

D'après (4.2), le logarithme de l'odds ratio est égal à $-\beta'\mathbf{Z}$, qui est constant au cours du temps, alors que le risque relatif est dépendant du temps. Le modèle est donc **à odds proportionnels** et **à risques non-proportionnels**.

La figure 4.1 (a) représente les courbes de survie théoriques de deux groupes d'individus définis par les niveaux associés à une covariable binaire. Les deux courbes ont été tracées avec un modèle log-logistique (i.e. $H_0(t) = t$), $e^{\beta} = 2$ pour le groupe 1 et $e^{\beta} = 1$ pour le groupe 2. La figure 4.1 (b), qui montre la courbe du log des odds en fonction du log-temps, permet de vérifier visuellement l'hypothèse des odds proportionnels (courbes parallèles).

FIGURE 4.1 – Courbes (a) de survie et (b) du log des odds théoriques pour deux groupes d'individus définis par les valeurs d'une covariable binaire, dans la cadre du modèle à odds proportionnels.



(a) (b)

Le calcul montre que le risque instantané $\lambda(t|\mathbf{Z})$ associé à $S(t|\mathbf{Z})$ dans (4.2) est donné par :

$$\lambda(t|\mathbf{Z}) = \frac{\lambda_0(t)}{1 - S_0(t) + S_0(t)e^{-\beta\mathbf{Z}}}$$

où $\lambda_0(t) = \lambda(t|0)$ est le risque instantané associé à $S_0(t)$.
Il s'ensuit que le rapport des risques instantanés de deux sujets de covariables Z_1 et Z_2 s'écrit

$$\frac{\lambda(t|Z_1)}{\lambda(t|Z_2)} = \frac{1 - S_0(t) + S_0(t)e^{-\beta Z_2}}{1 - S_0(t) + S_0(t)e^{-\beta Z_1}} = \frac{H_0(t) + e^{-\beta Z_2}}{H_0(t) + e^{-\beta Z_1}}$$

où $H_0(t) = \dfrac{1 - S_0(t)}{S_0(t)}$ est défini plus haut.
Il apparaît que ce rapport est une fonction croissante ou décroissante de temps selon le signe

de $Z_1 - Z_2$, valant $e^{\beta(Z_1-Z_2)}$ en $t = 0$ et 1 quand t tend vers $+\infty$. Par conséquent, **les risques convergent** dans le temps.

Remarquons enfin que le modèle log-logistique, de distribution $[1 + (t/\alpha)^\gamma]^{-1}$, avec $\alpha = e^{-betaZ}$, est un exemple de modèle à odds proportionnels. Mais, contrairement au modèle défini en 4.1 (sauf dans le cas particulier où $H_0(t)$ est linéaire), il est également à risques accélérés.

b. Estimation des paramètres

L'estimation des β par la méthode du maximum de vraisemblance avec $S_0(t)$ comme paramètre de nuisance ne peut être obtenue simplement. Plusieurs méthodes ont été proposées, reposant sur une estimation non paramétrique de $S_0(t)$, notamment dans les articles de Cheng *et al.* (1995), Cai *et al.* (2000), Yang et Prentice (1999) et Murphy *et al.* (1997). Ces méthodes sortent du cadre de ce travail et ne sont donc pas détaillées.

c. Interprétation en terme de fragilité

Le modèle à odds proportionnels peut s'interpréter comme un cas particulier du modèle marginal déduit du **modèle avec fragilité**.

Le concept de fragilité (en anglais : «frailty ») a été introduit par Vaupel *et al.* (1979) pour modéliser l'hétérogénéité intrinsèque individuelle ou partagée (e.g. familiale) de la population étudiée. Le modèle de fragilité prend en compte cette hétérogénéité en considérant un effet aléatoire dans la fonction de risque. Les modèles avec fragilité individuelle ont également été utilisés pour modéliser l'effet de variables cachées ou oubliées (Hougaard, 1995).

Le modèle avec fragilité individuelle considéré ici exprime le risque instantané conditionnel à la variable de fragilité ξ sous la forme à risques proportionnels. Pour un sujet i de covariable Z_i, ce risque instantané s'écrit :

$$\lambda(t|Z_i;\xi_i) = \lambda_0(t)\xi_i \exp(\beta' Z_i)$$

où ξ_i est une variable aléatoire attachée au sujet $i, i = 1, \cdots, n$ de fonction de répartition F_ξ et les $\xi_i, i = 1, \cdots, n$ sont supposées indépendantes.

La fonction de survie correspondante s'écrit

$$S(t|Z_i;\xi_i) = \exp\left\{-\int_0^t \xi_i \exp(\beta' Z_i)\lambda_0(u)du\right\} = \exp\left\{-\xi_i \exp(\beta' Z_i)H_0(t)\right\}$$
$$\text{en notant} \quad H_0(t) = \int_0^t \lambda_0(s)ds$$

et la fonction de survie marginale s'écrit

$$S(t|Z_i) = \int_0^\infty \exp\left\{-u \exp(\beta' Z_i)H_0(t)\right\} dF_\xi(u)$$

Cette quantité est la transformée de Laplace de ξ prise au point $\exp(\beta' Z_i)H_0(t)$.
Dans le cas où ξ_i suit une loi gamma de moyenne égale à 1 et de variance θ, que l'on note $\xi_i \sim \Gamma(1/\theta, 1/\theta)$, la fonction de survie devient

$$S(t|\mathbf{Z}) = \left\{1 + \theta e^{\beta' Z_i} H_0(t)\right\}^{-1/\theta}$$

Il s'agit de la loi de Pareto (voir Clayton et Cuzick, 1986).
Lorsque $\theta = 1$, on retrouve le modèle à odds proportionnels :

$$S(t|Z_i) = \frac{1}{1 + e^{\beta' Z_i} H_0(t)}$$

Le modèle à odds proportionnels peut donc être interprété comme un cas particulier du modèle marginal déduit d'un **modèle conditionnel avec une fragilité** de distribution gamma de moyenne 1 et de variance 1 de type « à risques proportionnels ».

4.1.2 Un modèle à risques non-proportionnels dont les risques se croisent

Le modèle le plus utilisé pour analyser les données d'expression en analyse de survie est le modèle de Cox à risques proportionnels. Ce dernier est largement employé car il permet de résumer l'effet moyen d'une variable sur le risque de base et la vraisemblance partielle permet une estimation simple des paramètres. Cependant, le modèle de Cox n'est pas adapté dans des situations où les risques instantanés se croisent, car l'effet de la variable explicative s'inverse au cours du temps et son effet moyen est proche de zéro.

Pour cette raison, nous proposons de considérer un modèle permettant d'identifier des facteurs dont l'effet s'inverse au cours du temps. Une interprétation de ce modèle est donné dans le paragraphe b..

a. Définition et Propriétés

Le modèle considéré ici est une version simplifiée de celui proposé par Quantin *et al.* (1996).

Pour un sujet $i, i = 1, \cdots, n$, on considère le **modèle semi-paramétrique à risques non-proportionnels** défini par la fonction de survie suivante :

$$S(t|\mathbf{Z}) = \exp\left\{-A_0(t)^{e^{\beta' \mathbf{Z}}}\right\} \tag{4.3}$$

où $A_0(t)$ est une fonction de risque cumulé arbitraire croissant de 0 à l'infini en fonction du temps, et $\beta = (\beta_1, \cdots, \beta_p)^T$ est un vecteur colonne de dimension p de paramètres à estimer. Cette expression est équivalente à

$$A(t\mathbf{Z}) = -\log\{S(t|\mathbf{Z})\} = A_0(t)^{e^{\beta' \mathbf{Z}}}$$

Le modèle est appartient aux modèle dits à « risques porportionnels exponentiels », de type $A(t|\mathbf{Z}) = \{A_0(t)\}^{g(\mathbf{Z})}$ avec $g(\mathbf{Z}) = \exp(\beta \mathbf{Z})$ (Wu, 2007; Devarajan et Ebrahimi, 2011).

Comme dans le cas du modèle de Cox, il s'agit ici d'un modèle **semi-paramétrique** car il est composé d'une partie paramétrique $\exp(\beta'\mathbf{Z})$ et d'une partie non spécifiée $A_0(t)$.

Pour un individu $i, i = 1, \cdots, n$, le modèle peut aussi s'écrire à l'aide du risque instantané de la manière suivante :

$$\lambda(t|Z_i) = \lambda_0(t)e^{\beta' Z_i} A_0(t)^{(e^{\beta' Z_i - 1})} \tag{4.4}$$

Comme précédemment, $\lambda_0(t)$ est la fonction de risque de base telle que $\int_0^t \lambda_0(s)ds = A_0(t)$.

Le rapport des risques instantanés, $\dfrac{\lambda(t|Z_i)}{\lambda(t|Z_j)} = e^{\beta(Z_i - Z_j)} A_0(t)^{(e^{\beta Z_i} - e^{\beta Z_j})}$, n'est pas constant au cours du temps ; les risques sont donc bien **non-proportionnels**.
De plus, pour $Z_i \geq Z_j$, si $\beta \geq 0$, le risque relatif est croissant de 0 vers $+\infty$ et est égal à 1 pour $\tau = A_0^{-1}\left[\exp\left\{-\dfrac{\beta(Z_i - Z_j)}{e^{\beta Z_i} - e^{\beta Z_j}}\right\}\right]$. Dans ce cas, le risque d'apparition de l'événement d'intérêt est plus petit pour le sujet i que pour le sujet j pour $0 \leq t \leq \tau$, et la tendance s'inverse pour $t \geq \tau$. Lorsque $\beta \leq 0$, le risque relatif est décroissant de $+\infty$ à 0 et est égal à 1 pour le même τ que défini précédemment. Le risque d'apparition de l'événement est donc plus grand pour le sujet i que pour le sujet j quand $0 \leq t \leq \tau$ et devient plus petit pour $t \geq \tau$. Par conséquent, **les risques se croisent** donc au temps τ.

Les courbes de survie théoriques de deux groupes d'individus définis par les deux valeurs d'une variable binaire sont représentées sur la figure 4.2 (a), avec $A_0(t) = t$, $e^{\beta} = 3$ pour le groupe 1 et $e^{\beta} = 1$ pour le groupe 2. Les deux courbes de survie se croisent au cours du temps (condition suffisante mais pas nécessaire pour que les risques se croisent). La courbe du log-risque cumulé en fonction du log-temps de la figure 4.2 (b) permet de vérifier la non-proportionnalité du modèle (courbes non parallèles).

b. Une interprétation biologique du croisement des risques

Dans ce paragraphe, nous montrons comment un effet, désigné sous le terme de **modulation**, entre deux marqueurs binaires $Z^{(1)}$ et $Z^{(2)}$ peut conduire à des **fonctions de risque marginales qui se croisent**. Dans la suite, la modulation fait référence au cas où l'un des marqueurs a un effet uniquement lorsque le second marqueur est présent, par exemple lorsque l'effet d'un traitement est modifié par une mutation génétique. L'effet du marqueur 1 (traitement) dépend de celui du marqueur 2 (mutation).

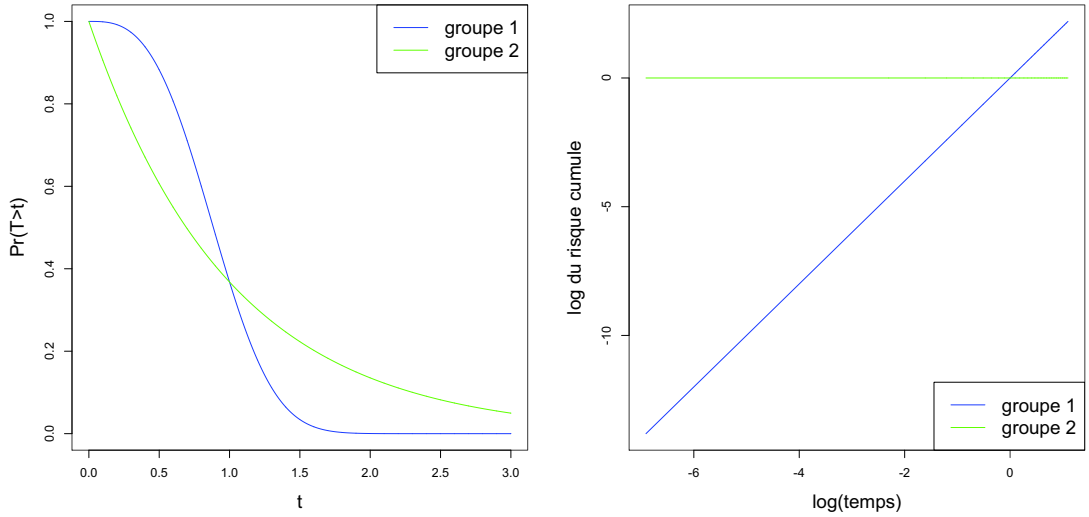
Soit la distribution jointe de $Z^{(1)}$ et $Z^{(2)}$ définie par :

$$p_{jj'} = \Pr\{Z^{(1)} = j; Z^{(2)} = j'\} \quad (j, j') \in \{0, 1\} \times \{0, 1\}$$

On suppose que la fonction de risque du sujet i avec $Z_i^{(1)} = j$ et $Z_i^{(2)} = j'$ est donnée par

$$\lambda(t|Z_i^{(1)} = j; Z_i^{(2)} = j') = \lambda_0(t) \exp\{(\alpha j + \gamma)j'\} \tag{4.5}$$

FIGURE 4.2 – Courbes (a) de survie et (b) du log-risque cumulé théoriques pour deux groupes d'individus définis par les deux valeurs d'une covariable binaire, dans le cadre du modèle à risques non-proportionnels.



(a) (b)

où $\lambda_0(t)$ est une fonction de risque de base arbitraire, et α et γ sont des coefficients de régression inconnus.

Le modèle (4.5) décrit l'effet de modulation entre les deux marqueurs, $Z^{(1)}$ et $Z^{(2)}$, dans lequel $Z^{(2)}$ a un effet multiplicatif sur le risque et $Z^{(1)}$ a un effet multiplicatif seulement si $Z^{(2)}$ est égal à un. Les fonctions de risques correspondantes en fonction des valeurs de $Z^{(1)}$ et $Z^{(2)}$ sont reportées dans le tableau 4.1.

TABLEAU 4.1 – Fonction de risque

	$Z^{(1)}$	
$Z^{(2)}$	0	1
0	$\lambda_0(t)$	$\lambda_0(t)$
1	$\lambda_0(t)e^{\gamma}$	$\lambda_0(t)e^{\alpha+\gamma}$

En supposant que le modèle (4.5) est vrai, les conséquences de l'omission de la variable $Z^{(2)}$ sur la formulation du risque relatif lié à $Z^{(1)}$ sont décrites ci-après.

L'expression du modèle (4.5) en terme de fonction de survie conditionnelle sachant $(Z_i^{(1)}, Z_i^{(2)})$ s'écrit :

$$S(t|Z_i^{(1)} = j, Z_i^{(2)} = j') = S_0(t)^{\exp\{(\alpha j+\gamma)j'\}}$$

où $S_0(t)$ est la fonction de survie correspondant à la fonction de risque de base $\lambda_0(t)$. La fonction de survie sachant $(Z_i^{(1)} = j)$ s'obtient directement à partir du théorème de Bayes :

$$S(t|Z_i^{(1)} = j) = \Pr(Z_i^{(2)} = 1|Z_i^{(1)} = j)S_0(t)^{\exp\{\alpha j+\gamma\}} + \Pr(Z_i^{(2)} = 0|Z_i^{(1)} = j)S_0(t)$$

La densité correspondante est

$$f(t|Z_i^{(1)}=j) = \Pr(Z_i^{(2)}=1|Z_i^{(1)}=j)\lambda_0(t)e^{\alpha j+\gamma}S_0(t)^{\exp\{\alpha j+\gamma\}} + \Pr(Z_i^{(2)}=0|Z_i^{(1)}=j)\lambda_0(t)S_0(t)$$

On en déduit la fonction de risque sachant $(Z_i^{(1)}=j)$:

$$\lambda(t|Z_i^{(1)}=j) = \frac{f(t|Z_i^{(1)}=j)}{S(t|Z_i^{(1)}=j)} = \lambda_0(t)\left[\frac{S_0(t)p_{j0} + e^{\alpha j+\gamma}S_0(t)^{e^{\alpha j+\gamma}}p_{j1}}{S_0(t)p_{j0} + S_0(t)^{e^{\alpha j+\gamma}}p_{j1}}\right]$$

Cette expression peut également être obtenue en prenant l'espérance de l'équation (4.5) conditionnellement à $Z^{(2)}$ étant donné le processus à risque. Enfin, le risque relatif du groupe de sujets tel que $(Z_i^{(1)}=1)$ par rapport au groupe tel que $(Z_i^{(1)}=0)$ s'écrit :

$$\frac{\lambda(t|Z_i^{(1)}=1)}{\lambda(t|Z_i^{(1)}=0)} = \left(\frac{p_{11}e^{\alpha+\gamma}S_0(t)^{e^{\alpha+\gamma}} + p_{10}S_0(t)}{p_{11}S_0(t)^{e^{\alpha+\gamma}} + p_{10}S_0(t)}\right) \times \left(\frac{p_{01}S_0(t)^{e^{\gamma}} + p_{00}S_0(t)}{p_{01}e^{\gamma}S_0(t)^{e^{\gamma}} + p_{00}S_0(t)}\right) \tag{4.6}$$

A partir de cette expression, on voit que **les risques peuvent se croiser au cours du temps**. Plus précisément, on montre que, lorsque α et γ sont positifs et en supposant une distribution jointe équilibrée pour le couple $(Z^{(1)}, Z^{(2)})$, les **risques relatifs s'inversent** pour un temps donné sur $]0;+\infty[$ (voir preuve ci-après). L'analyse marginale de l'effet de $Z^{(1)}$ par un modèle à risques proportionnels risque de conclure à l'absence d'effet de $Z^{(1)}$, alors que celle-ci a un effet modulé par $Z^{(2)}$.

Preuve. Le but est de démontrer que, sous certaines conditions, le risque relatif (4.6)

$$\mathrm{HR}(t) = \frac{\lambda(t|Z_i^{(1)}=1)}{\lambda(t|Z_i^{(1)}=0)} = \left(\frac{p_{11}e^{\alpha+\gamma}S_0(t)^{e^{\alpha+\gamma}} + p_{10}S_0(t)}{p_{11}S_0(t)^{e^{\alpha+\gamma}} + p_{10}S_0(t)}\right) \times \left(\frac{p_{01}S_0(t)^{e^{\gamma}} + p_{00}S_0(t)}{p_{01}e^{\gamma}S_0(t)^{e^{\gamma}} + p_{00}S_0(t)}\right)$$

vaut un pour un temps donné t_0 dans $]0;+\infty[$, et qu'il est supérieur à 1 pour $t < t_0$ et inférieur à 1 pour $t > t_0$.

Dans la suite, on suppose que $p_{11} = p_{10} = p_{01} = p_{00=1/4}$ et que $\alpha > 0$, et $\gamma > 0$. Les notations suivantes seront utilisées : $k = \alpha/\gamma$, $a = \exp(\gamma)$ avec $a > 1$, et $X = S_0(t)$ où X est croissant sur $]0,1]$ lorsque t est décroissant de $+\infty$ à 0. Pour $X > 0$, HR devient :

$$\begin{aligned}\mathrm{HR}(X) &= \left(\frac{a^{k+1}X^{(a^{k+1}-1)} + 1}{X^{(a^{k+1}-1)} + 1}\right)\left(\frac{X^{a-1} + 1}{aX^{a-1} + 1}\right) \\ &= \frac{a^{k+1}X^{(a^{k+1}+a-2)} + a^{k+1}X^{(a^{k+1}-1)} + X^{a-1} + 1}{aX^{(a^{k+1}+a-2)} + X^{(a^{k+1}-1)} + aX^{a-1} + 1} = \frac{N(X)}{D(X)}\end{aligned}$$

Notons que si X tend vers zéro (i.e. si t tend vers l'infini), $\mathrm{HR}(X)$ tend vers 1 (et $\mathrm{HR}(t)$ tend vers 1). Pour montrer l'existence et l'unicité de $0 < X_0 < 1$ tel que $\mathrm{HR}(X_0) = 1$, on s'intéresse à la différence $(N(X) - D(X))$, où $D(X) > 0$ pour $0 \leq X \leq 1$. Plus précisément, on note :

$$N(X) - D(X) = f(X)X^{a-1} \tag{4.7}$$

$$\text{avec } f(X) = (a^{k+1} - a)X^{(a^{k+1}-1)} + (a^{k+1} - 1)X^{a^{k+1}-a} + (1 - a) = 0.$$

Il est évident que la solution X_0 de $\mathrm{HR}(X) = 1$ est, si elle existe, également solution de l'équation $f(X) = 0$.
La dérivée première de f par rapport à X est égal à

$$\frac{\partial f(X)}{\partial X} = (a^{k+1} - a)(a^{k+1} - 1)\left[X^{(a^{k+1}-2)} + X^{a^{k+1}-a-1}\right]$$

Elle est strictement positive sur $]0;1[$ puisque $a > 1$ et $k > 0$, et donc f est strictement croissante sur $]0;1[$. Comme $f(0) < 0$ et $f(1) > 0$, l'équation $f(X) = 0$ possède une solution unique X_0 sur $]0;1[$. De plus, il s'en suit que, pour $0 < X < X_0$ (respectivement $X > X_0$), la fonction $f(X)$ est négative (resp. positive) et donc $(N(X) - D(X))$ est négative (resp. positive) comme le montre la formule (4.7) ci-dessus. Comme $(N(X) - D(X)) > 0$ (resp. > 0) est équivalent à $HR(X) > 1$, on en déduit qu'il existe un unique temps $t_0 = S_0^{-1}(X_0)$ avec $0 < t_0 < +\infty$ tel que $HR(t)$ est plus grand que un pour $t < t_0$, et plus petit que un pour $t > t_0$. On remarque que la fonction $HR(t)$ n'est pas monotone, puisque $HR(t)$ tends vers un quand t tends vers $+\infty$.

Les résultats précédents peuvent être résumés dans le tableau 4.2.

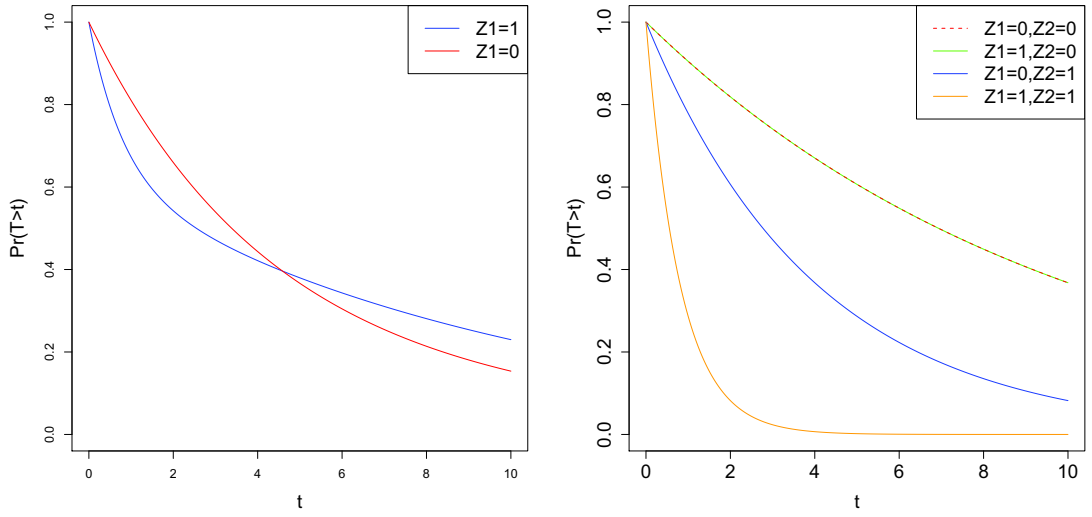
TABLEAU 4.2 – Résumé des signes de f et HR

X	0		X_0		1
$f(X)$	$1 - a < 0$	$-$	0	$+$	$2a^{k+1} - 2a > 0$
t	$+\infty$		t_0		0
$\mathrm{HR}(t)$	1^-	< 1	1	> 1	$\dfrac{e^{\gamma(k+1)} + 1}{e^{\gamma} + 1} > 1$

□

La figure 4.3 (a) représente les fonctions de survie marginales théoriques en fonction de la valeur de $Z^{(1)}$. Elle montre clairement que le risque relatif n'est pas monotone.
La figure 4.3 (b) représente les courbes de survies théoriques en fonction des groupes définis par la combinaison des valeurs de $Z^{(1)}$ et $Z^{(2)}$, avec $\alpha = \log(5)$, $\gamma = \log(2.5)$ et $\lambda_0(t) = 0.1$. Lorsque $Z^{(2)} = 0$, les deux courbes correspondants aux individus pour lesquels $Z^{(1)} = 0$ et $Z^{(1)} = 1$ se recouvrent entièrement. Lorsque $Z^{(2)} = 1$, on voit une différence entre les groupes définis par $Z^{(1)}$: le groupe pour lequel $Z^{(1)} = 0$ a une meilleure survie que le groupe pour lequel $Z^{(1)} = 1$. L'effet de $Z^{(1)}$ n'est perceptible que si $Z^{(2)}$ est présent.

FIGURE 4.3 – Courbes de survie théoriques en fonction (a) de la variable $Z^{(1)}$; (b) des groupes définis par la combinaison des valeurs de $Z^{(1)}$ et $Z^{(2)}$.



(a) (b)

Utiliser un modèle de Cox à risque proportionnels en ne considérant que la variable $Z^{(1)}$ conclurait à un non-rejet de l'hypothèse nulle $H_0 = \{\alpha = 0\}$, alors que $Z^{(1)}$ a un effet sur la survie à travers la variable omise $Z^{(2)}$. **Le modèle de Cox ne permet donc pas d'identifier des facteurs ayant un tel effet de modulation**, à la différence du modèle introduit précédemment.

4.1.3 Une écriture du score commune aux différents modèles

Dans le modèle de Cox, rappelons que la log-vraisemblance partielle s'écrit comme suit

$$\log \mathcal{L}(\beta) = \sum_{i=1}^{n} \int_0^\tau \left[\beta' Z_i - \log S^{(0)}(\beta; s)\right] dN_i(s)$$

A partir de cette expression, on peut aisément calculer le score en prenant la dérivée partielle par rapport à β. De la même façon, nous obtenons facilement la formule du score pour les modèles à odds proportionnels et à risques qui se croisent.

Ainsi, pour ces trois modèles, on peut écrire une **formule générale pour le score** calculé sous l'hypothèse nulle :

$$U(0) = \sum_{i=1}^{n} U_i(0) = \sum_{i=1}^{n} \int_0^\tau \left(Z_i - \frac{\sum_{l=1}^n Y_l(s) Z_l}{\sum_{l=1}^n Y_l(s)} \right) \omega(s) dN_i(s) \tag{4.8}$$

où $\omega(s)$ est une pondération qui dépend du modèle. La pondération $\omega(t)$ vaut 1 pour le modèle à risques proportionnels, $S_0(t)$ pour le modèle à odds proportionnels et $(1 + \log\{A_0(t)\})$ pour le modèle à risques non-proportionnels.

Il est intéressant de noter que, dans le cas des modèles à risques et à odds proportionnels, la pondération appartient à la famille des pondération proposées par Harrington et Fleming (1982), i.e. $\widehat{S}(t)^\rho, \rho > 0$, utilisées pour la construction d'une statistique du log-rang pondérée.

Il est important de noter que les U_i **ne sont pas indépendants**. En effet, on a

$$\mathbb{E}(U(0)) = 0$$

et

$$\mathbb{E}(n^{-1/2} U(0)^{\otimes 2}) = \frac{1}{n} \mathbb{E} \sum_{i=1}^{n} \int_0^\tau \left[Z_i \omega(s) - \frac{\sum_{l=1}^n Y_l(s) Z_l \omega(s)}{\bar{Y}(s)} \right]^{\otimes 2} Y_i(u) \lambda_0(s) ds$$

Et par conséquent la covariance des U_i n'est pas nulle.

En notant $\widehat{\omega}(t_i)$ un estimateur de $\omega(t_i)$ sous l'hypothèse nulle $\mathcal{H}_0$, un **estimateur du score** (que l'on notera $\widehat{U}$ à la place de $\widehat{U}(0)$ pour alléger l'écriture) est donné par

$$\widehat{U} = \widehat{U}(0) = \sum_{i=1}^{n} \widehat{U}_i(0) = \sum_{i=1}^{n} \delta_i \widehat{\omega}(t_i) \left(Z_i(t_i) - \frac{\sum_{l=1}^n Y_l(t_i) Z_l(t_i)}{\bar{Y}(t_i)} \right) \tag{4.9}$$

En pratique, pour le modèle à risques non-proportionnels dont les risques se croisent, $\widehat{\omega}(t) = 1 + \log(\widehat{A}_0(t))$ où $\widehat{A}_0(t)$ est l'estimateur de Nelson-Aalen (équation 3.2 p. 48). Pour le modèle à odds proportionnels, $\widehat{\omega}(t) = \widehat{S}_{KM}(t)$ est l'estimateur de Kaplan-Meier (équation 3.3, p. 48).

En utilisant les résultats des travaux de Pierce (1982), qui sont rappelés ci-après, et de la même façon que dans l'article de Broët *et al.* (2001), on montre que $\widehat{U}$ est un **estimateur convergent** de $U(0)$, de **distribution asymptotique normale**.
Soit T_n^ν une statistique dépendant d'un paramètre de nuisance ν et $\widehat{\nu}_n$ un estimateur de ν. La statistique obtenue en remplaçant ν par $\widehat{\nu}_n$ est notée $T_n^{\widehat{\nu}_n}$.
On suppose que T_n^ν possède une distribution normale et que : (i) pour tout ν, $n^{1/2} T_n^\nu$ et $n^{1/2}(\widehat{\nu}_n - \nu)$ convergent conjointement en loi vers une distribution normale ; (ii) on peut écrire un développement limité de T_n^ν tel que $n^{1/2} T_n^{\widehat{\nu}_n} = n^{1/2} T_n^{\nu_n} + B n^{1/2}(\widehat{\nu}_n - \nu) + 0_P(1)$, où $B = \lim \mathbb{E}(\partial T_n^{\nu_n} / \partial \nu)$; (iii) $\widehat{\nu}_n$ est asymptotiquement efficace.
Alors, d'après Pierce (1982), la distribution limite de $n^{1/2} T_n^{\widehat{\nu}_n}$ est une loi normale de moyenne nulle et de matrice de variance asymptotique $V_{11} - B V_{22} B^T$, où V_{11} et V_{22} sont les matrices de dispersion respectives de $T_n^{\nu_n}$ et $\widehat{\nu}_n$.

Ces résultats sont appliqués à $T_n^{\nu_n} = \widehat{U}(0)$ et $\widehat{\nu}_n = \widehat{\omega}(t)$. Dans notre cas, les conditions sont vérifiées pour les estimateurs de Kaplan-Meier de la fonction de survie et de Nelson-Aalen du risque cumulé, et donc pour $\widehat{\omega}(t) = \widehat{S}_0(t)$ dans le cas du modèle à odds proportionnels et pour $\widehat{\omega}(t) = 1 + \log \widehat{A}_0(t)$ dans le cas du modèle à risques qui se croisent (Kalbfleisch et Prentice, 2002). Par conséquent, $\widehat{U}(0)$ est un estimateur convergent de $U(0)$ pour les trois modèles considérés, et converge vers une loi normale de moyenne nulle.

4.2 Indice de séparabilité

4.2.1 Présentation de l'indice

L'indice est basé sur une interprétation originale du score calculé sous l'hypothèse nulle. En réarrangeant l'expression (4.8), on peut écrire, pour $i = 1, \cdots, n$:

$$U_i(0) = \int_0^\tau \left[\frac{\bar{Y}(s) - 1}{\bar{Y}(s)} \left(Z_i - \frac{\sum_{l=1; l \neq i}^n Y_l(s) Z_l}{\bar{Y}(s) - 1} \right) \right] \omega(s) dN_i(s)$$

On voit que les U_i peuvent s'écrire sous la forme d'une différence pondérée entre la valeur de la covariable du patient décédé en s et la moyenne des covariables du groupe de patients non encore décédés en s, que l'on appelle « **séparabilité** ». Ainsi, les U_i constituent une **mesure de séparabilité** entre les deux groupes de patients au temps s selon la valeur de la covariable. Des différences proches de zéro indiquent une séparabilité faible ou nulle ; des différences importantes indiquent que les deux groupes sont bien séparés. La figure 4.4 donne une représentation simplifiée de la notion de séparabilité, sur laquelle apparaissent, à chaque temps, la différence entre l'individu qui décèdent et les individus non-décédés.

Pour des raisons qui apparaîtront ci-après, au lieu des U_i, on utilise les composantes du **score robuste** W_i initialement proposées par Lin et Wei (1989) dans le cadre du modèle de Cox. Leur expression est la suivante :

$$W_i(0) = \int_0^\tau \left[Z_i - \frac{s^{(0)}(s)}{s^{(1)}(s)} \right] \omega(s) dN_i(s)$$

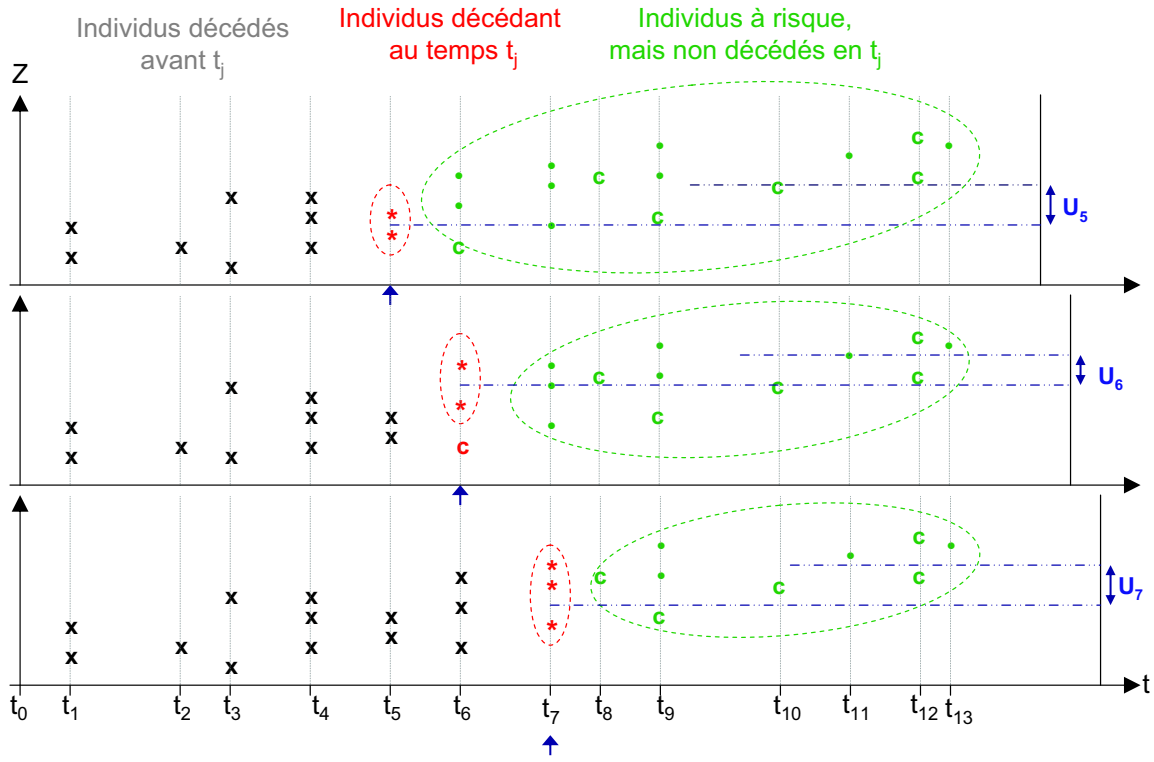
où

$$\begin{aligned} s^{(r)}(t) &= \mathbb{E}[S^{(r)}(t)],\ r = 0, 1 \\ S^{(0)}(t) &= \sum_{l=1}^n Y_i(t) \\ S^{(1)}(t) &= \sum_{i=1}^n Y_i(t) Z_i \end{aligned}$$

Une estimation des composantes du score robuste (que l'on notera $\widehat{W}_i$ au lieu de $\widehat{W}_i(0)$) est donnée par la relation suivante :

$$\begin{aligned} \widehat{W}_i &= \widehat{U}_i - \widehat{E}\widehat{U}_i \\ &= \delta_i \widehat{\omega}(t_i) \left(Z_i(t_i) - \frac{\sum_{l=1}^n Y_l(t_i) Z_l(t_i)}{\bar{Y}(t_i)} \right) - \sum_{l=1}^n \frac{Y_i(t_l) \delta_l \widehat{\omega}(t_l)}{\bar{Y}(t_l)} \left(Z_i(t_l) - \frac{\sum_{r=1}^n Y_r(t_l) Z_r(t_l)}{\bar{Y}(t_l)} \right) \end{aligned}$$

FIGURE 4.4 – Représentation schématique de la notion de séparabilité.
Sur le schéma, les individus déjà décédés ou censurés sont représentés par une croix. Les individus qui décèdent au temps considéré sont représentés par une étoile et ceux qui décéderont à un temps ultérieur par un point. Enfin, les individus censurés (au temps considéré ou ultérieurement) sont représentés par la lettre 'c'. Aux temps t_5, t_6 et t_7, on voit apparaître la différence entre les covariables des individus qui décèdent et celles des individus non-décédés.



Le terme $\widehat{E}\widehat{U}_i$ est une moyenne pondérée du score calculé aux temps t_l de décès précédant le temps t_i ($t_l \leq t_i$). La somme des $\widehat{W}_i$ est égale à la somme des $\widehat{U}_i$ (voir ci-dessous), mais les W_i sont iid alors que les U_i ne le sont pas. De plus, la somme des W_i constitue une statistique globale dont la valeur est grande si les groupes d'individus « décédés »/« non-décédés » sont bien séparés.
La **convergence de l'estimateur** $\widehat{W}$ se montre, comme dans le cas du score, en utilisant les travaux de Pierce (1982).

Il est facile de montrer que $\sum_i \widehat{W}_i = \sum_i \widehat{U}_i$, en montrant que $\sum_i \widehat{E}\widehat{U}_i = 0$:

$$\begin{aligned}
\widehat{E}\widehat{U}_i &= \sum_{i=1}^{n}\sum_{l=1}^{n} \frac{Y_i(t_l)\delta_l\widehat{\omega}(t_l)}{\bar{Y}(t_l)} \left(Z_i(t_l) - \frac{\sum_{r=1}^{n} Y_r(t_l)Z_r(t_l)}{\bar{Y}(t_l)} \right) \\
&= \sum_{l=1}^{n} \frac{\delta_l\widehat{\omega}(t_l)}{\bar{Y}(t_l)} \sum_{i=1}^{n} Y_i(t_l)Z_i(t_l) - \sum_{l=1}^{n} \frac{\delta_l\widehat{\omega}(t_l)}{\bar{Y}(t_l)} \sum_{i=1}^{n} Y_i(t_l) \frac{\sum_{r=1}^{n} Y_r(t_l)Z_r(t_l)}{\bar{Y}(t_l)} \\
&= \sum_{l=1}^{n} \frac{\delta_l\widehat{\omega}(t_l)}{\bar{Y}(t_l)} \sum_{i=1}^{n} Y_i(t_l)Z_i(t_l) - \sum_{l=1}^{n} \frac{\delta_l\widehat{\omega}(t_l)}{\bar{Y}(t_l)} \bar{Y}(t_l) \frac{\sum_{r=1}^{n} Y_r(t_l)Z_r(t_l)}{\bar{Y}(t_l)} \\
&= \sum_{l=1}^{n} \frac{\delta_l\widehat{\omega}(t_l)}{\bar{Y}(t_l)} \sum_{i=1}^{n} Y_i(t_l)Z_i(t_l) - \sum_{l=1}^{n} \frac{\delta_l\widehat{\omega}(t_l)}{\bar{Y}(t_l)} \sum_{r=1}^{n} Y_r(t_l)Z_r(t_l) \\
&= 0
\end{aligned}$$

Dans l'Annexe A.1 se trouve la preuve de Lin et Wei visant à montrer que, dans le cadre du modèle de Cox, $n^{-1/2}U(\beta^*)$ (β^* étant la vraie valeur de β) est **asymptotiquement équivalent** à $n^{-1/2}W(\beta^*)$, où $W(\beta^*)$ s'écrit comme une somme de n termes iid. La preuve pour les autres modèles s'obtient de manière similaire.

Finalement, l'indice est

$$\boxed{\mathbf{D}_0 = \frac{1}{k} \left(\sum_{i=1}^{n} \widehat{W}_i \right)^T \mathbf{\Sigma}^{-1} \left(\sum_{i=1}^{n} \widehat{W}_i \right) = \frac{1}{k} \left(\sum_{i=1}^{n} \widehat{U}_i \right)^T \mathbf{\Sigma}^{-1} \left(\sum_{i=1}^{n} \widehat{U}_i \right)}$$

où $\mathbf{\Sigma} = \sum_{i=1}^{n} \widehat{W}_i^T \widehat{W}_i$ et k est le nombre d'individus décédés non censurés ($k = \sum_{i=1}^{n} \delta_i$). L'indice $\mathbf{D}_0$ est égal à la **statistique du score robuste divisé par** k.

L'indice $\mathbf{D}_0$ est compris entre 0 et 1 (voir § 4.2.2) et s'interprète en termes de **pourcentage de séparabilité** au cours du temps entre les groupes où l'événement a ou n'a pas lieu.

Le tableau 4.3 résume l'écriture de l'indice sous les différents modèles considérés.

4.2.2 Propriétés de l'indice

Dans ce paragraphe sont exposées les propriétés de l'indice déduites de sa définition. Sa distribution sous l'hypothèse nulle est tout d'abord rappelée. Puis il est démontré, de façon formelle, qu'il est compris entre 0 et 1. Enfin, différentes interprétations possibles de $\mathbf{D}_0$ sont présentées.

TABLEAU 4.3 – Écriture de l'indice sous les différents modèles

Modèle	Risque instantané	Indice
PH	$\lambda_0(t)e^{\beta' Z_i}$	$\mathbf{D}_0^{(\mathrm{PH})} = \frac{1}{k} W^T \mathbf{\Sigma}^{-1} W$ avec $\widehat{\omega}(t) = 1$
PO	$\dfrac{\lambda_0(t)}{1 + S_0(t)\,(e^{\beta' Z_i} - 1)}$	$\mathbf{D}_0^{(\mathrm{PO})} = \frac{1}{k} W^T \mathbf{\Sigma}^{-1} W$ avec $\widehat{\omega}(t) = \widehat{S}_0(t)$
NPH	$\lambda_0(t)e^{\beta Z_i(1+\log(A_0(t)))}$	$\mathbf{D}_0^{(\mathrm{NPH})} = \frac{1}{k} W^T \mathbf{\Sigma}^{-1} W$ avec $\widehat{\omega}(t) = 1 + \log\{\widehat{A}_0(t)\}$

PH : modèle à risques proportionnels ; PO : modèle à odds proportionnels ; NPH : modèle à risques non-proportionnels dont les risques se croisent

a. Distribution de l'indice sous $\mathcal{H}_0$

L'indice est égal à la **statistique du score robuste divisé par k**.

On montre que sous $\mathcal{H}_0$, $\mathbf{D}_0$ suit asymptotiquement une **loi gamma** de paramètres de forme $a = p/2$ et d'échelle $\theta = 2/k$, noté $\mathbf{\Gamma}\left(\frac{p}{2}; \frac{2}{k}\right)$, puisque le score robuste suit une loi du chi-deux à p degrés de liberté.

On connaît donc les moyenne et variance asymptotiques de l'indice sous $\mathcal{H}_0$: $\mathbb{E}(\mathbf{D}_0) \overset{\mathcal{H}_0}{=} \frac{p}{k}$ et $\mathrm{V}(\mathbf{D}_0) \overset{\mathcal{H}_0}{=} \frac{2p}{k^2}$.

La figure 4.5 représente la distribution de l'indice sous l'hypothèse nulle pour $p = 1$ covariable et $k = 100$. L'histogramme correspond à un échantillon de la loi gamma de taille 500 et la courbe en rouge représente la distribution théorique.

La distribution asymptotique de notre indice étant connue, on pourrait calculer des quantités telles que le False Discovery Rate (FDR), le Family Wise Error Rate (FWER) ou le local FDR (lFDR).

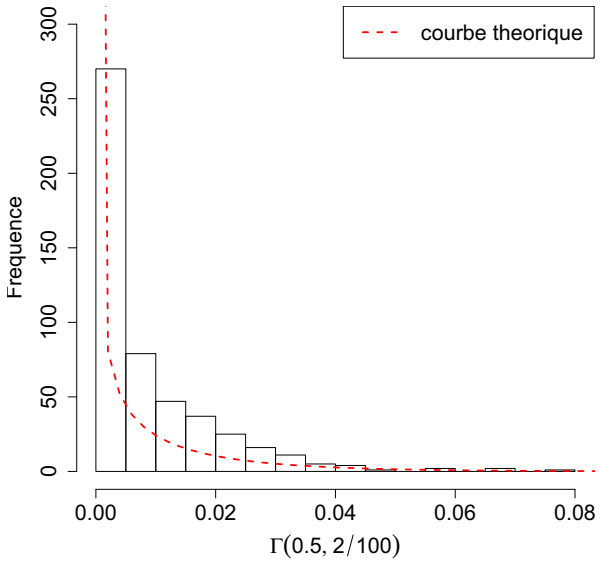
b. Bornes de l'indice

L'indice $\mathbf{D_0}$ est compris entre 0 et 1 (cf. proposition 4.1) ce qui permet de l'interpréter en termes de **pourcentage de séparabilité**.

Proposition 4.1 *Dans le cadre de p covariables, on a*

$$0 \leq \mathbf{D}_0 \leq 1.$$

FIGURE 4.5 – Histogramme d'un échantillon de distribution $\Gamma(0.5, 2/100)$.



Preuve. Les composantes du vecteur du score robuste déduit de la log-vraisemblance partielle sous $\mathcal{H}_0$ s'écrivent, pour un individu $i, i = 1, \cdots, n$ et une covariable $j, j = 1, \cdots, p$

$$\widehat{W}_{ij} = \delta_i \widehat{\omega}(t_i) \left(Z_{ij}(t_i) - \frac{\sum_{l=1}^{n} Y_l(t_i) Z_{lj}(t_i)}{\bar{Y}(t_i)} \right) - \sum_{l=1}^{n} \frac{Y_i(t_l) \delta_l \widehat{\omega}(t_l)}{\bar{Y}(t_l)} \left(Z_{ij}(t_l) - \frac{\sum_{r=1}^{n} Y_r(t_l) Z_{rj}(t_l)}{\bar{Y}(t_l)} \right)$$

Nous avons montré que $\sum_{i=1}^{n} \widehat{W}_{ij} = \sum_{i=1}^{n} \widehat{U}_{ij}$, car $\sum_{i=1}^{n} \widehat{E}\widehat{U}_{ij}$. La somme $\sum_{i=1}^{n} \widehat{U}_{ij}$ est composée de k termes, puisque δ_i vaut 1 si l'individu i décède au temps t_i et vaut 0 s'il est censuré en t_i. Donc $\sum_{i=1}^{n} \widehat{W}_{ij}$ est une somme de k termes. Cette propriété est utilisée plus loin pour écrire les variance et covariance des W_i. Dans la suite, le symbole " $\widehat{}$ " a été supprimé pour faciliter la lecture.

(i) Montrons que $\mathbf{D}_0 \geq 0$

On a

$$\mathbf{D}_0 = \frac{W^T \boldsymbol{\Sigma}^{-1} W}{k}$$

avec

$$\underset{(1\times p)}{W^T} = \left(\sum_{i=1}^{n} W_{i1}, \cdots, \sum_{i=1}^{n} W_{ip} \right)$$

et

$$\underset{(p\times p)}{\boldsymbol{\Sigma}} = \begin{pmatrix} \sum_i W_{i1}^2 & \sum_i W_{i1} W_{i2} & \cdots & \sum_i W_{i1} W_{ip} \\ \sum_i W_{i2} W_{i1} & \sum_i W_{i2}^2 & \cdots & \sum_i W_{i2} W_{ip} \\ \vdots & & & \vdots \\ \sum_i W_{ip} W_{i1} & \cdots & \cdots & \sum_i W_{ip}^2 \end{pmatrix}$$

$\mathbf{\Sigma}$ est la matrice de somme des carrés et des produits croisés. Elle est semi-définie positive. Par conséquent, $\mathbf{\Sigma}^{-1}$ est également semi-définie positive.
Donc, par définition, $W^T\mathbf{\Sigma}^{-1}W \geq 0$ et $\boxed{\mathbf{D}_0 \geq 0}$.

(ii) Montrons que $\mathbf{D}_0 \leq 1$

Soit $\mathbf{\Sigma}^*$ la matrice de variance-covariance des $W_g = \sum_{i=1}^n W_{ig}, g = 1, \cdots, p$. La matrice $\mathbf{\Sigma}^*(p \times p)$ peut être partitionnée comme suit $\mathbf{\Sigma}^* = \begin{pmatrix} \sigma_{11} & \sigma'_{21} \\ \sigma_{21} & \mathbf{\Sigma_{22}} \end{pmatrix}$, où :
σ_{11} est la variance de W_1

$$\underset{(1\times 1)}{\sigma_{11}} = \mathrm{Var}(W_1) = \sum_{i=1}^n W_{i1}^2 - \frac{1}{k}\left(\sum_{i=1}^n W_{i1}\right)^2$$

${\sigma_{21}}^T$ est le vecteur des covariances entre W_1 et $W^{(2)} = (W_2, \cdots, W_p)$:

$$\underset{1\times(p-1)}{{\sigma_{21}}^T} = \left(\sum_{i=1}^n W_{i1}W_{i2} - \frac{1}{k}\sum_{i=1}^n W_{i1}\sum_{i=1}^n W_{i2}, \cdots, \sum_{i=1}^n W_{i1}W_{ip} - \frac{1}{k}\sum_{i=1}^n W_{i1}\sum_{i=1}^n W_{ip}\right)$$

et $\mathbf{\Sigma_{22}}$ est la matrice de variance-covariance de $W^{(2)} = (W_2, \cdots, W_p)$

$$\underset{(p-1)\times(p-1)}{\mathbf{\Sigma_{22}}} = \begin{pmatrix} \sum_i W_{i2}^2 - \frac{1}{k}\left(\sum_i W_{i2}\right)^2 & \cdots & \sum_i W_{i2}W_{ip} - \frac{1}{k}\left(\sum_i W_{i2}\sum_i W_{ip}\right) \\ \sum_i W_{i3}W_{i2} - \frac{1}{k}\left(\sum_i W_{i3}\sum_i W_{i2}\right) & \cdots & \cdots \\ \vdots & & \vdots \\ \sum_i W_{ip}W_{i2} - \frac{1}{k}\left(\sum_i W_{ip}\sum_i W_{i2}\right) & \cdots & \sum_i W_{ip}^2 - \frac{1}{k}\left(\sum_i W_{ip}\right)^2 \end{pmatrix}$$

Le coefficient de corrélation multiple entre W_1 et $W^{(2)}$ est défini par Mardia *et al.* (1979)

$$r_{1.2,\cdots,p} = \left(\frac{{\sigma_{21}}^T{\mathbf{\Sigma_{22}}}^{-1}\sigma_{21}}{\sigma_{11}}\right)^{1/2}$$

Le carré du coefficient de corrélation multiple est noté r^2 et vaut

$$\boxed{r^2 = \frac{{\sigma_{21}}^T{\mathbf{\Sigma_{22}}}^{-1}\sigma_{21}}{\sigma_{11}}} \quad , \quad 0 \leq r^2 \leq 1 \tag{4.10}$$

On considère la partition de $\mathbf{\Sigma}$ suivante

$$\mathbf{\Sigma} = \begin{pmatrix} s_{11} & s_{12} \\ s_{21} & \mathbf{S_{22}} \end{pmatrix}$$

avec

$$\underset{(1\times 1)}{s_{11}} = \sum_{i=1}^n W_{i1}^2$$

et

$$\underset{1\times(p-1)}{s_{12}} = {s_{21}}^T = \left(\sum_{i=1}^n W_{i1}W_{i2}, \cdots, \sum_{i=1}^n W_{i1}W_{ip}\right)$$

$\mathbf{S_{22}}$ est la matrice de la somme des carrés et des produits croisés de $(W_2, \cdots, W_p)$. On a :

$$\underset{1\times 1}{\sigma_{11}} = s_{11} - \frac{1}{k}W_1^2$$

$$\underset{(p-1)\times 1}{\sigma_{21}} = s_{21} - \frac{1}{k}W_{(2)}W_1$$

$$\underset{1\times(p-1)}{\sigma_{12}} = \underset{1\times(p-1)}{\sigma_{21}}^T = {s_{21}}^T - \frac{1}{k}W_1{W_{(2)}}^T$$

$$\underset{(p-1)\times(p-1)}{\mathbf{\Sigma_{22}}} = \mathbf{S_{22}} - \frac{1}{k}W_{(2)}{W_{(2)}}^T$$

Pour calculer $\mathbf{\Sigma_{22}}^{-1}$, on utilise la propriété suivante Mardia *et al.* (1979) :

Proposition 4.2 *Si les inverses des matrices existent, alors pour $A(p\times p)$, $B(p\times n)$, $C(n\times n)$ et $D(n\times p)$, on a*

$$(A+BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1}+DA^{-1}B)^{-1}DA^{-1}$$

On applique cette propriété à $\underset{(p-1)\times(p-1)}{A} = \mathbf{S_{22}}$, $\underset{(p-1)\times 1}{B} = -W_{(2)}$, $\underset{1\times 1}{C} = \frac{1}{k}$ et $\underset{1\times(p-1)}{D} = {W_{(2)}}^T$, conduisant à

$$\mathbf{\Sigma_{22}}^{-1} = \mathbf{S_{22}}^{-1} + \mathbf{S_{22}}^{-1}W_{(2)}\left(k - {W_{(2)}}^T\mathbf{S_{22}}^{-1}W_{(2)}\right)^{-1}{W_{(2)}}^T\mathbf{S_{22}}^{-1}$$

où $E = \left(k - {W_{(2)}}^T\mathbf{S_{22}}^{-1}W_{(2)}\right)^{-1}$ est un scalaire positif. Ainsi

$$\mathbf{\Sigma_{22}}^{-1} = \mathbf{S_{22}}^{-1} + E\mathbf{S_{22}}^{-1}W_{(2)}{W_{(2)}}^T\mathbf{S_{22}}^{-1}$$

En développant les produits, on peut montrer que $\dfrac{{\sigma_{21}}^T\mathbf{\Sigma_{22}}^{-1}\sigma_{21}}{\sigma_{11}} \leq 1$ est équivalent à l'inégalité suivante :

$$\begin{aligned} &k{s_{21}}^T\mathbf{S_{22}}^{-1}s_{21} - {s_{21}}^T\mathbf{S_{22}}^{-1}s_{21}\cdot{W_{(2)}}^T\mathbf{S_{22}}^{-1}W_{(2)} + {s_{21}}^T\mathbf{S_{22}}^{-1}W_{(2)}\cdot{W_{(2)}}^T\mathbf{S_{22}}^{-1}s_{21} \\ &-W_1\cdot{W_{(2)}}^T\mathbf{S_{22}}^{-1}s_{21} - W_1\cdot{s_{21}}^T\mathbf{S_{22}}^{-1}W_{(2)} - k\cdot s_{11} + W_1^2 + s_{11}\cdot{W_{(2)}}^T\mathbf{S_{22}}^{-1}W_{(2)} \leq 0 \end{aligned} \tag{4.11}$$

On utilise cette inégalité pour montrer que $\mathbf{D}_0 = \dfrac{W^T\mathbf{\Sigma}^{-1}W}{k}$ est inférieur à 1. Pour calculer $\mathbf{\Sigma}^{-1}$ on utilise la propriété suivante Mardia *et al.* (1979)

Proposition 4.3 *Si les inverses des matrices existent, alors pour une partition $A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$ les éléments de $A^{-1} = \begin{pmatrix} A^{11} & A^{12} \\ A^{21} & A^{22} \end{pmatrix}$ sont*

$$\begin{aligned} A^{11} &= (A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1}, & A^{12} &= -A^{11}A_{12}A_{22}^{-1}, \\ A^{21} &= -A_{22}^{-1}A_{12}A^{11}, & A^{22} &= (A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1}. \end{aligned}$$

Ainsi, $\mathbf{\Sigma}^{-1}$ s'écrit

$$\mathbf{\Sigma}^{-1} = \begin{pmatrix} (s_{11} - {s_{21}}^T \mathbf{S_{22}}^{-1} s_{21})^{-1} & -D{s_{21}}^T \mathbf{S_{22}}^{-1} \\ -\mathbf{S_{22}}^{-1} s_{21} D & (\mathbf{S_{22}} - s_{21} s_{11}^{-1} {s_{21}}^T)^{-1} \end{pmatrix}$$

avec $D = (s_{11} - {s_{21}}^T \mathbf{S_{22}}^{-1} s_{21})^{-1}$, D est un scalaire positif.
En développant $W^T \mathbf{\Sigma}^{-1} W$, on obtient

$$W^T \mathbf{\Sigma}^{-1} W = W_1^2 D - W_1 D \cdot {W_{(2)}}^T \mathbf{S_{22}}^{-1} s_{21} - W_1 D \cdot {s_{21}}^T \mathbf{S_{22}}^{-1} W_{(2)} + {W_{(2)}}^T (\mathbf{S_{22}} - s_{21} s_{11}^{-1} {s_{21}}^T)^{-1} W_{(2)}$$

La matrice $B = (\mathbf{S_{22}} - s_{21} s_{11}^{-1} {s_{21}}^T)^{-1}$ peut être développée en utilisant la propriété 4.2 comme suit

$$B = \mathbf{S_{22}}^{-1} + \mathbf{S_{22}}^{-1} s_{21} (s_{11} - {s_{21}}^T \mathbf{S_{22}}^{-1} s_{21})^{-1} {s_{21}}^T \mathbf{S_{22}}^{-1} = \mathbf{S_{22}}^{-1} + D \cdot \mathbf{S_{22}}^{-1} s_{21} {s_{21}}^T \mathbf{S_{22}}^{-1}$$

de telle sorte que

$$W^T \mathbf{\Sigma}^{-1} W = W_1^2 D - W_1 D \cdot {W_{(2)}}^T \mathbf{S_{22}}^{-1} s_{21} - W_1 D \cdot {s_{21}}^T \mathbf{S_{22}}^{-1} W_{(2)} + {W_{(2)}}^T \mathbf{S_{22}}^{-1} W_{(2)}$$
$$+ D \cdot {W_{(2)}}^T \mathbf{S_{22}}^{-1} s_{21} \cdot {s_{21}}^T \mathbf{S_{22}}^{-1} W_{(2)}$$

Enfin, montrer que $\dfrac{W^T \mathbf{\Sigma}^{-1} W}{k} \leq 1$ est équivalent à montrer que

$$W_1^2 - k \cdot s_{11} + k {s_{21}}^T \mathbf{S_{22}}^{-1} s_{21} - W_1 \cdot {W_{(2)}}^T \mathbf{S_{22}}^{-1} s_{21} - W_1 \cdot {s_{21}}^T \mathbf{S_{22}}^{-1} W_{(2)}$$
$$+ s_{11} \cdot {W_{(2)}}^T \mathbf{S_{22}}^{-1} W_{(2)} - {s_{21}}^T \mathbf{S_{22}}^{-1} s_{21} \cdot {W_{(2)}}^T \mathbf{S_{22}}^{-1} W_{(2)} + {s_{21}}^T \mathbf{S_{22}}^{-1} W_{(2)} \cdot {W_{(2)}}^T \mathbf{S_{22}}^{-1} s_{21} \quad \leq 0$$

Cette inégalité à déjà été démontrée précédemment (4.11). Par conséquent, on a $\boxed{\mathbf{D_0} \leq 1}$. □

c. Interprétations de l'indice

• L'indice $\mathbf{D_0}$ constitue un indice de type **pseudo $\mathbf{R^2}$** dans le cadre de l'analyse de survie. Comme décrit au chapitre 2, l'une des d'interprétations du R^2 dans le modèle de régression linéaire repose sur la statistique du score divisé par le nombre d'observations (équation (2.10) p. 30). L'indice proposé dans ce travail dans le cadre de l'analyse de survie repose sur le score robuste divisé par le nombre d'individus non censurés, qui généralise donc la mesure proposée par Magee (1990).

L'avantage de cet indice par rapport à d'autres indices de type R^2 est qu'il ne nécessite pas d'estimer les paramètres du modèle. Ceci lui confère un atout particulièrement intéressant lorsque l'estimation des paramètres n'est pas évidente comme pour le modèle à risques non-proportionnels.

• L'indice $\mathbf{D_0}$ peut également être relié au **déterminant d'un estimateur de la matrice de variance-covariance** comme suit (Prop. 4.4).

Proposition 4.4

$$\mathbf{D}_0 = \frac{\det(\mathbf{\Sigma}) - \det(\mathbf{\Sigma}^*)}{\det \mathbf{\Sigma}}$$

avec

$$\underset{(p\times p)}{\mathbf{\Sigma}} = \begin{pmatrix} \sum_i \widehat{W}_{i1}^2 & \sum_i \widehat{W}_{i1}\widehat{W}_{i2} & \cdots & \sum_i \widehat{W}_{i1}\widehat{W}_{ip} \\ \sum_i \widehat{W}_{i2}\widehat{W}_{i1} & \sum_i \widehat{W}_{i2}^2 & \cdots & \sum_i \widehat{W}_{i2}\widehat{W}_{ip} \\ \vdots & & & \vdots \\ \sum_i \widehat{W}_{ip}\widehat{W}_{i1} & \cdots & \cdots & \sum_i \widehat{W}_{ip}^2 \end{pmatrix}$$

et

$$\underset{(p\times p)}{\mathbf{\Sigma}^*} = \begin{pmatrix} \sum_i \widehat{W}_{i1}^2 - \frac{1}{k}\left(\sum_i \widehat{W}_{i1}\right)^2 & \cdots & \sum_i \widehat{W}_{i1}\widehat{W}_{ip} - \frac{1}{k}\sum_i \widehat{W}_{i1}\sum_i \widehat{W}_{ip} \\ \sum_i \widehat{W}_{i2}\widehat{W}_{i1} - \frac{1}{k}\sum_i \widehat{W}_{i1}\sum_i \widehat{W}_{i2} & \cdots & \sum_i \widehat{W}_{i2}\widehat{W}_{ip} - \frac{1}{k}\sum_i \widehat{W}_{i2}\sum_i \widehat{W}_{ip} \\ \vdots & & \vdots \\ \sum_i \widehat{W}_{ip}\widehat{W}_{i1} - \frac{1}{k}\sum_i \widehat{W}_{i1}\sum_i \widehat{W}_{ip} & \cdots & \sum_i \widehat{W}_{ip}^2 - \frac{1}{k}\left(\sum_i \widehat{W}_{ip}\right)^2 \end{pmatrix}$$

La preuve de cette proposition se trouve dans l'Annexe A.2

4.2.3 Ajustement de l'indice

Comme les mesures de type R^2 du modèle linéaire, notre indice augmente lorsqu'on ajoute une ou plusieurs covariables qui n'ont pas d'effet sur la survie, comme le montre des simulations. Pour corriger cette augmentation, un indice ajusté $\mathbf{D}_{0(adj)}$ est proposé.

a. Présentation de l'indice ajusté

Nous rappelons que dans le cadre du modèle linéaire, le R^2 ajusté se calcule de la façon suivante (équation (2.3) p. 24) :

$$\mathrm{R}^2_{adj} = 1 - (1 - \mathrm{R}^2)\frac{n-1}{n-p-1}$$

Dans le cadre de l'analyse de survie, Royston et Sauerbrei (2004) ont proposé de calculer le pseudo-$\mathcal{R}^2$ ajusté de la manière suivante :

$$\mathcal{R}^2_{adj} = 1 - (1 - \mathcal{R}^2)\left(\frac{1}{1-p/k}\right)$$

L'inconvénient de cette mesure est que lorsque $p = 1$, on ne retrouve pas $\mathcal{R}^2_{adj} = \mathcal{R}^2$.

Pour calculer le $\mathbf{D}_0$ **ajusté** la formule suivante est plutôt utilisée :

$$\boxed{\mathbf{D}_{0(adj)} = 1 - (1 - \mathbf{D}_0)\left(\frac{1-1/k}{1-p/k}\right)}$$

Lorsque $p = 1$, on a $\mathbf{D}_{0(adj)} = \mathbf{D}_0$.

b. Bornes de l'indice $\mathbf{D}_{0(adj)}$

L'indice ajusté est **majoré par 1**.

Mais, comme dans le cadre du modèle linéaire, **$\mathbf{D}_{0(adj)}$ peut être négatif**. La borne inférieure de $\mathbf{D}_{0(adj)}$ vaut $\frac{1-p}{k-p} \leq 0$. Elle est d'autant plus proche de 0 que k est grand et p est petit. Cette borne converge plus rapidement que celle de l'indice ajusté proposé par Royston et Sauerbrei (qui vaut $\frac{-p}{k-p}$ et est inférieure à $\frac{1-p}{k-p}$).

En pratique, pour identifier des facteurs avec un pouvoir explicatif suffisant, on ne s'intéresse pas aux valeurs négatives de l'indice mais aux valeurs supérieures à un seuil fixé. Les valeurs négatives de l'indice ajusté n'ont donc pas d'impact sur le résultat de la sélection. Les covariables avec un $\mathbf{D}_{0(adj)} < 0$ ne seront pas sélectionnées.

4.2.4 Prise en compte des ex-æquo

Dans les paragraphes précédents, nous avons fait l'hypothèse que les temps de survenue de l'événement étudié étaient distincts, avec $\Delta N_i(t) = N_i(t) - N_i(t^*)$ ne pouvant valoir que 0 ou 1, $\forall i = 1, \cdots, n$. Cependant, en pratique, le temps ne peut pas être mesuré de façon continue et les événements sont observés sur un intervalle de temps. Par exemple, dans de nombreuses études, un examen ou une procédure doit être effectué pour déterminer si l'événement a eu lieu et, dans ce cas, les événements ayant eu lieu pendant un intervalle de temps $[t_{j-1}, t_j)$ ne seront considérés comme tels qu'au temps de l'examen t_j. Deux événements ou un événement et une censure peuvent donc être observés simultanément.
Dans ce qui suit, nous exposons la méthode employée pour la gestion des ex-æquo.

Les notations introduites ci-après sont spécifiques à cette sous-section.
On note N le nombre de temps distincts (censurés ou non) et k le nombre de temps d'évènements distincts ($k \leq N \leq n$).
Pour $i = 1, \cdots, N$, on appelle

- $D(t_i)$ l'ensemble des individus décédés au temps t_i ;
- $R(t_i)$ l'ensemble des individus à risque en t_i (incluant les individus décédant en t_i) ;
- $E(t_i)$ l'ensemble des individus décédés ou censurés en t_i.

On note respectivement d_i, n_i et e_i les cardinaux de ces trois ensembles.
On définit également

- $R^*(t_i)$ l'ensemble des individus à risque au temps t_i excluant les individus décédés en t_i ($R(t_i) - D(t_i)$) ;
- $R^*(t_{l(-i)})$ ($t_l < t_i$) l'ensemble des individus à risque en t_l sans les sujets décédés ou censurés en t_i ($R(t_l) - E(t_i)$).

Soit $\delta_i = \mathbf{1}_{d_i \geq 1}$ l'indicatrice d'au moins un décès en $t_i, i = 1, \cdots, N$ (où $\mathbf{1}$ désigne la fonction indicatrice).

a. Cas du modèle de Cox

La vraisemblance partielle du modèle de Cox sur l'intervalle est

$$\mathcal{L}(\beta) = \prod_{t\leq\tau}\prod_{i=1}^{n}\left\{\frac{Y_i(t)\exp(\beta' Z_i)}{\sum_{l=1}^{n} Y_l(t)\exp(\beta' Z_l)}\right\}^{\Delta N_i(t)}$$

En présence d'ex-æquo, Breslow (1972) et Peto (1972) ont proposé la vraisemblance suivante :

$$\mathcal{L}(\beta) = \prod_{t\leq\tau}\prod_{i=1}^{n}\frac{Y_i(t)\exp(\beta' Z_i)}{[\sum_{l=1}^{n} Y_l(t)\exp(\beta' Z_l)]^{\Delta N_i(t)}}$$

qui peut être estimée par

$$\widehat{\mathcal{L}}(\beta) = \prod_{i=1}^{N}\left\{\frac{\exp\left(\sum_{l\in D(t_i)}\beta' Z_l(t_i)\right)}{\left(\sum_{l\in R(t_i)}\exp(\beta' Z_l(t_i))\right)^{d_i}}\right\}^{\delta_i}$$

La log-vraisemblance partielle est alors estimée par

$$\log\widehat{\mathcal{L}}(\beta) = \sum_{i=1}^{N}\delta_i\left[\beta'\sum_{l\in D(t_i)} Z_l(t_i) - d_i\log\left\{\sum_{l\in R(t_i)}\exp(\beta' Z_l(t_i))\right\}\right]$$

On en déduit l'expression du score en présence d'ex-æquo

$$\widehat{U}(\beta) = \sum_{i=1}^{N}\delta_i\left[\sum_{l\in D(t_i)} Z_l(t_i) - d_i\sum_{l\in R(t_i)}\frac{Z_l(t_i)\exp(\beta' Z_l(t_i))}{\sum_{r\in R(t_i)}\exp(\beta' Z_r(t_i))}\right]$$

Et donc, sous $\mathcal{H}_0$, on a

$$\begin{aligned}\widehat{U}(0) &= \sum_{i=1}^{N}\widehat{U}_i = \sum_{i=1}^{N}\delta_i\left[\sum_{l\in D(t_i)} Z_l(t_i) - d_i\sum_{l\in R(t_i)}\frac{Z_l(t_i)}{n_i}\right]\\ &= \sum_{i=1}^{N}\frac{\delta_i d_i(n_i-d_i)}{n_i}\left[\sum_{l\in D(t_i)}\frac{Z_l(t_i)}{d_i} - \sum_{l\in R^*(t_i)}\frac{Z_l(t_i)}{n_i-d_i}\right]\end{aligned} \tag{4.12}$$

Sous cette forme, on retrouve l'interprétation des U_i. Ces derniers s'expriment comme des différences pondérées entre la moyenne des covariables du groupe $D(t_i)$ de patients décédés en t_i et la moyenne des covariables du groupe $R^*(t_i)$ de patients non encore décédés en t_i. Les U_i constituent donc une mesure de séparabilité entre les deux groupes de patients $D(t_i)$ et $R^*(t_i)$ au temps t_i.

Enfin, **la formule suivante est proposée pour le score robuste** :

$$\begin{aligned}\widehat{W}_i &= \widehat{U}_i - \widehat{E}\widehat{U}_i\\ &= \sum_{i=1}^{N}\delta_i\left[\sum_{l\in D(t_i)} Z_l(t_i) - d_i\sum_{l\in R(t_i)}\frac{Z_l(t_i)}{n_i}\right] - \sum_{l=1}^{i}\frac{\delta_l d_l}{n_l}\left[\sum_{r\in E(t_i)} Z_r(t_l) - \frac{e_i}{n_l}\sum_{r\in R(t_l)} Z_r(t_l)\right]\end{aligned}$$

que l'on peut réécrire pour faire apparaître les différences entre les groupes de patients décédés/non-décédés :

$$\widehat{W}_i = c_i \left[\sum_{l \in D(t_i)} \frac{Z_l(t_i)}{d_i} - \sum_{l \in R^*(t_i)} \frac{Z_l(t_i)}{n_i - d_i} \right] - \sum_{l=1}^{i} v_{il} \left[\sum_{r \in E(t_i)} \frac{Z_r(t_l)}{e_i} - \sum_{r \in R^*(t_{l(-i)})} \frac{Z_r(t_l)}{n_l - e_i} \right]$$

avec

$$c_i = \frac{\delta_i d_i (n_i - d_i)}{n_i} \text{ et } v_{il} = \frac{\delta_l d_l}{n_l} \times \frac{(n_l - e_i) e_i}{n_l}$$

L'indice s'écrit comme précédemment : $\mathbf{D}_0 = \frac{1}{k} \left(\sum_{i=1}^{n} \widehat{W}_i \right)^T \left(\sum_{i=1}^{n} \widehat{W}_i^T \widehat{W}_i \right)^{-1} \left(\sum_{i=1}^{n} \widehat{W}_i \right)$.

Avec cette écriture, la démonstration $0 \leq \mathbf{D}_0 \leq 1$ reste inchangée, mais elle requiert de montrer au préalable que $\sum_{i=1}^{N} \widehat{W}_i$ est une somme de k termes.

Preuve. Montrons que la somme des $W_j = \sum_{i=1}^{N} W_{ij}$ peut se décomposer en une somme de k termes.

Les composantes du vecteur du score robuste déduit de la log-vraisemblance partielle sous $\mathcal{H}_0$ s'écrivent pour un individu $i, i = 1, \cdots, n$ et une covariable $j, j = 1, \cdots, p$

$$\begin{aligned} \widehat{W}_{ij} &= \widehat{U}_{ij} - \widehat{E}\widehat{U}_{ij} \\ &= \sum_{i=1}^{N} \left[\delta_i \left(\sum_{l \in D(t_i)} Z_{lj}(t_i) - \frac{d_i}{n_i} \sum_{l \in R(t_i)} Z_{lj}(t_i) \right) - \sum_{l=1}^{i} \frac{\delta_l d_l}{n_l} \left(\sum_{r \in E(t_i)} Z_{rj}(t_l) - \frac{e_i}{n_l} \sum_{r \in R(t_l)} Z_{rj}(t_l) \right) \right] \end{aligned}$$

La somme $\sum_{i=1}^{N} \widehat{U}_{ij}$ est composée de k termes, puisque δ_i vaut 1 si au moins un individu décède au temps t_i et vaut 0 si tous les individus sont censurés en t_i.
On suppose que $\delta_i = 1$, au moins un individu décède au temps t_i, et que $\delta_{i+1} = \cdots = \delta_q = 0, i < q \leq N$, les individus aux temps t_{i+1} à t_q sont censurés. L'expression des $\widehat{E}\widehat{U}_{ij}$ est

$$\widehat{E}\widehat{U}_{ij} = \sum_{l=1}^{i} \frac{\delta_l d_l}{n_l} \left(\sum_{r \in E(t_i)} Z_{rj}(t_l) - \frac{e_i}{n_l} \sum_{r \in R(t_l)} Z_{rj}(t_l) \right)$$

L'expression des $\widehat{E}\widehat{U}_{i+1,j}$, où les individus sont censurés, est la suivante

$$\widehat{E}\widehat{U}_{i+1,j} = \sum_{l=1}^{i+1} \frac{\delta_l d_l}{n_l} \left(\sum_{r \in E(t_{i+1})} Z_{rj}(t_l) - \frac{e_{i+1}}{n_l} \sum_{r \in R(t_l)} Z_{rj}(t_l) \right)$$

Comme tous les individus en t_{i+1} sont censurés, $\delta_{i+1} = 0$, et l'expression précédente devient

$$\widehat{E}\widehat{U}_{i+1,j} = \sum_{l=1}^{i} \frac{\delta_l d_l}{n_l} \left(\sum_{r \in E(t_{i+1})} Z_{rj}(t_l) - \frac{e_{i+1}}{n_l} \sum_{r \in R(t_l)} Z_{rj}(t_l) \right)$$

Par conséquent, la somme des $\widehat{E}\widehat{U}_{ij}$ et des $\widehat{E}\widehat{V}_{i+1,j}$ vaut

$$\widehat{E}\widehat{U}_{ij}+\widehat{E}\widehat{V}_{i+1,j} = \sum_{l=1}^{i} \frac{\delta_l \omega_l d_l}{n_l} \left[\left(\sum_{r\in E(t_i)} Z_{rj}(t_l) + \sum_{r\in E(t_{i+1})} Z_{rj}(t_l) \right) - \frac{e_i + e_{i+1}}{n_l} \sum_{r\in R_l} Z_{jr}(t_l) \right]$$

De la même façon, si les individus aux temps $i+1, \cdots, q \quad (i+1 \leq q)$ sont censurés, on peut écrire

$$\widehat{E}\widehat{U}_{ij}+\cdots+\widehat{E}\widehat{U}_{iq} = \sum_{l=1}^{i} \frac{\delta_l \omega_l d_l}{n_l} \left[\left(\sum_{r\in E(t_i)} Z_{rj}(t_l) + \cdots + \sum_{r\in E(t_q)} Z_{rj}(t_l) \right) - \frac{e_i + \cdots + e_q}{n_l} \sum_{r\in R_l} Z_{rj}(t_l) \right]$$

Les temps de censure peuvent donc être regroupés avec les temps non-censurés. La somme $\sum_{i=1}^{N} \widehat{E}\widehat{U}_{ij}$ peut s'écrire comme une somme de k termes, et il en est donc de même pour la somme $\sum_{i=1}^{N} \widehat{W}_{ij}$.

□

b. Les autres modèles

Dans le cadre du modèle de Cox, un poids égal à 1 a été attribué à chacun des individus. Pour des modèles plus complexes (à odds proportionnels et dont les risques se croisent), les individus ne peuvent pas être pondérés de la même manière puisque les risques ne sont pas proportionnels. La méthode de Breslow ne peut donc pas s'appliquer. Il en est de même pour les méthodes de Cox (1972) et celle d'Efron (1977).

A notre connaissance, très peu d'auteurs ont étudié ce problème. Pour ces modèles, es ex-æquo sont alors ordonnés de façon arbitraire (par ajout d'un bruit blanc gaussien de variance très faible).

4.3 Conclusions sur les méthodes

Dans ce chapitre, un **nouvel indice de capacité de prédiction**, ou pseudo-R^2, a été proposé dans le cadre de l'analyse de survie. Cet indice est relié à la statistique du score robuste et peut se calculer sous différents modèles, tels que le modèle de Cox, le modèle à odds proportionnels et le modèle à risques non-proportionnels qui se croisent. Il est compris entre 0 et 1 et possède une interprétation en terme de **pourcentage de séparabilité**. Sa distribution sous l'hypothèse nulle est connue avec la possibilité de calculer des mesures de type FDR, FWER ou lFDR. Un pseudo-R^2 ajusté peut être calculé dans un cadre multivarié, mais peut, dans ce cas, prendre des valeurs négatives. Enfin, la prise en compte des ex-æquo a été présentée dans le cadre du modèle de Cox.

Chapitre 5

ÉTUDE PAR SIMULATIONS DES PROPRIÉTÉS DE L'INDICE

Contenu

L'objectif global de cette étude par simulations est d'étudier les propriétés statistiques et pratiques de l'indice sous différents modèles et de les comparer à celles des indices de la littérature.

Dans un premier temps, les propriétés statistiques de notre indice sont étudiées sous les différents modèles de survie que sont le modèle de Cox, le modèle à odds proportionnels et le modèle à risques qui se croisent. Dans un deuxième temps, les propriétés pratiques de notre indice et sa capacité à identifier correctement des gènes ayant un impact pronostique sur le délai de survenue de l'évènement d'intérêt, sont étudiées dans les trois modèles. Enfin, la capacité de notre indice à détecter des interactions potentielles entre variables génétiques est étudiée dans le cadre du modèle à risques qui se croisent.

5.1 Simulations en vue d'évaluer les propriétés statistiques de l'indice

Dans cette section, nous étudions le comportement de l'indice par rapport à plusieurs paramètres. Dans un premier temps, le schéma général des simulations est décrit ainsi que les spécifications liées à chaque modèle. Dans un deuxième temps, les résultats des simulations sont exposés.

Dans la suite, nous notons
- $\mathbf{D}_0^{(\mathrm{PH})}$ l'indice calculé dans le cadre du modèle de Cox à risques proportionnels
- $\mathbf{D}_0^{(\mathrm{PO})}$ l'indice calculé dans le cadre du modèle à odds proportionnels
- $\mathbf{D}_0^{(\mathrm{NPH})}$ l'indice calculé dans le cadre du modèle à risques non-proportionnels qui se croisent.

5.1.1 Schéma de simulation

Des simulations ont été réalisées en vue d'évaluer le comportement de l'indice, sous les différents modèles de survie, en faisant varier la valeur des coefficients de régression β, la distribution des variables explicatives, la taille d'échantillon et le schéma de censure.

Pour un individu $i, i = 1, \cdots, n$ donné, une variable explicative Z a été considérée, avec une distribution soit discrète (Bernoulli $\mathcal{B}(0.5)$), soit continue (uniforme $\mathcal{U}[0, \sqrt{(}3)]$ ou normale $\mathcal{N}(0, 1/4)$). Ces trois distributions ont été standardisées pour avoir la même variance. Les temps de survie X ont été générés avec la fonction $S(x|Z)$ dont la forme dépend du modèle considéré. Le tableau 5.1 donne l'expression de la fonction de survie utilisée en fonction du modèle.

Pour le modèle de Cox à risques proportionnels, la variable X a été générée avec une distribution exponentielle $\mathcal{E}(e^{\beta Z})$, dont la fonction de survie s'écrit

$$S(x|Z) = \exp(-\lambda x) \quad \text{avec} \quad \lambda = e^{\beta Z}$$

TABLEAU 5.1 – Fonction de survie utilisée pour les simulations en fonction du modèle

Nom du modèle	Fonction de survie	Fonction de survie simulée
PH	$S(x\|Z) = S_0(x)^{\exp(\beta Z)}$	$S(x\|Z) = \exp\left(-xe^{\beta Z}\right)$
POM	$S(x\|Z) = \dfrac{1}{1+H_0(x)e^{\beta Z}}$	$S(x\|Z) = \dfrac{1}{1+xe^{\beta Z}}$
NPH	$S(x\|Z) = \exp\left(-A_0(x)^{e^{\beta Z}}\right)$	$S(x\|Z) = \exp\left(-x^{e^{\beta Z}}\right)$

Pour le modèle à odds proportionnels, la variable survie a été généré avec une distribution log-logistique, telle que

$$S(x|Z) = \frac{1}{1+(x/\alpha)^{\gamma}} \quad \text{avec} \quad \alpha = e^{-\beta Z} \quad \text{et} \quad \gamma = 1$$

Enfin, pour le modèle à risques non-proportionnels, il s'agit d'une distribution de Weibull $\mathcal{W}(\alpha, \eta)$:

$$S(x|Z) = \exp\{-(x/\eta)^{\alpha}\} \quad \text{avec} \quad \alpha = e^{\beta Z} \quad \text{et} \quad \eta = 1$$

Pour les trois modèles, nous avons fait varier la valeur du paramètre de régression β tel que e^{β}=1, 1.25, 1.5, 1.75, 2, 3, 4 ou 5.
La taille d'échantillon n a été prise égale à 50, 100, 500 et 1000.
Le mécanisme de la censure C a été supposé indépendant de X sachant Z et la distribution de la variable de censure $C_i, i = 1, \cdots, n$ était soit uniforme $\mathcal{U}[0, r]$ soit exponentielle $\mathcal{E}(\gamma)$. Le calcul des paramètres r et γ en fonction du pourcentage de censure p_c est détaillé dans l'Annexe B.1. Le pourcentage de censure a été pris égal à 0%, 25% et 50%.

Les données ont été générées comme suit. Pour chaque sujet $i, i = 1, \cdots, n$, une valeur de la covariable Z_i a été générée. Étant donné cette valeur, un temps de survie X_i a été généré à partir d'une distribution soit exponentielle, soit log-logistique, soit de Weibull (se reporter au tableau 5.1). La variable de censure C_i a été généré de façon indépendante et le temps de survie observé T_i a été calculé comme le minimum entre X_i et C_i.

Pour chaque configuration, 1000 répétitions ont été générées.

Notre indice calculé sous les différents modèles a été comparé aux indices suivants basés sur la vraisemblance partielle (voir § a., p. 61 et suivantes) :

– l'indice d'Allison, noté ρ_k^2 (Allison, 1995)
– la version modifiée de l'indice d'Allison, notée ρ_n^2 (O'Quigley *et al.*, 2005)

– l'indice de Nagelkerke, noté R^2_N (Nagelkerke, 1991)
– l'indice de Xu et O'Quigley, noté ρ^2_{XOQ} (Xu et O'Quigley, 1999)

Ces indices ont été utilisés pour la comparaison car ils s'inscrivent dans le même contexte que notre indice, puisqu'ils reposent sur une statistique visant à tester l'hypothèse nulle $\mathcal{H}_0 = \{\beta = 0\}$. Les indices listés ci-dessus sont issus de la statistique de log-rapport de vraisemblance partielle, tandis que $\mathbf{D}_0$ est basé sur la statistique du score robuste (voir chapitre 4).
Dans le cadre du modèle à risques non-proportionnels, nous avons également comparé notre indice, $\mathbf{D}_0^{(\mathrm{NPH})}$, à celle calculée dans le cadre du modèle de Cox, $\mathbf{D}_0^{(\mathrm{PH})}$.

5.1.2 Résultats des simulations

a. Simulations sous un modèle à risques proportionnels

Les tableaux B.1, B.2 et B.3 dans l'annexe B montrent les résultats des simulations pour $\mathbf{D}_0^{(\mathrm{PH})}$ pour les différentes distributions de la variable explicative, respectivement $Z \sim \mathcal{B}(1/2)$, $Z \sim \mathcal{U}[0, \sqrt(3)]$ et $Z \sim \mathcal{N}(0, 1/4)$.

Les figures 5.1, 5.2 et 5.3, permettent de comparer notre indice avec ceux de la littérature dans le cas où $n = 500$. Les figures pour les autres tailles d'échantillon sont données dans l'Annexe B (figures B.1 à B.9).

Influence de la valeur des coefficients. Les tableaux montrent que lorsque $\beta = 0$, notre indice est proche de 0 : la séparabilité est proche de 0. Il en est de même pour les indices de la littérature. L'indice $\mathbf{D}_0^{(\mathrm{PH})}$ atteint des valeurs moyennes au plus égales à 4.5% pour $n = 50$, 50% de censure et $Z \sim \mathcal{N}(0, 1/4)$.
En réalité, notre indice a une espérance égale à $1/k$ sous l'hypothèse nulle, puisqu'il suit une loi gamma $\Gamma(\frac{1}{2}; \frac{2}{k})$ sous $\mathcal{H}_0$. Sous l'hypothèse nulle, notre indice tend donc asymptotiquement vers 0. Lorsque $n = 50$ (tableaux B.1, B.2 et B.3), $\mathbf{D}_0^{(\mathrm{PH})}$ a une espérance de 2% en l'absence de censure et de 4% lorsque la censure vaut 50%.

La valeur moyenne de l'indice $\mathbf{D}_0^{(\mathrm{PH})}$ augmente vers 1 lorsque $|\beta|$ augmente, la séparabilité augmente donc avec l'effet de la variable explicative. Elle passe par exemple de près de 0% a plus de 40% lorsque le risque relatif passe de 1 à 5, pour une distribution uniforme de la covariable. La même remarque peut être faite pour les indices de la littérature.

Lorsque la covariable est distribuée selon une loi de Bernoulli ou uniforme, notre indice prend des valeurs supérieures aux autres indices. Lorsque la distribution de la covariable est normale, notre indice prend des valeurs proches des indices ρ^2_k et ρ^2_{XOQ}.
Les trois graphiques montrent que les indices d'Allison et de Nagelkerke d'un coté et les indices d'O'Quigley et d'Allison dans sa version modifiée d'autre part, ont un comportement similaire.

Influence de la taille d'échantillon. Nous avons vu précédemment que $\mathbf{D}_0^{(\mathrm{PH})}$ est sensible à la taille de l'échantillon, et plus particulièrement, au nombre d'individus non-censurés, lorsque le risque relatif vaut 1.
Lorsque $\beta \neq 0$, les tableaux de l'annexe B montrent que la valeur moyenne de $\mathbf{D}_0^{(\mathrm{PH})}$ dépend peu de la taille de l'échantillon pour des effets modérés à élevés ($e^\beta \geq 1.5$). Les valeurs moyennes de notre indice pour $n = 50$ à 500 sont proches de celles obtenues pour $n = 1000$, qui sont supposées approcher sa limite asymptotique. Ainsi, pour une distribution de Bernoulli, la valeur moyenne de $\mathbf{D}_0^{(\mathrm{PH})}$ varie de 33.9% à 32.6% pour $e^\beta = 3$ lorsque n passe de 50 à 1000. Les écarts-types de $\mathbf{D}_0^{(\mathrm{PH})}$ (indiqués entre parenthèses) sont d'autant plus petits que la taille de l'échantillon est grande, que le pourcentage de censure est faible et que les effets sont importants.

Les indices de la littérature sont tous peu affectés par la taille de l'échantillon (voir Annexe B, figures B.1 à B.9, pour les résultats avec $n = 50$, $n = 100$ et $n = 1000$).

Influence de la censure. Notre indice est sensiblement affecté par le pourcentage de censure, en particulier pour des effets importants de la covariable. Cette sensibilité est plus importante lorsque Z suit une loi normale : pour $n = 100$ (tableau B.3), $e^\beta = 5$, la valeur moyenne de l'indice fluctue entre 30 et 40%.
En revanche, la distribution de censure, uniforme ou exponentielle, n'a pas d'effet sur la valeur moyenne de l'indice.

De même, les indices de la littérature sont tous sensiblement affectés par la censure. Les indices ρ_n^2 et R_N^2 diminuent lorsque le pourcentage de censure augmente, tandis que les valeurs de ρ_k^2 et ρ_{XOQ}^2 augmentent. La distribution de censure ne change pas la nature des résultats obtenus (voir Annexe B, figures B.10 à B.18, pour les résultats avec une censure exponentielle)

Influence de la distribution des covariables. Enfin, $\mathbf{D}_0^{(\mathrm{PH})}$ est dépendant de la distribution de la covariable. Ceci est peu surprenant étant donné qu'il se calcule directement à partir des valeurs des covariables en chaque temps.
La valeur moyenne de l'indice est plus petite lorsque Z a une distribution normale, que lorsque Z suit une loi uniforme. De même, la valeur moyenne de l'indice est plus petite lorsque Z a une distribution uniforme que lorsque Z a une distribution de Bernoulli.

Les indices ρ_n^2, R_N^2, ρ_k^2 et ρ_{XOQ}^2 ne sont, contrairement à $\mathbf{D}_0^{(\mathrm{PH})}$, pas sensibles à la distribution des covariables.

FIGURE 5.1 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{PH})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à risques proportionnels, pour différents risques relatifs et différents pourcentages de censure, calculés pour une variable Z de Bernoulli $\mathcal{B}(1/2)$, une censure uniforme et $n = 500$ (1000 répétitions).

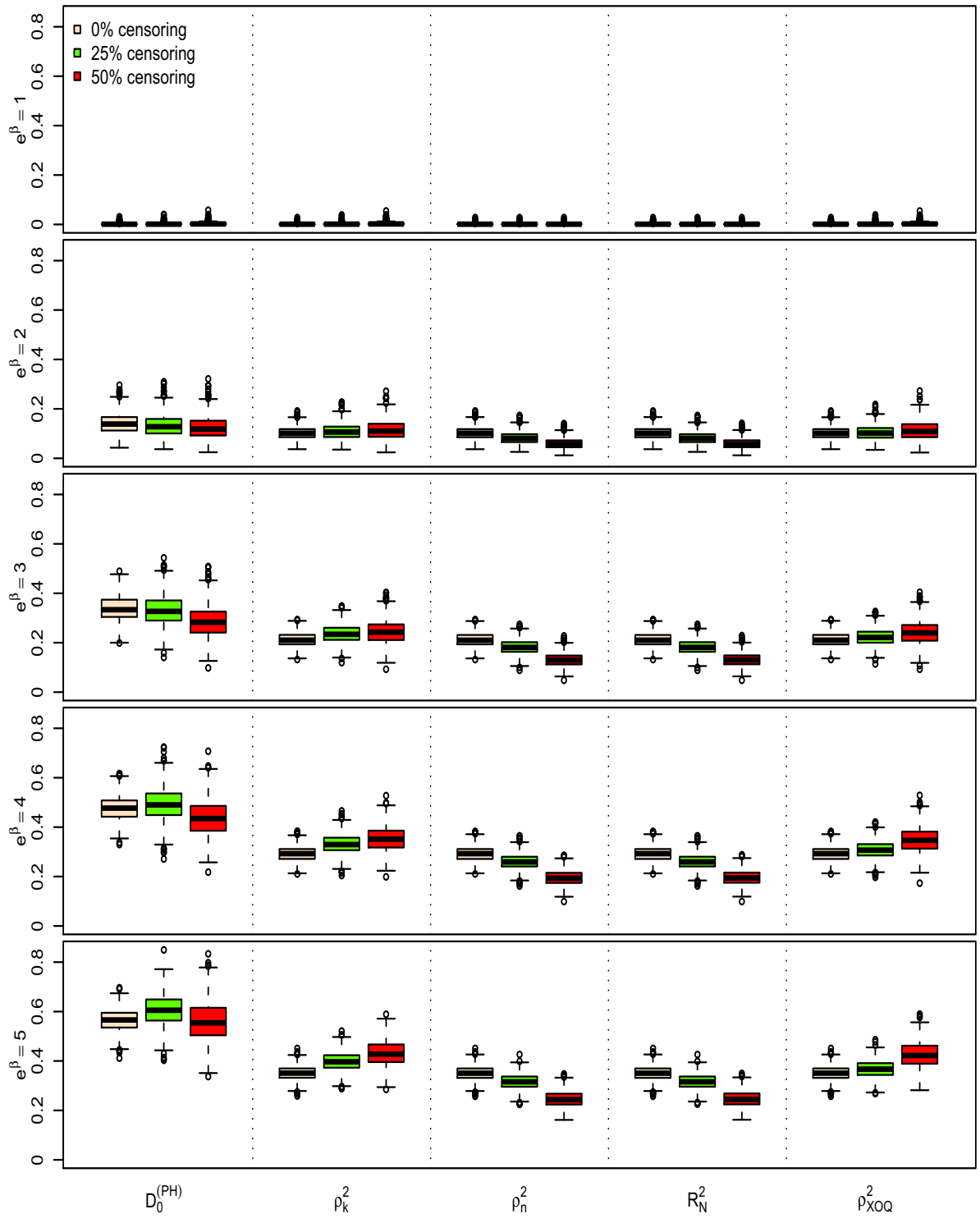


FIGURE 5.2 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{PH})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à risques proportionnels, pour différents risques relatifs et différents pourcentages de censure, calculés pour une variable Z uniforme $\mathcal{U}[0, \sqrt{(3)}]$, une censure uniforme et $n = 500$ (1000 répétitions).

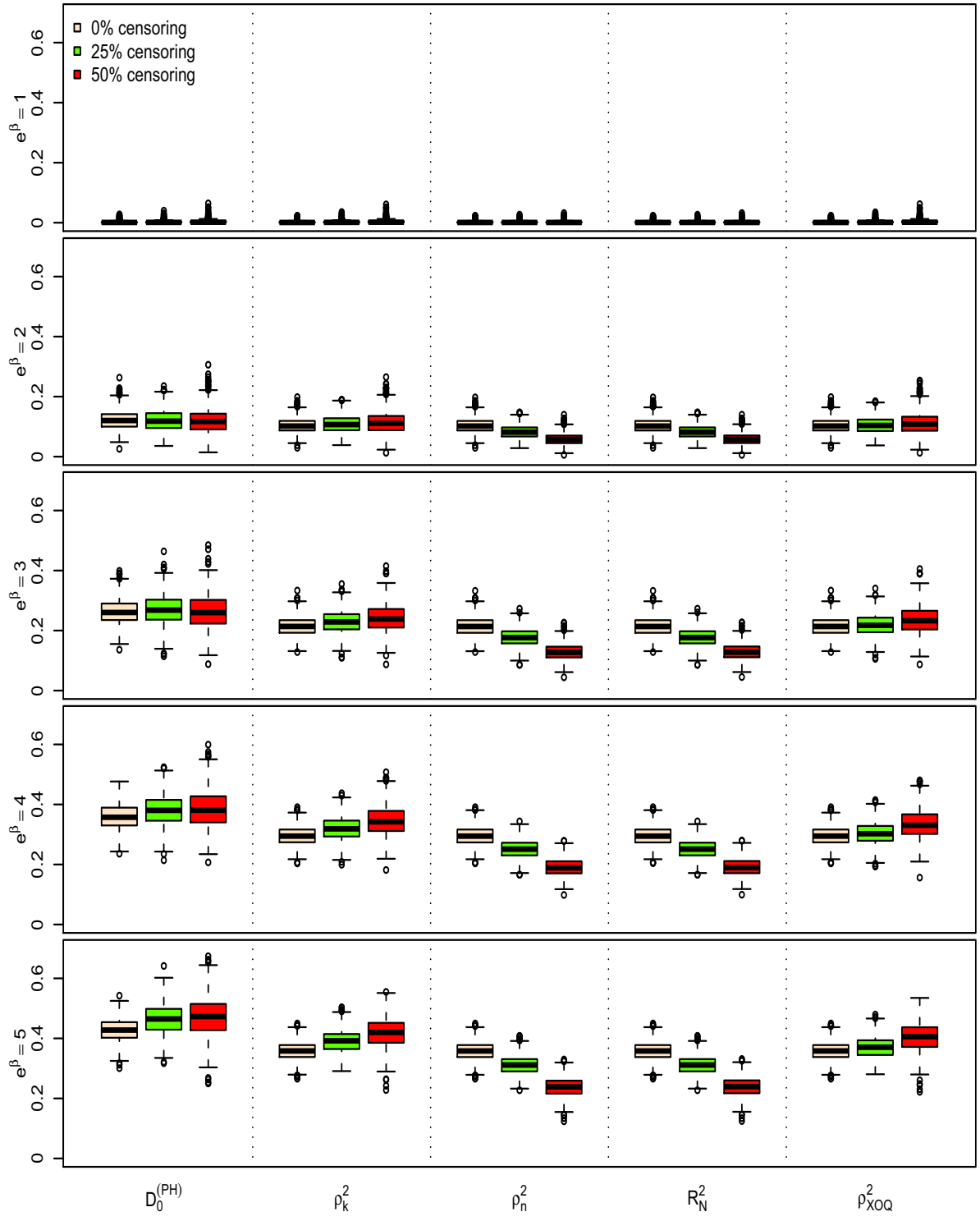
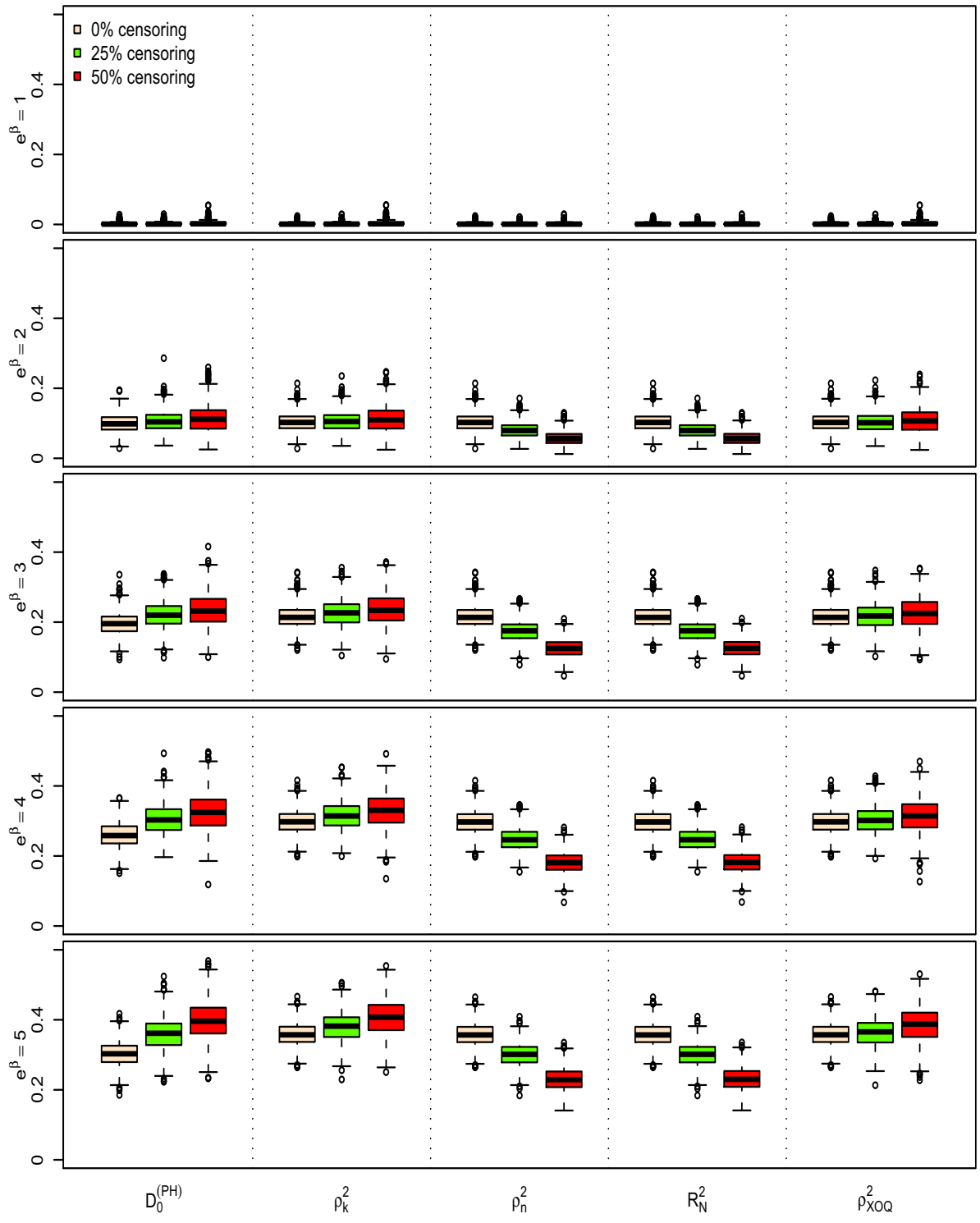


FIGURE 5.3 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{PH})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à risques proportionnels, pour différents risques relatifs et différents pourcentages de censure, calculés pour une variable Z normale $\mathcal{N}(0, 1/4)$, une censure uniforme et $n = 500$ (1000 répétitions).



b. Simulations sous un modèle à odds proportionnels

Les tableaux B.4, B.5 et B.6 dans l'annexe B donnent les résultats des simulations pour un modèle à odds proportionnels. Notre indice peut être comparé aux indices de la littérature grâce aux figures 5.4, 5.5 et 5.6, ainsi qu'aux figures B.19 à B.36 de l'Annexe B. Les commentaires concernant l'influence de la valeur des coefficients, la taille de l'échantillon et la censure sont similaires au cas du modèle à risques proportionnels.

Influence de la valeur des coefficients. Notre indice est proche de 0 quand $\beta = 0$, puis augmente avec le risque relatif. Par exemple, lorsque Z suit une loi de Bernoulli, la valeur moyenne de $\mathbf{D}_0^{(\mathrm{PO})}$ varie de 0 à plus de 20% lorsque e^β passe de 1 à 5 (tableau B.4).

Les indices de la littérature prennent également des valeurs quasi-nulles pour $\beta = 0$ et augmentent avec la valeur de $|\beta|$. Les indices d'Allison et de Nagelkerke possèdent des valeurs moyennes très proches, mais atteignent des valeurs relativement faibles, même pour des valeurs de coefficients élevés.

Influence de la taille d'échantillon. $\mathbf{D}_0^{(\mathrm{PO})}$ est peu sensible à la taille de l'échantillon pour $e^1 > 1$. Ainsi, lorsque $Z \sim \mathcal{N}(0, 1/4)$, pour $e^\beta = 4$ et $p_c = 25\%$, l'indice vaut en moyenne 17.0% si $n = 50$ et 15.2% si $n = 1000$ (tableau B.6).

Les indices de la littérature sont également peu sensibles à la taille d'échantillon.

Influence de la censure. $\mathbf{D}_0^{(\mathrm{PO})}$ est sensiblement affecté par le pourcentage de censure. En effet, pour une distribution uniforme de Z (tableau B.5), $n = 1000$ et $e^\beta = 5$, l'indice varie entre 17 et 27% pour 0 et 50% de censure, respectivement. Le mécanisme de censure, quant-à-lui, n'influence pas la valeur moyenne de l'indice.

Les indices d'Allison et de Nagelkerke sont peu sensibles à la censure, contrairement aux indices d'Allison dans sa version modifiée et de Xu et O'Quigley qui sont sensiblement affectés par la censure.

Influence de la distribution des covariables. La différence par rapport aux simulations dans le cadre du modèle à risques proportionnels est la faible sensibilité de l'indice $\mathbf{D}_0^{(\mathrm{PO})}$ à la distribution de la covariable. Par exemple dans le tableau B.4, pour $e^\beta = 4$, $n = 500$, en l'absence de censure, la valeur moyenne de $\mathbf{D}_0^{(\mathrm{PO})}$ varie de 12.3% pour $Z \sim \mathcal{N}(0, 1/4)$, à 13.6% pour $Z \sim \mathcal{U}[0, \sqrt{(3)}]$ et 14.9% pour $Z \sim \mathcal{B}(1/2)$.

Il est en de même pour les indices de la littérature.

D'une manière générale, notre indice a un comportement très proche de celui d'Allison dans sa version modifiée et celui de Xu et O'Quigley.

FIGURE 5.4 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{PO})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à odds proportionnels, pour différents risques relatifs et différents pourcentages de censure, calculés pour une variable Z de Bernoulli $\mathcal{B}(1/2)$, une censure uniforme et $n = 500$ (1000 répétitions).

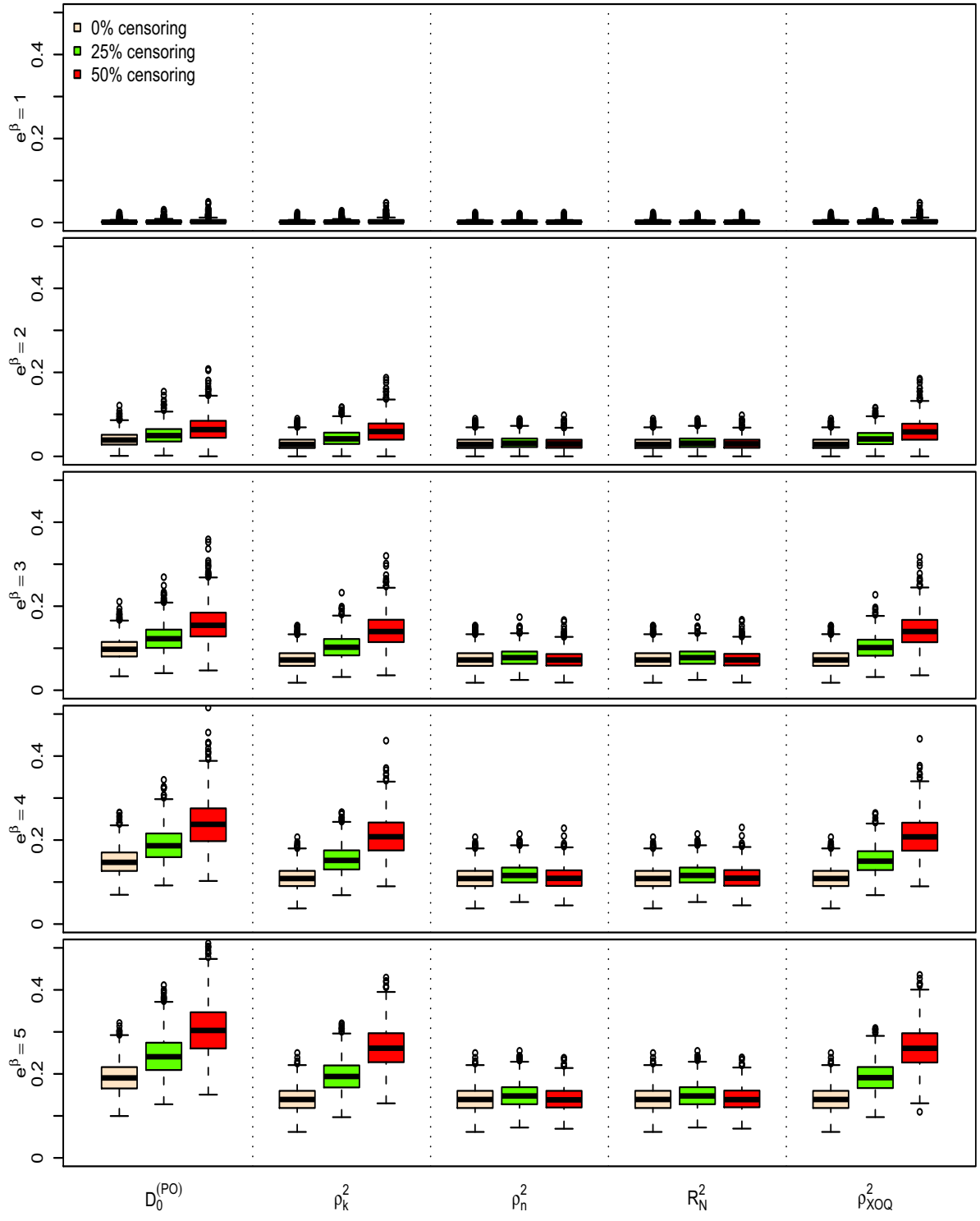


FIGURE 5.5 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{PO})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à odds proportionnels, pour différents risques relatifs et différents pourcentages de censure, calculés pour une variable Z uniforme $\mathcal{U}[0, \sqrt{(3)}]$, une censure uniforme et $n = 500$ (1000 répétitions).

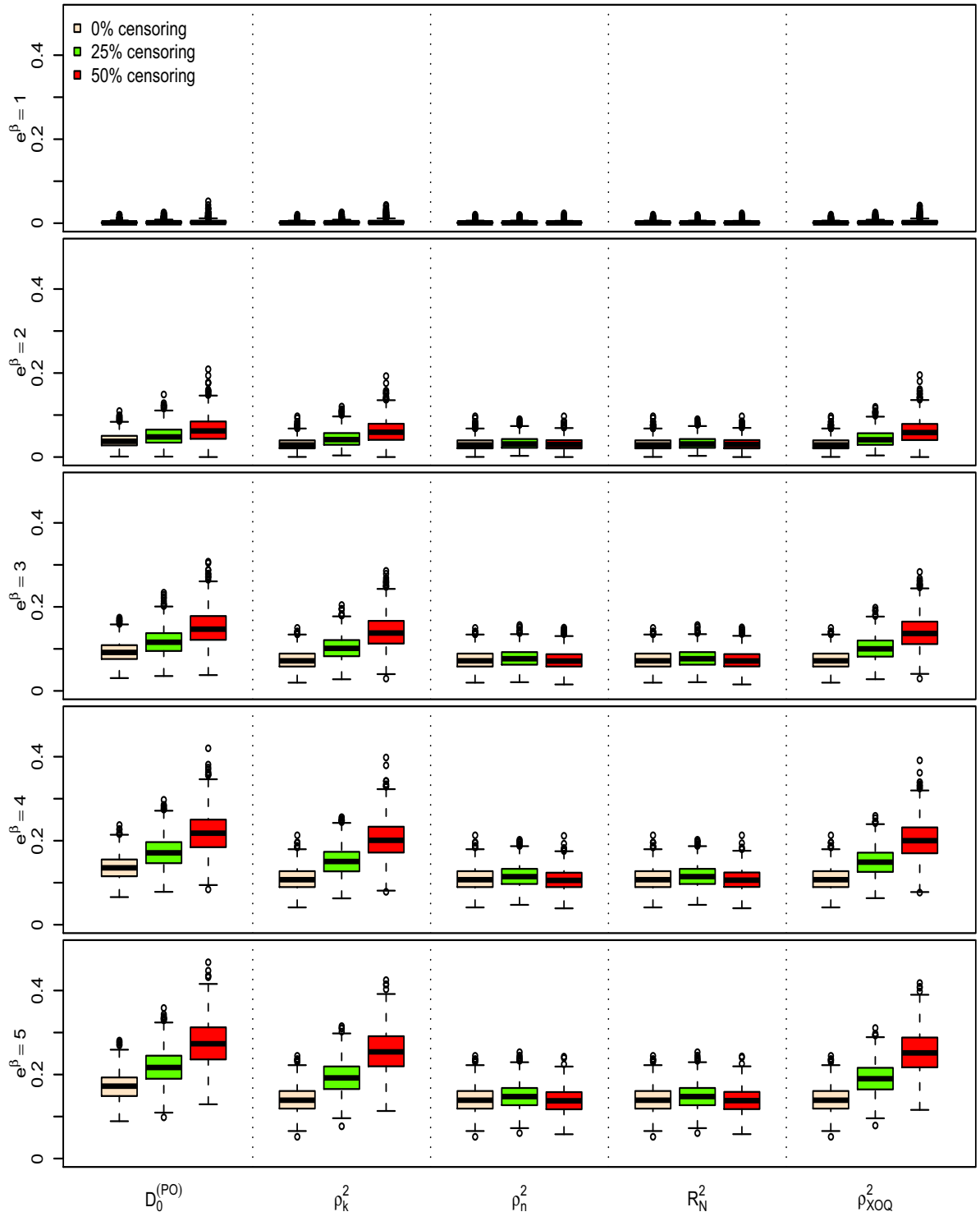
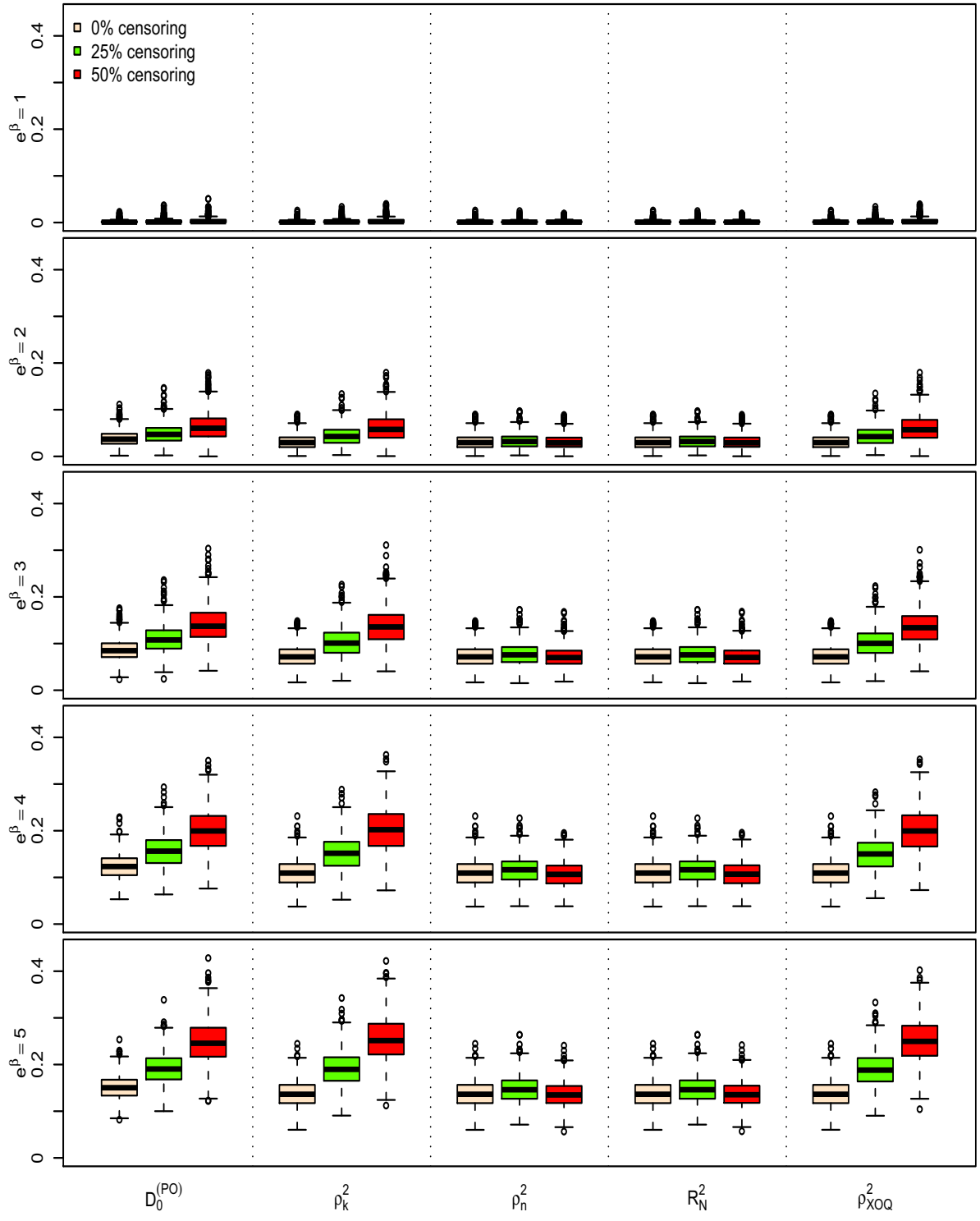


FIGURE 5.6 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{PO})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à odds proportionnels, pour différents risques relatifs et différents pourcentages de censure, calculés pour une variable Z normale $\mathcal{N}(0, 1/4)$, une censure uniforme et $n = 500$ (1000 répétitions).



c. Simulations sous un modèle à risques non-proportionnels

Les résultats des simulations sont présentés dans les tableaux B.7, B.8 et B.9 dans l'Annexe B, ainsi que sur les figures 5.7, 5.8 et 5.9 ci-après et B.37 à B.54 dans l'Annexe B.

Influence de la valeur des coefficients. L'indice $\mathbf{D}_0^{(\mathrm{NPH})}$ est proche de 0 lorsque $\beta = 0$. Il augmente avec la valeur de e^β : il passe de 0 à plus de 20% lorsque e^β varie de 1 à 5 (tableaux de l'annexe B).

Contrairement à $\mathbf{D}_0^{(\mathrm{NPH})}$, la valeur moyenne des indices de la littérature, ρ_k^2, ρ_k^2, R_N^2 et ρ_{XOQ}^2, ainsi que de $\mathbf{D}_0^{(\mathrm{PH})}$ n'augmentent pas lorsque e^β augmente.

Influence de la taille d'échantillon. Notre indice n'est pas affecté par la taille de l'échantillon. Par exemple, il vaut en moyenne 25.5% pour $n = 50$ et 29.1% pour $n = 1000$ lorsque Z a une distribution uniforme, pour $e^\beta = 4$ et $p_c = 0\%$ (tableau B.8).

Influence de la censure. Contrairement aux modèles à risques proportionnels et à odds proportionnels, notre indice n'est pas affecté par le pourcentage de censure. Lorsque Z suit une loi de Bernoulli, $e^\beta = 4$ et $n = 500$, il vaut en moyenne 38% pour $p_c = 0\%$ et 38.6% pour $p_c = 50\%$ (tableau B.7).

Lorsque Z suit une loi normale, les indices ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2 et $\mathbf{D}_0^{(\mathrm{PH})}$ ont un comportement très instable vis-à-vis de la censure, variant de 0 à 5% en moyenne pour ces censures égales à 0 et 50%, respectivement.

Influence de la distribution des covariables. La valeur moyenne de notre indice est influencée par la distribution des covariables. On voit que pour $e^\beta = 3$, $n = 500$ et en l'absence de censure, $\mathbf{D}_0^{(\mathrm{NPH})}$ vaut respectivement 31.2%, 24.4% et 17.9% pour une distribution de Bernoulli, uniforme ou normale (voir annexe B).

FIGURE 5.7 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{NPH})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à risques non-proportionnels qui se croisent, pour différents risques relatifs et différents pourcentages de censure, calculés pour une variable Z de Bernoulli $\mathcal{B}(1/2)$ et $n = 500$, une censure uniforme (1000 répétitions).

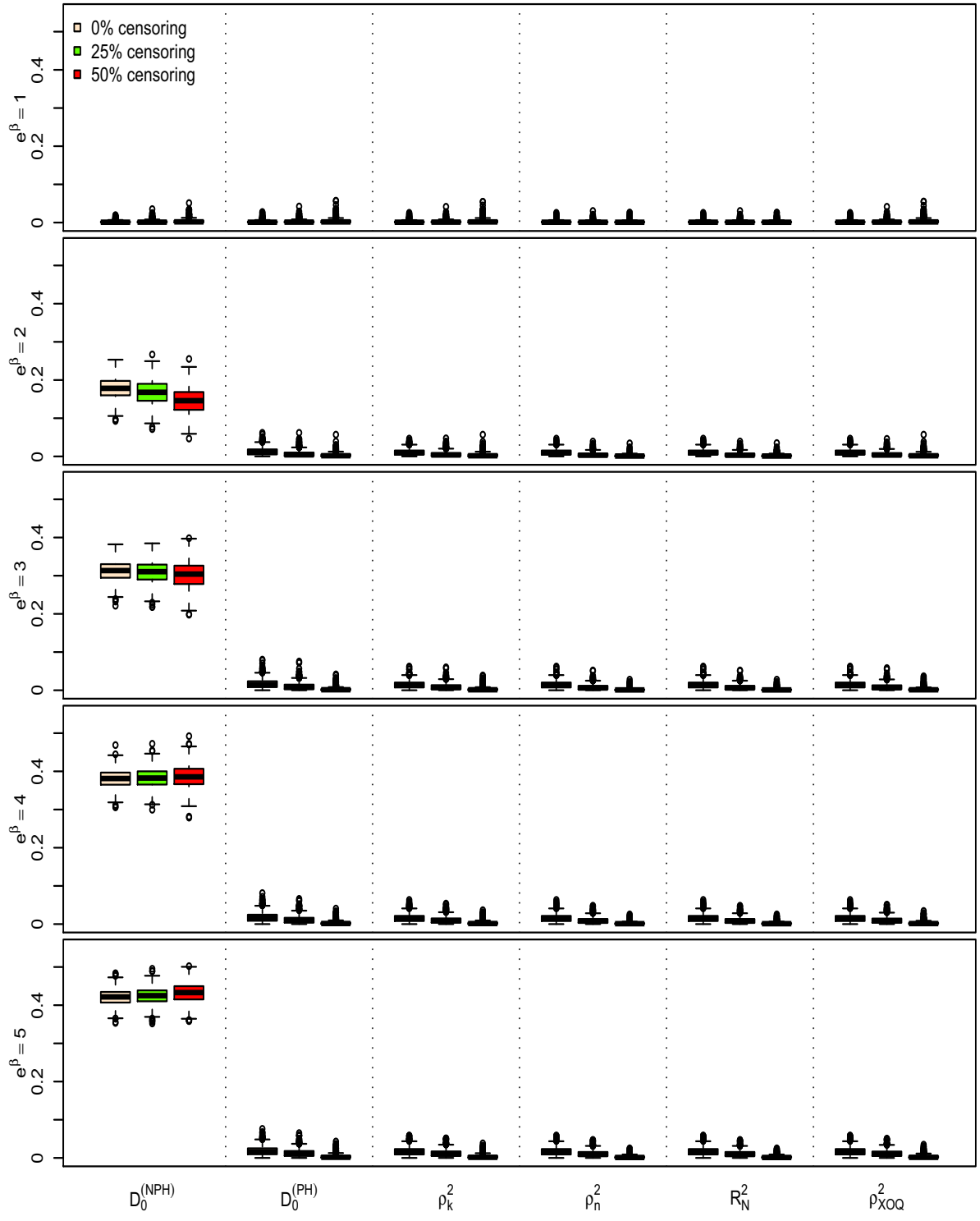


FIGURE 5.8 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{NPH})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à risques non-proportionnels qui se croisent, pour différents risques relatifs et différents pourcentages de censure, calculés pour une variable Z uniforme $\mathcal{U}[0, \sqrt{(3)}]$, une censure uniforme et $n = 500$ (1000 répétitions).

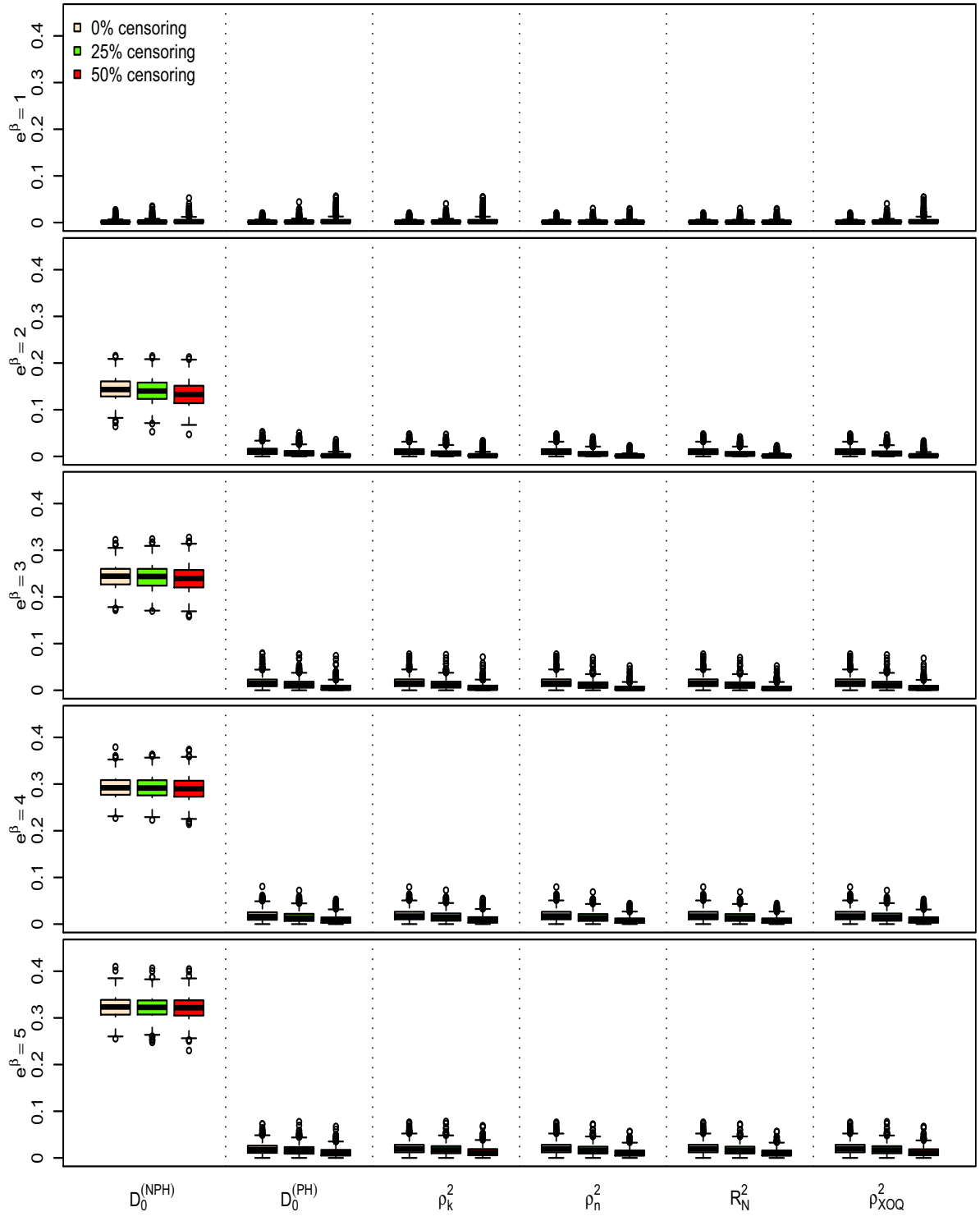
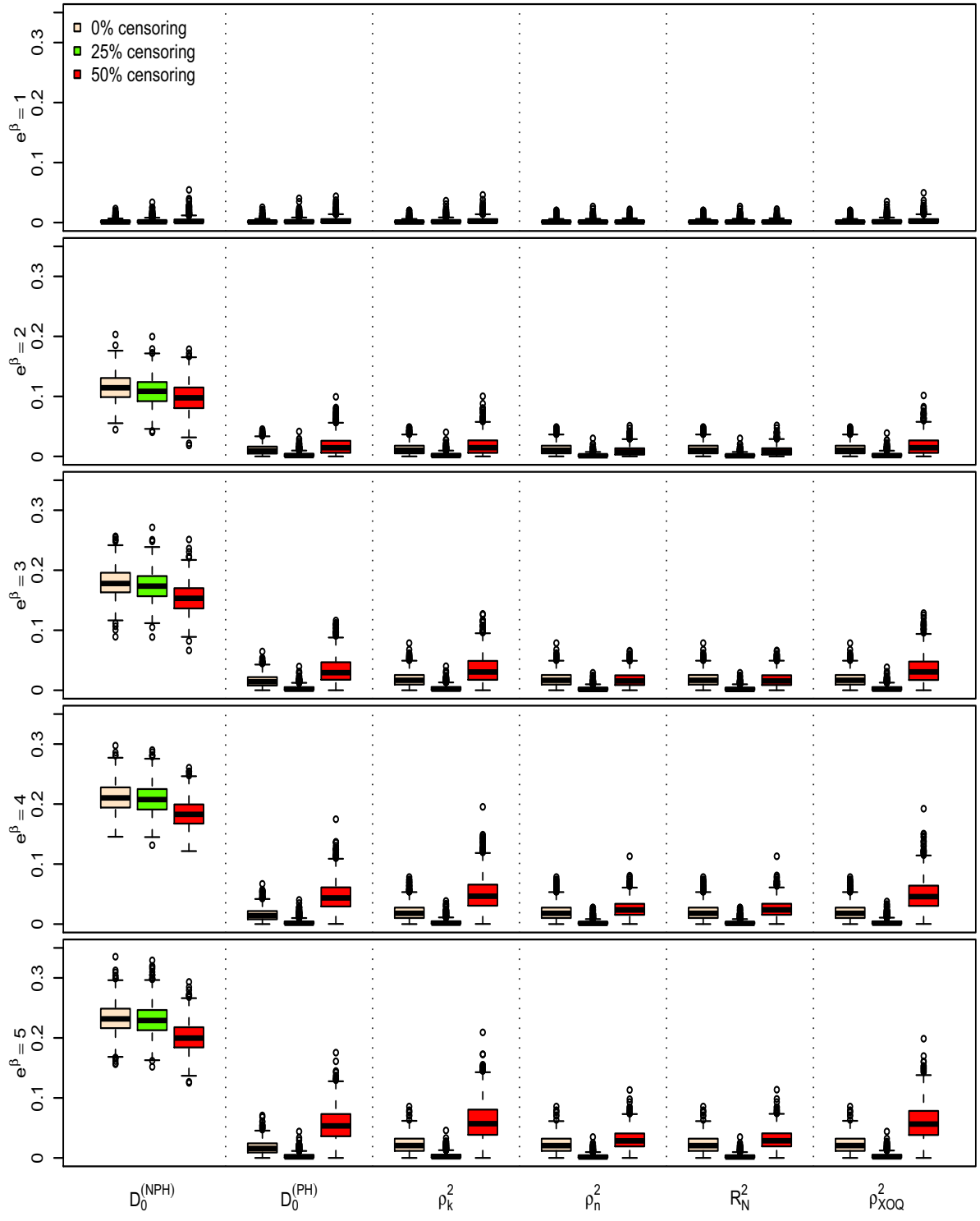


FIGURE 5.9 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{NPH})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à risques non-proportionnels qui se croisent, pour différents risques relatifs et différents pourcentages de censure, calculés pour une variable Z normale $\mathcal{N}(0, 1/4)$, une censure uniforme et $n = 500$ (1000 répétitions).



5.2 Simulations en vue d'étudier les propriétés pratiques de l'indice

Dans cette section, nous étudions les propriétés pratiques de l'indice, sur un exemple simulé, pour sélectionner des facteurs génétiques ayant un impact pronostic sur le délai de survenue de l'évènement d'intérêt.

5.2.1 Schéma de simulation

La méthode utilisée dans ce paragraphe est inspirée de l'article de Bair et Tibshirani (2004). Elle a été modifiée afin de simuler des données proches de données d'expressions génétiques réelles.

Dans la suite, nous considérons un jeu de données simulées composé de 1000 mesures d'expression génétiques. A chaque individu $i(i = 1, \cdots, n); n = 50, 100$ ou 200 a été associé un temps de survie X_i, un temps de censure C_i et un vecteur de 1000 valeurs $Z_i = \left\{ Z_i^{(g)}; g = 1, \cdots, 1000 \right\}$ correspondant aux mesures d'expression.

• Simulation des données de survie.

En fonction du modèle considéré, nous avons simulé des données de survie selon une distribution exponentielle, log-logistique ou de Weibull, comme dans le tableau 5.2.

TABLEAU 5.2 – Fonction de survie utilisée pour les simulations en fonction du modèle

Nom du modèle	Fonction de survie	Fonction de survie simulée
PH	$S(x\|Z) = S_0(x)^{\exp(\beta Z)}$	$S(x\|Z) = \exp\left(-xe^{\nu}\right)$
POM	$S(x\|Z) = \dfrac{1}{1 + H_0(x)e^{\beta Z}}$	$S(x\|Z) = \dfrac{1}{1 + xe^{\nu}}$
NPH	$S(x\|Z) = \exp\left(-A_0(x)^{e^{\beta Z}}\right)$	$S(x\|Z) = \exp\left(-x^{e^{\nu}}\right)$

Pour les individus tels que $(n/2 + 1) \leq i \leq n$, la paramètre ν a été pris égal à 0. Pour les individus i tels que $1 \leq i \leq n/2$ ($n = 50, 100$ ou 200), e^{ν} a été pris égal à 3 et 5 pour les modèles de Cox et à odds proportionnels et à 2 et 3 pour le modèle à risques qui se croisent. Les individus $i = (n/2 + 1)$ à n appartiennent au groupe de patients à bas risque de survenue de l'évènement d'intérêt, et les individus $i = 1$ à $n/2$ au groupe de patients à haut risque de survenue de l'évènement.
Pour chaque jeu de données, les temps de censure C_i $(i = 1, \cdots, n)$ ont été supposés indépendants de X sachant Z, de distribution uniforme sur $[0, r]$, r ayant été choisi pour avoir un pourcentage

de censure proche de 25 et 50%.
Le temps de survie observé T_i a été calculé comme le minimum entre X_i et C_i.

• **Simulation des données d'expression génétique.**

Pour chaque jeu de données et chaque individu i, 1000 mesures d'expression génétiques $Z_i^{*(g)}$ $(i = 1, \cdots, n; g = 1, \cdots, 1000)$ ont été simulées selon le schéma représenté dans le tableau 5.3.

TABLEAU 5.3 – Représentation de l'exemple simulé

$n = 50, 100$ or 200 patients

		Groupe de patients à haut risque $(n/2)$	*Groupe de patients à à bas risque* $(n/2)$
Distribution des temps de survie		$S(x) = xe^{-\nu}$ ou $S(x) = [1 + x \cdot e^{\nu}]^{-1}$ ou $S(x) = x^{\exp\{\nu\}}$	$S(x) = x$ ou $S(x) = [1 + x]^{-1}$ ou $S(x) = x$
Gènes Z^*	$g =$ 1-50	Log-$\mathcal{N}(4, 1.5)$	Log-$\mathcal{N}(0, 1.5)$
	$g =$ 51-100	Log-$\mathcal{N}(3, 1.5)$	Log-$\mathcal{N}(0, 1.5)$
	$g =$ 101-150	60% Log-$\mathcal{N}(0, 1.5)$ + 40% Log-$\mathcal{N}(1, 1.5)$	
	$g =$ 151-250	50% Log-$\mathcal{N}(0, 1.5)$ + 50% Log-$\mathcal{N}(0.5, 1.5)$	
	$g =$ 251-350	30% Log-$\mathcal{N}(0, 1.5)$ + 70% Log-$\mathcal{N}(0.1, 1.5)$	
	$g =$ 301-1000	Log-$\mathcal{N}(0, 1.5)$	

Les mesures d'expression des gènes $g = 1$ à 50 et pour les individus $i = 1$ à $n/2$ suivent

une distribution Log-$\mathcal{N}(\mu = 4, \sigma = 1.5)$, avec $\mathbb{E}(Z^*) = e^{\mu+0.5\sigma^2}$ et $\mathbb{V}(Z^*) = e^{2\mu\sigma^2}(e^{\sigma^2} - 1)$. Pour le reste des individus ($i = n/2 + 1, \cdots, n$), les mesures d'expression de ces gènes suivent une distribution Log-$\mathcal{N}(0, 1.5)$. Les mesures d'expression des gènes $g = 51$ à 100 et pour les individus $i = 1$ à $n/2$ suivent une distribution log-normale de paramètres $\mu = 3$ et $\sigma = 1.5$, notée Log-$\mathcal{N}(\mu = 3, \sigma = 1.5)$. Pour le reste des individus ($i = n/2 + 1, \cdots, n$), les mesures d'expression de ces gènes suivent une distribution Log-$\mathcal{N}(0, 1.5)$. Pour les mesures d'expression des gènes $g = 101$ à 150 et pour 40% individus tirés au hasard parmi les n, les $Z_i^{*(g)}$ suivent une distribution log-normale Log-$\mathcal{N}(1, 1.5)$, alors que, pour les individus restant, ils suivent un distribution log-normale Log-$\mathcal{N}(0, 1.5)$. Pour les mesures d'expression des gènes $g = 151$ à 250 et pour 50% individus tirés au hasard parmi les n, les $Z_i^{*(g)}$ suivent une distribution log-normale Log-$\mathcal{N}(0.5, 1.5)$, alors que, pour les individus restant, ils suivent un distribution log-normale Log-$\mathcal{N}(0, 1.5)$. Pour les mesures d'expression des gènes $g = 251$ à 350 et pour 70% individus tirés au hasard parmi les n, les $Z_i^{*(g)}$ suivent une distribution log-normale Log-$\mathcal{N}(0.1, 1.5)$, alors que, pour les individus restant, ils suivent un distribution log-normale Log-$\mathcal{N}(0, 1.5)$. Enfin, pour les mesures d'expression des gènes $g = 351$ à 1000, les $Z_i^{*(g)}$ ($i = 1, \cdots, n; g = 351, \cdots, 1000$) suivent une distribution log-normale Log-$\mathcal{N}(0, 1.5)$ pour l'ensemble des individus.

Les gènes d'une même voie biologique étant susceptibles d'interagir, nous avons introduit des corrélations entre les mesures d'expression par groupes (les mesures d'expression génetique sont dépendantes par petits groupes mais chaque groupe est indépendant des autres). Le protocole suivant a été appliqué (Dalmasso *et al.*, 2005; Qiu *et al.*, 2005). Pour chaque groupe de 10 gènes, indexés par $l, l = 1, \cdots, 100$, un vecteur aléatoire $A = a_{il}, i = 1, \cdots, n$ a été généré à partir d'une distribution Log-$\mathcal{N}(0, 1)$. La matrice des données a alors été construite de sorte que $Z_{il}^{(g)} = \sqrt{\rho} \cdot A_{il} + \sqrt{1-\rho} \cdot Z_{il}^{*(g)}$, avec $\rho = 0.5$.

Enfin, pour montrer le comportement de notre indice dans des situtations proches d'analyse de données de génomique réelles, nous avons standardisé les données par une normalisation quantile classique (Bolstad *et al.*, 2003).

Dans ce schéma de simulation, les cent premiers gènes sont différentiellement exprimés entre les groupes de patients à haut et bas risque. Les 250 gènes suivants ne sont pas liés au statut haut risque ou bas risque des patients, mais sont différentiellement exprimés selon un facteur binaire non relié au statut haut risque/bas rique. Les autres gènes ne sont pas liés au risque d'apparition de l'évènement.

• Evaluation des simulations.

Pour un seuil donné, nous avons calcué le nombre de gènes différentiellement exprimé obtenus avec notre indice ($\mathbf{D}_0^{(\mathrm{PH})}$, $\mathbf{D}_0^{(\mathrm{POM})}$ ou $\mathbf{D}_0^{(\mathrm{NPH})}$ en fonction du modèle considéré), les indices d'Allison dans sa version initiale ρ_n^2 et modifiée ρ_k^2, de Nagelkerke R_N^2, de Xu et O'Quigley ρ_{XOQ}^2, le FDR calculé à partir du score robuste du modeèle considéré FDR_{sc} ou du log-rapport de vraisemblance du modèle de Cox FDR_L, pour les différentes distribution de survie, les différentes tailles déchantillon, les différents pourcentages de censure et les différentes valeurs du paramètre

ν. Nous avons estimé les taux de vrais positifs (nombre de vrais positifs selectionnés divisé par le nombre de gènes ayant un impact réel sur la survie) et de vrais négatifs (nombre de vrais négatifs sélectionnés divisé par le nombre de gènes n'ayant pas d'impact sur la survie), obtenus avec les cinq indices et les deux FDR, en fonction du seuil choisi. Ces critères ont été respectivement estimés à partir de la moyenne sur 100 itérations de : (i) la proportion de gènes correctement sélectionnés (i.e. gènes g appartiennant à $\{1, \cdots, 100\}$; (ii) la proportion de gènes correctement « non-selectionnés » (i.e. gènes g appartenant à $\{101, \cdots, 1000\}$).

5.2.2 Résultats des simulations

a. Simulations sous un modèle à risques proportionnels

La figure 5.10 représente le taux de vrais positifs en fonction du taux de faux négatifs dans quatre configurations : $e^{\nu} = 2$ ou 3 et un pourcentage de censure égal à 25 ou 50%, pour n=50. Les résultats pour n=100 et 200 sont donnés en Annexe B, figures B.55 et B.56, p. 225 et 226.

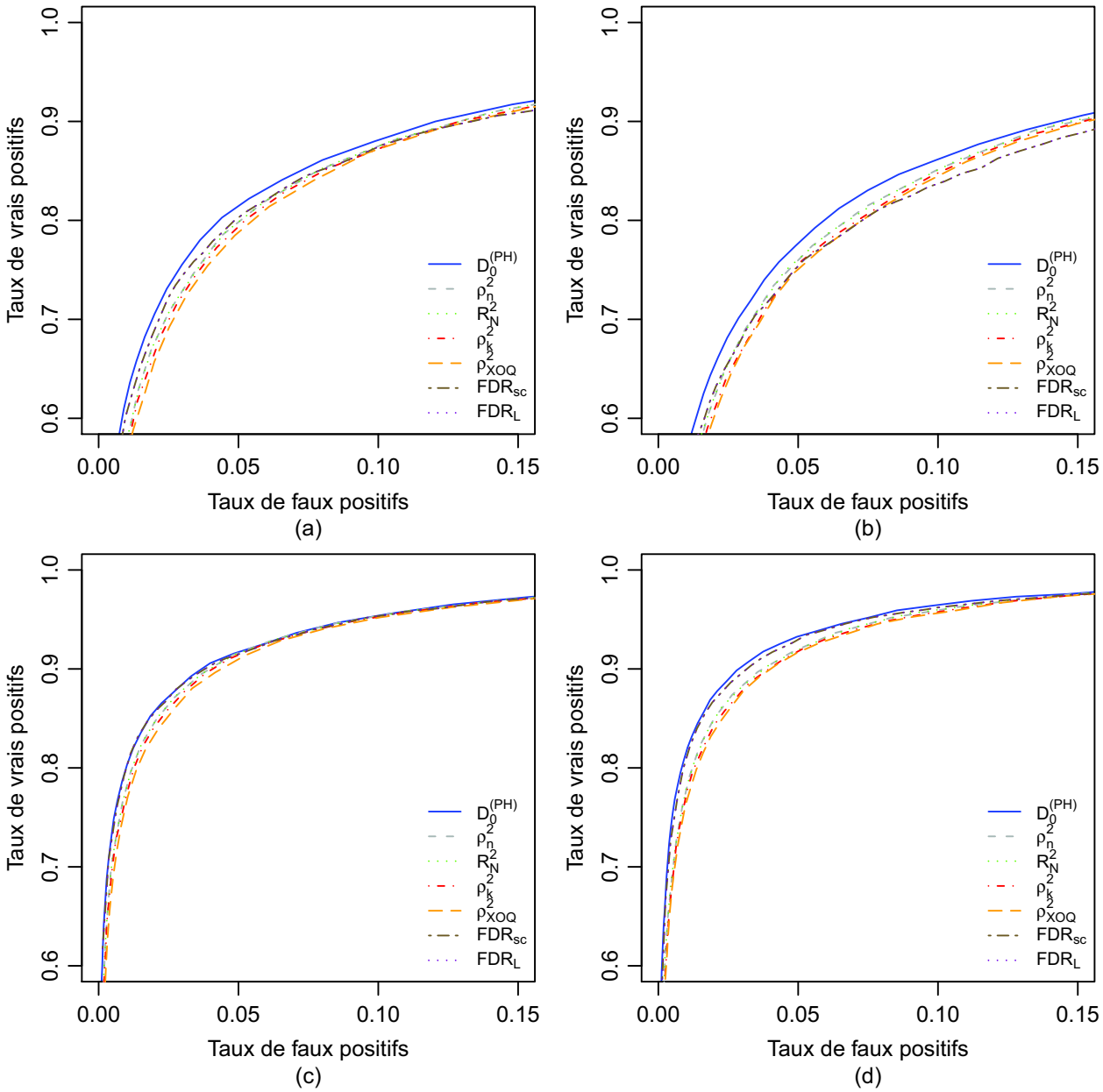
Pour $n = 50$, les figures montrent que notre indice possède de meilleurs caractéristiques que les indices de la littérature. Lorsque les pourcentage de censure est proche de 25%, notre indice est également meilleur que le FDR calculé à partir du score robuste ou du log-rapport de vraisemblance. Dans le cas où $p_c = 50\%$, $\mathbf{D}_0^{(\mathrm{PH})}$ et FDR_{sc} donnent des résultats similaires.

Dans le cas $n = 100$, notre indice donne également de meilleurs résultats que les indices de la littérature, avec un comportement proche du FDR calculé à partir du score robuste.

Lorsque $n = 200$, l'ensemble des indices et mesures de FDR ont des caractéristiques similaires, les courbes atteignant quasiment le point $(0, 1)$. Dans ce cas de figure, le comportement des différentes mesures est proche d'un comportement asymptotique.

Notre indice présente donc un avantage pour de faibles tailles d'échantillons, des pourcentages de censure plus élevés et des effets plus petits.

FIGURE 5.10 – Courbe du taux de vrais positifs en fonction du taux de faux négatifs de $\mathbf{D}_0^{(PH)}$, ρ_n^2, R_N^2, ρ_k^2, ρ_{XOQ}^2, FDR_{sc} et FDR_L, dans le cadre du modèle de Cox, pour n=50, avec (a) $e^\nu = 2$ et $p_c = 0.25$, (b) $e^\nu = 2$ et $p_c = 0.50$, (c) $e^\nu = 3$ et $p_c = 0.25$ et (d) $e^\nu = 3$ et $p_c = 0.50$.



b. Simulations sous un modèle à odds proportionnels

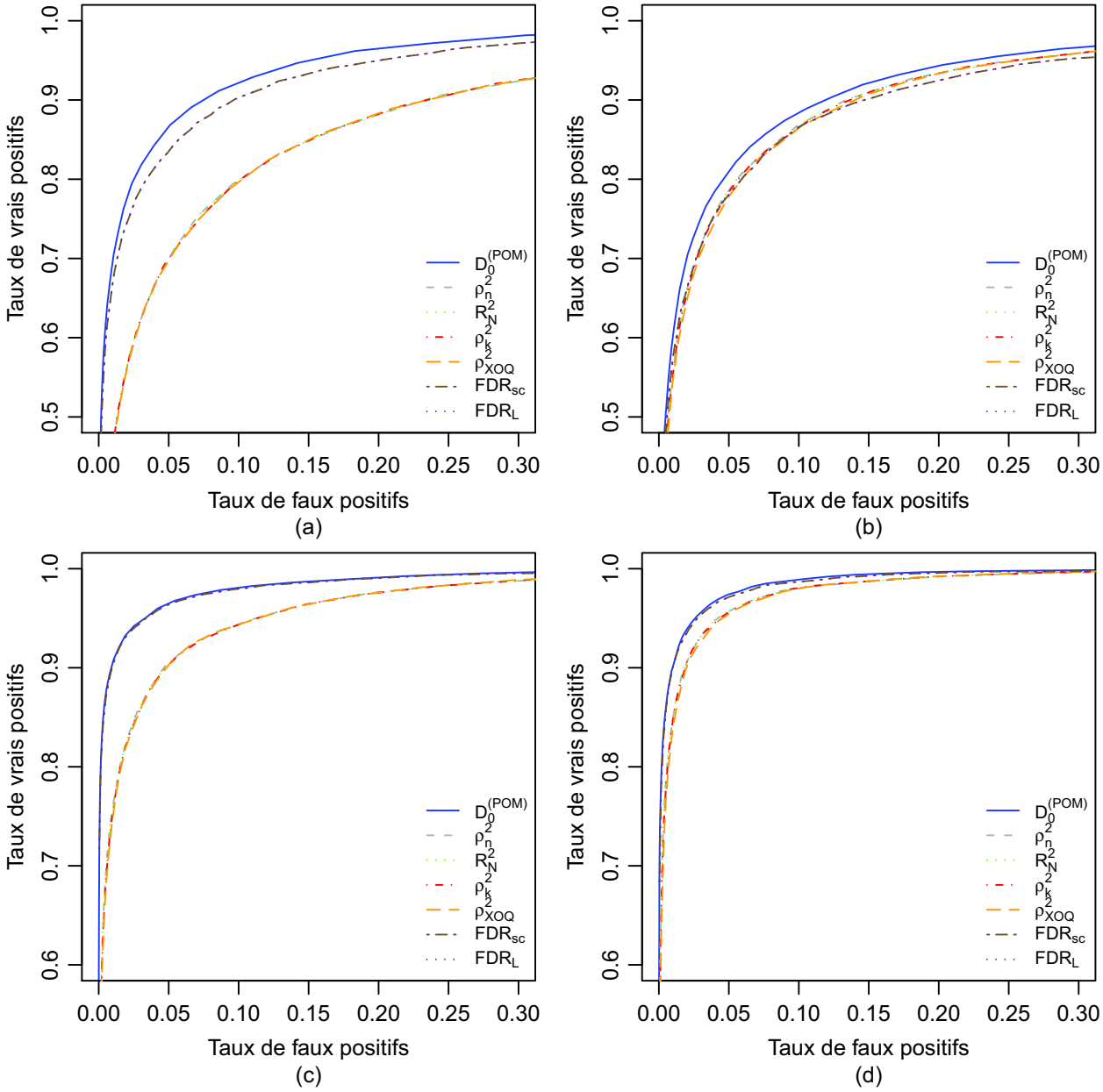
La figure 5.11 représente le taux de vrais positifs en fonction du taux de faux négatifs dans quatre configurations : $e^{\nu} = 3$ ou 5 et un pourcentage de censure égal à 25 ou 50%, pour n=50. Les résultats pour n=100 et 200 sont donnés en Annexe B, figures B.57 et B.58, p. 227 et 228.

La figure 5.11 montre que notre indice possède les meilleures caractéristiques pour $n = 50$. Dans tous les cas de figure, il donne des meilleurs résultats que les indices de la littérature et que le FDR calculé à partir du log-rapport de vraisemblance. Son comportement est très proche du FDR calculé à partir du score robuste, et légèrement meilleur pour $e^{\nu} = 3$ et $p_c = 0.25$.

Lorsque $n = 100$ et $n = 200$, notre indice est également meilleur que les autres indices et que FDR_L, sauf dans le cas d'un effet tel que $e^{\nu} = 5$ et pour un pourcentage de censure de 50%, où les courbes des différentes mesures se superposent. D'une manière générale, les courbes correspondant à $\mathbf{D}_0^{(\text{POM})}$ et FDR_{sc} sont confondues.

Comme dans le cadre du modèle de Cox, l'indice calculé sous le modèle à odds proportionnels présente un intérêt pour des faibles tailles d'échantillon et des effets petits.

FIGURE 5.11 – Courbe du taux de vrais positifs en foncion du taux de faux négatifs de $\mathbf{D}_0^{(\mathrm{PH})}$, ρ_n^2, R_N^2, ρ_k^2, ρ_{XOQ}^2, FDR_{sc} et FDR_L, dans le cadre du modèle à odds proportionnels, pour n=50, avec (a) $e^\nu = 3$ et $p_c = 0.25$, (b) $e^\nu = 3$ et $p_c = 0.50$, (c) $e^\nu = 5$ et $p_c = 0.25$ et (d) $e^\nu = 5$ et $p_c = 0.50$.



c. Simulations sous un modèle à risques non-proportionnels

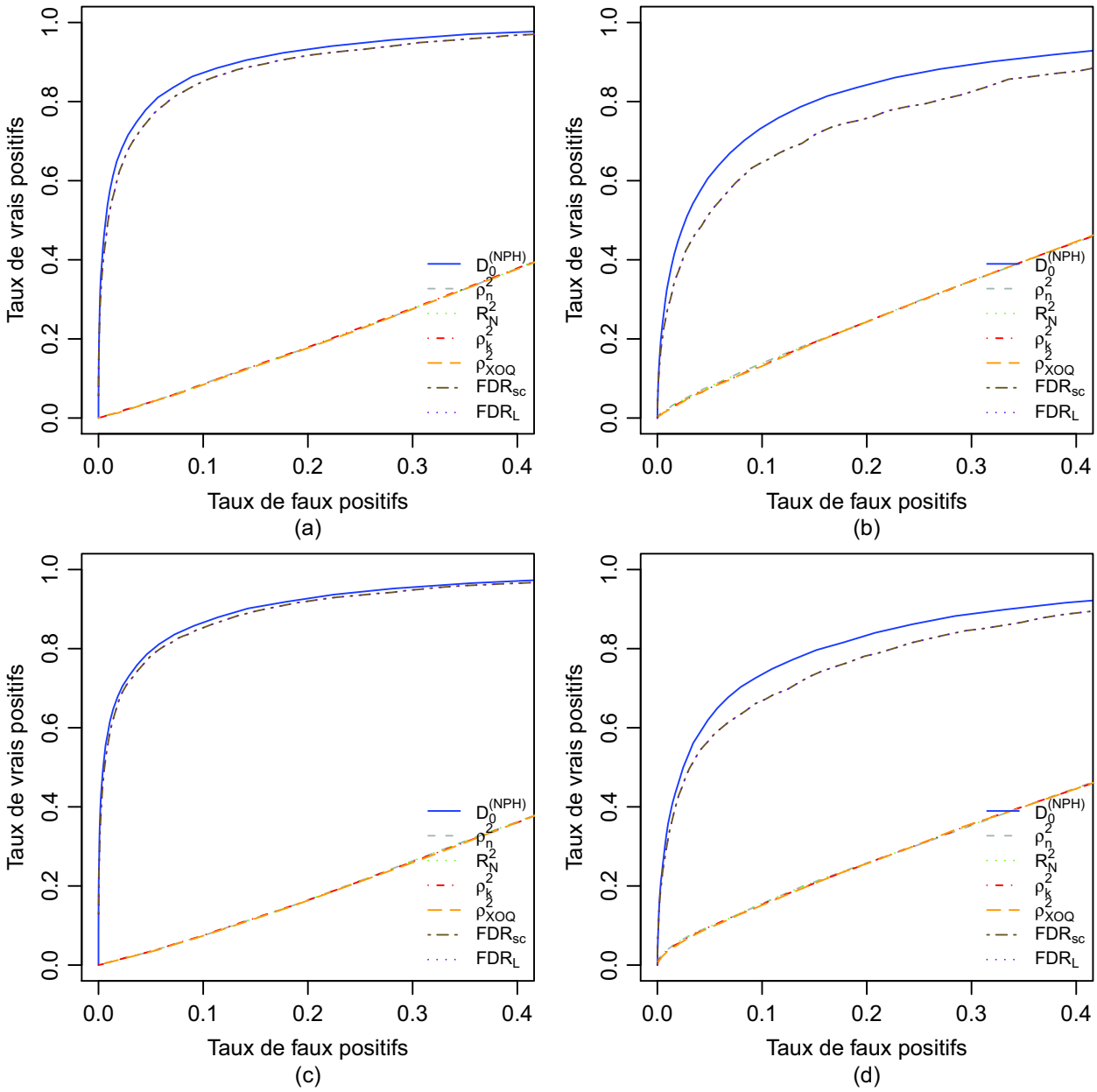
La figure 5.12 représente le taux de vrais positifs en fonction du taux de faux négatifs dans quatre configurations : $e^{\nu} = 2$ ou 3 et un pourcentage de censure égal à 25 ou 50%, pour n=50. Les résultats pour n=100 et 200 sont donnés en Annexe B, figures B.59 et B.60, p. 229 et 230.

Comme le montre la figure 5.12, l'indice $\mathbf{D}_0^{(\text{NPH})}$ présente des meilleures caractéristiques que les autres mesures. Les courbes des indices ρ_n^2, R_N^2, ρ_k^2, ρ_{XOQ}^2 ainsi que FDR_L sont proches de la diagonale, indiquant leur faible capacité à identifier des gènes dont le rapport des risques instantanés s'inverse au cours du temps. Le FDR calculé à partir du score robuste possède des caractéristiques assez proche de $\mathbf{D}_0^{(\text{NPH})}$, mais un peu inférieures.

Lorsque $n = 100$ et $n = 200$, les courbes des indices ρ_n^2, R_N^2, ρ_k^2, ρ_{XOQ}^2 et de FDR_L sont, comme précédemment, proches de la diagonale. Notre indice se comporte de la même façon que le FDR calculé à partir du score robuste.

Dans cette simulation, l'indice $\mathbf{D}_0^{(\text{NPH})}$ présente un net avantage sur les indices de la littérature. Il possède un intérêt par rapport au FDR pour des petites tailles d'échantillon et des effets petits.

FIGURE 5.12 – Courbe du taux de vrais positifs en foncion du taux de faux négatifs de $\mathbf{D}_0^{(\mathrm{PH})}$, ρ_n^2, R_N^2, ρ_k^2, ρ_{XOQ}^2, FDR_{sc} et FDR_L, dans le cadre du modèle à risque qui se croisent, pour n=50, avec (a) $e^\nu = 2$ et $p_c = 0.25$, (b) $e^\nu = 2$ et $p_c = 0.50$, (c) $e^\nu = 3$ et $p_c = 0.25$ et (d) $e^\nu = 3$ et $p_c = 0.50$.



5.3 Simulations dans le cas particulier d'effets modulateurs

Cette section vise à étudier les propriétés pratiques de l'indice sous le modèle à risques non-proportionnels entraînant un croisement des fonctions de risque instantanés. Le but est de montrer que notre indice se comporte mieux que les indices de la littérature pour sélectionner des variables dont le rapport des risques instantanés s'inverse au cours du temps, et potentiellement modulées par une variable cachée, non-mesurée.

5.3.1 Schéma de simulation

Nous reprenons ici le modèle (4.5) introduit p.82 décrivant l'effet modulateur d'une variable cachée sur une variable observée.

On considère deux covariables de Bernoulli $Z^{(1)}$ et $Z^{(2)}$ valant 0 ou 1. Les temps de survie X ont été générés à partir du modèle suivant :

$$\lambda(x|Z^{(1)}; Z^{(2)}) = \lambda_0(x) \exp\{(\alpha Z^{(1)} + \gamma)Z^{(2)}\} \tag{5.1}$$

où $\lambda_0(x)$ a été pris égal à 0.01, de sorte que x varie à peu près entre 0 et 150. Dans ce modèle, la covariable $Z^{(1)}$ a un effet uniquement si $Z^{(2)}$ est présent (effet modulateur). L'effet de $Z^{(1)}$ a été pris égal à $\alpha = \log(1), \log(2), \log(3), \log(4), \log(5), \log(6), \log(7), \log(8), \log(9), \log(10)$. Quand $Z^{(1)} = 1$, l'effet de $Z^{(2)}$, $\gamma = 0.5\alpha$, a été choisi de sorte que le croisement s'effectue autour du temps médian.

Le mécanisme de censure C a été supposé indépendant de X sachant $Z^{(1)}$ et $Z^{(2)}$. et la distribution de la censure était uniforme sur $[0, r]$, r étant choisi de telle sorte que le pourcentage de censure soit environ de 30%.

La distribution jointe de $Z^{(1)}$ et $Z^{(2)}$ était déterminée par les proportions suivantes : $\Pr(Z^{(1)} = 0; Z^{(2)} = 0) = 0.15$, $\Pr(Z^{(1)} = 0; Z^{(2)} = 1) = 0.45$, $\Pr(Z^{(1)} = 1; Z^{(2)} = 1) = 0.15$ et $\Pr(Z^{(1)} = 1; Z^{(2)} = 0) = 0.25$.

Les données ont été générées de la manière suivante. Pour chaque sujet $i, i = 1, \cdots, n$, une valeur de la covariable Z_i a été générée. Étant donné cette valeur, un temps de survie X_i a été généré selon le modèle (5.1). La variable de censure C_i a été générée de façon indépendante et le temps de survie observé T_i a été pris égal à $\min(X_i, C_i)$.

Un total de 200 répétitions a été généré.

5.3.2 Résultats des simulations

Le but est de vérifier que notre indice calculé sous le modèle à risques non-proportionnels est capable de détecter la séparabilité due à $Z^{(1)}$ en analyse marginale (i.e. lorsque $Z^{(2)}$ est omise du modèle) et de comparer son comportement avec celui d'autres indices. Ainsi, pour chaque

répétition, les valeurs de $\mathbf{D}_0^{(\mathrm{NPH})}$, $\mathbf{D}_0^{(\mathrm{PH})}$, ρ_N^2,ρ_k^2, R_N^2 et ρ_{XOQ}^2 ont été calculées à partir du modèle n'incluant que la variable $Z^{(1)}$.

La figure 5.13 montre les boxplot de $\mathbf{D}_0^{(\mathrm{NPH})}$, $\mathbf{D}_0^{(\mathrm{PH})}$, ρ_N^2,ρ_k^2, R_N^2 et ρ_{XOQ}^2 calculés conditionnellement à la valeur de $Z^{(1)}$, pour différentes valeurs de α.

La moyenne de $\mathbf{D}_0^{(\mathrm{NPH})}$ est proche de 0 lorsque $\alpha = \gamma = 0$. Elle augmente avec α et γ, tandis que la moyenne des autres indices n'augmente pas. Les valeurs moyennes de notre indice sont toujours supérieures à celles des autres indices lorsque α et γ sont non nuls.

Notre indice est donc capable de détecter des interactions potentielles à l'inverse des autres indices basés sur le modèle de Cox.

5.4 Conclusions des simulations

En résumé, les simulations montrent que notre indice possède les qualités d'une bonne mesure de capacité de prédiction. En effet, il remplit les critères définis dans la section 3.3 :

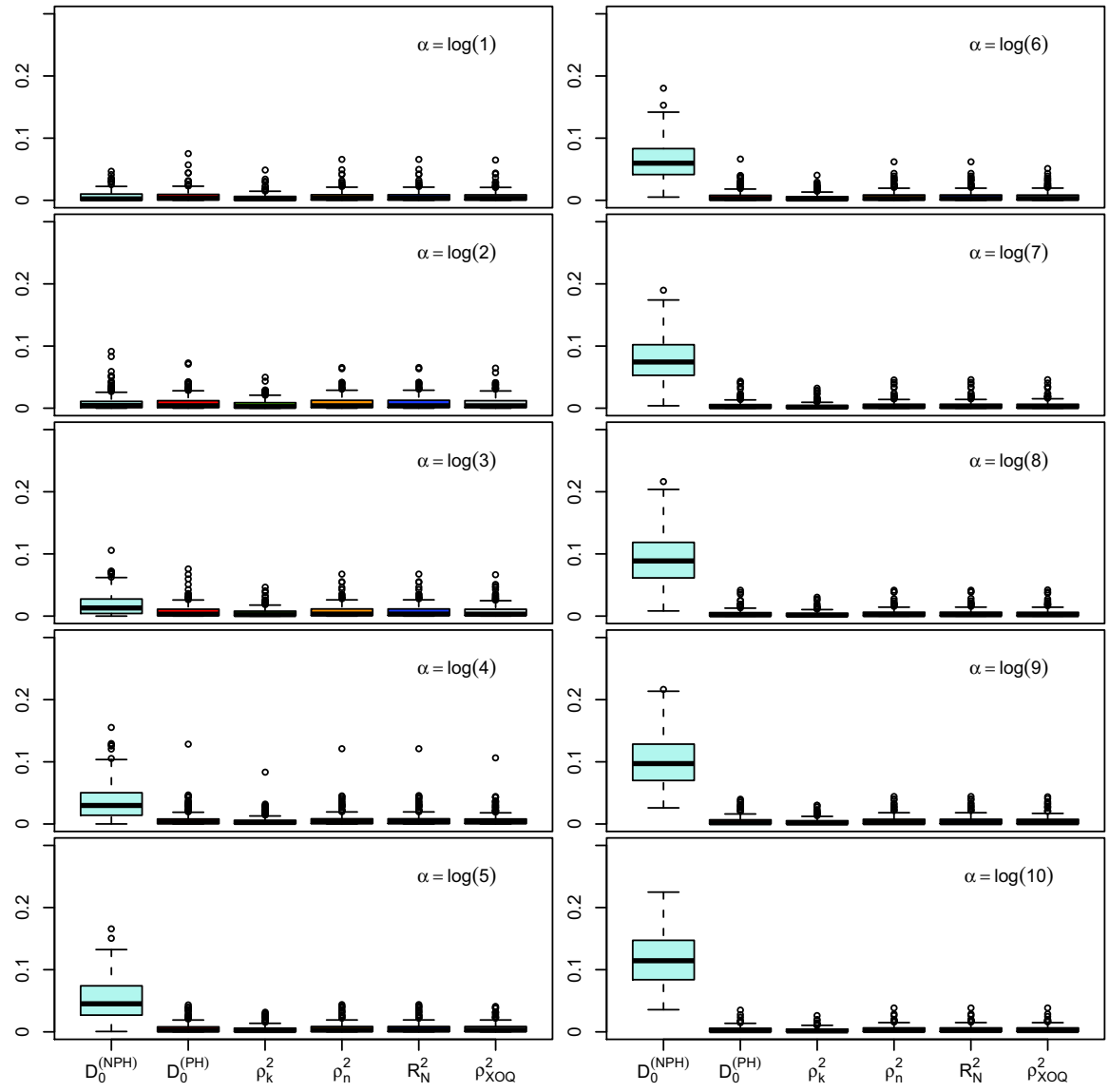
(1) $\mathbf{D}_0$ est compris entre 0 et 1.
(2) Lorsque les coefficients de régression sont égaux à 0, $\mathbf{D}_0$ est proche de 0.
(3) $\mathbf{D}_0$ est une fonction croissante de la valeur absolue des coefficients de régression.
(4) $\mathbf{D}_0$ exploite l'intervalle $[0, 1]$ entièrement, sans posséder de majorant strictement inférieur à 1 ou de minorant strictement supérieur à 0.
(5) La valeur de $\mathbf{D}_0$ est peu modifiée en présence de censure.
(6) La valeur de $\mathbf{D}_0$ n'est pas affecté par la taille de l'échantillon.
(7) $\mathbf{D}_0$ n'est pas Être affectée par une transformation monotone sur l'échelle des temps.
(8) $\mathbf{D}_0$ a une interprétation intuitive.
(9) $\mathbf{D}_0$ est facilement transposable à des modèles de survie à risques non-proportionnels.

Le respect des critères (1), (4), (7), (8) et (9) découle directement de la définition même de l'indice. Les critères (2), (3), (5), (6) sont vérifiés à travers les simulations.

Les indices de la littérature auxquels notre indice a été comparé (indices d'Allison et sa version modifiée, de Nagelkerke et de Xu et O'Quigley) ne remplissent pas les critères 8 et 9. Ils ne possèdent pas d'interprétation intuitive car ils reposent sur une transformation de la statistique du log-rapport de vraisemblance. La transposition à des modèles plus complexes que le modèle de Cox n'est pas évidente et nécessite l'estimation des paramètres de régression β, qui n'est pas immédiate. De plus, l'indice d'Allison ne remplit pas le critère 4, comme nous l'avons évoqué précédemment (voir section 3.3).

L'inconvénient majeur de notre indice est sa sensibilité à la distribution des covariables. Ceci peut Être remédié en travaillant sur les rangs des covariables, ou bien en transformant les données de sorte qu'elles aient la même distribution. Dans les études de génomique, les données sont généralement transformées et standardisées par une normalisation quantile classique (fonction

FIGURE 5.13 – Boxplot de $\mathbf{D}_0^{(\mathrm{NPH})}$, $\mathbf{D}_0^{(\mathrm{PH})}$, ρ_N^2,ρ_k^2, R_N^2 et ρ_{XOQ}^2 calculés conditionnellement à la valeur de $Z^{(1)}$, pour différentes valeurs de α



normalize.quantiles(preprocessCore) dans R). La distribution des covariables a donc un impact mineur pour une utilisation de l'indice en génomique.

D'un point de vue pratique, notre indice possède de meilleures caractéristiques que les indices d'Allison et sa version modifiée, de Nagelkerke et de Xu et O'Quigley, ainsi que le FDR (calculé à partir de la log-vraisemblance partielle ou du score robuste), en particulier pour des petites tailles d'échantillons. Notre possède donc une meilleure capacité à identifier correctement les gènes ayant un impact sur le délai de survenue de l'évènement. Enfin, nous avons vu que $\mathbf{D}_0^{(\mathrm{NPH})}$ était le seul indice capable de détecter des variables à risques croisant et de sélectionner des variables potentiellement modulées par d'autres variables omises du modèle.

Enfin, notons que les codes des fonctions R permettant de calculer $\mathbf{D}_0^{(\mathrm{PH})}$, $\mathbf{D}_0^{(\mathrm{PO})}$ et $\mathbf{D}_0^{(\mathrm{NPH})}$ sont présentés dans l'Annexe D.

Chapitre 6

EXEMPLES D'UTILISATION DE L'INDICE

Contenu

6.1 Introduction à l'oncogénomique

Épidémiologie du cancer

Le cancer est une **cause majeure de décès** dans le monde : c'est la deuxième cause de décès dans les pays industrialisés, après les maladies cardio-vasculaires. D'après les données de l'Organisation Mondiale de la Santé (OMS), en 2008, le nombre de cas incident de cancer a été estimé à 12.4 millions, 6 672 000 chez les hommes et 5 779 000 chez les femmes, et le nombre de décès a été estimé à 7.6 millions, 4 293 000 chez les hommes et 3 300 000 chez les femmes. D'une manière générale, le cancer du poumon est la forme la plus incidente et la plus mortelle chez l'homme. Chez la femme, il s'agit du cancer du sein. L'OMS prévoit qu'en 2030, l'incidence dépassera les 20 millions de cas.

Importance de l'oncogénomique

L'**oncogénomique** (ou **cancérologie génomique**) a pour objectif principal l'implémentation des techniques de la génomique en vue d'un apport de connaissances étiologiques, diagnostiques, pronostiques et thérapeutiques des pathologies tumorales permettant à terme l'amélioration de la prise en charge des patients atteints de cancer. Le cancer est un processus multi-étapes comportant une accumulation d'anomalies génétiques (amplifications, délétions, mutations,...) aboutissant à des modifications cellulaires. Les cellules tumorales prolifèrent de manière excessive, échappent au système immunitaire et envahissent les tissus voisins ou à distance formant des métastases. L'accumulation d'anomalies du génome concerne en particulier les gènes contrôlant la prolifération cellulaire, la réparation de l'ADN, la mort cellulaire « programmée » (ou apoptose), la relation avec les cellules voisines (migration cellulaire) et la capacité d'échappement au système immunitaire. On distingue classiquement les oncogènes, dont la sur-expression contribue à la cancérogenèse, les gènes suppresseurs de tumeurs dont la sous-expression induit la cancérogenèse, et les gènes dits « modificateurs » dont la fonction peut modifier le risque d'apparition de la maladie dû à une exposition, environnementale ou génétique, particulière (e.g. l'alcool).

Les technologies récentes de génomique à haut débit permettent l'analyse simultanée de l'ensemble du génome (analyse pan-génomique) de nombreuses tumeurs conduisant à l'identification d'associations entre variations génomiques et modifications du phénotype clinique d'apparition ou d'évolution d'une pathologie. Ces études conduisent à une meilleure connaissance des voies biologiques impliquées, des interactions gènes-environnement et des biomarqueurs spécifiques à la pathologie étudiée.

L'un des objectifs de la médecine génomique est l'identification de sous-groupes de patients pouvant bénéficier de traitements ciblés. Cette médecine dite personnalisée pour objectif de déterminer les patients susceptibles de bénéficier de thérapeutiques ciblant une voie biologique précise impliquée dans la cancérogenèse. Cette démarche a également a pour objectif de mieux

discriminer les patients guéris ou à faible risque de récidive permettant d'éviter la prescription de thérapeutiques présentant de forts effets secondaires.

6.2 Exemple 1 : sélection de variables génomiques dans différents types de cancer dans le cadre du modèle de Cox

6.2.1 Objectif

Certains oncogènes et gènes suppresseurs de tumeurs sont altérés dans plusieurs types de cancer, indépendamment de l'organe ou de la cause du cancer. Ces gènes « acteurs multiples » sont potentiellement des gènes majeurs du métabolisme cellulaire dont l'altération est une étape importante de la cancérogenèse. Dans ce chapitre, nous utilisons l'indice que nous avons développé pour l'identification de facteurs pronostiques transcriptomiques communs à huit études correspondant à cinq tumeurs solides différentes : carcinome mammaire, carcinome bronchopulmonaire, carcinome urothélial (vessie), gliome (système nerveux central) et mélanome malin (peau). Notre indice est comparé à ceux de la littérature, ainsi qu'à des méthodes de sélection plus classiques, reposant sur l'estimation de quantités liées aux statistiques de tests d'hypothèses (les q-values).

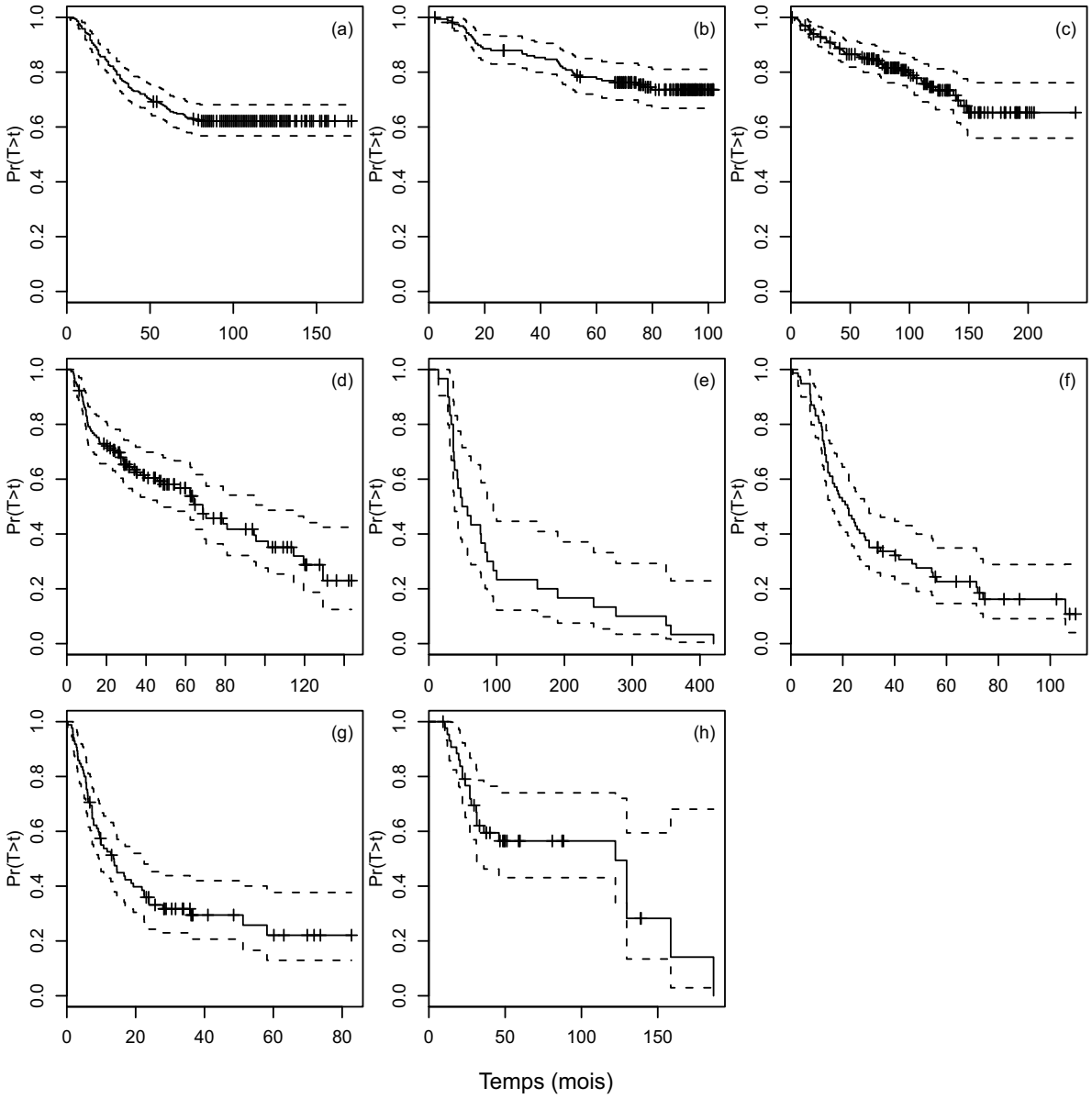
6.2.2 Présentation des données

Les données consistent en huit études génomiques indépendantes (Wang *et al.*, 2005; Pawitan *et al.*, 2005; Schmidt *et al.*, 2008; Raponi *et al.*, 2006; Als *et al.*, 2007; Phillips *et al.*, 2006; Freije *et al.*, 2004; Bogunovic *et al.*, 2009), de taille et d'évènement d'intérêt différents, ayant été analysées sur une même plateforme génomique (Affymetrix HU133 Plus 2.0 ou HU133A ; Affymetrix, Santa Clara, CA, USA). Les jeux de données sont disponibles publiquement sur le site de GEO (*http ://www.ncbi.nlm.nih.gov/geo/*) sous les numéros de référence GSE2034, GSE1456, GSE11121, GSE4573, GSE5287, GSE4271, GSE4412 et GSE19234, et sont brièvement décrits dans ce qui suit.

Cohorte GSE2034, cancer du sein (Wang *et al.*, 2005). Cette série comprend 286 patientes atteintes de carcinome mammaire sans envahissement lymphatique, parmi lesquels 106 ont développé des métastases, qui est l'évènement d'intérêt dans cette étude. Le critère principal est la survie sans métastase définie par le délai entre la date du traitement initial et la date d'une progression métastatique documentée ou du décès. La survie médiane est de 80 mois. La survie à deux ans sans métastase est de 83.9%[79.8%; 88.3%], et la survie à cinq ans est de 66.7%[61.4%; 72.4%].

Cohorte GSE1456, cancer du sein (Pawitan *et al.*, 2005). Cette série (référencée sous le nom « Cohorte de Stockholm ») est composée de 159 patientes atteintes de cancer du sein primitif non métastatique. Le critère principal est la survie sans métastase définie par le délai

FIGURE 6.1 – Courbes de survie de Kaplan-Meier des différentes études : (a) Cohorte GSE2034, cancer du sein (Wang *et al.*, 2005) ; (b) Cohorte GSE1456, cancer du sein (Pawitan *et al.*, 2005) ; (c) Cohorte GSE11121, cancer du sein (Schmidt *et al.*, 2008) ; (d) Cohorte GSE4573, cancer broncho-pulmonaire (Raponi *et al.*, 2006) ; (e) Cohorte GSE5287, cancer urothélial (vessie) (Als *et al.*, 2007) ; (f) Cohorte GSE4271, gliome (Phillips *et al.*, 2006) ; (g) Cohorte GSE4412, gliome (Freije *et al.*, 2004) ; (h) Cohorte GSE19234, mélanome (Bogunovic *et al.*, 2009).



entre la date du traitement initial et la date d'une progression métastatique documentée ou du décès. La survie médiane est de 80 mois. La survie à deux ans est de 87.9%[83.0%, 93.2%], et la survie à cinq ans de 77.6%[71.3%, 84.4%].

Cohorte GSE11121, cancer du sein (Schmidt *et al.*, 2008). Cette série est composée de 200 patientes atteintes d'un cancer du sein sans envahissement lymphatique et sans chimiothérapie adjuvante. Le critère principal est la survie sans métastase définie par le délai entre la date du traitement initial et la date d'une progression métastatique documentée ou du décès. La survie médiane est de 149 mois. La survie à deux ans est de 92.9%[89.3%; 96.5%], et la survie à cinq ans de 85.4%[80.6%; 90.6%].

Cohorte GSE4573, cancer broncho-pulmonaire (Raponi *et al.*, 2006). Cette série comprend 129 patients atteints de carcinomes épidermoïdes traités par exérèse chirurgicale réglée (lobectomie ou pneumonectomie). Le critère principal est la survie globale définie par le délai entre la date du traitement initial et la date du décès. La survie médiane est de 63 mois. La survie à deux ans est de 70.5%[63.1%; 78.9%], et la survie à cinq ans de 56.8%[48.3%; 66.7%].

Cohorte GSE5287, cancer urothélial (vessie) (Als *et al.*, 2007). Cette série est composée de 30 patients atteints de tumeur urothéliale, traités par chimiothérapie. Le critère principal est la survie globale définie par le délai entre la date de la première chimiothérapie et la date du décès. La survie médiane est de 47 mois. La survie à deux ans est de 96.7%[90.5%; 100%], et la survie à cinq ans de 46.7%[31.8%; 68.4%].

Cohorte GSE4271, gliome (Phillips *et al.*, 2006). Cette série comprend 77 patients atteints d'astrocytomes de grades élevés (III et IV) traités par résection chirurgicale de la tumeur. Le critère principal est la survie globale définie par le délai entre la date du traitement initial et la date du décès. La survie médiane est de 21 mois. La survie à deux ans est de 45.5%[35.6%, 58.1%], et la survie à cinq ans de 22.6%[14.7%, 34.9%].

Cohorte GSE4412, gliome (Freije *et al.*, 2004). Cette série comprend 85 patients souffrant de gliome de stade III ou IV, tous types histologiques confondus (glioblastomes, astrocytomes anaplasiques, oligo-dendrogliomes anaplasiques, oligo-astrocytomes anaplasiques mixtes). Le critère principal est la survie globale définie par le délai entre la date de la chirurgie et la date du décès. La survie médiane est de 13 mois. La survie à deux ans est de 33.2%[24.3%, 45.3%], et la survie à cinq ans de 22.1%[12.9%, 37.7%].

Cohorte GSE19234, mélanome (Bogunovic *et al.*, 2009). Les auteurs considèrent 44 échantillons de tissus provenant de mélanomes métastatiques. Le critère principal est la survie

globale définie par le délai entre l'exérèse de la lésion métastatique et le décès. La survie médiane est de 46 mois. La survie à deux ans est de $76.7\%[65.0\%, 90.4\%]$, et la survie à cinq ans de $56.5\%[43.1\%, 74\%]$.

Pour l'ensemble de ces études, les hybridations ont été réalisées à partir de matériel congelé sur des puces Affymetrix HU133A, excepté pour la série « mélanome », pour laquelle, elles ont été faites sur la puce HU133 Plus 2.0 (HU133A+HU133B). Pour chaque patient, nous disposons donc de l'information issue de 22 283 transcrits (puce HU133A).
La figure 6.1 représente les courbes de survie de ces différentes études.

6.2.3 Choix du seuil

Pour choisir un seuil pour opérer la sélection, nous avons utilisé la procédure introduite par Blangiardo et Richardson (2007). Les principales étapes de la procédure sont les suivantes. Dans un premier temps, les gènes sont rangés selon la mesure étudiée. Pour chaque étude, nous avons considéré les sous-ensembles des transcrits associés au critère principal de jugement (i.e. dont la valeur de la mesure est supérieure/inférieure à un certain seuil) et étudié la cardinalité de l'ensemble des transcrits communs aux différentes études. Ce nombre a ensuite été comparé au nombre de gènes communs attendus, calculé sous l'hypothèse nulle d'indépendance entre les études considérés. Le rapport entre les valeurs observées et attendues a été calculé pour tous les seuils possibles. Pour notre étude, le seuil retenu est celui tel que le rapport soit supérieur à 2 et avec une séparabilité supérieure à 0.07 (différence cliniquement « utile »), ce qui correspond, d'après nos simulations, à un risque relatif d'environ 1.5 .

Nous avons utilisé cette procédure avec les critères suivants : (1) notre index $\mathbf{D}_0^{(\mathrm{PH})}$ calculé sous un modèle de Cox à risque proportionnels ; (2) l'indice d'Allison ρ_n^2 ; (3) la version modifiée de l'indice d'Allison ρ_k^2 ; (4) l'indice de Nagelkerke R_N^2 ; (5) l'indice de Xu et O'Quigley ρ_{XOQ}^2 ; (6) la q-value associée au FDR (False Discovery Rate) calculé à partir de la statistique du score robuste et estimé selon une méthode non-paramétrique (Dalmasso *et al.*, 2005) ; (7) la q-value associée au FDR calculé à partir de la statistique du log-rapport de vraisemblance, estimé avec la même méthode.

6.2.4 Résultats de la sélection

L'indice proposé a été calculé pour les 22 283 mesures d'expression génétique pour les huit jeux de données. Le seuil calculé avec la procédure d'intersection de Blangiardo et Richardson (2007) était de 7% pour $\mathbf{D}_0^{(\mathrm{PH})}$.

Pour $\mathbf{D}_0^{(\mathrm{PH})} \geq 7\%$ (ce qui correspond, d'après nos simulations, à un risque relatif d'environ 1.5), nous avons sélectionné 5 transcrits correspondant à quatre gènes (tableau 6.1).

TABLEAU 6.1 – Gènes les plus différentiellement exprimés communs aux huit études pour $\mathbf{D}_0^{(\mathrm{PH})} \geq 7\%$

AffyID	Gène	Nom complet du gène	Cytobande	RR
211596-s-at	*LRIG1*	leucine-rich repeats and immunoglobulin-like domains 1	3p14	< 1
218355-at	*KIF4A*	kinesin family member 4A	Xq13.1	> 1
218726-at	*HJURP*	Holliday junction recognition protein	2q37.1	> 1
204817-at	*ESPL1*	extra spindle pole bodies homolog 1 (S. cerevisiae)	12q13.13	> 1
38158-at	*ESPL1*	extra spindle pole bodies homolog 1 (S. cerevisiae)	12q13.13	> 1

AffyID, identifiant Affymetrix pour chaque probe ; RR, risque relatif
Si $RR > 1$, la sur-expression du gène est associé à un mauvais pronostique. Si $RR < 1$, la sous-expression du gène est associée à un bon pronostique.

Parmi les gènes identifiés, les gènes *HJURP* et *LRIG1* jouent un rôle dans la cancérogenèse. *HJURP* code pour un facteur biologique nécessaire à la stabilité des chromosomes dans les cellules cancéreuses immortalisées. Il est sur-exprimé dans le cancer broncho-pulmonaire (Kato *et al.*, 2007). *LRIG1* code pour une protéine qui agit comme un suppresseur de tumeur ayant un rôle inhibiteur sur la croissance cellulaire et dont l'expression diminue dans le cancer du sein (Miller *et al.*, 2008). La majorité des tumeurs du sein ErbB2 (HER2) positives (sur-expression liée à l'amplification du gène) montrent une sous-expression de la protéine *LRIG1*. Dans notre exemple, l'augmentation de l'expression de *HJURP* et la diminution de l'expression *LRIG1* sont associées à un mauvais pronostique.

Notre processus de sélection a également permis d'identifier deux gènes impliqués dans la régulation du cycle cellulaire. Le gène *KIF4A* code pour une protéine importante pour la régulation de la mitose, notamment pour la condensation des chromosomes et l'organisation du fuseau de division. Il possède un lien fonctionnel (régulation en amont) et physique (site de liaison) avec le produit du gène *BRCA2* (breast cancer 2, early onset) (Wu *et al.*, 2008).

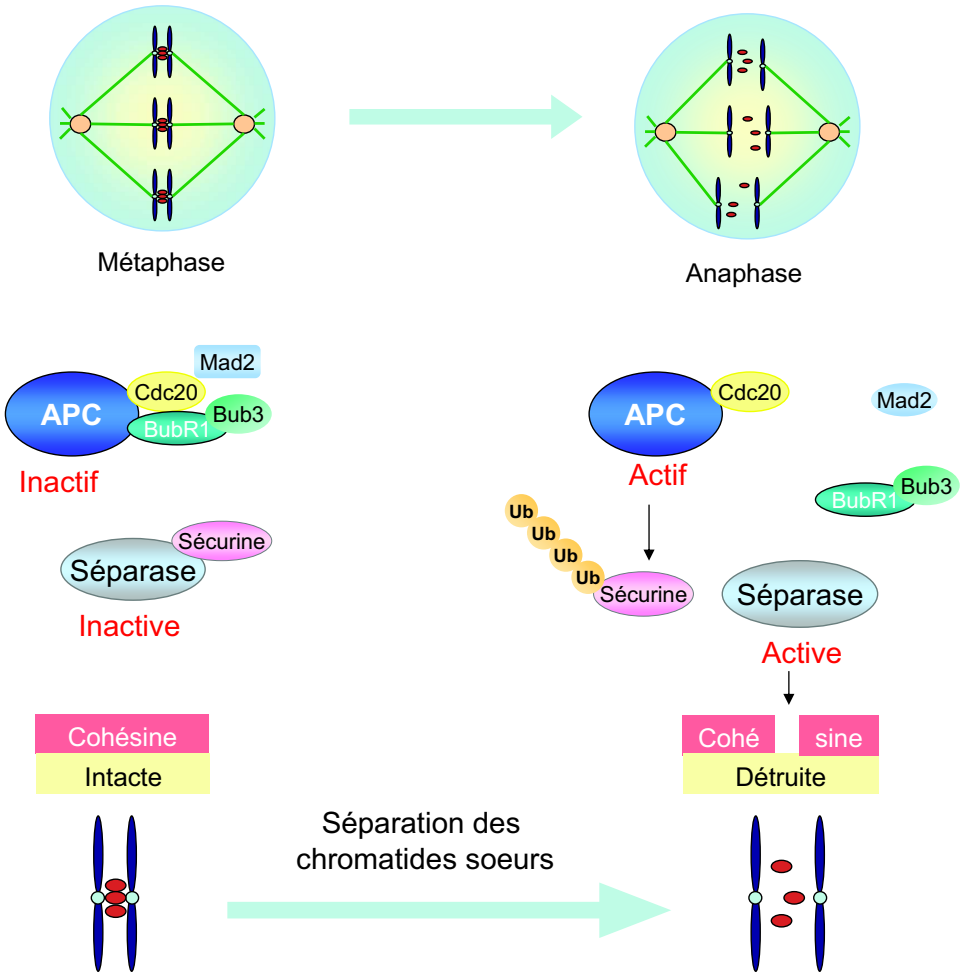
Le gène *ESPL1* joue un rôle central dans la ségrégation des chromosomes au moment de l'anaphase (voir figure 6.2 tirée de Bharadwaj et Yu (2004)). Sa sur-expression induit une aneuploidie et est associé à la cancérogenèse (Zhang *et al.*, 2008). L'article de Zhang *et al.* (2008) montre que le transcrit de *ESPL1* est sur-exprimé dans le cancer du sein chez l'homme. Il est important de noter que *ESPL1* et *KIF4A*, ont été précédemment identifiés dans une méta-analyse conduite par Carter *et al.* (2006). Pour ces deux gènes, la sur-expression, conduisant à une prolifération cellulaire, est associée à un mauvais pronostique.

Pour chacun des gènes, les risques relatifs sont concordants (« même sens ») pour l'ensemble des huit études.

Les courbes de survie relatives aux niveaux d'expression (haut et bas, définis selon les 1$^{\mathrm{er}}$, 2$^{\mathrm{ème}}$ ou 3$^{\mathrm{ème}}$ quartiles) de ces cinq transcrits sont représentées sur les figures C.1, C.2, C.3, C.4, C.5, de l'annexe C.

Avec le même seuil de 7%, l'indice d'Allison ρ_k^2 permet de sélectionner 3 transcrits correspondant aux gènes *KIF4A* et *ESPL1* et l'indice de Xu et O'Quigley's ρ_{XOQ}^2 permet de sélectionner

FIGURE 6.2 – D'après Bharadwaj et Yu (2004). Régulation de la séparation des chromatides au cours de l'anaphase. La séparase est maintenue sous forme inactive par association avec la sécurine. Le complexe APC (Anaphase Promoting Complex)/cdc20 induit l'ubiquitination de la sécurine, conduisant à sa destruction protéolytique par le protéasome. Cette destruction entraîne l'activation de la séparase, la dissociation du complexe cohésine et la séparation des chromatides.



2 transcrits correspondant au gène *ESPL1*. Les transcrits identifiés avec ces deux indices sont tous inclus dans notre sous-liste. Pour ρ_N^2 et R_N^2 avec un seuil de 7%, aucun transcrit n'est retenu. Aucun transcrit n'a non plus été sélectionné à partir de la q-value calculée sur le score robuste ou le log-rapport de vraisemblance avec un seuil de 40%.

6.3 Exemple 2 : sélection de variables génomiques dans une étude de cancer du poumon dans le cadre d'un modèle à risques non-proportionnels

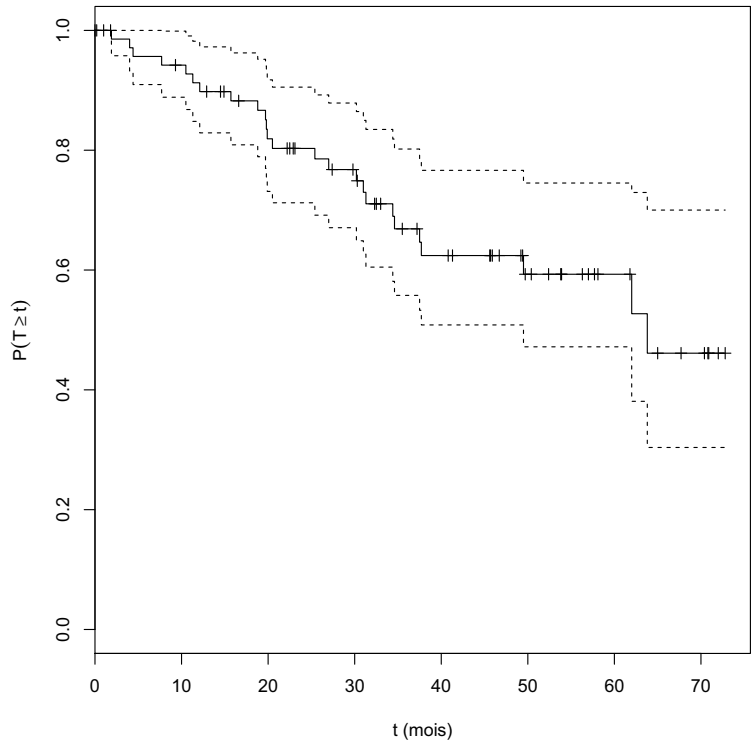
6.3.1 Objectif

La plupart des études visant à modéliser la relation entre des marqueurs biologiques et des données de survie reposent sur le modèle de Cox à risques proportionnels, qui suppose une risque relatif constant. Cependant, lorsqu'on s'intéresse à des données de haute dimensionnalité, l'hypothèse de proportionnalité des risques n'est pas vérifiée pour l'ensemble des gènes. Certains gènes peuvent, en effet, avoir un effet croisant en terme de risque de l'apparition de l'évènement étudié, tel que le rapport des risques instantanés s'inverse au cours du temps (>1 puis <1). L'objectif de cette sous-section est d'illustrer l'utilisation de notre mesure pour identifier des facteurs transcriptomiques ayant un tel effet (dénommé par la suite « crossing effect » ou « effet d'inversion » des risques instantanés) dans une étude sur des adénocarcinomes broncho-pulmonaires. La sélection obtenue à l'aide de l'indice de séparabilité dédié à cet « effet d'inversion » est comparée à celle obtenue à partir de l'indice calculé sous un modèle à risques proportionnels.

6.3.2 Présentation des données

Cohorte Merlion (Broët *et al.*, 2009). Cette série rétrospective est composée de 74 patients opérés (exérèse réglée) entre août 2000 et février 2004 à l'Hôtel-Dieu de Paris d'un adénocarcinome primitif broncho-pulmonaire de stade IB (pT2N0) ou cancer à grandes cellules (non neuro-endocrine) de localisation périphérique. L'évènement d'intérêt est la récidive tumorale. Le critère principal est la survie sans récidive définie par le délai entre la chirurgie initiale et la récurrence de la maladie (locale ou métastatique) ou le décès lié à la maladie. La survie médiane est de 63.8 mois. La survie à deux ans est de $80.3\%[71.2\%, 90.5\%]$, et la survie à cinq ans de $59.3\%[47.2\%, 74.5\%]$. Pour chaque patient, nous avons utilisé l'information contenue dans les 51 852 transcrits de la puce Affymetrix HU133 Plus 2.0 exprimés pour les gènes localisés sur les chromosomes autosomaux. La figure 6.3 montre la courbe de Kaplan-Meier de la survie pour la cohorte Merlion.

FIGURE 6.3 – Courbe de survie de Kaplan-Meier de la cohorte Merlion (Broët *et al.*, 2009)



6.3.3 Résultats de la sélection

Les gènes ont été rangés soit selon la valeur de $\mathbf{D}_0^{(\mathrm{NPH})}$, soit selon la valeur de $\mathbf{D}_0^{(\mathrm{PH})}$. Dans les deux cas, nous avons focalisé notre étude sur les 200 premiers transcrits. La séparabilité minimale était proche pour les deux indices (29.8% pour $\mathbf{D}_0^{(\mathrm{NPH})}$ et 29.1% pour $\mathbf{D}_0^{(\mathrm{PH})}$). Seule une petite proportion de transcrits (5%) était commun aux deux listes. Nous avons examiné les processus biologiques qui étaient significativement surreprésentés en utilisant le système de classification de PANTHER (Protein ANalysis THrough Evolutionary Relationships) (Thomas *et al.*, 2003). Les résultats montrent que les deux indices aboutissent à la sélection de gènes impliqués dans des processus biologiques différents (voir tableau 6.2).

Pour le cycle cellulaire, $\mathbf{D}_0^{(\mathrm{NPH})}$ permet de sélectionner 25 transcrits (nombre significativement plus grand que les 5% attendus par chance), tandis que $\mathbf{D}_0^{(\mathrm{PH})}$ ne permet d'en sélectionner que 16 (nombre non supérieur aux 5% attendus par chance). Les deux listes de gènes impliqués dans le cycle cellulaire sont présentées dans les tableaux 6.3 et 6.4.

TABLEAU 6.2 – Processus biologiques de l'étude de cancer broncho-pulmonaire en fonction de l'indice utilisé : (a) avec $\mathbf{D}_0^{(\mathrm{NPH})}$, (b) avec $\mathbf{D}_0^{(\mathrm{PH})}$

Biological process	♯ Genes NCBI	♯ observed	♯ expected	P-value
(a) Sélection avec $\mathbf{D}_0^{(\mathbf{NPH})}$				
nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	3825	53	30.74	$2.17 \cdot 10^{-5}$
metabolic process	8267	86	66.43	$1.20 \cdot 10^{-3}$
primary metabolic process	7950	83	63.88	$1.47 \cdot 10^{-3}$
cell cycle	1840	23	14.79	$2.25 \cdot 10^{-2}$
polyphosphate catabolic process	4	1	0.03	$3.16 \cdot 10^{-2}$
cellular component organization	1443	18	11.6	$4.24 \cdot 10^{-2}$
vesicle-mediated transport	1160	15	9.32	$4.75 \cdot 10^{-2}$
(b) Sélection avec $\mathbf{D}_0^{(\mathbf{PH})}$				
oxidative phosphorylation	76	3	0.57	$2.02 \cdot 10^{-2}$
intracellular signaling cascade	1568	19	11.81	$2.72 \cdot 10^{-2}$
lipid metabolic process	1119	3	8.43	$2.83 \cdot 10^{-2}$
muscle contraction	448	0	3.38	$3.29 \cdot 10^{-2}$
spermatogenesis	501	8	3.77	$3.68 \cdot 10^{-2}$

TABLEAU 6.3 – Liste des gènes impliqués dans le cycle cellulaire sélectionnés à partir de $\mathbf{D}_0^{(\mathrm{NPH})}$

ID Affy	Symbole du gène	Nom UniGene	Cytobande	$\mathbf{D}_0^{(\mathrm{NPH})}$
221326-s-at	*TUBD1*	tubulin, delta 1	chr17q23.1	37.9%
209661-at	*KIFC3*	kinesin family member C3	chr16q13-q21	35.9%
208228-s-at	*FGFR2*	fibroblast growth factor receptor 2	chr10q26	34.5%
206003-at	*CEP135*	centrosomal protein 135kDa	chr4q12	34.4%
225237-s-at	*MSI2*	musashi homolog 2 (Drosophila)	chr17q22	34.1%
1562139-a-at	*FOXP2*	forkhead box P2	chr7q31	34.0%
206113-s-at	*RAB5A*	RAB5A, member RAS oncogene family	chr3p24-p22	33.7%
200796-s-at	*MCL1*	myeloid cell leukemia sequence 1 (BCL2-related)	chr1q21	32.9%
221281-at	*SRC*	v-src sarcoma (Schmidt-Ruppin A-2) viral oncogene homolog (avian)	chr20q12-q13	32.7%
207319-s-at	*CDC2L5*	cell division cycle 2-like 5 (cholinesterase-related cell division controller)	chr7p13	32.5%
213023-at	*UTRN*	utrophin	chr6q24	32.1%
214908-s-at	*TRRAP*	transformation/transcription domain-associated protein	chr7q21.2-q22.1	32.1%
222563-s-at	*TNKS2*	tankyrase, TRF1-interacting ankyrin-related ADP-ribose polymerase 2	chr10q23.3	32.1%
1555346-at	*CDC20B*	cell division cycle 20 homolog B (S. cerevisiae)	chr5q11.2	31.9%
219944-at	*CLIP4*	CAP-GLY domain containing linker protein family, member 4	chr2p23.2	31.8%
239223-s-at	*FBXL20*	F-box and leucine-rich repeat protein 20	chr17q12	31.6%
222540-s-at	*RSF1*	remodeling and spacing factor 1	chr11q14.1	31.1%
203639-s-at	*FGFR2*	fibroblast growth factor receptor 2	chr10q26	31.0%
224010-at	*ANAPC11*	anaphase promoting complex subunit 11	chr17q25.3	31.0%
215739-s-at	*TUBGCP3*	tubulin, gamma complex associated protein 3	chr13q34	30.9%
206235-at	*LIG4*	ligase IV, DNA, ATP-dependent	chr13q33-q34	30.8%
243999-at	*SLFN5*	schlafen family member 5	chr17q12	30.5%
211103-at	*MYO7A*	myosin VIIA	chr11q13.5	30.4%
211401-s-at	*FGFR2*	fibroblast growth factor receptor 2	chr10q26	30.1%
212308-at	*CLASP2*	cytoplasmic linker associated protein 2	chr3p22.3	29.8%

TABLEAU 6.4 – Liste des gènes impliqués dans le cycle cellulaire sélectionnés à partir de $\mathbf{D}_0^{(\mathrm{PH})}$

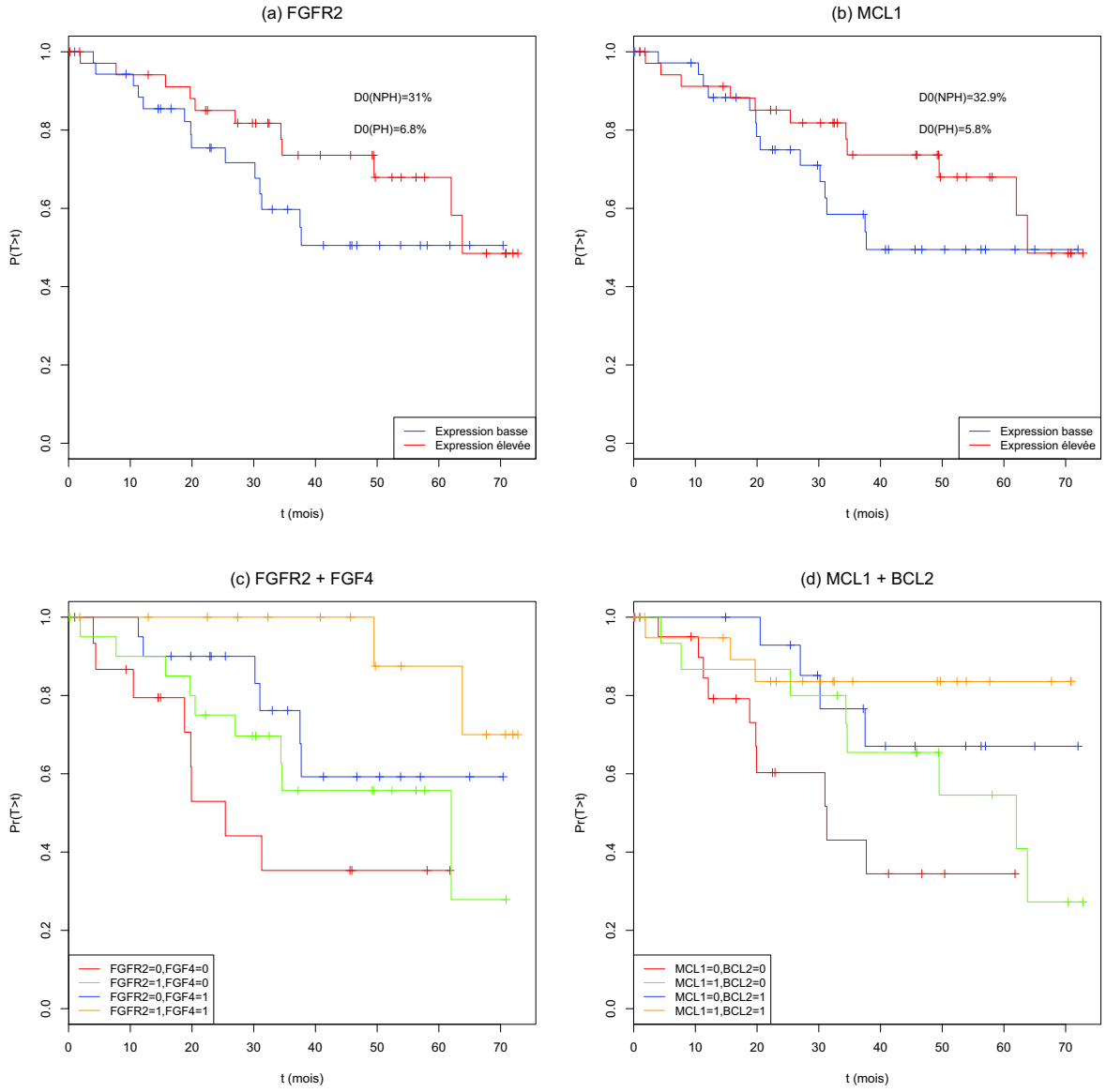
ID Affy	Symbole du gène	Nom UniGene	Cytobande	$\mathbf{D}_0^{(\mathrm{PH})}$
226765-at	*SPTBN1*	CDNA clone IMAGE :3897439	chr2p21	41.0%
1566851-at	*TRIM42*	tripartite motif-containing 42	chr3q23	37.1%
202257-s-at	*CD2BP2*	CD2 (cytoplasmic tail) binding protein 2	chr16p11.2	36.3%
203348-s-at	*ETV5*	ets variant 5	chr3q28	35.8%
210356-x-at	*MS4A1*	membrane-spanning 4-domains, subfamily A, member 1	chr11q12	34.3%
205861-at	*SPIB*	Spi-B transcription factor (Spi-1/PU.1 related)	chr19q13.3-q13.4	34.2%
233251-at	*STRBP*	Chromosome 9 open reading frame 45, mRNA	chr9q33.3	31.5%
216609-at	*TXN*	Full length insert cDNA clone YI46D09	chr9q31	31.4%
214114-x-at	*FASTK*	Fas-activated serine/threonine kinase	chr7q35	30.9%
202252-at	*RAB13*	RAB13, member RAS oncogene family	chr1q21.2	30.9%
202810-at	*DRG1*	developmentally regulated GTP binding protein 1	chr22q12.2	30.7%
202676-x-at	*FASTK*	Fas-activated serine/threonine kinase	chr7q35	30.7%
229664-at	*MAPK8*	mitogen-activated protein kinase 8	chr10q11.22	30.5%
224434-s-at	*MORG1*	mitogen-activated protein kinase organizer 1	chr19p13.13	30.3%
210187-at	*FKBP1A*	FK506 binding protein 1A, 12kDa	chr20p13	29.9%
210111-s-at	*KLHDC10*	KIAA0265 gene	chr7q32.2	29.7%

Parmi les 25 transcrits du cycle cellulaire sélectionnés avec $\mathbf{D}_0^{(\mathrm{NPH})}$, nous discutons le comportement de deux gènes, *FGFR2* et *MCL1*, connus pour être impliqués dans de multiples voies biologiques, et leur potentielle modulation par d'autres gènes. En particulier, nous avons cherché à savoir si l'« effet d'inversion » des risques instantanés de ces deux gènes (voir figures 6.5(a) et 6.5(b)) pouvait être associé à un effet modulateur lié à la présence d'autres gènes.

Le gène *FGFR2* (fibroblast growth factor receptor 2), est impliqué dans différents types de cancer (Katoh, 2008) et une faible expression a été reportée comme associée à une survie plus courte dans le cancer broncho-pulmonaire (Raponi *et al.*, 2006). L'analyse de l'expression du gène *FGFR2*, en tenant compte de l'expression du gène *FGF4*, qui est l'un de ses ligands, suggère un potentiel effet modulateur entre les deux gènes. Les risques relatifs (RR) calculés sous un modèle de Cox PH pour les quatre groupes résultant de la dichotomisation des deux variables d'expression de ces deux gènes à la médiane ont été calculés. Les courbes de Kaplan-Meier correspondantes sont représentées sur la figure 6.5(c). Comme le montre cette dernière, les patients avec une faible expression (inférieure à la médiane) de *FGFR2* et *FGF4* ont le plus mauvais pronostic (groupe de référence). Lorsque *FGFR2* est fortement exprimé (supérieur à la valeur médiane) et que *FGF4* est faiblement exprimé, la survie n'est pas significativement améliorée (RR=0.532 $[0.202, 1.399]$). Cependant, la sur-expression de *FGF4* améliore significativement la survie (RR = 0.329 $[0.112, 0.967]$). Enfin, les patients ayant une tumeur avec une forte expression des deux gènes *FGFR2* et *FGF4* ont le meilleur pronostic (RR= 0.103 $[0.021, 0.516]$).

Nous discutons ensuite l'interaction entre *MCL1* et *BCL2*, deux gènes anti-apoptotiques appartenant à la même famille (famille BCL-2). Initialement considérés comme des oncogènes, l'impact pronostique du couple *BCL2*/*MCL1* pour différents types de cancer reste l'objet de controverses associées à des résultats apparemment contradictoires. Ces derniers peuvent être expliqués par une fonction biologique complexe avec association d'un effet « négatif » sur la prolifération cellulaire et d'un effet « négatif » sur l'apoptose (voir Zinkel *et al.* (2006)pour une revue). L'effet anti-apoptotique est associé à une résistance à la chimiothérapie, permettant d'expliquer un rôle pronostique défavorable lié à la sur-expression de ces gènes dans certaines tumeurs telles que les lymphomes ou certaines tumeurs ovariennes avancées. A l'inverse, l'activité antiproliférative des gènes *MCL1* et *BCL2* pourrait expliquer en partie le rôle pronostique favorable, décrit dans certains carcinomes précoces, tels que les adénocarcinomes broncho-pulmonaires (Martin *et al.*, 2003). Dans notre étude, nous montrons par l'analyse combinée de l'expression des deux gènes *MCL1* et *BCL2* l'existence d'un potentiel effet modulateur *MCL1*/*BCL2*. Comme le montre la figure 6.5(d), dans notre série d'adénocarcinomes broncho-pulmonaires précoces, les patients avec une faible expression (inférieur à la médiane) de *MCL1* et *BCL2* ont le plus mauvais pronostique (groupe de référence). Lorsque *MCL1* est fortement exprimé (supérieur à la médiane) et *BCL2* est faiblement exprimé, le pronostique n'est pas significativement amélioré (RR= $0.533[0.202, 1.4103]$). Au contraire, les patients ayant une tumeur avec une faible expression de *MCL1* et une forte expression de *BCL2* ont une survie significativement meilleure (RR=$0.296[0.091, 0.962]$). Enfin, la sur-expression des deux gènes *MCL1* et *BCL2* est associée

FIGURE 6.4 – Courbe de Kaplan-Meier des groupes définis par les niveaux d'expression de (a) *FGFR2*, (b) *MCL1*, et par les groupes défini par les quatre combinaisons des niveaux d'expression de (c) *FGFR2* et *FGF4*, (d) *MCL1* et *BCL2* dans l'étude Merlion.



au meilleur pronostique (RR= $0.189[0.051, 0.700]$).

Dans ces deux exemples, on peut faire l'hypothèse que l'« effet d'inversion » des risques instantanés de *FGFR2* et *MCL1* observé dans l'analyse marginale est lié à l'effet modulant d'un deuxième gène, respectivement *FGF4* et *BLC2*. Cette hypothèse est concordante avec l'activité biologique connue de ces gènes. Le gène *FGFR2* code pour un récepteur, dont l'activation et l'activité nécessitent la présence de son ligand (i.e. *FGF4*). De même, *BCL2* et *MCL1* codent pour deux protéines de la même famille susceptibles d'agir ensembles sur l'apoptose et la prolifération cellulaire. Les sous-groupes de patient définis par une sur-expression de ces couples de gènes peut se révéler cliniquement important et pourrait faire l'objet d'investigations complémentaires.

6.4 Conclusion sur les exemples

L'utilisation pratique des indices $\mathbf{D}_0^{(\mathrm{PH})}$ et $\mathbf{D}_0^{(\mathrm{NPH})}$ permet l'**identification de marqueurs transcriptomiques** ayant une capacité de discrimination (ou séparabilité) basée sur le délai d'apparition de l'évènement élevée. Dans le cadre du modèle de Cox, l'indice $\mathbf{D}_0^{(\mathrm{PH})}$ permet de sélectionner un sous-ensemble de gènes communs à huit études de cancer de taille différentes qui sont potentiellement des acteurs majeurs de la carcinogenèse. L'utilisation de $\mathbf{D}_0^{(\mathrm{NPH})}$ dans le cadre de la sélection de facteurs génomiques entraînant un croisement des fonctions de risques instantanés (modèle à risques non-proportionnels) permet l'identification de marqueurs dont l'effet est potentiellement modulé par d'autres gènes. Les sous-ensembles et les processus biologiques identifiés par les deux indices sont différents montrant ainsi leur complémentarité.

Chapitre 7

DISCUSSION ET CONCLUSION

L'objectif de cette thèse était le développement d'un **indice de séparabilité** adapté aux **données de génomique** en **analyse de survie**.

Les avancées récentes de la génomique ont induit un nombre croissant de banques de données publiques (par exemple GEO (Barrett *et al.*, 2005), Oncomine (Rhodes *et al.*, 2007), ArrayExpress (Parkinson *et al.*, 2009)).) En oncologie génomique, l'utilisation des données déposées sur ces sites et produites par différents laboratoires de recherche permet la réalisation d'**études dites combinées** (portant sur le même ou sur différents types tumoraux). Les informations obtenues par ces analyses combinées ont pour objectif une meilleure connaissance à la fois des mécanismes communs aux différents types de cancer, mais également des spécificités liés aux différents cancers. Ces données sont cependant **hétérogènes** car produites sur des plateformes génomiques différentes, centrées sur des types de cancers variés et des critères principaux différents. Le nombre d'échantillons de ces études est également extrêmement variable allant de quelques dizaines à plusieurs centaines.

En analyse de survie, les approches classiques d'**identification de facteurs génomiques** potentiellement impliqués dans le cancer consistent à utiliser comme critère de sélection des quantités dérivées de statistiques de test d'hypothèse (liées au risque global d'erreur, Familywise error rate et « adjusted p-value », ou à l'espérance du taux de faux positif, False discovery rate et « q-value ») obtenues dans le cadre d'analyses univariées. Lors d'études combinées, ce type de sélection s'avère problématique car elle dépend fortement de la taille des différentes études considérées. Pour tenter de s'affranchir de ce problème, nous proposons l'utilisation des **mesures de capacité de prédiction** ou **pseudo-$\mathbf{R^2}$**.

Comme nous l'avons montré dans les chapitres 2 et 3, la généralisation de la notion de R^2 (coefficient de détermination), initialement définie dans le cadre du modèle linéaire, au modèle de régression logistique et à l'analyse de survie est délicate. De nombreuses mesures, issues des différentes interprétations possibles du R^2, ont été proposées. En analyse de survie, on distingue les mesures basées sur la notion de somme des écarts, les mesures dérivées de la notion de corrélation, les mesures de concordance et les mesures issues des statistiques de test, comme le

log-rapport de vraisemblance. Dans ce dernier contexte, aucune mesure n'a exploité la relation entre le R^2 et la statistique du score.

Dans ce travail, nous avons proposé un **nouveau pseudo-R^2** permettant de mesurer la capacité de « **séparabilité** » d'une variable sur le temps de survie (chapitre 4). Cet index est lié à la statistique du **score robuste**. Il est compris entre 0 et 1 et peut être interprété en termes de pourcentage de séparabilité. Différentes variantes adaptées aux différents modèles à risques proportionnels (modèle de Cox) et à risques non-proportionnels (modèles à odds proportionnels et à effet d'inversion des risques instantanés) ont été construites. On notera que ce pseudo-R^2 est obtenu **conditionnellement au modèle** et ne doit pas être utilisé comme mesure d'ajustement du modèle, mais comme mesure de quantification de l'effet d'une covariable sur le délai d'apparition d'un événement sous un modèle donné.

Les résultats des **simulations** du chapitre 5 ont montré que la valeur de l'indice $\mathbf{D}_0$ est proche de 0 en l'absence de covariable (séparabilité nulle). Son espérance est en fait égale à l'inverse du nombre d'individus non-censurés, qui tend asymptotiquement vers 0. La valeur de l'indice augmente avec la valeur absolue des coefficients de régression (i.e. la séparabilité augmente). De plus, $\mathbf{D}_0$ est peu affecté par la censure et la taille de l'échantillon, ce qui permet de comparer les résultats d'études de tailles différentes et avec des pourcentages de censure différents. Enfin, il peut être défini pour différents types de modèles, aussi bien à risques proportionnels que non-proportionnels. L'analyse des simulations a montré que cet indice est cependant sensible à la distribution des covariables dans le cadre des modèles considérés (modèle de Cox, à odds proportionnels et à effet d'inversion des risques instantanés). De manière pratique, nous proposons à l'utilisateur d'effectuer une transformation préalable des covariables se ramenant à une distribution unique (typiquement, une transformation log-normale est appliquée aux données génomiques). Le comportement de notre indice a également été comparé aux indices définis à partir de la log-vraisemblance partielle du modèle, i.e. les indices d'Allison et sa version modifiée, de Nagelkerke et de Xu et O'Quigley. Sous le modèle de Cox, notre indice possède une meilleure séparabilité que les indices de la littérature. Sous le modèle à odds proportionnels, notre indice a un comportement très proche de celui d'Allison dans sa version modifiée et de Xu et O'Quigley, et meilleur que les indices d'Allison dans sa version initiale et de Nagelkere. Enfin, notre indice est le seul à pouvoir détecter des variables ayant un effet d'inversion des risques instantanés.

L'**utilisation pratique** de l'indice en oncogénomique a permis l'identification de gènes pronostiques du temps de survie intéressants (chapitre 6). Dans le cadre du modèle de Cox, notre indice a permis de sélectionner un petit ensemble de gènes (*ESPL1*, *KIF4A*, *HJURP*, *LRIG1*) communs à différents types de tumeurs solides et impliqués dans le processus de carcinogenèse. Dans ce travail, nous avons considéré une méthode de sélection assez restrictive, reposant sur la simple intersection des résultats obtenus sur les différentes études. Si nécessaire, des méthodes moins restrictives peuvent être adoptées. Dans le cadre du modèle à risques non-proportionnels, notre indice a permis, dans une étude de cancer broncho-pulmonaire, l'identification de gènes impliqués dans des processus biologiques liés à l'évolution tumorale, n'ayant pas été sélectionnés

sous l'hypothèse de risques proportionnels. Parmi les gènes du cycle cellulaire de notre sélection, deux gènes, *FGFR2* and *MCL1*, ont plus particulièrement été étudiés, leurs effets pouvant être liés à un effet modulateur d'autres gènes de la même voie biologique. Pour cette étude, nous avons considéré uniquement l'effet modulateur d'un gène par rapport à un seul gène. Il s'agit bien évidemment d'une simplification de la réalité, les facteurs génomiques étant impliqués dans des réseaux beaucoup plus complexes, pouvant expliquer le phénomène d'effet d'inversion des risques instantanés. L'application de notre indice sur des données génomiques a donc permis de sélectionner des gènes dont il serait intéressant de valider ou d'invalider le rôle biologique réel sur la survie de manière plus approfondie.

Nos travaux pourraient être prolongés dans plusieurs directions. Un premier problème concerne l'écriture de l'indice en présence d'ex-æquo dans le cadre des modèles à risques non-proportionnels. A notre connaissance, très peu d'auteurs ont étudié ce problème. En règle générale, la solution adoptée est soit une application de la solution proposée par Breslow dans le cadre du modèle de Cox, soit l'ordonnancement des ex-æquo de manière arbitraire par tirage aléatoire. Un second problème concernerait l'utilisation du pseudo-R^2 dans un cadre multivarié. En effet, la valeur du pseudo-R^2 augmente avec le nombre de covariables, même si l'effet pronostique des covariables est nul. Il est donc difficile de comparer des modèles avec un nombre de variables explicatives différentes. Une des solutions serait l'utilisation d'un pseudo-R^2 ajusté (voir p. 96), celui-ci pouvant cependant être négatif. Le cas multivarié est donc problématique et nécessiterait une étude spécifique. Enfin, l'application pratique de notre indice ne se limite pas à l'oncogénomique. Notre indice pourrait être utilisé dans le cadre d'autres pathologies avec données de génomique à haut débit, telle que les maladies auto-immunes ou infectieuses.

En conclusion, nous avons proposé un nouvel indice de séparabilité adapté aux données de génomique qui semble être un outil prometteur, pouvant identifier des gènes ayant des effets biologiques complexes sur la survie, non mis en évidence par les techniques classiques.

ANNEXES

Annexe A

Résultats complémentaires sur l'indice

A.1 Preuve de Lin et Wei montrant que le score et le score robuste sont asymptotiquement équivalents.

Dans ce qui suit, nous reprenons les principales étapes de la démonstration de Lin et Wei (1989) visant à montrer, dans le cadre du modèle de Cox, la propriété suivante (voir p. 90) :

Proposition *$n^{-1/2}U(\beta^*)$ (β^* étant la vraie valeur de β) est asymptotiquement équivalent à $n^{-1/2}W(\beta^*)$, où $W(\beta^*)$ s'écrit comme une somme de n termes indépendants et identiquement distribués (iid).*

Preuve. Pour faciliter la lecture, nous reprenons les notations de Lin et Wei :

$$S^{(0)}(\beta;t) = \frac{1}{n}\sum_{i=1}^{n} Y_i(t)\exp(\beta' Z_i(t))$$

$$S^{(1)}(\beta;t) = \frac{1}{n}\sum_{i=1}^{n} Y_i(t)Z_i(t)\exp(\beta' Z_i(t))$$

$$s^{(r)}(\beta;t) = \mathbb{E}(S^{(r)}(\beta;t)) \text{ pour } r = 0,1$$

On suppose qu'il existe un voisinage $\mathcal{B}$ de β^* tel que pour $\tau < \infty$ et $r = 0, 1$,

$$\sup_{t\in[0,\tau];\beta\in\mathcal{B}} \left\| S^{(r)}(\beta;t) - s^{(r)}(\beta;t) \right\| \xrightarrow{P} 0 \text{ quand } n \to \infty$$

et quand $s^{(1)}(\beta;t)/s^{(0)}(\beta;t)$ est borné sur $\mathcal{B}\times[0;\tau]$. L'expression $\xrightarrow{P}$ signifie « tend en probabilité vers ».

Dans le cadre du modèle de Cox, on a

$$\begin{aligned} U(\beta) &= \sum_{i=1}^{n}\int_0^\infty Z_i(t)dN_i(t) - \int_0^\infty \frac{S^{(1)}(\beta;t)}{S^{(0)}(\beta;t)}d\bar{N}(t) \\ &= \sum_{i=1}^{n}\int_0^\infty Z_i(t)dN_i(t) - \int_0^\infty \frac{S^{(1)}(\beta;t)}{S^{(0)}(\beta;t)}\, n\cdot d\widetilde{F}_n(t) \end{aligned}$$

avec $\widetilde{F}_n(t) = \dfrac{\bar{N}(t)}{n}$.

On peut réécrire le score comme suit

$$\begin{aligned} U(\beta) &= \sum_{i=1}^{n} \int_0^\infty Z_i(t) dN_i(t) - \int_0^\infty \frac{S^{(1)}(\beta;t)}{S^{(0)}(\beta;t)}\, n \left(d\widetilde{F}_n(t) - d\widetilde{F}(t) \right) - \int_0^\infty \frac{S^{(1)}(\beta;t)}{S^{(0)}(\beta;t)}\, n \cdot d\widetilde{F}(t) \\ &+ \int_0^\infty \frac{s^{(1)}(\beta;t)}{s^{(0)}(\beta;t)}\, n \left(d\widetilde{F}_n(t) - d\widetilde{F}(t) \right) - \int_0^\infty \frac{s^{(1)}(\beta;t)}{s^{(0)}(\beta;t)}\, n \left(d\widetilde{F}_n(t) - d\widetilde{F}(t) \right) \end{aligned}$$

avec $\widetilde{F}(t) = \mathbb{E}\left[\widetilde{F}_n(t) \right]$.

D'où

$$\begin{aligned} & n^{-1/2} U(\beta^*) \\ &= n^{-1/2} \sum_{i=1}^{n} \int_0^\infty Z_i(t) dN_i(t) - n^{1/2} \int_0^\infty \frac{s^{(1)}(\beta^*;t)}{s^{(0)}(\beta^*;t)} \left(d\widetilde{F}_n(t) - d\widetilde{F}(t) \right) \\ &- n^{1/2} \int_0^\infty \frac{S^{(1)}(\beta^*;t)}{S^{(0)}(\beta^*;t)} d\widetilde{F}(t) - n^{1/2} \int_0^\infty \left(\frac{S^{(1)}(\beta^*;t)}{S^{(0)}(\beta^*;t)} - \frac{s^{(1)}(\beta^*;t)}{s^{(0)}(\beta^*;t)} \right) \left(d\widetilde{F}_n(t) - d\widetilde{F}(t) \right) \end{aligned} \tag{A.1}$$

Le terme $n^{1/2} \left(\widetilde{F}_n(t) - \widetilde{F}(t) \right)$ converge en distribution vers un processus gaussien de moyenne nulle. Par conséquent, le dernier terme de l'équation (A.1) ci-dessus est $O_p(1)$. Le troisième terme de (A.1) peut être développé de la façon suivante :

$$\begin{aligned} & n^{1/2} \int_0^\infty \frac{S^{(1)}(\beta^*;t)}{S^{(0)}(\beta^*;t)} d\widetilde{F}(t) \\ &= n^{1/2} \int_0^\infty \frac{1}{s^{(0)}(\beta^*;t)} \left[S^{(1)}(\beta^*;t) - \frac{s^{(1)}(\beta^*;t)}{s^{(0)}(\beta^*;t)} \left(S^{(0)}(\beta^*;t) - s^{(0)}(\beta^*;t) \right) \right] d\widetilde{F}(t) \\ &+ n^{1/2} \int_0^\infty \left[\frac{S^{(1)}(\beta^*;t)}{S^{(0)}(\beta^*;t)} - \frac{S^{(1)}(\beta^*;t)}{s^{(0)}(\beta^*;t)} + \frac{s^{(1)}(\beta^*;t) S^{(0)}(\beta^*;t)}{(s^{(0)}(\beta^*;t))^2} - \frac{s^{(1)}(\beta^*;t)}{s^{(0)}(\beta^*;t)} \right] d\widetilde{F}(t) \\ &= n^{1/2} \int_0^\infty \frac{1}{s^{(0)}(\beta^*;t)} \left[S^{(1)}(\beta^*;t) - \frac{s^{(1)}(\beta^*;t)}{s^{(0)}(\beta^*;t)} \left(S^{(0)}(\beta^*;t) - s^{(0)}(\beta^*;t) \right) \right] d\widetilde{F}(t) \\ &+ n^{1/2} \int_0^\infty \left(1 - \frac{S^{(0)}(\beta^*;t)}{s^{(0)}(\beta^*;t)} \right) \left(\frac{S^{(1)}(\beta^*;t)}{S^{(0)}(\beta^*;t)} - \frac{s^{(1)}(\beta^*;t)}{s^{(0)}(\beta^*;t)} \right) d\widetilde{F}(t) \\ &= n^{1/2} \int_0^\infty \frac{1}{s^{(0)}(\beta^*;t)} \left[S^{(1)}(\beta^*;t) - \frac{s^{(1)}(\beta^*;t)}{s^{(0)}(\beta^*;t)} \left(S^{(0)}(\beta^*;t) - s^{(0)}(\beta^*;t) \right) \right] d\widetilde{F}(t) \\ &+ O_p(1) \end{aligned}$$

On regroupe les termes :

$$\begin{aligned} n^{-1/2} U(\beta^*) &= n^{-1/2} \sum_{i=1}^{n} \int_0^\infty Z_i(t) dN_i(t) - n^{1/2} \int_0^\infty \frac{s^{(1)}(\beta^*;t)}{s^{(0)}(\beta^*;t)} \left(d\widetilde{F}_n(t) - d\widetilde{F}(t) \right) \\ &- n^{1/2} \int_0^\infty \left[\frac{S^{(1)}(\beta^*;t)}{s^{(0)}(\beta^*;t)} - \frac{s^{(1)}(\beta^*;t) S^{(0)}(\beta^*;t)}{(s^{(0)}(\beta^*;t))^2} + \frac{s^{(1)}(\beta^*;t)}{s^{(0)}(\beta^*;t)} \right] d\widetilde{F}(t) + O_p(1) \\ &= n^{-1/2} \sum_{i=1}^{n} \int_0^\infty Z_i(t) dN_i(t) - n^{1/2} \int_0^\infty \frac{s^{(1)}(\beta^*;t)}{s^{(0)}(\beta^*;t)} d\widetilde{F}_n(t) \\ &- n^{1/2} \int_0^\infty \frac{1}{s^{(0)}(\beta^*;t)} \left[S^{(1)}(\beta^*;t) - \frac{S^{(0)}(\beta^*;t) s^{(1)}(\beta^*;t)}{s^{(0)}(\beta^*;t)} \right] d\widetilde{F}(t) + O_p(1) \end{aligned}$$

Par conséquent, $n^{-1/2}U(\beta^*)$ est asymptotiquement équivalent à $n^{-1/2}\sum_i W_i(\beta^*)$, où

$$\begin{aligned} W_i(\beta) &= \int_0^\infty \left(Z_i(t) - \frac{s^{(1)}(\beta;t)}{s^{(0)}(\beta;t)}\right) dN_i(t) \\ &\quad - \int_0^\infty \frac{Y_i(t)\exp(\beta' Z_i(t))}{s^{(0)}(\beta;t)} \left(Z_i(t) - \frac{s^{(1)}(\beta;t)}{s^{(0)}(\beta;t)}\right) d\widetilde{F}(t) \\ &= \int_0^\infty \left(Z_i(t) - \frac{s^{(1)}(\beta;t)}{s^{(0)}(\beta;t)}\right) dM_i(t) \end{aligned}$$

Les $W_i(\beta^*)$ sont iid.

Un estimateur de $W_i(\beta)$ est obtenu en remplaçant $s^{(0)}(\beta;t)$, $s^{(1)}(\beta;t)$ et $\widetilde{F}(t)$ par $S^{(0)}(\beta;t)$, $S^{(1)}(\beta;t)$ et $\widetilde{F}_n(t)$, respectivement.

□

A.2 Preuve montrant la relation entre $\mathbf{D_0}$ et les déterminants des matrices Σ et Σ^*

La preuve qui suit vise à démontrer la proposition suivante (voir p. 96) :

Proposition

$$\mathbf{D}_0 = \frac{\det(\boldsymbol{\Sigma}) - \det(\boldsymbol{\Sigma}^*)}{\det \boldsymbol{\Sigma}}$$

avec

$$\underset{(p\times p)}{\boldsymbol{\Sigma}} = \begin{pmatrix} \sum_i \widehat{W}_{i1}^2 & \sum_i \widehat{W}_{i1}\widehat{W}_{i2} & \cdots & \sum_i \widehat{W}_{i1}\widehat{W}_{ip} \\ \sum_i \widehat{W}_{i2}\widehat{W}_{i1} & \sum_i \widehat{W}_{i2}^2 & \cdots & \sum_i \widehat{W}_{i2}\widehat{W}_{ip} \\ \vdots & & & \vdots \\ \sum_i \widehat{W}_{ip}\widehat{W}_{i1} & \cdots & \cdots & \sum_i \widehat{W}_{ip}^2 \end{pmatrix}$$

et

$$\underset{(p\times p)}{\boldsymbol{\Sigma}^*} = \begin{pmatrix} \sum_i \widehat{W}_{i1}^2 - \frac{1}{k}\left(\sum_i \widehat{W}_{i1}\right)^2 & \cdots & \sum_i \widehat{W}_{i1}\widehat{W}_{ip} - \frac{1}{k}\sum_i \widehat{W}_{i1}\sum_i \widehat{W}_{ip} \\ \sum_i \widehat{W}_{i2}\widehat{W}_{i1} - \frac{1}{k}\sum_i \widehat{W}_{i1}\sum_i \widehat{W}_{i2} & \cdots & \sum_i \widehat{W}_{i2}\widehat{W}_{ip} - \frac{1}{k}\sum_i \widehat{W}_{i2}\sum_i \widehat{W}_{ip} \\ \vdots & & \vdots \\ \sum_i \widehat{W}_{ip}\widehat{W}_{i1} - \frac{1}{k}\sum_i \widehat{W}_{i1}\sum_i \widehat{W}_{ip} & \cdots & \sum_i \widehat{W}_{ip}^2 - \frac{1}{k}\left(\sum_i \widehat{W}_{ip}\right)^2 \end{pmatrix}$$

Preuve. Cette preuve comporte de nombreuses similarités avec celle de la Proposition 4.1. Ici encore, le symbole $\widehat{\ }$ a été supprimé pour faciliter la lecture.

(i) Calcul des déterminants de $\boldsymbol{\Sigma}$ et $\boldsymbol{\Sigma}^*$.
a. Comme dans ce qui précède, on considère la partition de $\boldsymbol{\Sigma}$ suivante

$$\boldsymbol{\Sigma} = \begin{pmatrix} s_{11} & s_{12} \\ s_{21} & \mathbf{S_{22}} \end{pmatrix}$$

(dont les éléments sont donnés ci-avant).
Pour calculer le déterminant de $\boldsymbol{\Sigma}$, on utilise la propriété suivante Mardia *et al.* (1979) :

Proposition A.1

$$\begin{vmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{vmatrix} = |A_{11}||A_{22} - A_{21}A_{11}^{-1}A_{12}| = |A_{22}||A_{11} - A_{12}A_{22}^{-1}A_{21}|$$

Cas particulier :

$$\begin{vmatrix} A & x \\ x' & c \end{vmatrix} = |A|(c - x'A^{-1}x)$$

où x est un vecteur colonne et c un scalaire.

Ainsi

$$|\mathbf{\Sigma}| = |\mathbf{S_{22}}|(s_{11} - s_{21}^T \mathbf{S_{22}}^{-1} s_{21})$$

b. La matrice $\mathbf{\Sigma}^*(p \times p)$ peut être partitionnée comme suit $\mathbf{\Sigma}^* = \begin{pmatrix} \sigma_{11} & \sigma_{21}' \\ \sigma_{21} & \mathbf{\Sigma_{22}} \end{pmatrix}$ (voir ci-avant).
En utilisant la propriété A.1, on peut écrire

$$|\mathbf{\Sigma}^*| = |\mathbf{\Sigma_{22}}|(\sigma_{11} - \sigma_{21}^T \mathbf{\Sigma_{22}}^{-1} \sigma_{21})$$

avec

$$\sigma_{11} - \sigma_{21}^T \mathbf{\Sigma_{22}}^{-1} \sigma_{21} = \left(s_{11} - \frac{1}{k} W_1^2\right) - \left(s_{21}^T - \frac{1}{k} W_1 W_{(2)}^T\right) \left(\mathbf{S_{22}} - \frac{1}{k} W_{(2)} W_{(2)}^T\right)^{-1} \left(s_{21} - \frac{1}{k} W_{(2)} W_1\right)$$

L'inverse de $\mathbf{\Sigma_{22}}$ a déjà été calculé précédemment :

$$\mathbf{\Sigma_{22}}^{-1} = \mathbf{S_{22}}^{-1} + E \mathbf{S_{22}}^{-1} W_{(2)} {W_{(2)}}^T \mathbf{S_{22}}^{-1}$$

où $E = \left(k - {W_{(2)}}^T \mathbf{S_{22}}^{-1} W_{(2)}\right)^{-1}$ est un scalaire positif.
On développe :

$$\begin{aligned} \sigma_{11} - \sigma_{21}^T \mathbf{\Sigma_{22}}^{-1} \sigma_{21} = & \left(s_{11} - \frac{1}{k} W_1^2\right) - \left(s_{21}^T \mathbf{S_{22}}^{-1} + E \cdot s_{21}^T \mathbf{S_{22}}^{-1} W_{(2)} W_{(2)}^T \mathbf{S_{22}}^{-1} \right. \\ & \left. - \frac{1}{k} W_1 W_{(2)}^T \mathbf{S_{22}}^{-1} - \frac{1}{k} E \cdot W_1 W_{(2)}^T \mathbf{S_{22}}^{-1} W_{(2)} W_{(2)}^T \mathbf{S_{22}}^{-1}\right) \times \left(s_{21} - \frac{1}{k} W_{(2)} W_1\right) \\ = & \; s_{11} - \frac{1}{k} W_1^2 - s_{21}^T \mathbf{S_{22}}^{-1} s_{21} - E \cdot s_{21}^T \mathbf{S_{22}}^{-1} W_{(2)} W_{(2)}^T \mathbf{S_{22}}^{-1} s_{21} \\ & + \frac{1}{k} W_1 W_{(2)}^T \mathbf{S_{22}}^{-1} s_{21} + \frac{1}{k} E W_1 \cdot W_{(2)}^T \mathbf{S_{22}}^{-1} W_{(2)} W_{(2)}^T \mathbf{S_{22}}^{-1} s_{21} \\ & + \frac{1}{k} W_1 s_{21}^T \mathbf{S_{22}}^{-1} W_{(2)} + \frac{1}{k} E W_1 \cdot s_{21}^T \mathbf{S_{22}}^{-1} W_{(2)} W_{(2)}^T \mathbf{S_{22}}^{-1} W_{(2)} \\ & - \frac{1}{k^2} W_1^2 W_{(2)}^T \mathbf{S_{22}}^{-1} W_{(2)} - \frac{1}{k^2} E W_1^2 \cdot W_{(2)}^T \mathbf{S_{22}}^{-1} W_{(2)} W_{(2)}^T \mathbf{S_{22}}^{-1} W_{(2)} \end{aligned}$$

Après avoir mis E en facteur, développé et simplifié, on obtient :

$$\begin{aligned} \sigma_{11} - \sigma_{21}^T \mathbf{\Sigma_{22}}^{-1} \sigma_{21} = & \; E \cdot \Big[k s_{11} - s_{11} W_{(2)}^T \mathbf{S_{22}}^{-1} W_{(2)} - W_1^2 - k s_{21}^T \mathbf{S_{22}}^{-1} s_{21} \\ & + s_{21}^T \mathbf{S_{22}}^{-1} s_{21} W_{(2)}^T \mathbf{S_{22}}^{-1} W_{(2)} - s_{21}^T \mathbf{S_{22}}^{-1} W_{(2)} W_{(2)}^T \mathbf{S_{22}}^{-1} s_{21} \\ & + W_1 W_{(2)}^T \mathbf{S_{22}}^{-1} s_{21} + W_1 s_{21}^T \mathbf{S_{22}}^{-1} W_{(2)} \Big] \end{aligned}$$

Pour le calcul de $|\mathbf{\Sigma_{22}}| = |\mathbf{S_{22}} - \frac{1}{k} W_{(2)}^T W_{(2)}|$, on considère la propriété suivante Mardia *et al.* (1979) :

Proposition A.2 *Soit les matrice $B(p \times n)$, $B(n \times p)$ et $A(p \times p)$. Si l'inverse de A existe, alors*

$$|A + BC| = |A|^{-1}|I_p + A^{-1}BC| = |A^{-1}||I_n + CA^{-1}B|$$

où I_n fait référence à la matrice identité de dimension $(n \times n)$.

Cas particulier :

$$|A + bb'| = |A|(1 + b'A^{-1}b)$$

où b est un vecteur colonne.

On a donc

$$|\mathbf{\Sigma_{22}}| = |\mathbf{S_{22}}| \left(1 - \frac{1}{k} W_{(2)}^T \mathbf{S_{22}}^{-1} W_{(2)}\right)$$

or

$$E = \left(k - W_{(2)}^T \mathbf{S_{22}}^{-1} W_{(2)}\right)^{-1} = \frac{1}{k\left(1 - \frac{1}{k} W_{(2)}^T \mathbf{S_{22}}^{-1} W_{(2)}\right)}$$

d'où

$$\begin{aligned} |\mathbf{\Sigma}^*| = & \ |\mathbf{\Sigma_{22}}| \left(\sigma_{11} - \sigma_{21}^T \mathbf{\Sigma_{22}}^{-1} \sigma_{21}\right) \\ = & \ |\mathbf{S_{22}}| \Bigg(s_{11} - \frac{1}{k} s_{11} W_{(2)}^T \mathbf{S_{22}}^{-1} W_{(2)} - \frac{1}{k} W_1^2 - s_{21}^T \mathbf{S_{22}}^{-1} s_{21} + \frac{1}{k} s_{21}^T \mathbf{S_{22}}^{-1} s_{21} W_{(2)}^T \mathbf{S_{22}}^{-1} W_{(2)} \\ & - \frac{1}{k} s_{21}^T \mathbf{S_{22}}^{-1} W_{(2)} W_{(2)}^T \mathbf{S_{22}}^{-1} s_{21} + \frac{1}{k} W_1 W_{(2)}^T \mathbf{S_{22}}^{-1} s_{21} + \frac{1}{k} W_1 s_{21}^T \mathbf{S_{22}}^{-1} W_{(2)} \Bigg) \end{aligned}$$

c. Par conséquent, on a

$$\begin{aligned} \frac{|\mathbf{\Sigma}| - |\mathbf{\Sigma}^*|}{|\mathbf{\Sigma}|} = & \ \frac{1}{s_{11} - s_{21}^T \mathbf{S_{22}}^{-1} s_{21}} \Bigg(\frac{1}{k} s_{11} W_{(2)}^T \mathbf{S_{22}}^{-1} W_{(2)} + \frac{1}{k} W_1^2 - \frac{1}{k} s_{21}^T \mathbf{S_{22}}^{-1} s_{21} W_{(2)}^T \mathbf{S_{22}}^{-1} W_{(2)} \\ & + \frac{1}{k} s_{21}^T \mathbf{S_{22}}^{-1} W_{(2)} W_{(2)}^T \mathbf{S_{22}}^{-1} s_{21} - \frac{1}{k} W_1 W_{(2)}^T \mathbf{S_{22}}^{-1} s_{21} - \frac{1}{k} W_1 s_{21}^T \mathbf{S_{22}}^{-1} W_{(2)} \Bigg) \end{aligned}$$

(ii) Calcul de $\mathbf{D}_0$.

On a $\mathbf{D}_0 = \dfrac{W^T \mathbf{\Sigma}^{-1} W}{k}$ avec

$$\mathbf{\Sigma}^{-1} = \begin{pmatrix} (s_{11} - {s_{21}}^T \mathbf{S_{22}}^{-1} s_{21})^{-1} & -D {s_{21}}^T \mathbf{S_{22}}^{-1} \\ -\mathbf{S_{22}}^{-1} s_{21} D & (\mathbf{S_{22}} - s_{21} s_{11}^{-1} {s_{21}}^T)^{-1} \end{pmatrix}$$

et $D = (s_{11} - {s_{21}}^T \mathbf{S_{22}}^{-1} s_{21})^{-1}$, D est un scalaire positif.
En développant $W^T \mathbf{\Sigma}^{-1} W$, on obtient

$$W^T \mathbf{\Sigma}^{-1} W = W_1^2 D - W_1 D \cdot {W_{(2)}}^T \mathbf{S_{22}}^{-1} s_{21} - W_1 D \cdot {s_{21}}^T \mathbf{S_{22}}^{-1} W_{(2)} + {W_{(2)}}^T (\mathbf{S_{22}} - s_{21} s_{11}^{-1} {s_{21}}^T)^{-1} W_{(2)}$$

La matrice $B = (\mathbf{S_{22}} - s_{21} s_{11}^{-1} {s_{21}}^T)^{-1}$ peut Ătre développée comme ci-avant :

$$B = \mathbf{S_{22}}^{-1} + \mathbf{S_{22}}^{-1} s_{21} (s_{11} - {s_{21}}^T \mathbf{S_{22}}^{-1} s_{21})^{-1} {s_{21}}^T \mathbf{S_{22}}^{-1} = \mathbf{S_{22}}^{-1} + D \cdot \mathbf{S_{22}}^{-1} s_{21} {s_{21}}^T \mathbf{S_{22}}^{-1}$$

de telle sorte que

$$W^T\mathbf{\Sigma}^{-1}W = W_1^2 D - W_1 D \cdot {W_{(2)}}^T {\mathbf{S_{22}}}^{-1} s_{21} - W_1 D \cdot {s_{21}}^T {\mathbf{S_{22}}}^{-1} W_{(2)} + {W_{(2)}}^T {\mathbf{S_{22}}}^{-1} W_{(2)} + D \cdot {W_{(2)}}^T {\mathbf{S_{22}}}^{-1} s_{21} \cdot {s_{21}}^T {\mathbf{S_{22}}}^{-1} W_{(2)}$$

On a alors

$$\mathbf{D}_0 = \frac{W^T\mathbf{\Sigma}^{-1}W}{k} = \frac{D}{k}\Big[W_1^2 - W_1 W_{(2)}^T {\mathbf{S_{22}}}^{-1} s_{21} - W_1 s_{21}^T {\mathbf{S_{22}}}^{-1} W_{(2)} + s_{11} W_{(2)}^T {\mathbf{S_{22}}}^{-1} W_{(2)} - W_{(2)}^T {\mathbf{S_{22}}}^{-1} W_{(2)} s_{21}^T {\mathbf{S_{22}}}^{-1} s_{21} + W_{(2)}^T {\mathbf{S_{22}}}^{-1} s_{21} s_{21}^T {\mathbf{S_{22}}}^{-1} W_{(2)}\Big]$$

En conclusion,

$$\boxed{\mathbf{D}_0 = \frac{W^T\mathbf{\Sigma}^{-1}W}{k} = \frac{\det(\mathbf{\Sigma}) - \det(\mathbf{\Sigma}^*)}{\det(\mathbf{\Sigma})}}$$

□

La proposition 4.4 montre que $\mathbf{D}_0$ s'écrit comme la **différence relative entre un estimateur de la matrice de variance-covariance des W_i calculée sous l'hypothèse où les W_i sont d'espérance nulle, $\mathbf{\Sigma}$, et un estimateur de la matrice de variance-covariance des W_i calculée sous l'hypothèse où l'espérance des W_i n'est pas nulle, $\mathbf{\Sigma}^*$**. Le déterminant de la matrice de variance-covariance est parfois appelé « **variance généralisée** » car il mesure le degré de dispersion autour de la valeur moyenne $\bar{W}$.

Annexe B

Résultats complets des simulations

B.1 Calcul des paramètres des différents mécanismes de censure.

Soit C la variable de censure, de distribution soit uniforme $C \sim \mathcal{U}\{0, r\}$, soit exponentielle $C \sim \mathcal{E}(\gamma)$. On suppose que le mécanisme de censure est indépendant du temps de survie réel X sachant Z.

L'objectif est de déterminer les valeurs de r et γ pour que l'espérance de la censure soit égale à un pourcentage donné p_c (dans nos simulations 25 ou 50%). La méthodologie utilisée ici est celle de l'article de Lininger *et al.* (1979).

On note respectivement f_C, f_X et f_Z les densités de la variable de censure C, de la survie réelle X et de la variable explicative Z, et $\mathcal{C}$ et $\mathcal{Z}$ les domaines de définition de C et Z. Pour une censure uniforme, la densité de C est

$$f_C(c) = \frac{1}{r}\ \mathbf{1}_{[0,r]}(c)$$

Pour une censure exponentielle, la densité de C est

$$f_C(c) = \gamma e^{-\gamma c}$$

- Sous l'hypothèse nulle $\mathcal{H}_0 : \{\beta = 0\}$, la probabilité de censure s'écrit

$$\Pr(X > C) = \int_{\mathcal{C}} \int_c^{\infty} f_{X,C}(x; c)\ dx\ dc = \int_{\mathcal{C}} \int_c^{\infty} f_X(x) f_C(c)\ dx\ dc$$

où $f_{X,C}$, densité jointe de (X, C), est égale au produit des densités marginales de X et C puisque les deux variables sont indépendantes.

- Sous l'hypothèse alternative, la probabilité de censure vaut

$$\begin{aligned} \Pr(X > C; Z) &= \Pr(T > C|Z) \Pr(Z) \\ &= \int_{\mathcal{Z}} \int_{\mathcal{C}} \int_c^{\infty} f_X(x) f_C(c) f_Z(z)\ dx\ dc\ dz \end{aligned}$$

puisque C est indépendant de X sachant Z.

• Pour un modèle de Cox à risques proportionnels, la fonction de survie est générée avec une distribution exponentielle de paramètre $e^{\beta Z}$. Pour déterminer les valeurs de r et γ pour un pourcentage de censure espéré p_c fixé, nous avons résolu, de manière itérative, les équations suivantes dont r_p et γ_p sont les inconnues (qui sont, respectivement, les valeurs de r et γ pour un pourcentage de censure égal à p_c).

- Sous l'hypothèse nulle, les équations à résoudre sont

$$p_c = \frac{1}{r_p}\left(1 - e^{-r_p}\right) \text{ , pour une censure uniforme}$$

et

$$p_c = \frac{\gamma_p}{1+\gamma_p} \text{ , pour une censure exponentielle}$$

- Sous l'hypothèse alternative (i.e. en présence des covariables, $\beta \neq 0$), les équations à résoudre sont

$$p_c = \int_{\mathcal{Z}} \left\{ \frac{1}{r_p \exp(\beta z)} \left(1 - e^{-r_p \exp(\beta z)}\right) f_Z(z) \right\} dz \text{ , pour une censure uniforme}$$

et

$$p_c = \int_{\mathcal{Z}} \left\{ \frac{\gamma_p}{\gamma_p + \exp(\beta z)} f_Z(z) \right\} dz \text{ , pour une censure exponentielle.}$$

• Pour le modèle à odds proportionnels, la fonction de survie est générée avec une distribution log-logistique de densité $f(t) = e^{\beta z}/(1 + te^{\beta z})^2$.

- Dans ce cas, les équations à résoudre sous l'hypothèse nulle sont

$$p_c = \frac{\ln(1+r_p)}{r_p} \text{ , pour une censure uniforme}$$

et

$$p_c = \int_0^\infty \frac{\gamma_p e^{-\gamma_p c}}{1+c} dc \text{ , pour une censure exponentielle}$$

- Sous l'hypothèse alternative, les équations sont

$$p_c = \int_{\mathcal{Z}} \left\{ \frac{\ln(1 + r_p e^{\beta z})}{r_p} f_Z(z) \right\} dz \text{ , pour une censure uniforme}$$

et

$$p_c = \int_{\mathcal{Z}} \left\{ \int_0^\infty \left(\frac{\gamma_p e^{-\gamma_p c}}{1 + c e^{\beta z}} \right) dc \right\} f_Z(z) dz \text{ , pour une censure exponentielle.}$$

• Pour le modèle à risques non-proportionnels conduisant à un croisement des risques instantanés, la fonction de survie est générée avec une distribution de Weibull de densité $f(t) = e^{\beta Z} t^{(e^{\beta Z}-1)} e^{-te^{\beta Z}}$.

- Les équations à résoudre sous l'hypothèse nulle sont alors

$$p_c = \frac{1}{r_p}\left(1 - e^{-r_p}\right) \text{ , pour une censure uniforme}$$

et

$$p_c = \frac{\gamma_p}{1+\gamma_p} \text{ , pour une censure exponentielle}$$

- Sous l'hypothèse alternative, les équations sont

$$p_c = \frac{1}{r_p} \int_{\mathcal{Z}} \left\{ \int_0^{r_p} e^{-ce^{\beta Z}} \, dc \right\} f_Z(z) \, dz \text{ , pour une censure uniforme}$$

et

$$p_c = \int_{\mathcal{Z}} \left\{ \int_0^{\infty} e^{-ce^{\beta Z}} \gamma_p e^{-\gamma_p c} \, dc \right\} f_Z(z) \, dz \text{ , pour une censure exponentielle.}$$

• Pour les trois modèles, pour un individu i, nous avons donc généré deux variables de distribution soit uniforme $U^i_{0.25}$ et $U^i_{0.5}$, définies respectivement sur $\{0, r^i_{0.25}\}$ et $\{0, r^i_{0.5}\}$, soit exponentielle $\mathcal{E}^i_{0.25}$ et $\mathcal{E}^i_{0.5}$ de paramètres respectifs $\gamma^i_{0.25}$ et $\gamma^i_{0.5}$.
Les variables X^i, $\min(X^i, C^i_{0.25})$ et $\min(X^i, C^i_{0.5})$, représentent donc les temps de survie observés pour un individu i lorsque le pourcentage de censure espéré est respectivement égal à 0, 25 et 50%.

B.2 Tableaux et figures complémentaires

TABLEAU B.1 – Valeurs moyennes de $\mathbf{D}_0^{(\mathrm{PH})}$ sous un modèle à risques proportionnels, pour différents risques relatifs e^β, différents pourcentages de censure p_c, différentes tailles d'échantillon n, différents types de censure, calculées pour une variable de Bernoulli $Z \sim \mathcal{B}(1/2)$ (1000 répétitions). Les écarts-types sont indiqués entre parenthèses.

		$C \sim \mathcal{U}[0,r]$				$C \sim \mathcal{E}(\gamma)$			
β	p_c	$\mathbf{D}_0^{(\mathrm{PH})}$ ($n=50$)	$\mathbf{D}_0^{(\mathrm{PH})}$ ($n=100$)	$\mathbf{D}_0^{(\mathrm{PH})}$ ($n=500$)	$\mathbf{D}_0^{(\mathrm{PH})}$ ($n=1000$)	$\mathbf{D}_0^{(\mathrm{PH})}$ ($n=50$)	$\mathbf{D}_0^{(\mathrm{PH})}$ ($n=100$)	$\mathbf{D}_0^{(\mathrm{PH})}$ ($n=500$)	$\mathbf{D}_0^{(\mathrm{PH})}$ ($n=1000$)
1	0	0.0257(0.0431)	0.0114(0.0180)	0.0021(0.0031)	0.0010(0.0015)	0.0234(0.0368)	0.0113(0.0178)	0.0020(0.0028)	0.0010(0.0014)
	0.25	0.0314(0.0456)	0.0144(0.0211)	0.0026(0.0039)	0.0014(0.0020)	0.0308(0.0455)	0.0148(0.0226)	0.0028(0.0040)	0.0013(0.0019)
	0.50	0.0431(0.0620)	0.0199(0.0262)	0.0041(0.0057)	0.0020(0.0027)	0.0410(0.0582)	0.0210(0.0303)	0.0040(0.0056)	0.0021(0.0031)
1.25	0	0.0394(0.0550)	0.0236(0.0313)	0.0151(0.0111)	0.0142(0.0078)	0.0362(0.0537)	0.0262(0.0336)	0.0153(0.0116)	0.0137(0.0076)
	0.25	0.0474(0.0688)	0.0265(0.0342)	0.0154(0.0136)	0.0139(0.0083)	0.0447(0.0601)	0.0278(0.0368)	0.0152(0.0122)	0.0142(0.0088)
	0.50	0.0556(0.0760)	0.0359(0.0465)	0.0156(0.0148)	0.0138(0.0101)	0.0571(0.0742)	0.0343(0.0451)	0.0169(0.0157)	0.0147(0.0109)
1.5	0	0.0696(0.0758)	0.0583(0.0557)	0.0474(0.0225)	0.0468(0.0149)	0.0674(0.0805)	0.0585(0.0527)	0.0477(0.0226)	0.0464(0.0152)
	0.25	0.0772(0.0926)	0.0604(0.0571)	0.0443(0.0221)	0.0438(0.0163)	0.0758(0.0870)	0.0582(0.0584)	0.0454(0.0236)	0.0437(0.0161)
	0.50	0.0865(0.1084)	0.0609(0.0641)	0.0441(0.0253)	0.0425(0.0179)	0.0871(0.1051)	0.0619(0.0683)	0.0447(0.0272)	0.0431(0.0182)
1.75	0	0.1169(0.1072)	0.0992(0.0694)	0.0919(0.0324)	0.0903(0.0224)	0.117(0.1092)	0.1014(0.0751)	0.0918(0.0328)	0.0904(0.0223)
	0.25	0.1177(0.1159)	0.0996(0.0772)	0.0867(0.0331)	0.0849(0.0236)	0.1142(0.1114)	0.1019(0.0795)	0.0869(0.0337)	0.0860(0.0244)
	0.50	0.1268(0.1276)	0.1011(0.0892)	0.0817(0.0362)	0.0802(0.0247)	0.1162(0.1274)	0.1008(0.0912)	0.0809(0.0367)	0.0803(0.0258)
2	0	0.1612(0.1229)	0.1536(0.0854)	0.1414(0.0408)	0.1425(0.0280)	0.1604(0.1187)	0.1493(0.0843)	0.1414(0.0412)	0.1435(0.0279)
	0.25	0.1635(0.1400)	0.1467(0.0939)	0.1320(0.0433)	0.1320(0.0296)	0.1542(0.1267)	0.1428(0.0924)	0.1333(0.0405)	0.1327(0.0299)
	0.50	0.1639(0.1500)	0.1383(0.0977)	0.1235(0.0454)	0.1203(0.0306)	0.1657(0.1548)	0.1331(0.1017)	0.1252(0.0458)	0.1214(0.0314)
3	0	0.3338(0.1563)	0.3427(0.1165)	0.3377(0.0526)	0.3396(0.0365)	0.3323(0.1525)	0.3364(0.1131)	0.3372(0.0510)	0.3398(0.0356)
	0.25	0.3390(0.1903)	0.3227(0.1296)	0.3294(0.0612)	0.3261(0.0457)	0.3309(0.1762)	0.3219(0.1341)	0.3197(0.0613)	0.3231(0.0420)
	0.50	0.3232(0.2080)	0.3052(0.1394)	0.2866(0.0659)	0.2858(0.0460)	0.3059(0.2066)	0.3008(0.1466)	0.2974(0.0686)	0.2931(0.0472)
4	0	0.4505(0.1427)	0.4653(0.1085)	0.4766(0.0476)	0.4796(0.0363)	0.4537(0.1416)	0.4656(0.1098)	0.4772(0.0473)	0.4785(0.0348)
	0.25	0.4721(0.1847)	0.4767(0.1460)	0.4915(0.0660)	0.4868(0.0441)	0.4643(0.1917)	0.4726(0.1396)	0.4742(0.0628)	0.4742(0.0445)
	0.50	0.4654(0.2380)	0.4349(0.1711)	0.4366(0.0742)	0.4308(0.0513)	0.4581(0.2321)	0.4391(0.1602)	0.4395(0.0775)	0.4393(0.0524)
5	0	0.5327(0.1439)	0.5495(0.0987)	0.5656(0.0433)	0.5665(0.0302)	0.5389(0.1441)	0.5481(0.0962)	0.5641(0.0432)	0.5675(0.0299)
	0.25	0.5832(0.1930)	0.5927(0.1430)	0.6057(0.0627)	0.6094(0.0458)	0.5661(0.1876)	0.5782(0.1330)	0.5853(0.0605)	0.5847(0.0439)
	0.50	0.5702(0.2363)	0.5610(0.1762)	0.5578(0.0793)	0.5514(0.0546)	0.5508(0.2280)	0.5609(0.1716)	0.5554(0.0765)	0.5569(0.0543)

TABLEAU B.2 – Valeurs moyennes de $\mathbf{D}_0^{(\mathrm{PH})}$ sous un modèle à risques proportionnels, pour différents risques relatifs e^β, différents pourcentages de censure p_c, différentes tailles d'échantillon n, différents types de censure, calculées pour une variable uniforme $Z \sim \mathcal{U}[0, \sqrt{(3)}]$ (1000 répétitions). Les écarts-types sont indiqués entre parenthèses.

		$C \sim \mathcal{U}[0,r]$				$C \sim \mathcal{E}(\gamma)$			
β	p_c	$\mathbf{D}_0^{(\mathrm{PH})}$ ($n=50$)	$\mathbf{D}_0^{(\mathrm{PH})}$ ($n=100$)	$\mathbf{D}_0^{(\mathrm{PH})}$ ($n=500$)	$\mathbf{D}_0^{(\mathrm{PH})}$ ($n=1000$)	$^{(\mathrm{PH})}$ ($n=50$)	$\mathbf{D}_0^{(\mathrm{PH})}$ ($n=100$)	$\mathbf{D}_0^{(\mathrm{PH})}$ ($n=500$)	$\mathbf{D}_0^{(\mathrm{PH})}$ ($n=1000$)
1	0	0.0243(0.0349)	0.0115(0.0159)	0.0021(0.0032)	0.0011(0.0015)	0.0196(0.0283)	0.0115(0.0169)	0.0022(0.0033)	0.0011(0.0015)
	0.25	0.0294(0.0439)	0.0135(0.0199)	0.0028(0.0040)	0.0013(0.0019)	0.0291(0.0433)	0.0141(0.0203)	0.0027(0.0039)	0.0014(0.0018)
	0.50	0.0440(0.0591)	0.0212(0.0308)	0.0040(0.0061)	0.0021(0.0027)	0.0474(0.0703)	0.0223(0.0321)	0.0043(0.0065)	0.0022(0.0029)
1.25	0	0.0368(0.0457)	0.0241(0.0276)	0.0154(0.0114)	0.0140(0.0073)	0.0363(0.0473)	0.0252(0.0300)	0.0149(0.0112)	0.0135(0.0073)
	0.25	0.0427(0.0576)	0.0277(0.0334)	0.0152(0.0124)	0.0137(0.0079)	0.0430(0.0584)	0.0271(0.0359)	0.0151(0.0123)	0.0140(0.0085)
	0.50	0.0562(0.0752)	0.0339(0.0435)	0.0172(0.0158)	0.0137(0.0099)	0.0570(0.0778)	0.0327(0.0445)	0.0152(0.0145)	0.0142(0.0102)
1.5	0	0.0659(0.0666)	0.0558(0.0472)	0.0442(0.0195)	0.0435(0.0138)	0.0659(0.0691)	0.0515(0.0448)	0.0452(0.0198)	0.0431(0.0135)
	0.25	0.0702(0.0767)	0.0533(0.0511)	0.0438(0.0215)	0.0418(0.0153)	0.0723(0.0829)	0.0569(0.0528)	0.0437(0.0215)	0.0426(0.0151)
	0.50	0.0831(0.0948)	0.0605(0.0618)	0.0444(0.0254)	0.0416(0.0173)	0.0846(0.1015)	0.0635(0.0649)	0.0450(0.0270)	0.0426(0.0180)
1.75	0	0.1033(0.0842)	0.0897(0.0598)	0.0808(0.0263)	0.0819(0.0190)	0.0996(0.0827)	0.0902(0.0592)	0.0819(0.0270)	0.0806(0.0195)
	0.25	0.1065(0.0983)	0.0921(0.0684)	0.0802(0.0307)	0.0794(0.0209)	0.1051(0.1021)	0.0896(0.0676)	0.0815(0.0295)	0.0800(0.0211)
	0.50	0.1163(0.1153)	0.1001(0.0865)	0.0804(0.0348)	0.0788(0.0245)	0.1151(0.1217)	0.1021(0.0862)	0.0790(0.0353)	0.0776(0.0244)
2	0	0.1344(0.0937)	0.1267(0.0669)	0.1222(0.0314)	0.1215(0.0227)	0.1388(0.1005)	0.1298(0.0717)	0.1207(0.0334)	0.1202(0.0227)
	0.25	0.1475(0.1242)	0.1310(0.0796)	0.1209(0.0356)	0.1202(0.0267)	0.1401(0.1123)	0.1324(0.0846)	0.1185(0.0354)	0.1182(0.0260)
	0.50	0.1546(0.1368)	0.1332(0.0962)	0.1193(0.0415)	0.1161(0.0291)	0.1511(0.1369)	0.1302(0.0922)	0.1183(0.0415)	0.1151(0.0285)
3	0	0.2712(0.1278)	0.2604(0.0884)	0.2624(0.0406)	0.2618(0.0287)	0.2623(0.1216)	0.2648(0.0857)	0.2611(0.0395)	0.2614(0.0288)
	0.25	0.2798(0.1485)	0.2718(0.1068)	0.2697(0.0486)	0.2700(0.0354)	0.2678(0.1408)	0.2661(0.0993)	0.2661(0.0452)	0.2653(0.0348)
	0.50	0.2922(0.1883)	0.2752(0.1318)	0.2626(0.0565)	0.2646(0.0394)	0.2782(0.1644)	0.2746(0.1286)	0.2630(0.0559)	0.2637(0.0412)
4	0	0.3526(0.1226)	0.3544(0.0910)	0.3585(0.0416)	0.3597(0.0282)	0.3511(0.1238)	0.3511(0.0890)	0.3607(0.0401)	0.3603(0.0289)
	0.25	0.3863(0.1637)	0.3795(0.1147)	0.3803(0.0508)	0.3835(0.0367)	0.3729(0.1542)	0.3744(0.1141)	0.3750(0.0509)	0.3753(0.0365)
	0.50	0.4072(0.2012)	0.3900(0.1391)	0.3838(0.0633)	0.3810(0.0466)	0.4035(0.1956)	0.3748(0.1370)	0.3747(0.0624)	0.3798(0.0442)
5	0	0.4169(0.1214)	0.4201(0.0854)	0.4274(0.0382)	0.4259(0.0284)	0.4072(0.1180)	0.4210(0.0864)	0.4258(0.0394)	0.4270(0.0281)
	0.25	0.4673(0.1674)	0.4604(0.1156)	0.4643(0.0500)	0.4640(0.0371)	0.4566(0.1538)	0.4572(0.1151)	0.4563(0.0514)	0.4533(0.0368)
	0.50	0.4928(0.1925)	0.4726(0.1451)	0.4721(0.0649)	0.4714(0.0470)	0.4758(0.1889)	0.4676(0.1434)	0.4608(0.0647)	0.4615(0.0446)

TABLEAU B.3 – Valeurs moyennes de $\mathbf{D}_0^{(\mathrm{PH})}$ sous un modèle à risques proportionnels, pour différents risques relatifs e^{β}, différents pourcentages de censure p_c, différentes tailles d'échantillon n, différents types de censure, calculées pour une variable normale $Z \sim \mathcal{N}(0, 1/4)$ (1000 répétitions). Les écarts-types sont indiqués entre parenthèses.

β	p_c	$C \sim \mathcal{U}[0,r]$				$C \sim \mathcal{E}(\gamma)$			
		$\mathbf{D}_0^{(\mathrm{PH})}$ ($n=50$)	$\mathbf{D}_0^{(\mathrm{PH})}$ ($n=100$)	$\mathbf{D}_0^{(\mathrm{PH})}$ ($n=500$)	$\mathbf{D}_0^{(\mathrm{PH})}$ ($n=1000$)	$\mathbf{D}_0^{(\mathrm{PH})}$ ($n=50$)	$\mathbf{D}_0^{(\mathrm{PH})}$ ($n=100$)	$\mathbf{D}_0^{(\mathrm{PH})}$ ($n=500$)	$\mathbf{D}_0^{(\mathrm{PH})}$ ($n=1000$)
1	0	0.0235(0.0335)	0.0106(0.0149)	0.0021(0.0031)	0.0011(0.0015)	0.0249(0.0340)	0.0116(0.0163)	0.0021(0.0030)	0.0010(0.0014)
	0.25	0.0294(0.0406)	0.0142(0.0189)	0.0025(0.0035)	0.0012(0.0018)	0.0312(0.0414)	0.0136(0.0200)	0.0027(0.0037)	0.0014(0.0020)
	0.50	0.0428(0.0582)	0.0203(0.0273)	0.0041(0.0057)	0.0020(0.0029)	0.0449(0.0610)	0.0226(0.0304)	0.0040(0.0058)	0.0020(0.0030)
1.25	0	0.034(0.0429)	0.0234(0.0284)	0.0138(0.0099)	0.0138(0.0071)	0.0345(0.0445)	0.023(0.0261)	0.0146(0.0108)	0.0135(0.0074)
	0.25	0.0440(0.0551)	0.0262(0.0325)	0.0148(0.0115)	0.0134(0.0081)	0.0427(0.0547)	0.0262(0.0314)	0.0147(0.0117)	0.0132(0.0081)
	0.50	0.0544(0.0711)	0.0308(0.0406)	0.0159(0.0152)	0.0139(0.0098)	0.0567(0.0751)	0.0315(0.0396)	0.0167(0.0159)	0.0143(0.0099)
1.5	0	0.0569(0.0576)	0.0478(0.0394)	0.0395(0.0170)	0.0388(0.0122)	0.0599(0.0604)	0.0476(0.0387)	0.0405(0.0174)	0.0393(0.0125)
	0.25	0.0706(0.0753)	0.0555(0.0517)	0.0417(0.0212)	0.0409(0.0140)	0.0664(0.0732)	0.0528(0.0488)	0.0417(0.0205)	0.0399(0.0140)
	0.50	0.0772(0.0893)	0.0598(0.0598)	0.0430(0.0249)	0.0402(0.0173)	0.0858(0.0990)	0.0589(0.0600)	0.0433(0.0249)	0.0411(0.0176)
1.75	0	0.09(0.0749)	0.0778(0.0526)	0.0706(0.0223)	0.0683(0.0159)	0.0894(0.077)	0.0777(0.0498)	0.0693(0.0209)	0.0695(0.0156)
	0.25	0.0976(0.0879)	0.0837(0.0599)	0.0748(0.0258)	0.0739(0.0184)	0.0953(0.0888)	0.0833(0.0614)	0.0721(0.0262)	0.0721(0.0182)
	0.50	0.1067(0.1093)	0.0917(0.0809)	0.0761(0.0324)	0.0749(0.0226)	0.1077(0.1088)	0.0877(0.0728)	0.0743(0.0321)	0.0740(0.0226)
2	0	0.1135(0.0816)	0.1104(0.0577)	0.1003(0.0251)	0.0992(0.0181)	0.1212(0.0845)	0.1105(0.0590)	0.0989(0.0244)	0.0985(0.0179)
	0.25	0.1248(0.0981)	0.1172(0.0705)	0.1061(0.0301)	0.1063(0.0215)	0.1289(0.1055)	0.1178(0.0686)	0.1060(0.0308)	0.1061(0.0217)
	0.50	0.1368(0.1236)	0.1226(0.0841)	0.1135(0.0387)	0.1094(0.0271)	0.1495(0.1300)	0.1253(0.0882)	0.1097(0.0382)	0.1084(0.0265)
3	0	0.2090(0.0969)	0.2078(0.0687)	0.1963(0.0329)	0.1930(0.0236)	0.2157(0.0963)	0.2070(0.0675)	0.1950(0.0315)	0.1917(0.0228)
	0.25	0.2357(0.1270)	0.2324(0.0916)	0.2213(0.0396)	0.2181(0.0281)	0.2350(0.1250)	0.2240(0.0860)	0.2141(0.0389)	0.2163(0.0279)
	0.50	0.2619(0.1578)	0.2412(0.1100)	0.2349(0.0483)	0.2360(0.0360)	0.2498(0.1561)	0.2486(0.1100)	0.2324(0.0499)	0.2313(0.0344)
4	0	0.2803(0.1035)	0.2698(0.0763)	0.2599(0.0356)	0.2572(0.0251)	0.2710(0.1015)	0.2702(0.0732)	0.2576(0.0343)	0.2571(0.0256)
	0.25	0.3162(0.1330)	0.3097(0.0977)	0.3036(0.0434)	0.2998(0.0312)	0.3088(0.1309)	0.3017(0.0919)	0.2944(0.0429)	0.2941(0.0299)
	0.50	0.3564(0.1772)	0.3366(0.1163)	0.3254(0.0548)	0.3276(0.0384)	0.3457(0.1711)	0.3316(0.1210)	0.3249(0.0539)	0.3198(0.0375)
5	0	0.3237(0.1049)	0.3090(0.0755)	0.3030(0.0352)	0.2999(0.0262)	0.3214(0.1021)	0.3162(0.0738)	0.3018(0.0370)	0.2986(0.0261)
	0.25	0.3752(0.1320)	0.3669(0.0966)	0.3599(0.0452)	0.3601(0.0323)	0.3646(0.1330)	0.3547(0.0931)	0.3497(0.0438)	0.3516(0.0299)
	0.50	0.4064(0.1718)	0.4078(0.1194)	0.3975(0.0568)	0.3954(0.0400)	0.4013(0.1698)	0.3875(0.1203)	0.3858(0.0540)	0.3856(0.0392)

FIGURE B.1 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{PH})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à risques proportionnels, pour différents risques relatifs et différents pourcentages de censure, calculés pour une variable Z de Bernoulli $\mathcal{B}(1/2)$, une censure uniforme et $n = 50$ (1000 répétitions).

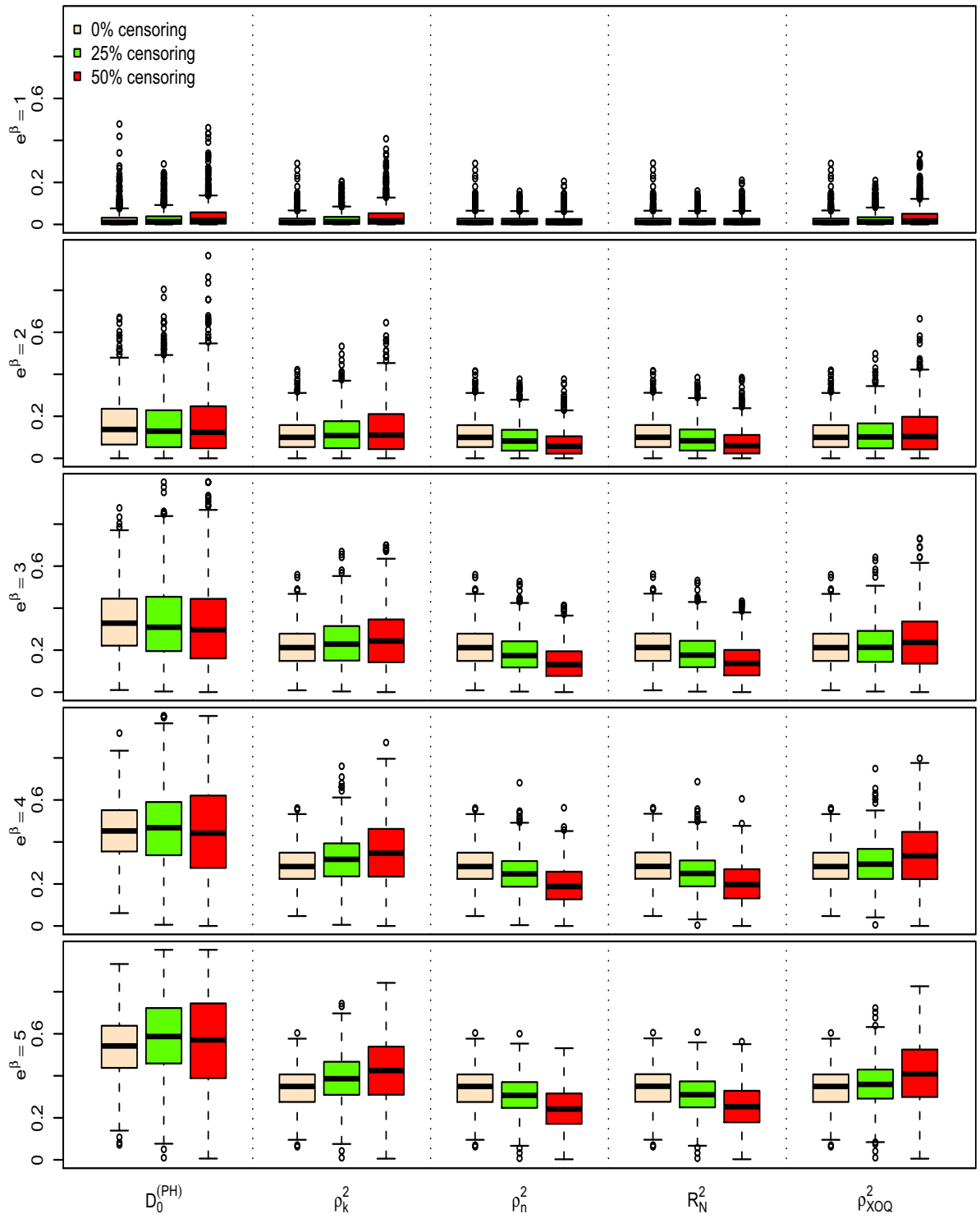


FIGURE B.2 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{PH})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à risques proportionnels, pour différents risques relatifs et différents pourcentages de censure, calculés pour une variable Z uniforme $\mathcal{U}[0, \sqrt{(3)}]$, une censure uniforme et $n = 50$ (1000 répétitions).

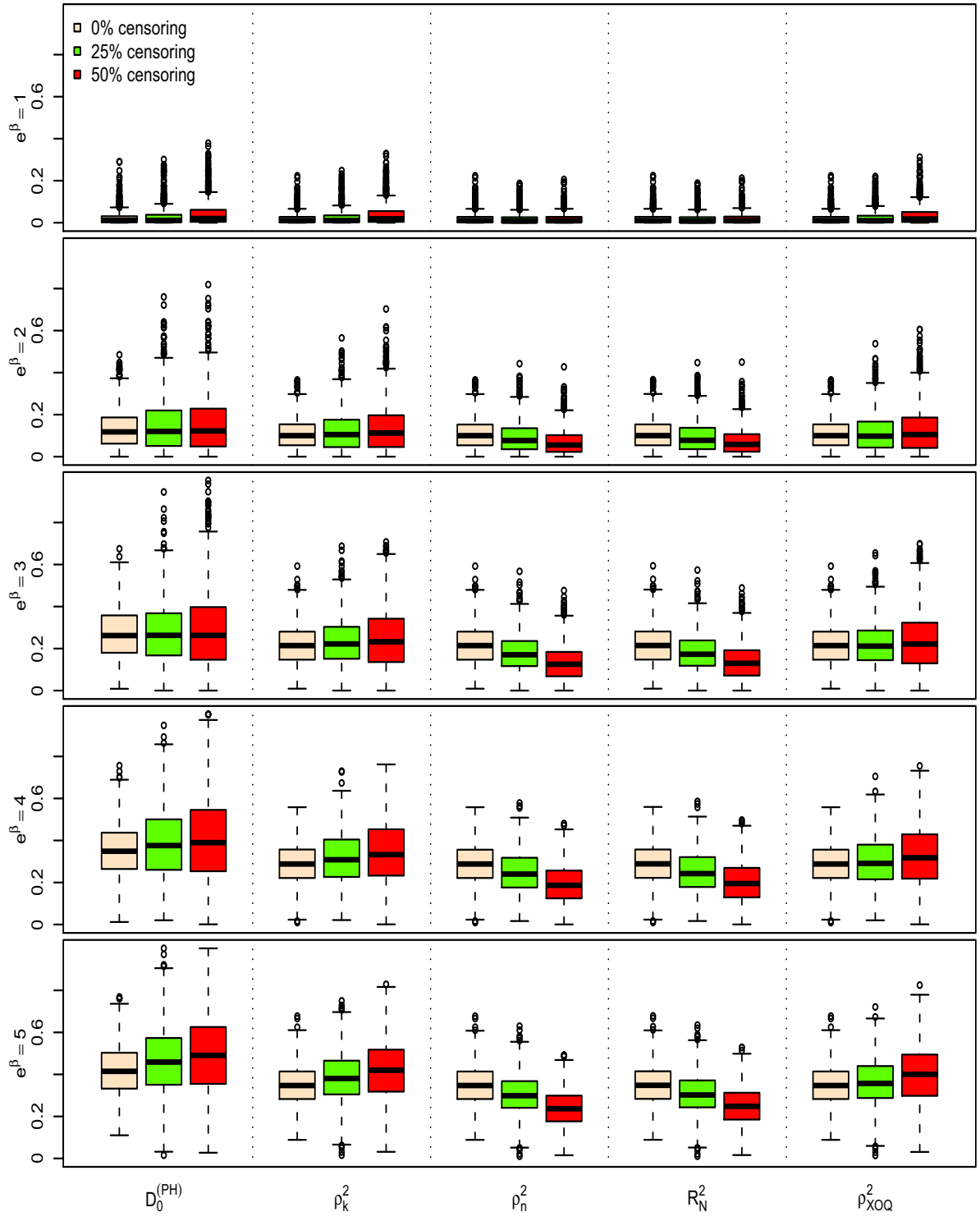


FIGURE B.3 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{PH})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à risques proportionnels, pour différents risques relatifs et différents pourcentages de censure, calculés pour une variable Z normale $\mathcal{N}(0, 1/4)$, une censure uniforme et $n = 50$ (1000 répétitions).

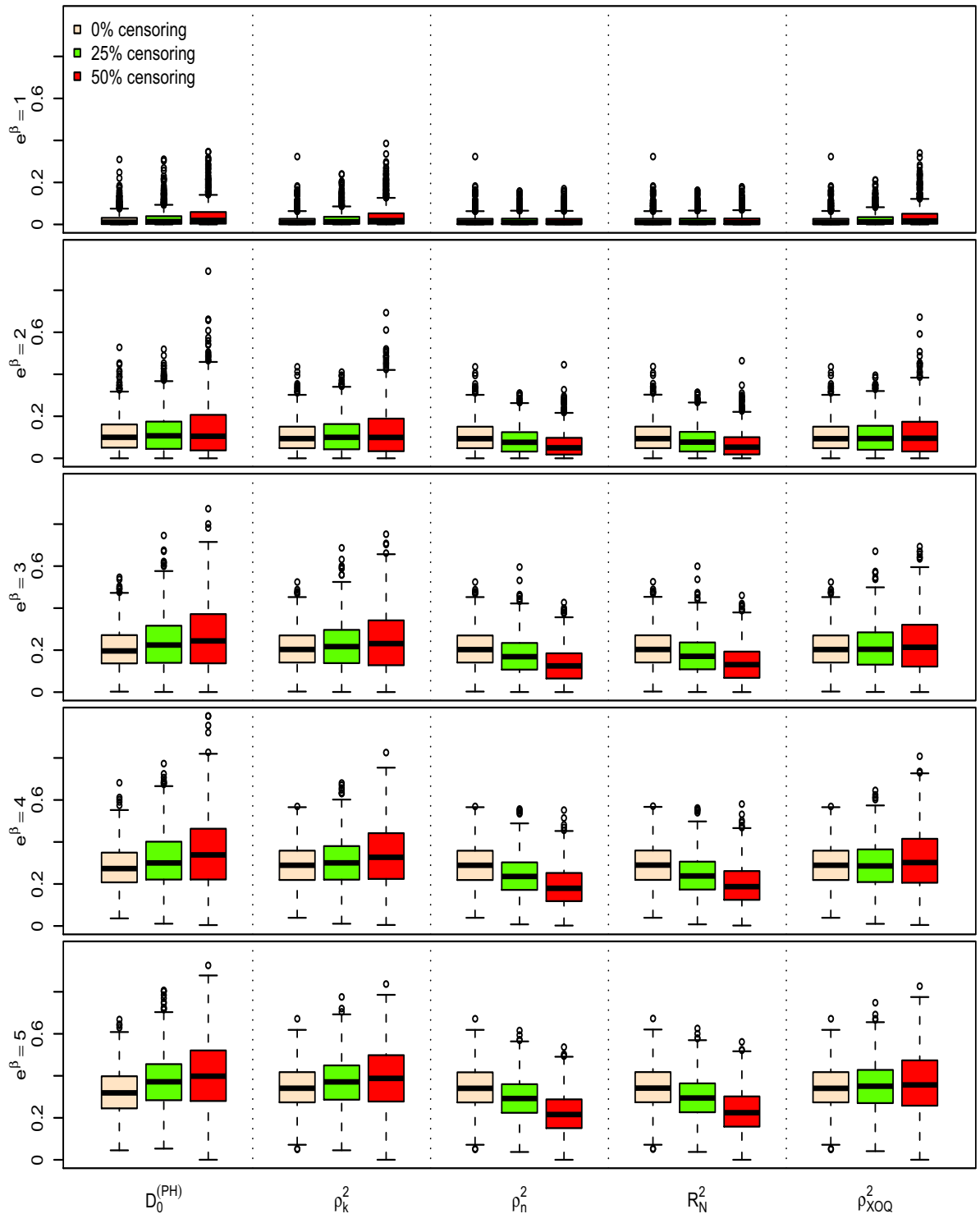


FIGURE B.4 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{PH})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à risques proportionnels, pour différents risques relatifs et différents pourcentages de censure, calculés pour une variable Z de Bernoulli $\mathcal{B}(1/2)$, une censure uniforme et $n = 100$ (1000 répétitions).

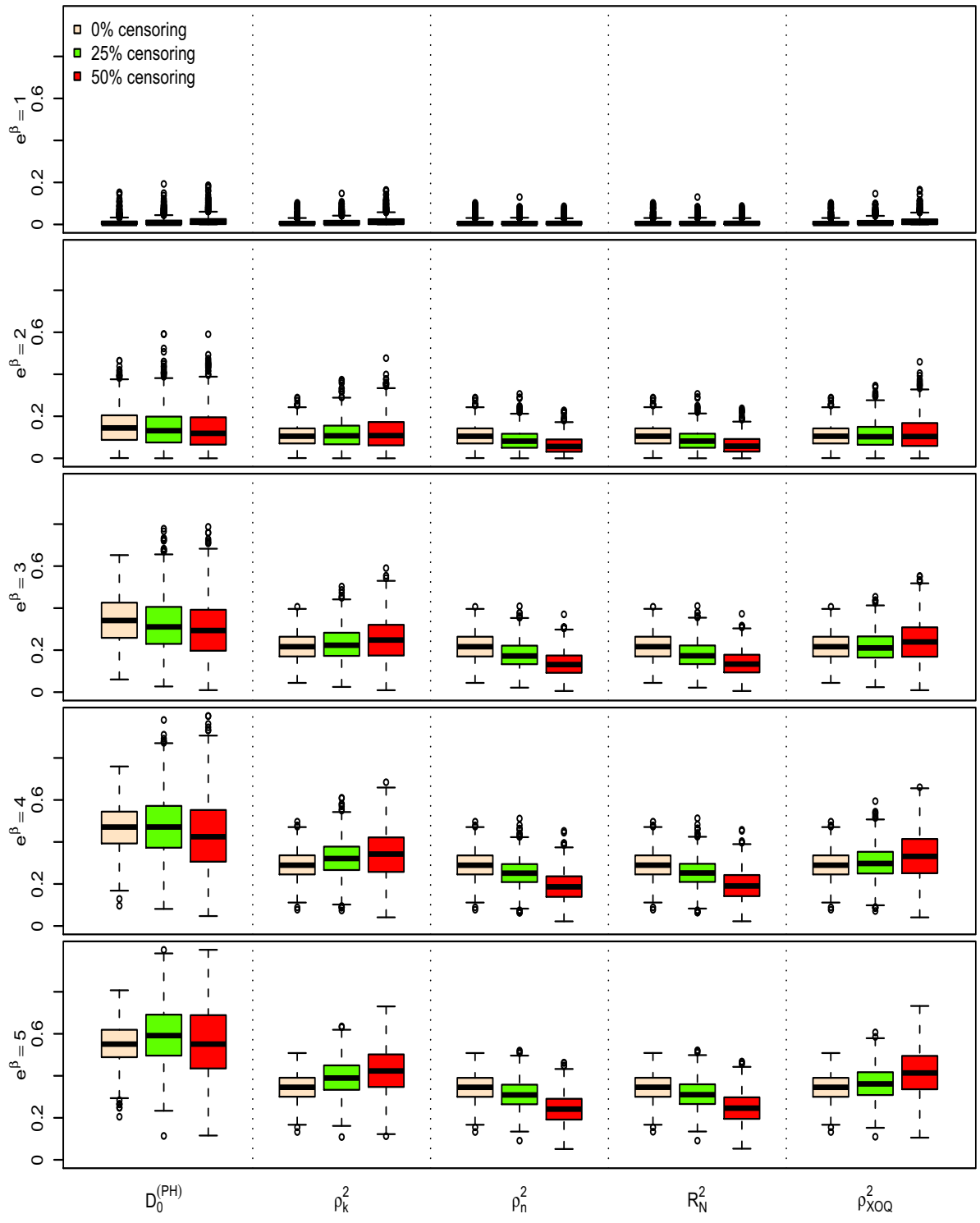


FIGURE B.5 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{PH})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à risques proportionnels, pour différents risques relatifs et différents pourcentages de censure, calculés pour une variable Z uniforme $\mathcal{U}[0, \sqrt{(3)}]$, une censure uniforme et $n = 100$ (1000 répétitions).

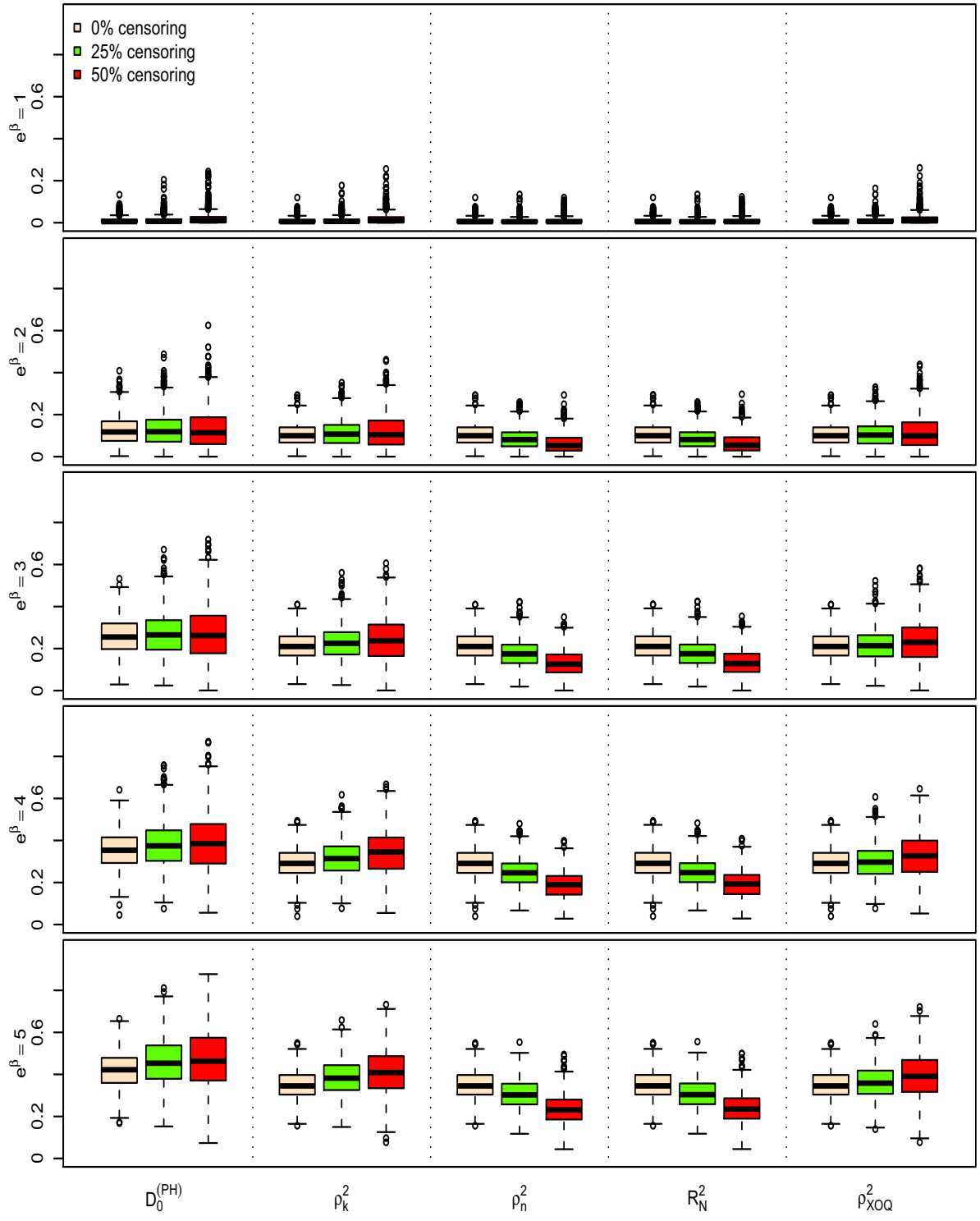


FIGURE B.6 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{PH})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à risques proportionnels, pour différents risques relatifs et différents pourcentages de censure, calculés pour une variable Z normale $\mathcal{N}(0, 1/4)$, une censure uniforme et $n = 100$ (1000 répétitions).

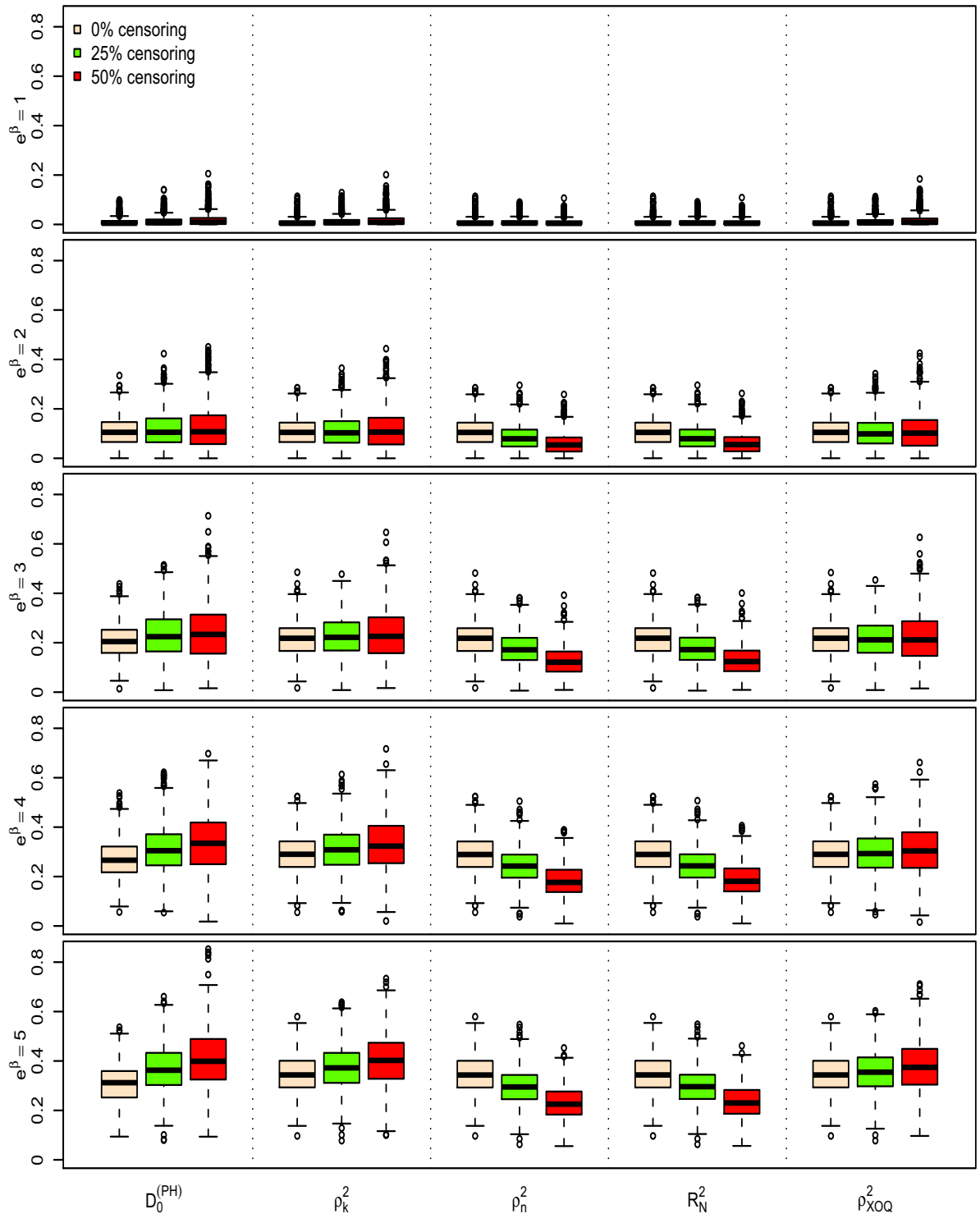


FIGURE B.7 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{PH})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à risques proportionnels, pour différents risques relatifs et différents pourcentages de censure, calculés pour une variable Z de Bernoulli $\mathcal{B}(1/2)$, une censure uniforme et $n = 1000$ (1000 répétitions).

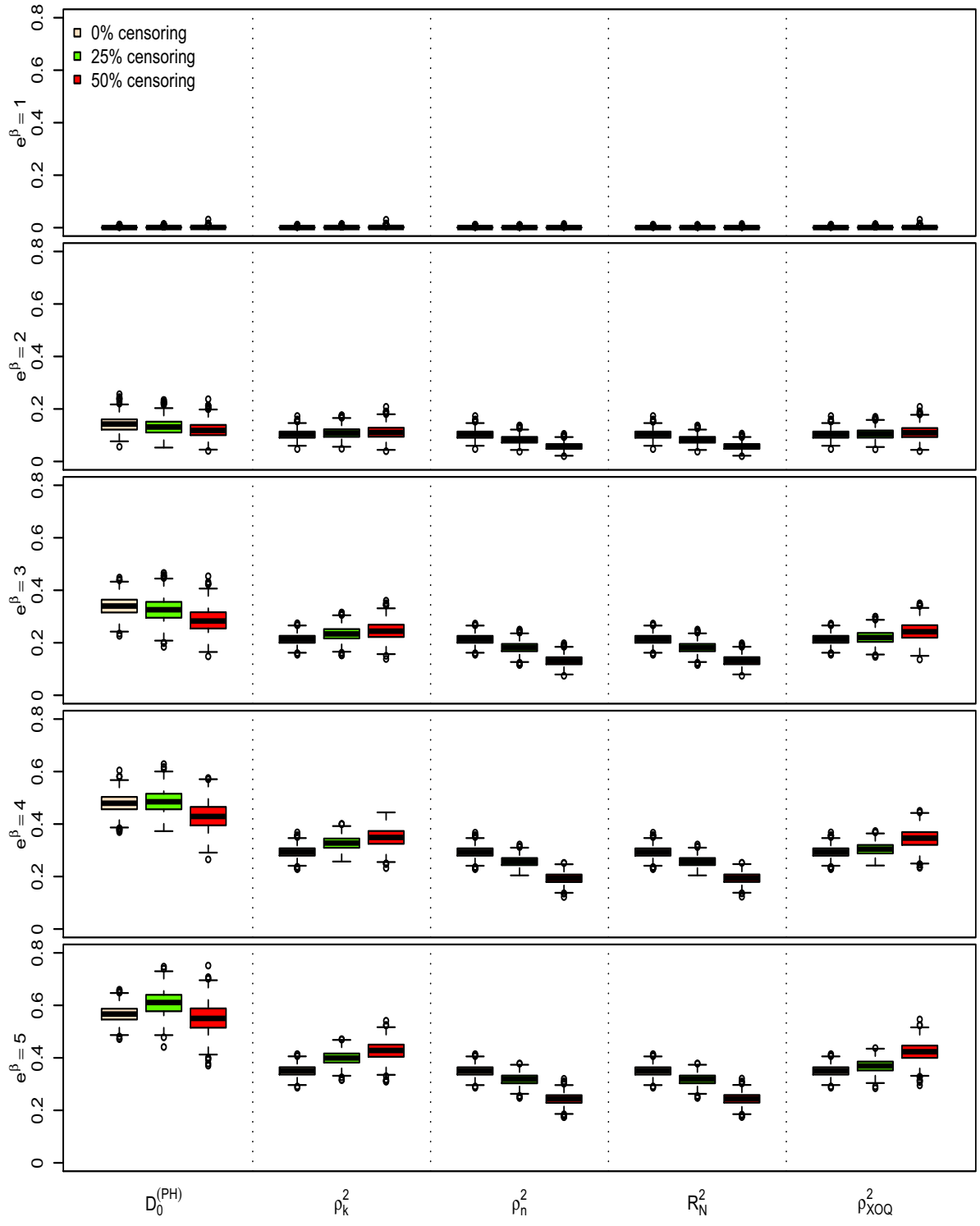


FIGURE B.8 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{PH})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à risques proportionnels, pour différents risques relatifs et différents pourcentages de censure, calculés pour une variable Z uniforme $\mathcal{U}[0, \sqrt{(3)}]$, une censure uniforme et $n = 1000$ (1000 répétitions).

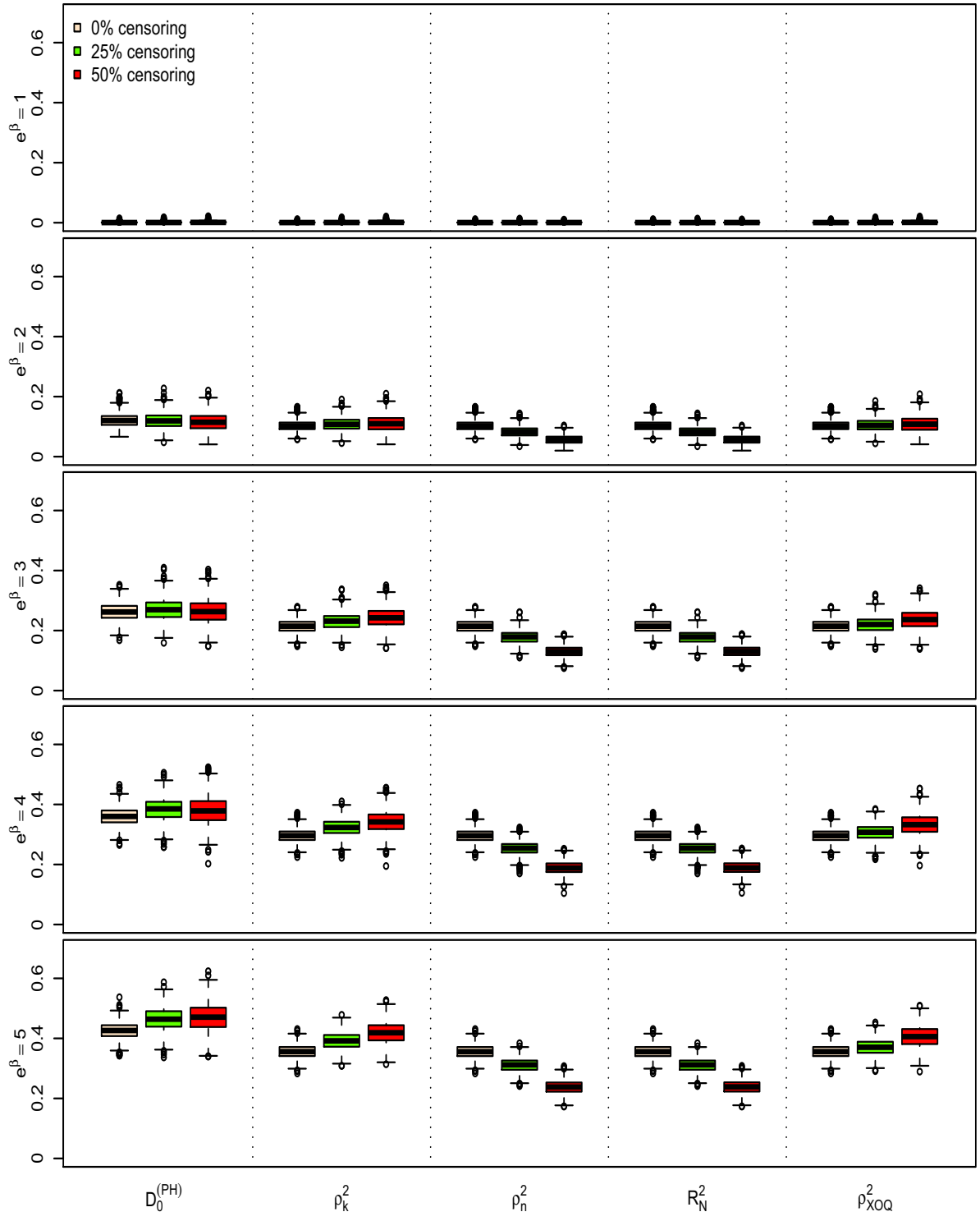


FIGURE B.9 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{PH})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à risques proportionnels, pour différents risques relatifs et différents pourcentages de censure, calculés pour une variable Z normale $\mathcal{N}(0, 1/4)$, une censure uniforme et $n = 1000$ (1000 répétitions).

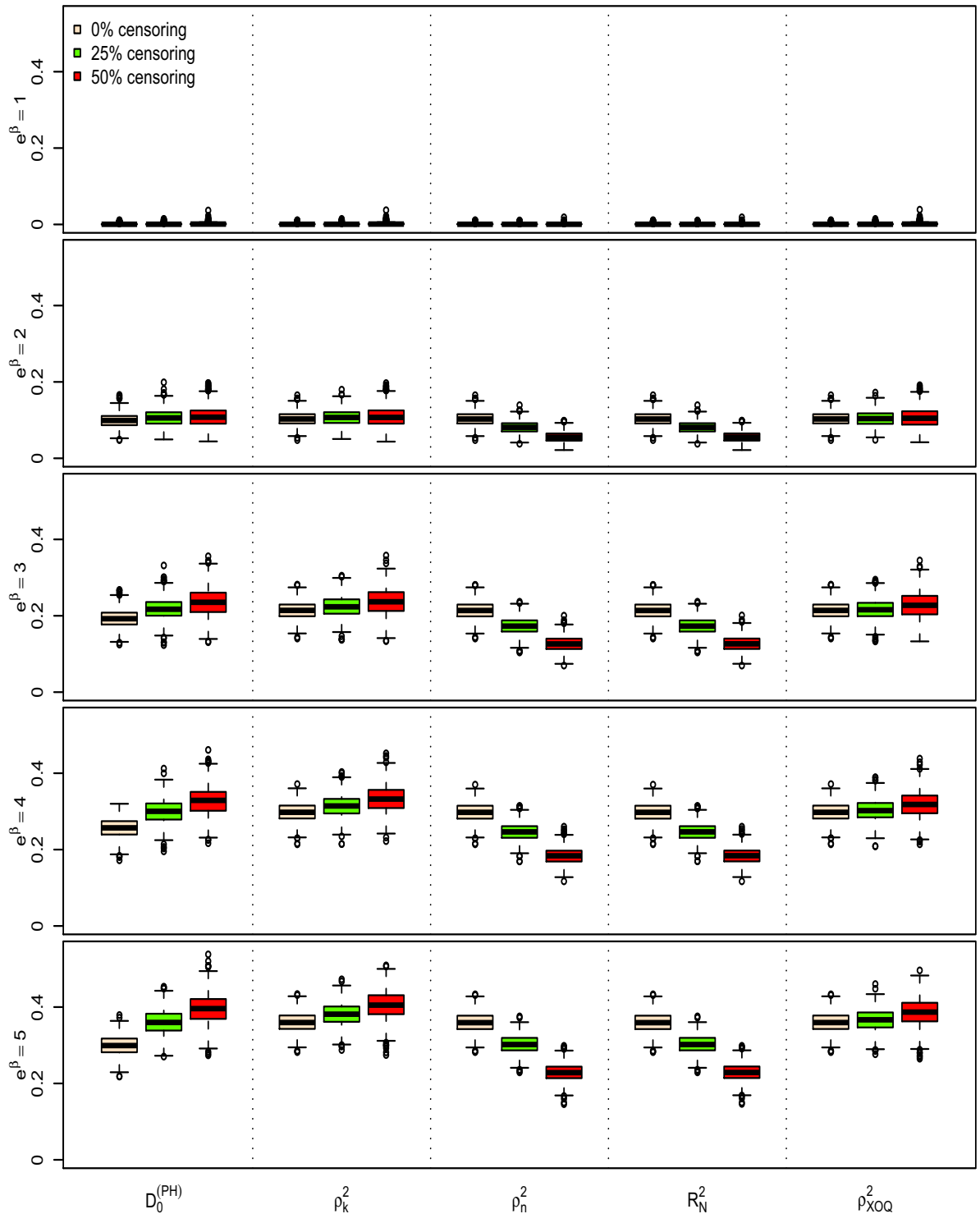


FIGURE B.10 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{PH})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à risques proportionnels, pour différents risques relatifs et différents pourcentages de censure, calculés pour une variable Z de Bernoulli $\mathcal{B}(1/2)$, une censure exponentielle et $n = 50$ (1000 répétitions).

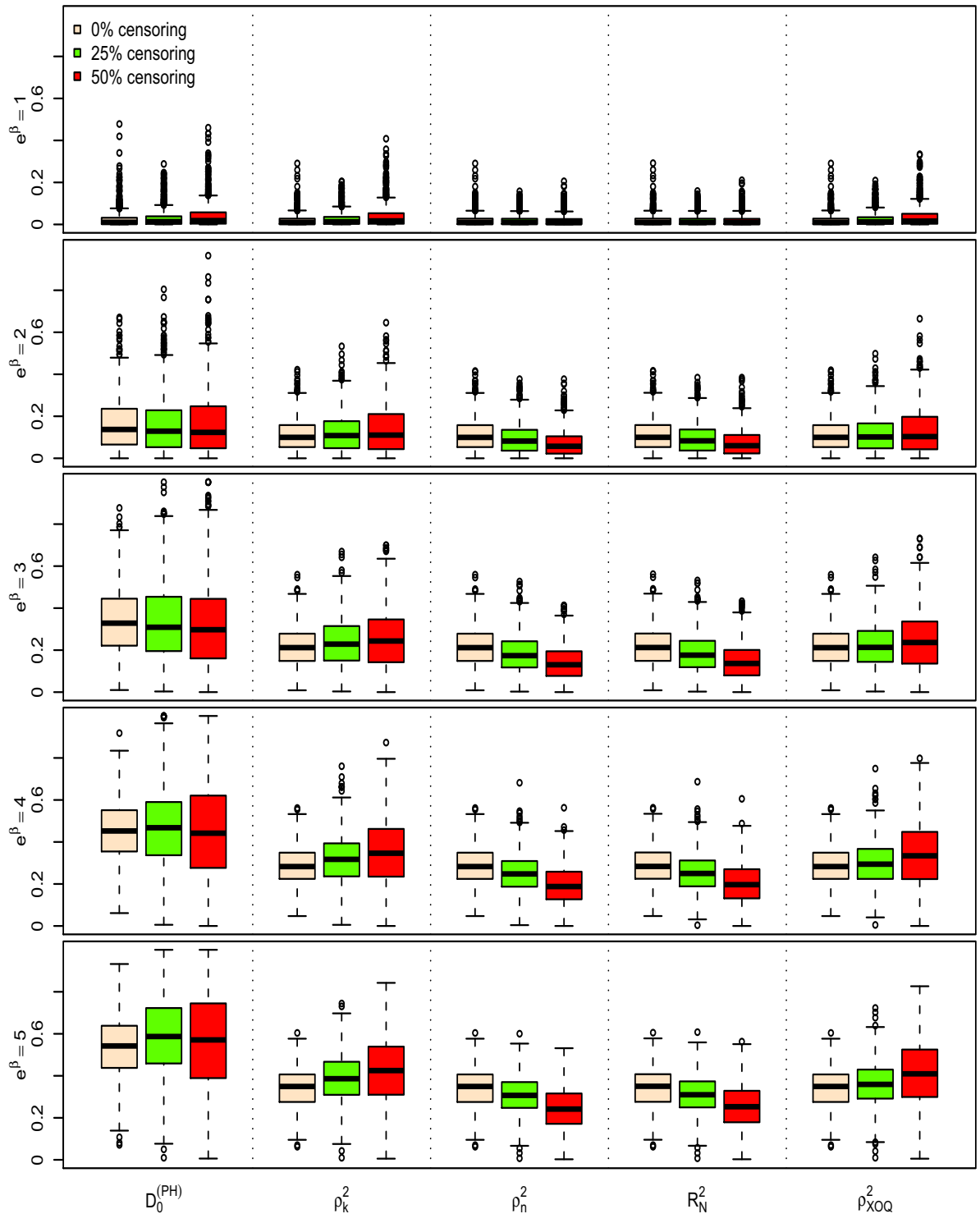


FIGURE B.11 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{PH})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à risques proportionnels, pour différents risques relatifs et différents pourcentages de censure, calculés pour une variable Z uniforme $\mathcal{U}[0, \sqrt{(3)}]$, une censure exponentielle et $n = 50$ (1000 répétitions).

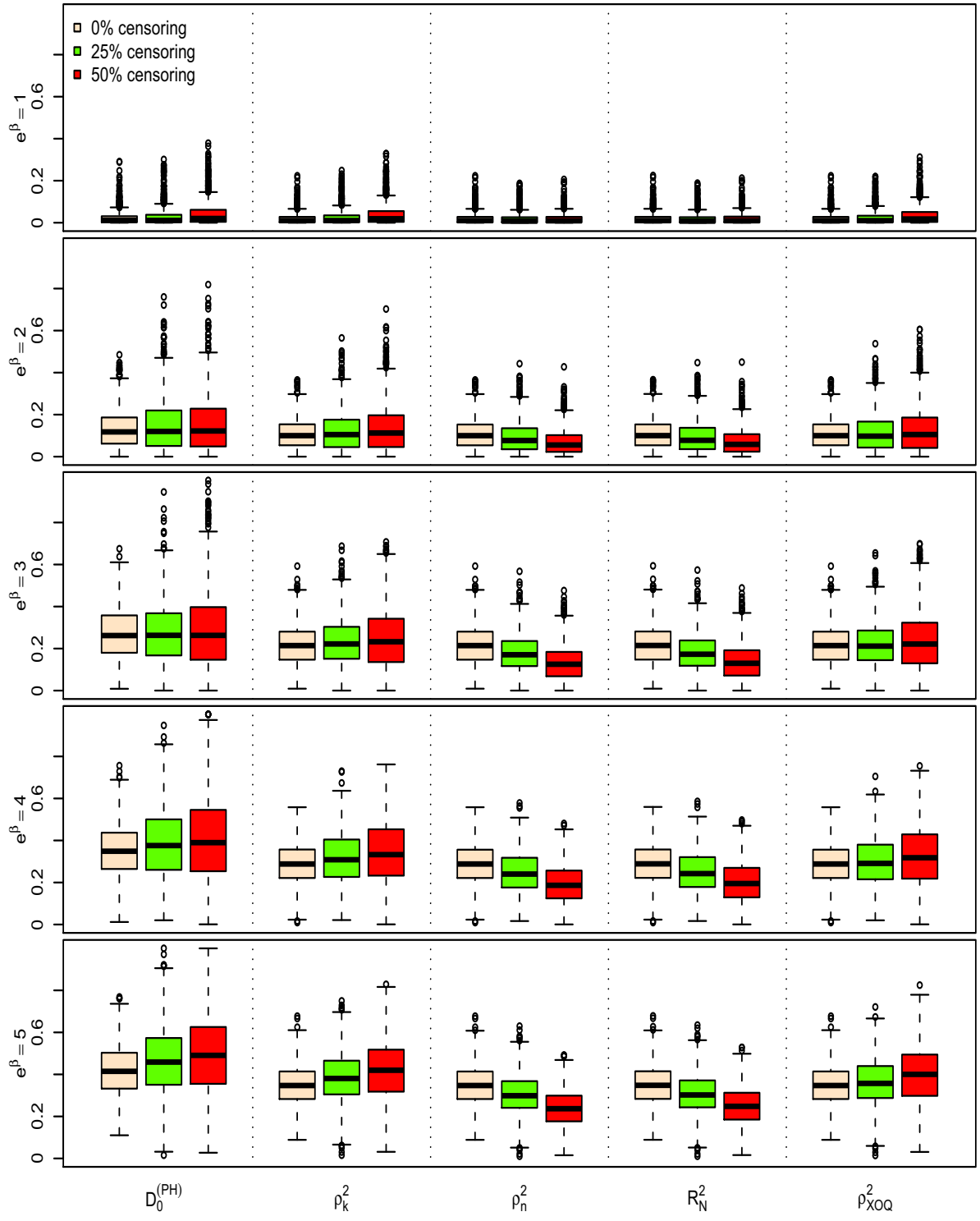


FIGURE B.12 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{PH})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à risques proportionnels, pour différents risques relatifs et différents pourcentages de censure, calculés pour une variable Z normale $\mathcal{N}(0, 1/4)$, une censure exponentielle et $n = 50$ (1000 répétitions).

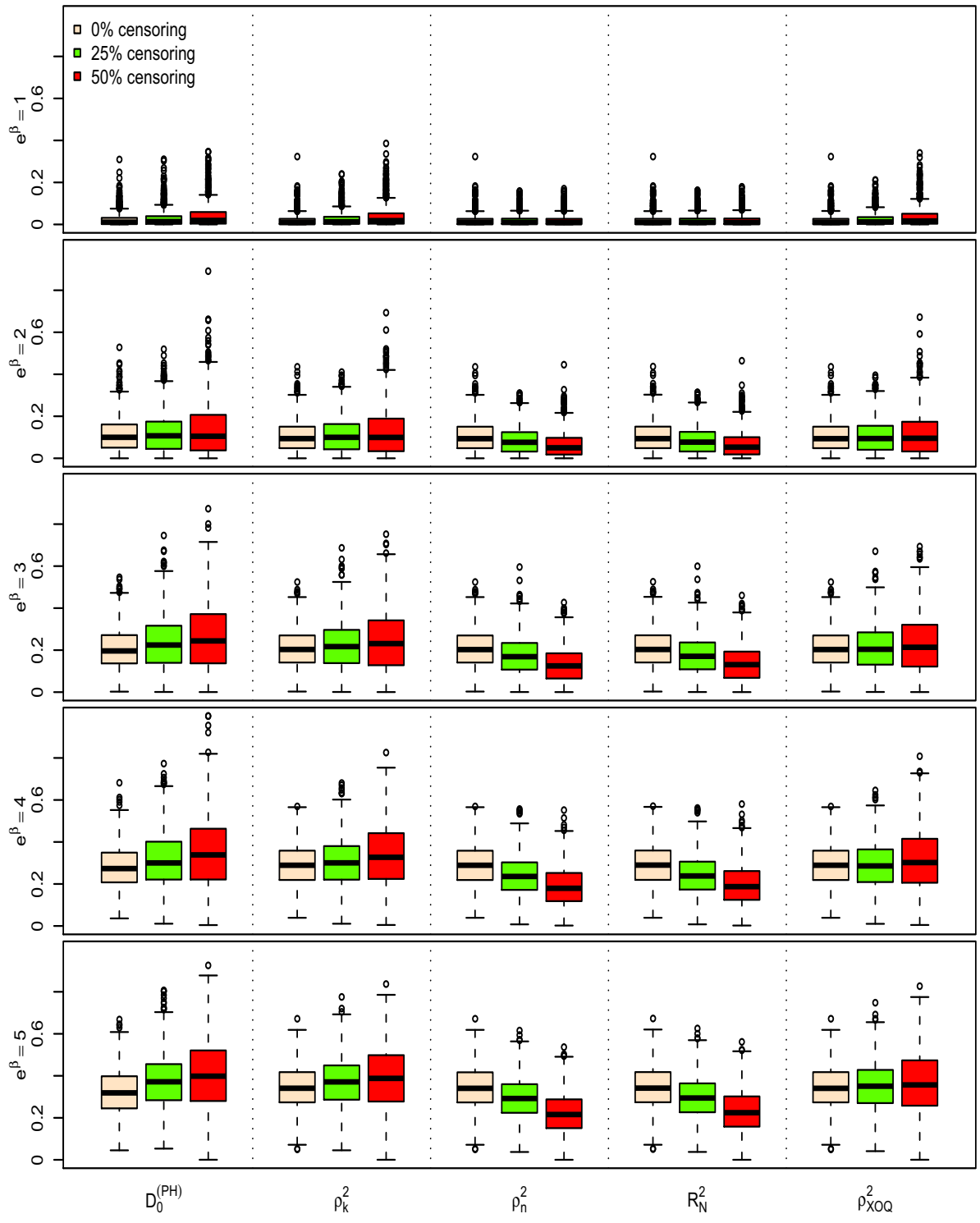


FIGURE B.13 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{PH})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à risques proportionnels, pour différents risques relatifs et différents pourcentages de censure, calculés pour une variable Z de Bernoulli $\mathcal{B}(1/2)$, une censure exponentielle et $n = 100$ (1000 répétitions).

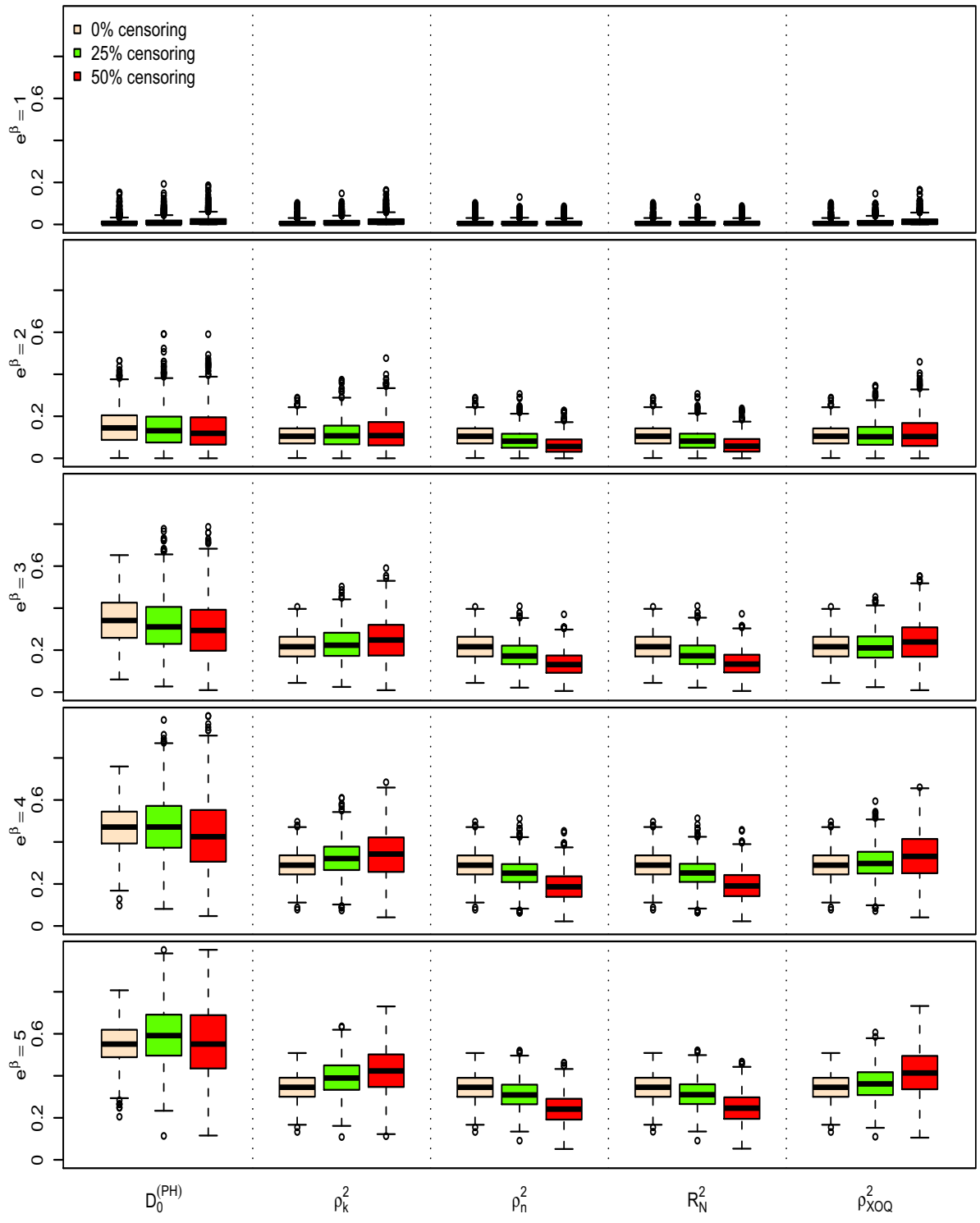


FIGURE B.14 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{PH})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à risques proportionnels, pour différents risques relatifs et différents pourcentages de censure, calculés pour une variable Z uniforme $\mathcal{U}[0, \sqrt{(3)}]$, une censure exponentielle et $n = 100$ (1000 répétitions).

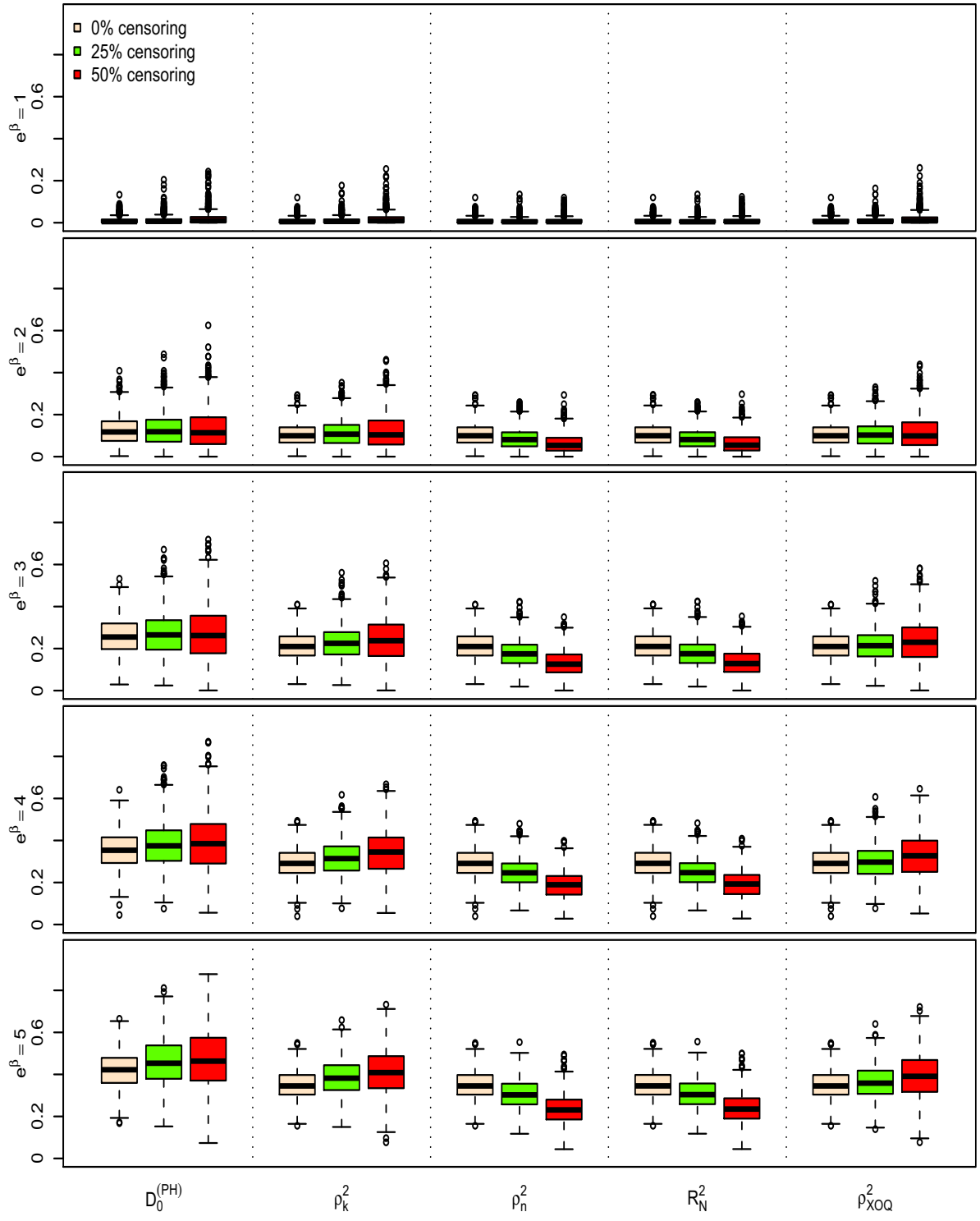


FIGURE B.15 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{PH})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à risques proportionnels, pour différents risques relatifs et différents pourcentages de censure, calculés pour une variable Z normale $\mathcal{N}(0, 1/4)$, une censure exponentielle et $n = 100$ (1000 répétitions).

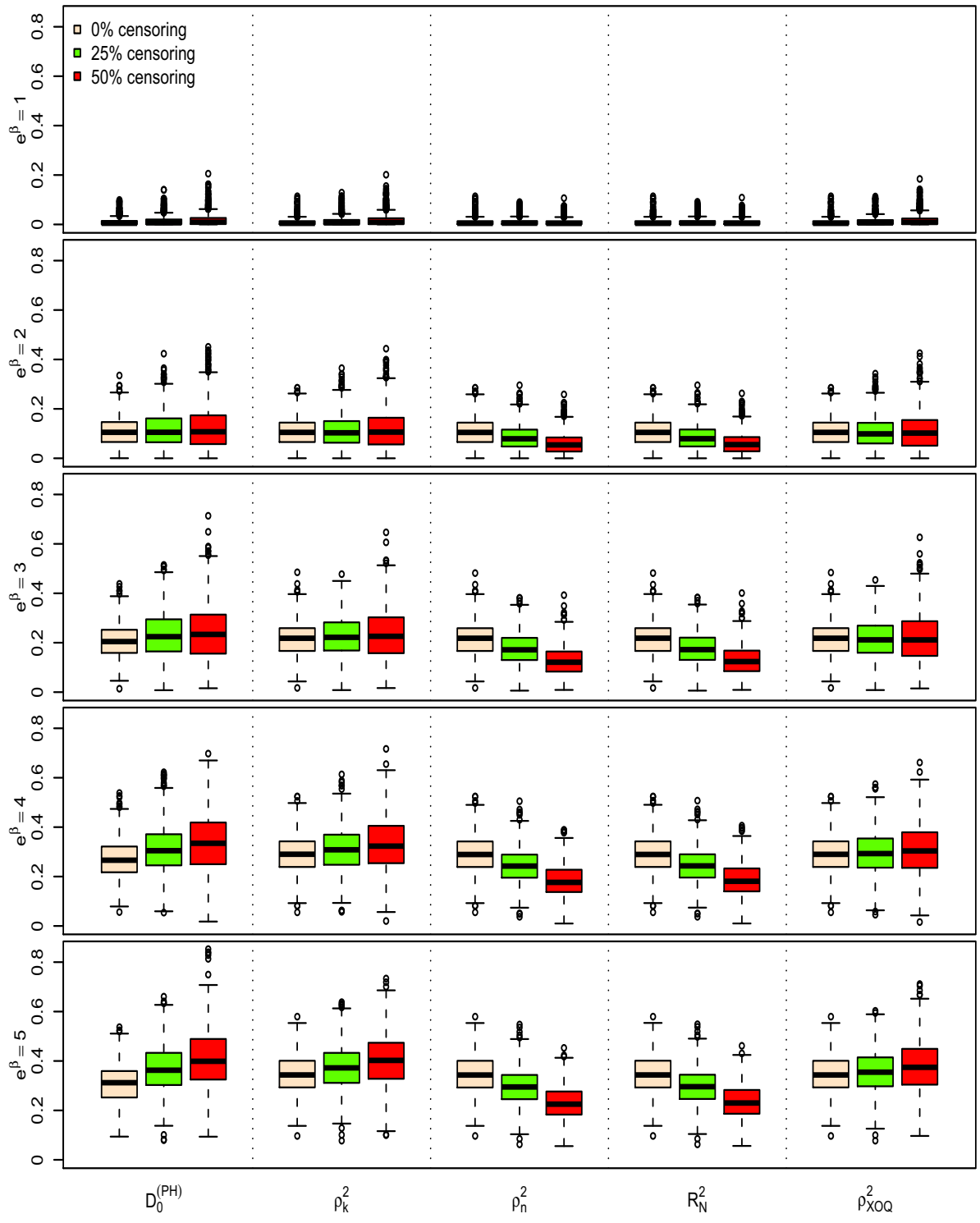


FIGURE B.16 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{PH})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à risques proportionnels, pour différents risques relatifs et différents pourcentages de censure, calculés pour une variable Z de Bernoulli $\mathcal{B}(1/2)$, une censure exponentielle et $n = 1000$ (1000 répétitions).

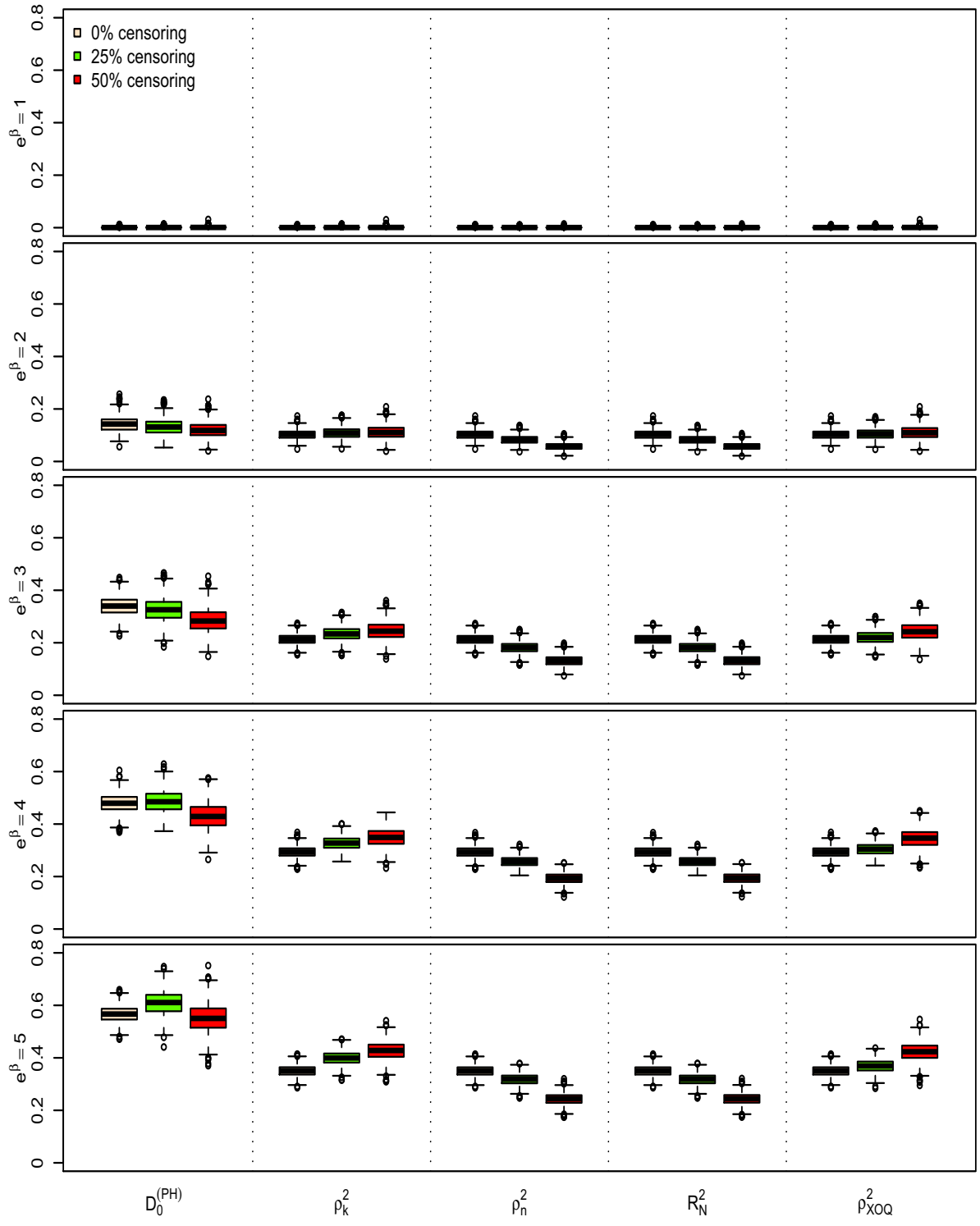


FIGURE B.17 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{PH})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à risques proportionnels, pour différents risques relatifs et différents pourcentages de censure, calculés pour une variable Z uniforme $\mathcal{U}[0, \sqrt{(3)}]$, une censure exponentielle et $n = 1000$ (1000 répétitions).

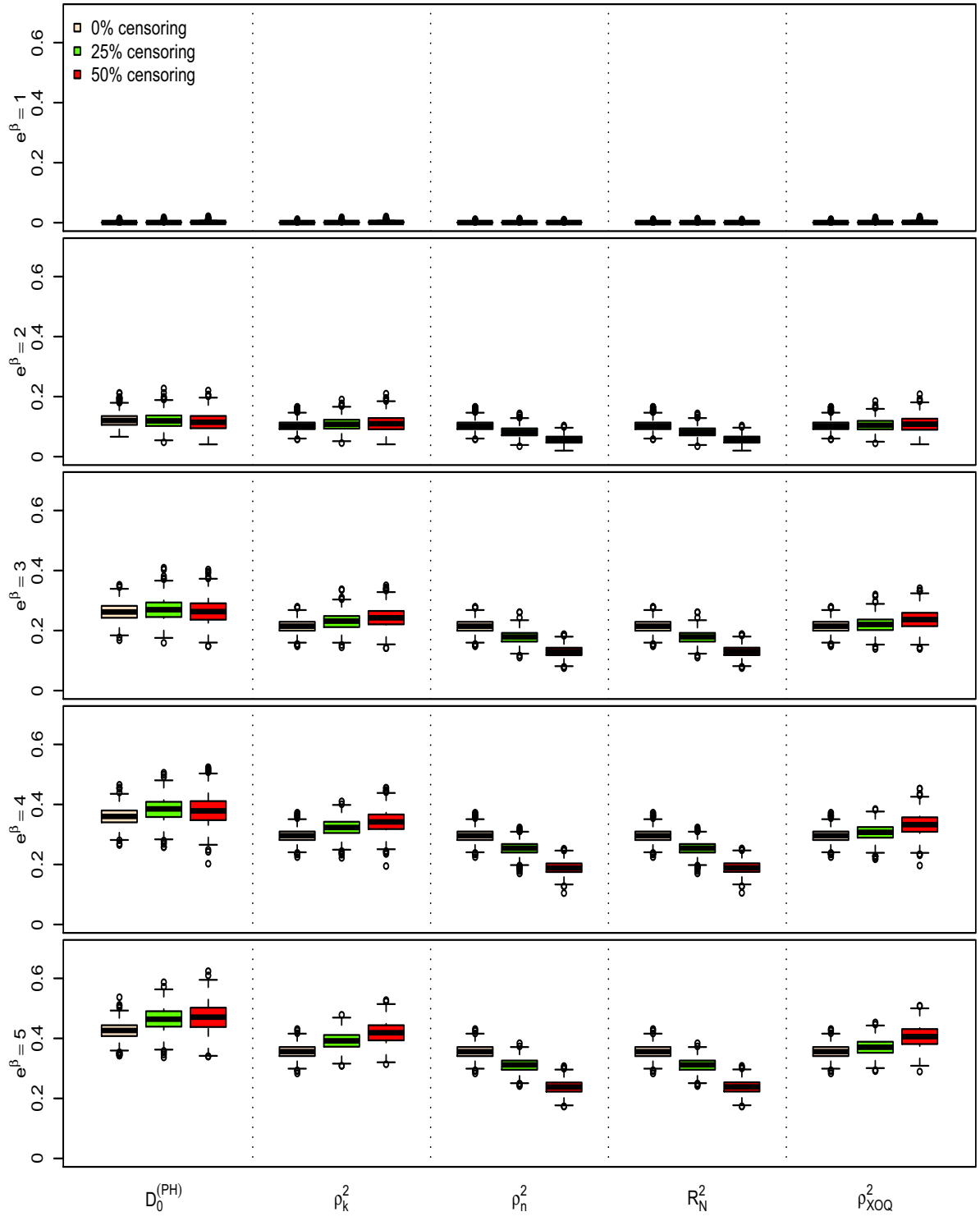


FIGURE B.18 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{PH})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à risques proportionnels, pour différents risques relatifs et différents pourcentages de censure, calculés pour une variable Z normale $\mathcal{N}(0, 1/4)$, une censure exponentielle et $n = 1000$ (1000 répétitions).

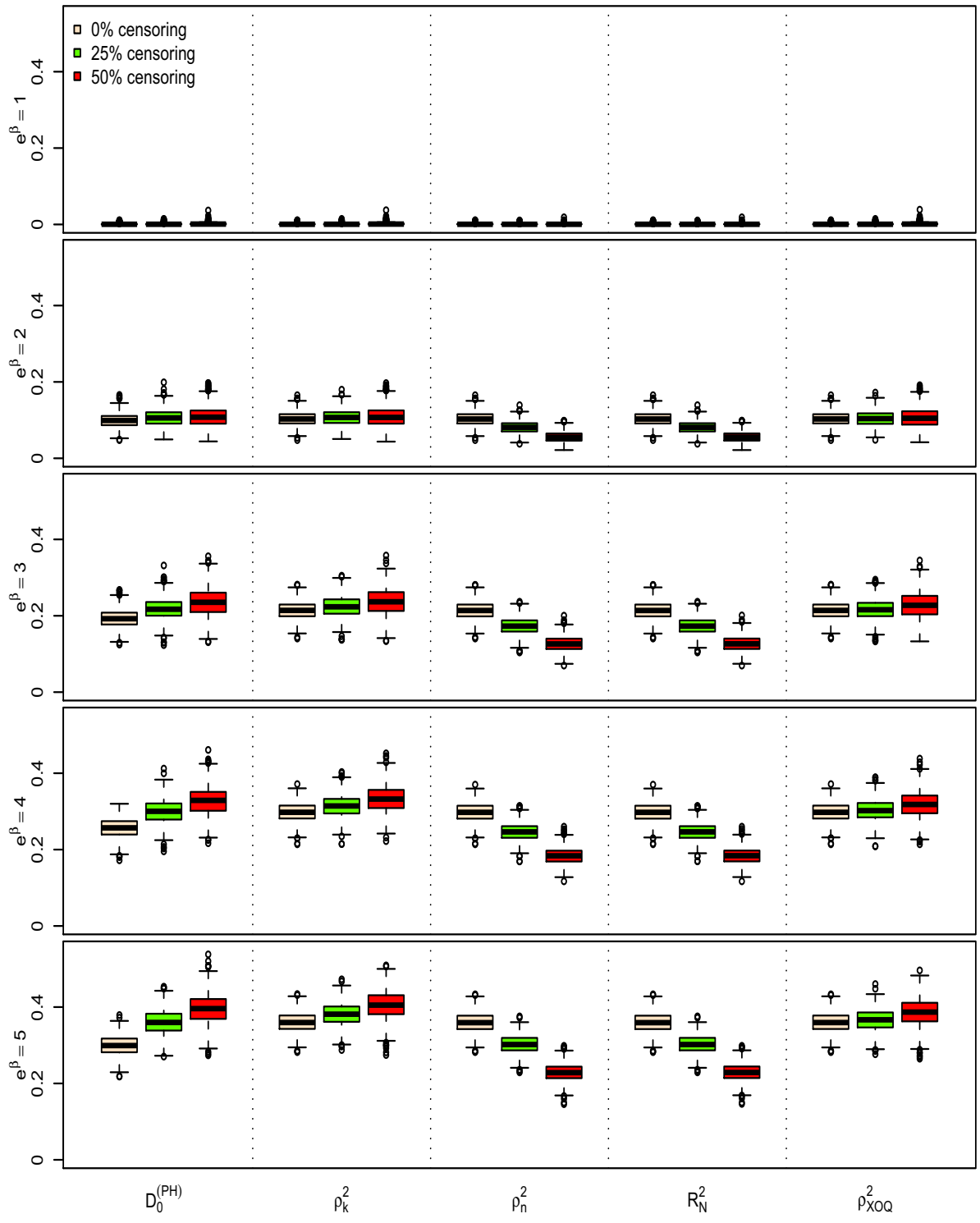


TABLEAU B.4 – Valeurs moyennes de $\mathbf{D}_0^{(\mathrm{PO})}$ sous un modèle à odds proportionnels, pour différentes valeurs de e^β, différents pourcentages de censure p_c, différentes tailles d'échantillon n, différents types de censure, calculées pour une variable de Bernoulli $Z \sim \mathcal{B}(1/2)$ (1000 répétitions). Les écarts-types sont indiqués entre parenthèses.

β	p_c	$C \sim \mathcal{U}[0,r]$				$C \sim \mathcal{E}(\gamma)$			
		$\mathbf{D}_0^{(\mathrm{PO})}$ ($n=50$)	$\mathbf{D}_0^{(\mathrm{PO})}$ ($n=100$)	$\mathbf{D}_0^{(\mathrm{PO})}$ ($n=500$)	$\mathbf{D}_0^{(\mathrm{PO})}$ ($n=1000$)	$\mathbf{D}_0^{(\mathrm{PO})}$ ($n=50$)	$\mathbf{D}_0^{(\mathrm{PO})}$ ($n=100$)	$\mathbf{D}_0^{(\mathrm{PO})}$ ($n=500$)	$\mathbf{D}_0^{(\mathrm{PO})}$ ($n=1000$)
1	0	0.0211(0.0293)	0.0104(0.0145)	0.0020(0.0029)	0.0010(0.0014)	0.0216(0.0326)	0.0105(0.0144)	0.0019(0.0025)	0.0010(0.0014)
	0.25	0.0266(0.0380)	0.0141(0.0190)	0.0027(0.0038)	0.0013(0.0018)	0.0267(0.0384)	0.0129(0.0177)	0.0026(0.0035)	0.0014(0.0020)
	0.50	0.0407(0.0537)	0.0208(0.0295)	0.0040(0.0058)	0.0019(0.0029)	0.0403(0.0570)	0.0196(0.0293)	0.0039(0.0053)	0.0021(0.0029)
1.25	0	0.0239(0.0339)	0.0135(0.0187)	0.0065(0.0071)	0.0052(0.0042)	0.0260(0.0376)	0.0140(0.0196)	0.0062(0.0068)	0.0052(0.0045)
	0.25	0.0348(0.0477)	0.0204(0.0269)	0.0084(0.0095)	0.0067(0.0055)	0.0335(0.0450)	0.0195(0.0252)	0.0079(0.0088)	0.0067(0.0058)
	0.50	0.0472(0.0635)	0.0252(0.0356)	0.0113(0.0129)	0.0090(0.0080)	0.0494(0.0681)	0.0268(0.0382)	0.0104(0.0119)	0.0087(0.0081)
1.5	0	0.0363(0.0481)	0.0248(0.0276)	0.0158(0.0112)	0.0148(0.0075)	0.0352(0.0446)	0.0246(0.0273)	0.0155(0.0104)	0.0146(0.0074)
	0.25	0.0433(0.0545)	0.0292(0.0346)	0.0201(0.0148)	0.0187(0.0098)	0.0483(0.0620)	0.0311(0.0363)	0.0195(0.0136)	0.0184(0.0094)
	0.50	0.0613(0.0795)	0.0425(0.0485)	0.0266(0.0206)	0.0242(0.0133)	0.0642(0.0865)	0.0425(0.0512)	0.0251(0.0187)	0.0237(0.0132)
1.75	0	0.0436(0.0486)	0.0358(0.0353)	0.0282(0.0149)	0.0265(0.0100)	0.0466(0.0525)	0.0363(0.0345)	0.0276(0.0145)	0.0265(0.0103)
	0.25	0.0625(0.0735)	0.0474(0.0480)	0.0359(0.0193)	0.0335(0.0130)	0.0580(0.0700)	0.0468(0.0473)	0.0348(0.0189)	0.0334(0.0133)
	0.50	0.0800(0.0890)	0.0624(0.0639)	0.0469(0.0269)	0.0432(0.0177)	0.0818(0.0985)	0.0622(0.0630)	0.0446(0.0259)	0.0419(0.0179)
2	0	0.0590(0.0632)	0.0499(0.0422)	0.0405(0.0174)	0.0407(0.0122)	0.0597(0.0629)	0.0494(0.0425)	0.0417(0.0181)	0.0399(0.0125)
	0.25	0.0768(0.0832)	0.0616(0.0536)	0.0514(0.0226)	0.0513(0.0158)	0.0765(0.0834)	0.0642(0.0557)	0.0523(0.0235)	0.0499(0.0163)
	0.50	0.1018(0.1107)	0.0839(0.0709)	0.0665(0.0308)	0.0659(0.0218)	0.0984(0.1106)	0.0777(0.0701)	0.0658(0.0313)	0.0624(0.0220)
3	0	0.1139(0.0855)	0.1060(0.0618)	0.0989(0.0267)	0.0965(0.0185)	0.1145(0.0912)	0.1061(0.0626)	0.0958(0.0265)	0.0957(0.0186)
	0.25	0.1458(0.1160)	0.1335(0.0794)	0.1246(0.0346)	0.1220(0.0239)	0.1448(0.1154)	0.1286(0.0775)	0.1193(0.0337)	0.1191(0.0241)
	0.50	0.1835(0.1473)	0.1781(0.1130)	0.1588(0.0462)	0.1556(0.0328)	0.1851(0.1502)	0.1606(0.1049)	0.1479(0.0441)	0.1485(0.0319)
4	0	0.1658(0.1078)	0.1581(0.0729)	0.1496(0.0326)	0.1478(0.0226)	0.1683(0.1045)	0.1584(0.0742)	0.1493(0.0309)	0.1483(0.0230)
	0.25	0.2089(0.1416)	0.1955(0.0936)	0.1887(0.0420)	0.1858(0.0293)	0.2099(0.1321)	0.1946(0.0906)	0.1862(0.0397)	0.1845(0.0291)
	0.50	0.2724(0.1840)	0.2481(0.1244)	0.2388(0.0576)	0.2359(0.0396)	0.2525(0.1747)	0.2418(0.1223)	0.2313(0.0520)	0.2282(0.0387)
5	0	0.2084(0.1137)	0.2010(0.0791)	0.1927(0.0366)	0.1947(0.0257)	0.2179(0.1135)	0.2016(0.0806)	0.1927(0.0358)	0.1933(0.0246)
	0.25	0.2621(0.1411)	0.2567(0.1036)	0.2439(0.0471)	0.2462(0.0335)	0.2574(0.1430)	0.2500(0.1004)	0.2407(0.0464)	0.2409(0.0320)
	0.50	0.3366(0.1873)	0.3232(0.1344)	0.3065(0.0625)	0.3099(0.0442)	0.3371(0.1953)	0.3105(0.1318)	0.2949(0.0600)	0.2958(0.0412)

TABLEAU B.5 – Valeurs moyennes de $\mathbf{D}_0^{(\mathrm{PO})}$ sous un modèle à odds proportionnels, pour différentes valeurs de e^β, différents pourcentages de censure p_c, différentes tailles d'échantillon n, différents types de censure, calculées pour une variable uniforme $Z \sim \mathcal{U}[0, \sqrt{(3)}]$ (1000 répétitions). Les écarts-types sont indiqués entre parenthèses.

β	p_c	$C \sim \mathcal{U}[0,r]$				$C \sim \mathcal{E}(\gamma)$			
		$\mathbf{D}_0^{(\mathrm{PO})}$ ($n=50$)	$\mathbf{D}_0^{(\mathrm{PO})}$ ($n=100$)	$\mathbf{D}_0^{(\mathrm{PO})}$ ($n=500$)	$\mathbf{D}_0^{(\mathrm{PO})}$ ($n=1000$)	$\mathbf{D}_0^{(\mathrm{PO})}$ ($n=50$)	$\mathbf{D}_0^{(\mathrm{PO})}$ ($n=100$)	$\mathbf{D}_0^{(\mathrm{PO})}$ ($n=500$)	$\mathbf{D}_0^{(\mathrm{PO})}$ ($n=1000$)
1	0	0.0190(0.0269)	0.0103(0.0138)	0.0020(0.0028)	0.0010(0.0014)	0.0204(0.0280)	0.0098(0.0129)	0.0020(0.0027)	0.0010(0.0014)
	0.25	0.0291(0.0404)	0.0143(0.0190)	0.0027(0.0038)	0.0014(0.0020)	0.0271(0.0362)	0.0131(0.0187)	0.0027(0.0037)	0.0014(0.0020)
	0.50	0.0396(0.0546)	0.0199(0.0281)	0.0039(0.0058)	0.0020(0.0030)	0.0380(0.0507)	0.0205(0.0293)	0.0040(0.0057)	0.0020(0.0028)
1.25	0	0.0248(0.0330)	0.0136(0.0177)	0.0064(0.0068)	0.0053(0.0044)	0.0257(0.0355)	0.0144(0.0180)	0.0062(0.0068)	0.0052(0.0044)
	0.25	0.0337(0.0427)	0.0194(0.0268)	0.0083(0.0090)	0.0067(0.0057)	0.0339(0.0480)	0.0181(0.0245)	0.0079(0.0089)	0.0067(0.0058)
	0.50	0.0491(0.0683)	0.0244(0.0327)	0.0111(0.0121)	0.0090(0.0081)	0.0517(0.0654)	0.0273(0.0349)	0.0106(0.0122)	0.0087(0.0079)
1.5	0	0.0340(0.0415)	0.0248(0.0300)	0.0154(0.0109)	0.0146(0.0071)	0.0342(0.0460)	0.0237(0.0275)	0.0154(0.0101)	0.0145(0.0074)
	0.25	0.0449(0.0558)	0.0310(0.0361)	0.0196(0.0143)	0.0184(0.0093)	0.0437(0.0557)	0.0322(0.0371)	0.0193(0.0130)	0.0182(0.0093)
	0.50	0.0598(0.0794)	0.0433(0.0524)	0.0261(0.0201)	0.0241(0.0130)	0.0627(0.0792)	0.0381(0.0461)	0.0248(0.0182)	0.0234(0.0131)
1.75	0	0.0414(0.0467)	0.0352(0.0325)	0.0278(0.0149)	0.0264(0.0100)	0.0471(0.0532)	0.0349(0.0337)	0.0267(0.0138)	0.0259(0.0102)
	0.25	0.0576(0.0641)	0.0418(0.0409)	0.0350(0.0193)	0.0332(0.0128)	0.0576(0.0659)	0.0441(0.0433)	0.0336(0.0180)	0.0325(0.0131)
	0.50	0.0842(0.1023)	0.0596(0.0634)	0.0465(0.0273)	0.0428(0.0175)	0.0793(0.0886)	0.0566(0.0580)	0.0429(0.0245)	0.0409(0.0172)
2	0	0.0553(0.0569)	0.0475(0.0414)	0.0399(0.0170)	0.0399(0.0119)	0.0568(0.0540)	0.0456(0.0373)	0.0406(0.0170)	0.0390(0.0119)
	0.25	0.0753(0.0776)	0.0590(0.0486)	0.0506(0.0223)	0.0504(0.0155)	0.0751(0.0755)	0.0610(0.0514)	0.0506(0.0219)	0.0487(0.0155)
	0.50	0.1001(0.1038)	0.0818(0.0723)	0.0657(0.0308)	0.0651(0.0221)	0.1009(0.1061)	0.0845(0.0738)	0.0639(0.0294)	0.0611(0.0208)
3	0	0.1064(0.0793)	0.0977(0.0575)	0.0933(0.0249)	0.0909(0.0172)	0.1092(0.0792)	0.0978(0.0570)	0.0898(0.0235)	0.0904(0.0171)
	0.25	0.1366(0.1053)	0.1269(0.0714)	0.1177(0.0321)	0.1152(0.0224)	0.1382(0.1020)	0.1275(0.0718)	0.1119(0.0299)	0.1128(0.0224)
	0.50	0.1820(0.1410)	0.1631(0.0980)	0.1504(0.0430)	0.1476(0.0306)	0.1770(0.1434)	0.1580(0.0973)	0.1396(0.0406)	0.1413(0.0304)
4	0	0.1533(0.0918)	0.1451(0.0653)	0.1366(0.0289)	0.1355(0.0204)	0.1527(0.0930)	0.1425(0.0632)	0.1366(0.0283)	0.1356(0.0205)
	0.25	0.1835(0.1187)	0.1823(0.0865)	0.1724(0.0373)	0.1706(0.0264)	0.1945(0.1188)	0.1792(0.0835)	0.1707(0.0370)	0.1690(0.0260)
	0.50	0.2397(0.1569)	0.2337(0.1119)	0.2193(0.0505)	0.2173(0.0357)	0.2346(0.1531)	0.2293(0.1122)	0.2141(0.0481)	0.2103(0.0341)
5	0	0.1835(0.0990)	0.1766(0.0678)	0.1726(0.0316)	0.1736(0.0222)	0.1812(0.0958)	0.1752(0.0676)	0.1734(0.0313)	0.1731(0.0216)
	0.25	0.2381(0.1297)	0.2307(0.0895)	0.2183(0.0411)	0.2194(0.0288)	0.2300(0.1251)	0.2252(0.0876)	0.2167(0.0407)	0.2154(0.0279)
	0.50	0.2979(0.1676)	0.2928(0.1241)	0.2757(0.0548)	0.2777(0.0381)	0.2865(0.1662)	0.2742(0.1122)	0.2690(0.0539)	0.2679(0.0364)

TABLEAU B.6 – Valeurs moyennes de $\mathbf{D}_0^{(\mathrm{PO})}$ sous un modèle à odds proportionnels, pour différentes valeurs de e^β, différents pourcentages de censure p_c, différentes tailles d'échantillon n, différents types de censure, calculées pour une variable normale $Z \sim \mathcal{N}(0, 1/4)$ (1000 répétitions). Les écarts-types sont indiqués entre parenthèses.

β	p_c	$C \sim \mathcal{U}[0,r]$				$C \sim \mathcal{E}(\gamma)$			
		$\mathbf{D}_0^{(\mathrm{PO})}$ ($n=50$)	$\mathbf{D}_0^{(\mathrm{PO})}$ ($n=100$)	$\mathbf{D}_0^{(\mathrm{PO})}$ ($n=500$)	$\mathbf{D}_0^{(\mathrm{PO})}$ ($n=1000$)	$\mathbf{D}_0^{(\mathrm{PO})}$ ($n=50$)	$\mathbf{D}_0^{(\mathrm{PO})}$ ($n=100$)	$\mathbf{D}_0^{(\mathrm{PO})}$ ($n=500$)	$\mathbf{D}_0^{(\mathrm{PO})}$ ($n=1000$)
1	0	0.0209(0.0282)	0.0100(0.0147)	0.0020(0.0028)	0.0010(0.0014)	0.0217(0.0286)	0.0105(0.0144)	0.0020(0.0029)	0.0010(0.0014)
	0.25	0.0269(0.0364)	0.0141(0.0205)	0.0027(0.0040)	0.0013(0.0018)	0.0276(0.0380)	0.0142(0.0207)	0.0026(0.0040)	0.0013(0.0019)
	0.50	0.0453(0.0573)	0.0198(0.0266)	0.0041(0.0057)	0.0020(0.0029)	0.0443(0.0587)	0.0199(0.0263)	0.0040(0.0059)	0.0019(0.0028)
1.25	0	0.0246(0.0316)	0.0146(0.0191)	0.0060(0.0062)	0.0050(0.0041)	0.0238(0.0328)	0.0143(0.0194)	0.0063(0.0065)	0.0051(0.0045)
	0.25	0.0329(0.0423)	0.0201(0.0264)	0.0077(0.0082)	0.0064(0.0055)	0.0336(0.0436)	0.0177(0.0233)	0.0081(0.0083)	0.0065(0.0059)
	0.50	0.0488(0.0671)	0.0260(0.0342)	0.0105(0.0118)	0.0087(0.0079)	0.0469(0.0577)	0.0274(0.0378)	0.0111(0.0118)	0.0084(0.0078)
1.5	0	0.0335(0.0420)	0.0230(0.0256)	0.0154(0.0107)	0.0142(0.0074)	0.0311(0.0385)	0.0224(0.0247)	0.0150(0.0104)	0.0142(0.0072)
	0.25	0.0419(0.0511)	0.0304(0.0369)	0.0195(0.0138)	0.0180(0.0097)	0.0434(0.0555)	0.0282(0.0318)	0.0189(0.0136)	0.0178(0.0092)
	0.50	0.0581(0.0790)	0.0396(0.0455)	0.0261(0.0193)	0.0237(0.0135)	0.0614(0.0765)	0.0394(0.0443)	0.0245(0.0187)	0.0229(0.0127)
1.75	0	0.0442(0.0467)	0.0343(0.0334)	0.0264(0.0139)	0.0260(0.0098)	0.0440(0.0483)	0.0340(0.0321)	0.0267(0.0139)	0.0259(0.0097)
	0.25	0.0576(0.0639)	0.0416(0.0417)	0.0333(0.0180)	0.0329(0.0126)	0.0549(0.0625)	0.0422(0.0410)	0.0337(0.0182)	0.0322(0.0125)
	0.50	0.0821(0.0914)	0.0562(0.0547)	0.0430(0.0250)	0.0425(0.0171)	0.0796(0.0880)	0.0535(0.0537)	0.0430(0.0248)	0.0411(0.0171)
2	0	0.0584(0.0567)	0.0474(0.0389)	0.0384(0.0162)	0.0374(0.0115)	0.0563(0.0555)	0.0447(0.0368)	0.0384(0.0160)	0.0375(0.0116)
	0.25	0.0728(0.0724)	0.0568(0.0472)	0.0491(0.0212)	0.0474(0.0150)	0.0715(0.0717)	0.0564(0.0468)	0.0479(0.0209)	0.0467(0.0149)
	0.50	0.0984(0.0982)	0.0782(0.0663)	0.0637(0.0297)	0.0612(0.0203)	0.0952(0.0938)	0.0773(0.0679)	0.0603(0.0283)	0.0592(0.0205)
3	0	0.0996(0.0731)	0.0917(0.0500)	0.0859(0.0225)	0.0844(0.0159)	0.0966(0.0697)	0.0880(0.0508)	0.0844(0.0219)	0.0844(0.0160)
	0.25	0.1254(0.0932)	0.1152(0.0657)	0.1093(0.0296)	0.1076(0.0207)	0.1236(0.0962)	0.1164(0.0659)	0.1058(0.0288)	0.1057(0.0207)
	0.50	0.1650(0.1289)	0.1479(0.0899)	0.1404(0.0399)	0.1375(0.0281)	0.1681(0.1322)	0.1481(0.0886)	0.1337(0.0393)	0.1338(0.0281)
4	0	0.1322(0.0789)	0.1262(0.0566)	0.1234(0.0264)	0.1200(0.0189)	0.1334(0.0816)	0.1279(0.0552)	0.1220(0.0257)	0.1218(0.0170)
	0.25	0.1702(0.1014)	0.1686(0.0751)	0.1559(0.0344)	0.1522(0.0246)	0.1686(0.1037)	0.1585(0.0690)	0.1533(0.0325)	0.1530(0.0220)
	0.50	0.2220(0.1441)	0.2076(0.1015)	0.2001(0.0455)	0.1964(0.0325)	0.2222(0.1440)	0.2017(0.0965)	0.1938(0.0428)	0.1932(0.0301)
5	0	0.1637(0.0820)	0.1603(0.0584)	0.1514(0.0259)	0.1514(0.0189)	0.1611(0.0821)	0.1602(0.0598)	0.1520(0.0261)	0.1509(0.0186)
	0.25	0.2109(0.1127)	0.1945(0.0759)	0.1925(0.0342)	0.1915(0.0246)	0.2075(0.1075)	0.1972(0.0788)	0.1899(0.0336)	0.1889(0.0243)
	0.50	0.2697(0.1506)	0.2531(0.1060)	0.2482(0.0464)	0.2460(0.0322)	0.2677(0.1446)	0.2472(0.1016)	0.2393(0.0450)	0.2376(0.0329)

FIGURE B.19 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{PO})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à odds proportionnels, pour différentes valeurs de e^β et différents pourcentages de censure, calculés pour une variable Z de Bernoulli $\mathcal{B}(1/2)$, une censure uniforme et $n = 50$ (1000 répétitions).

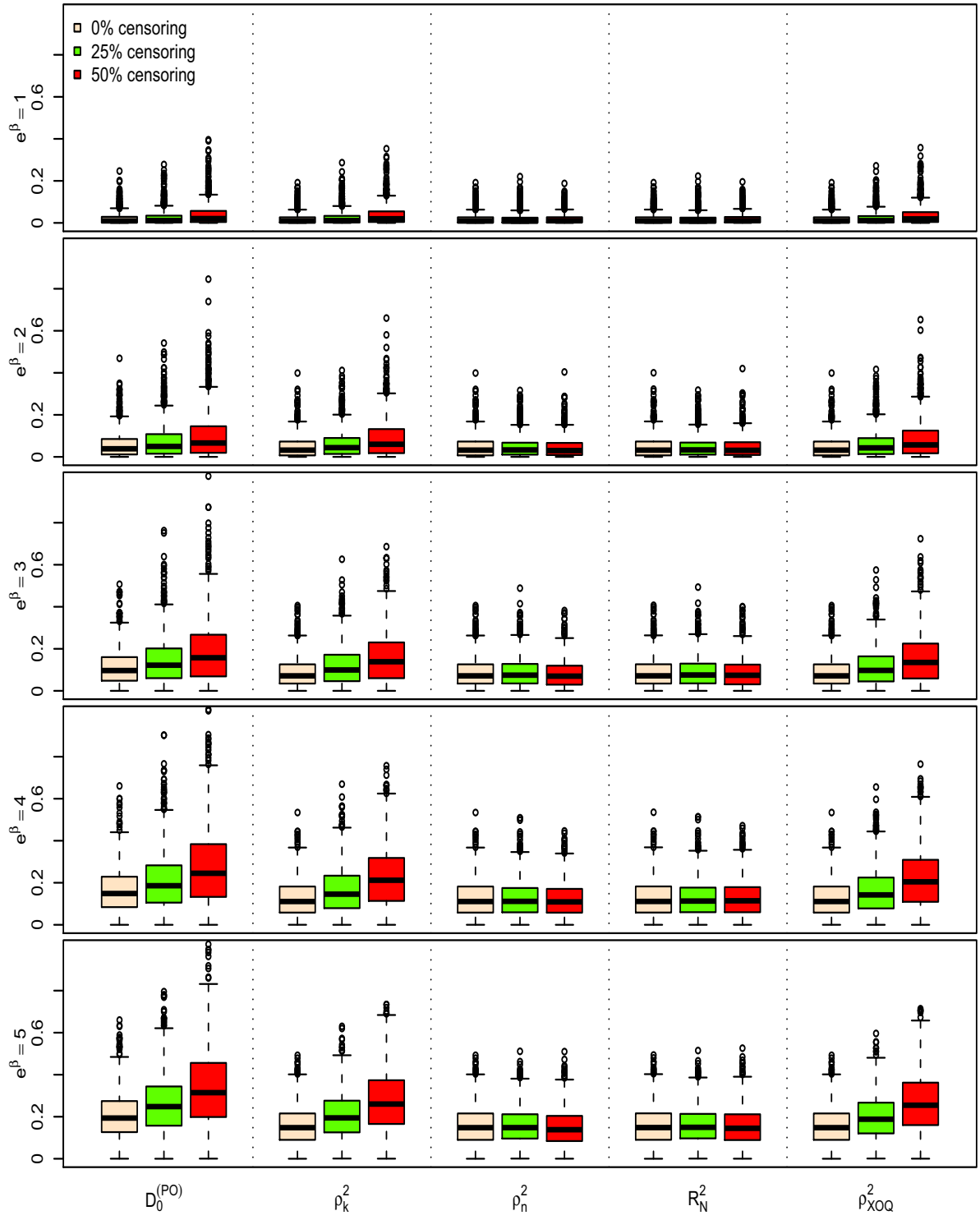


FIGURE B.20 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{PO})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à odds proportionnels, pour différentes valeurs de e^β et différents pourcentages de censure, calculés pour une variable Z uniforme $\mathcal{U}[0, \sqrt{(3)}]$, une censure uniforme et $n = 50$ (1000 répétitions).

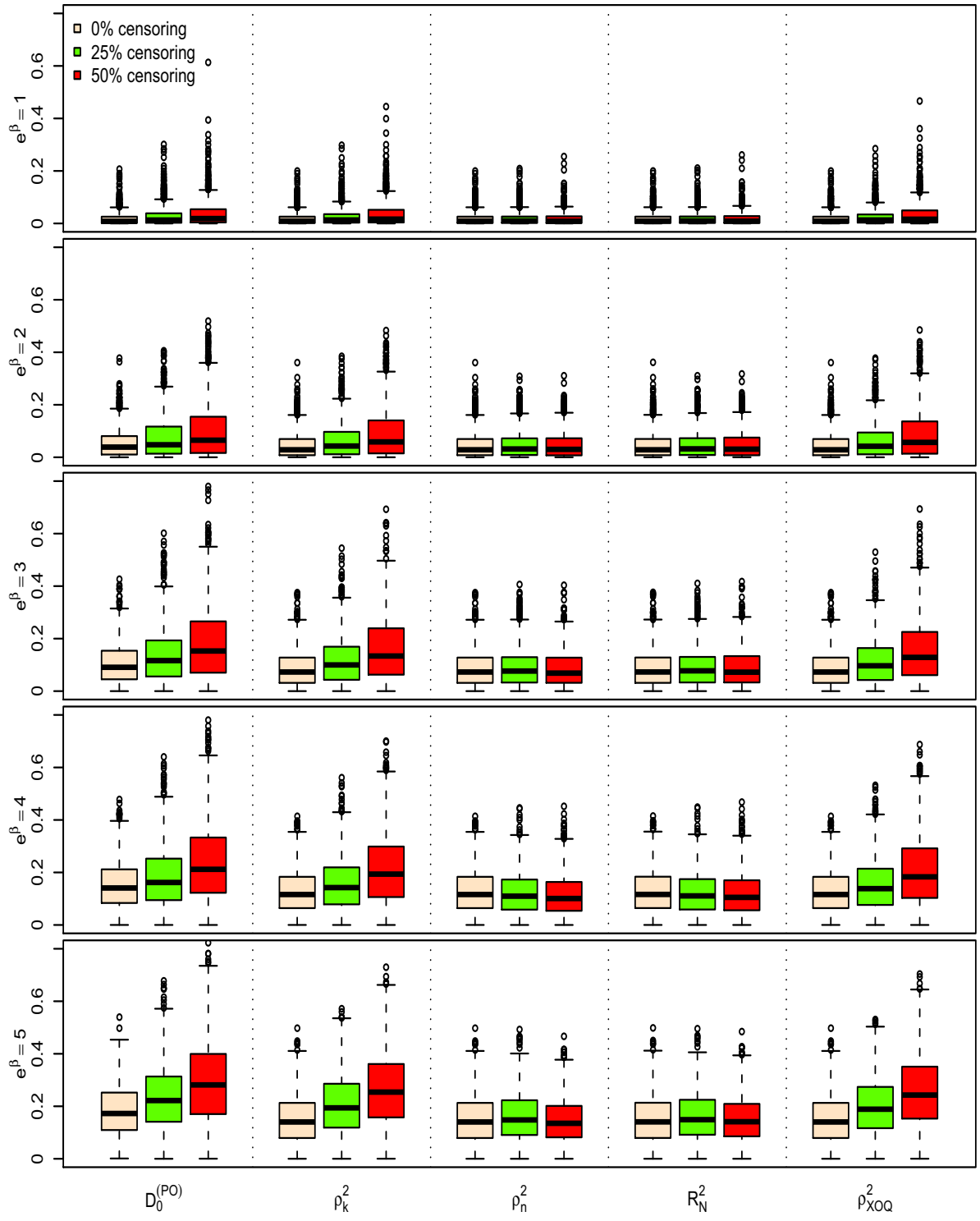


FIGURE B.21 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{PO})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à odds proportionnels, pour différentes valeurs de e^β et différents pourcentages de censure, calculés pour une variable Z normale $\mathcal{N}(0, 1/4)$, une censure uniforme et $n = 50$ (1000 répétitions).

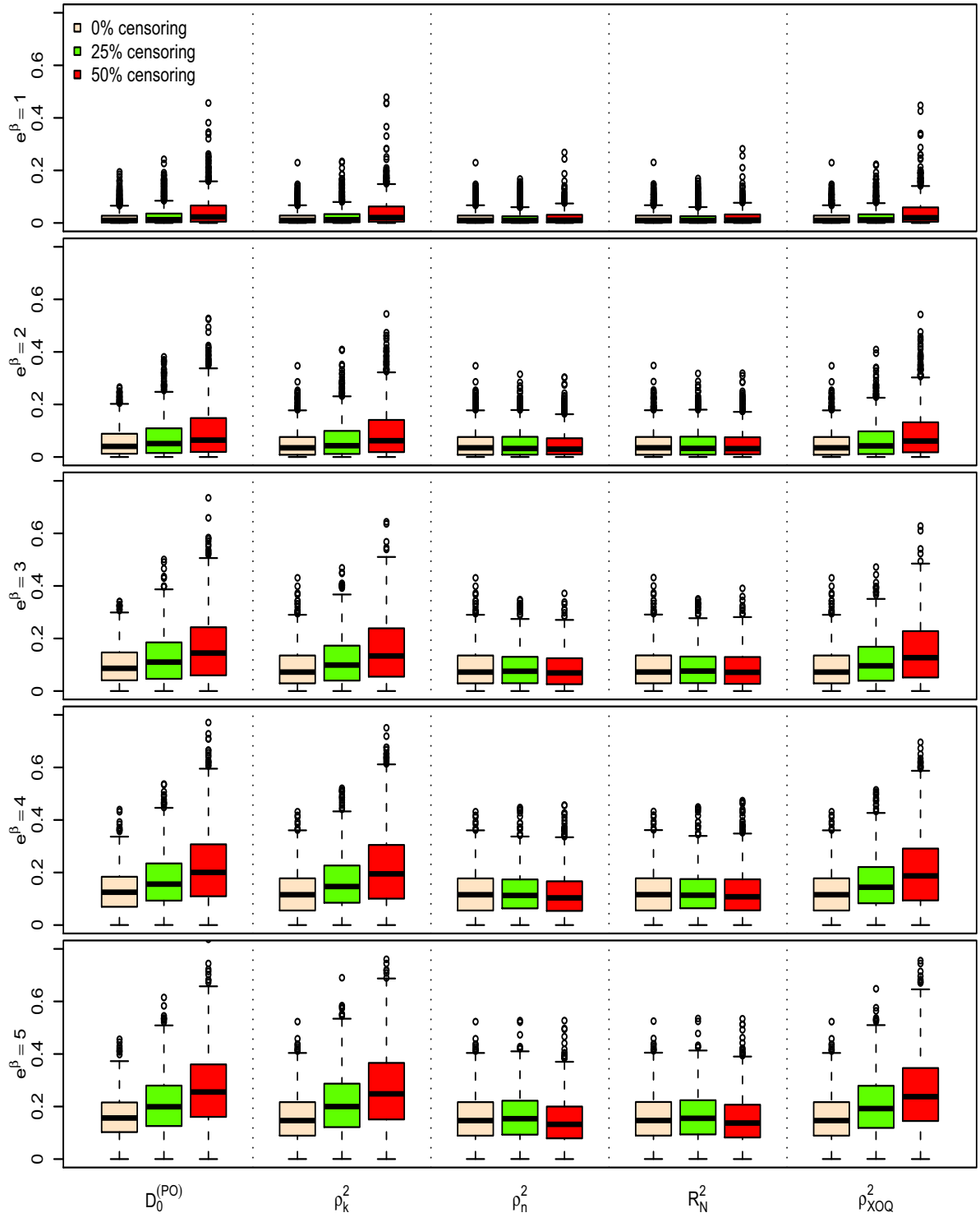


FIGURE B.22 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{PO})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à odds proportionnels, pour différentes valeurs de e^β et différents pourcentages de censure, calculés pour une variable Z de Bernoulli $\mathcal{B}(1/2)$, une censure uniforme et $n = 100$ (1000 répétitions).

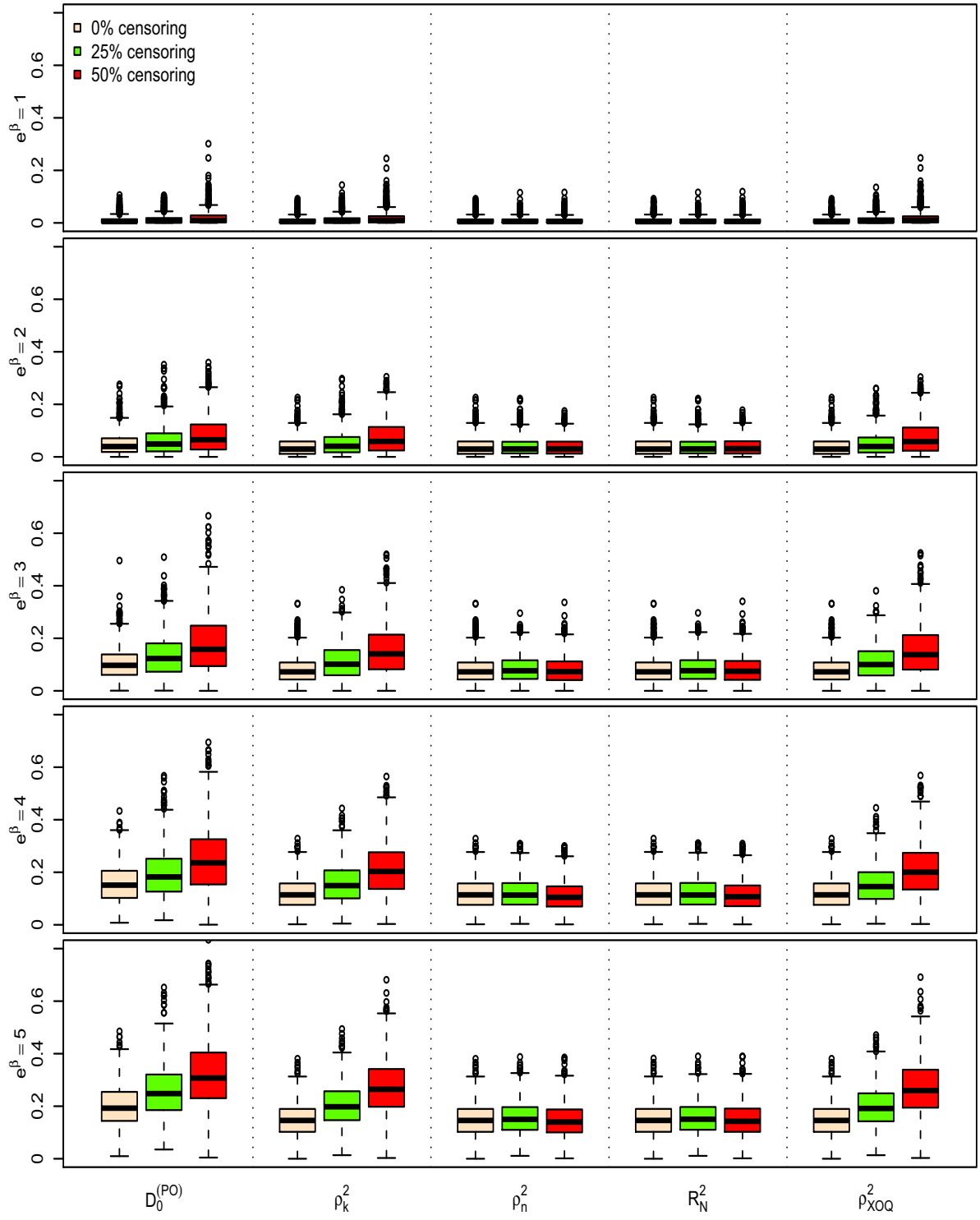


FIGURE B.23 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{PO})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à odds proportionnels, pour différentes valeurs de e^β et différents pourcentages de censure, calculés pour une variable Z uniforme $\mathcal{U}[0, \sqrt{(3)}]$, une censure uniforme et $n = 100$ (1000 répétitions).

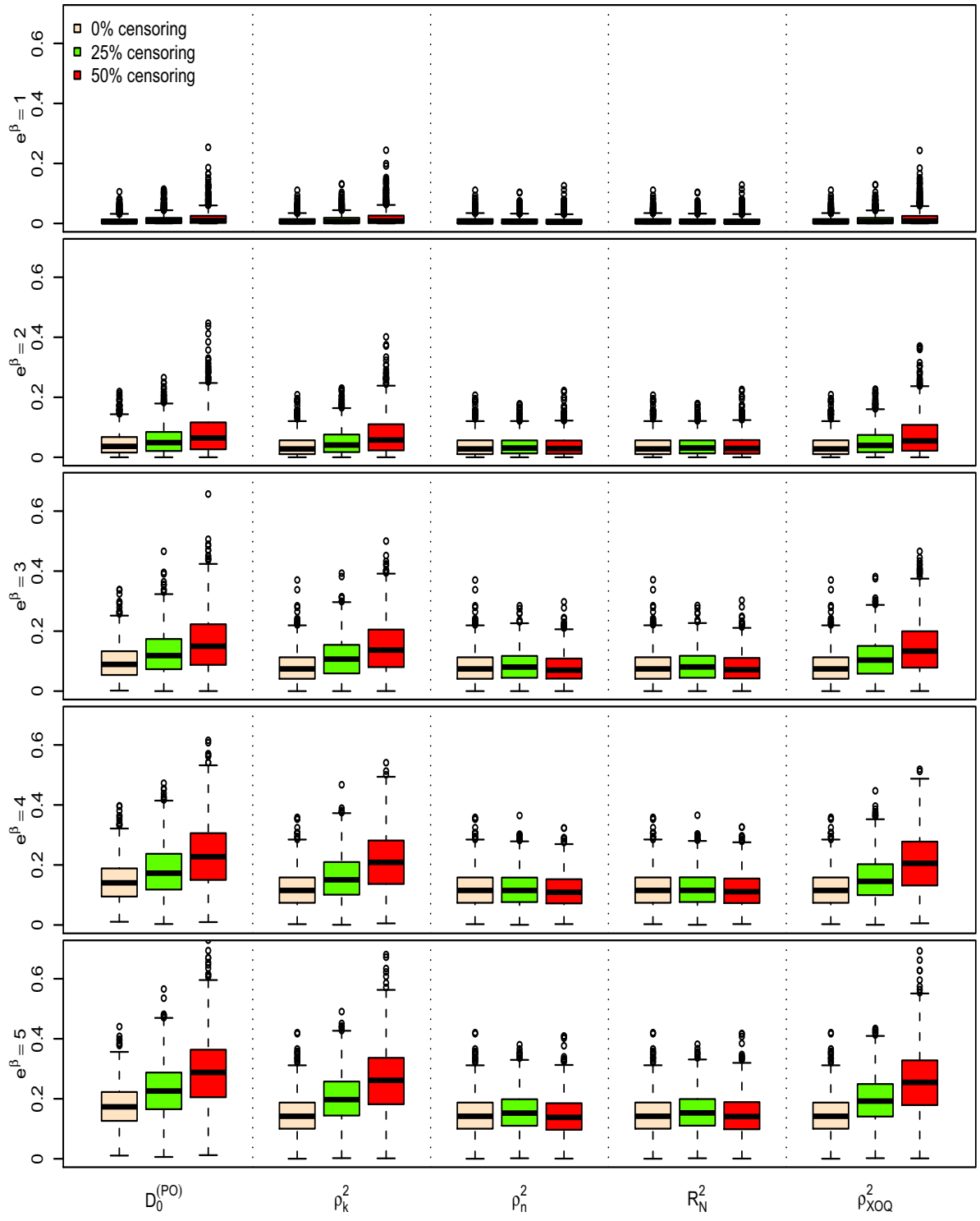


FIGURE B.24 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{PO})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à odds proportionnels, pour différentes valeurs de e^β et différents pourcentages de censure, calculés pour une variable Z normale $\mathcal{N}(0, 1/4)$, une censure uniforme et $n = 100$ (1000 répétitions).

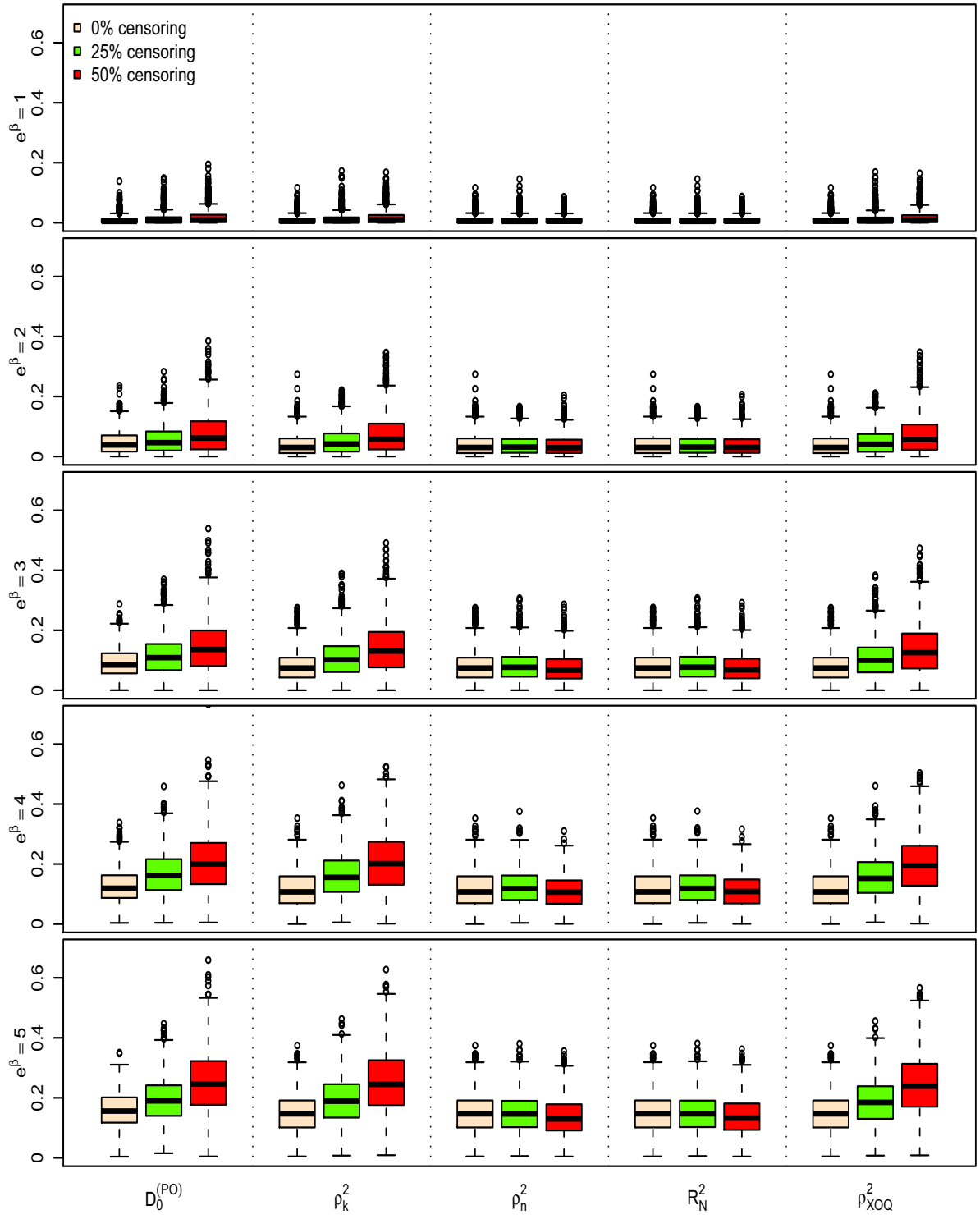


FIGURE B.25 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{PO})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à odds proportionnels, pour différentes valeurs de e^β et différents pourcentages de censure, calculés pour une variable Z de Bernoulli $\mathcal{B}(1/2)$, une censure uniforme et $n = 1000$ (1000 répétitions).

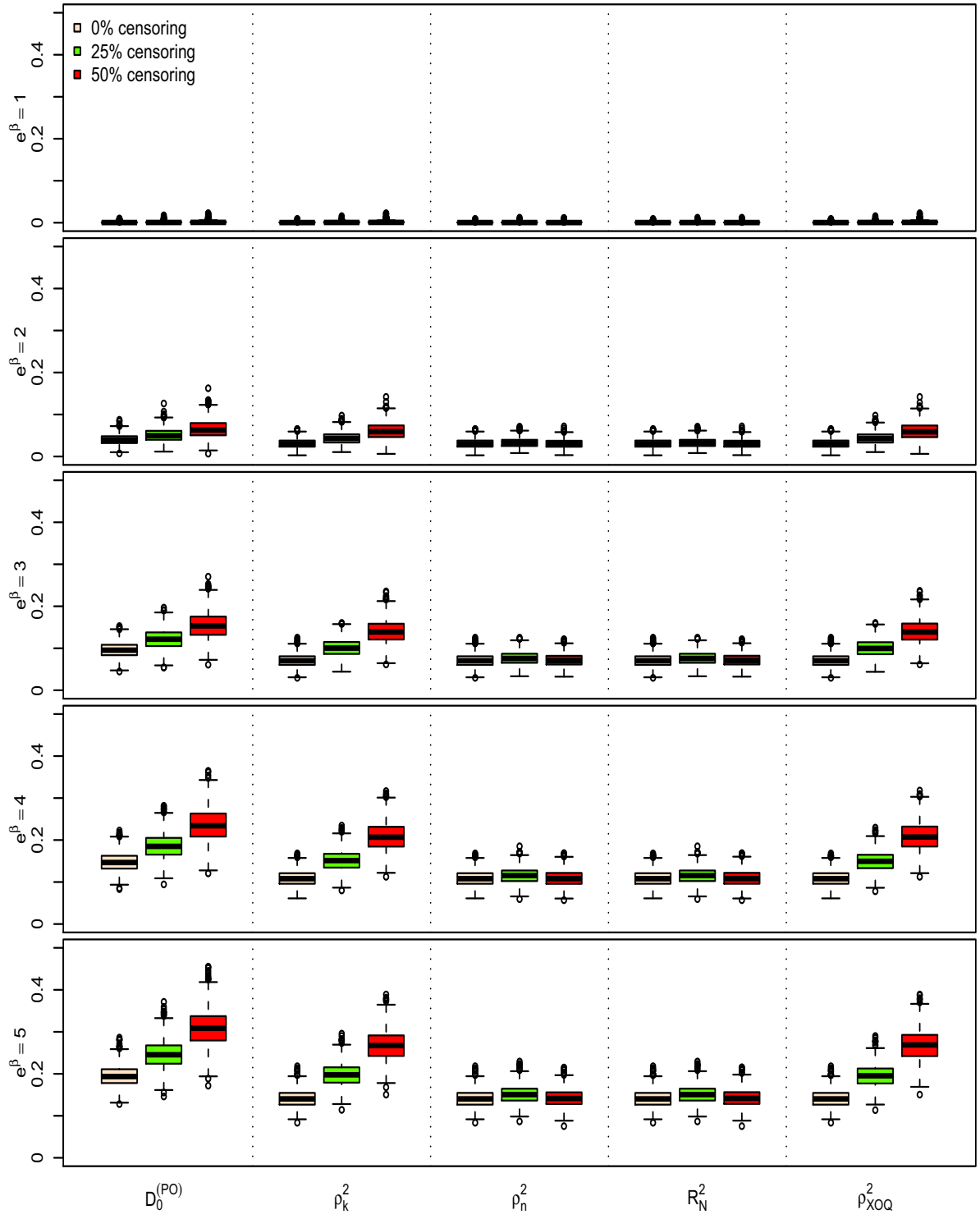


FIGURE B.26 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{PO})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à odds proportionnels, pour différentes valeurs de e^β et différents pourcentages de censure, calculés pour une variable Z uniforme $\mathcal{U}[0, \sqrt{(3)}]$, une censure uniforme et $n = 1000$ (1000 répétitions).

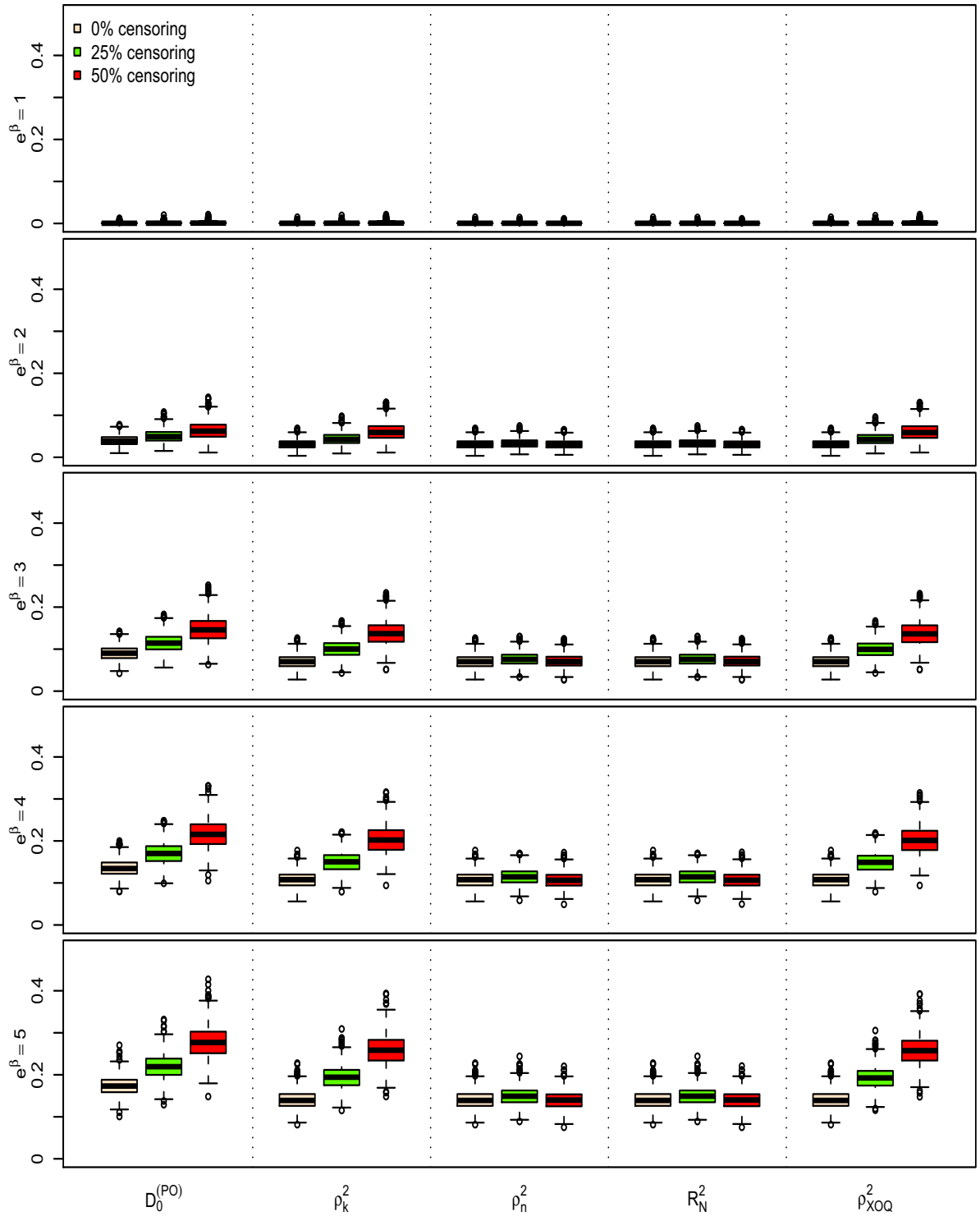


FIGURE B.27 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{PO})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à odds proportionnels, pour différentes valeurs de e^β et différents pourcentages de censure, calculés pour une variable Z normale $\mathcal{N}(0, 1/4)$, une censure uniforme et $n = 1000$ (1000 répétitions).

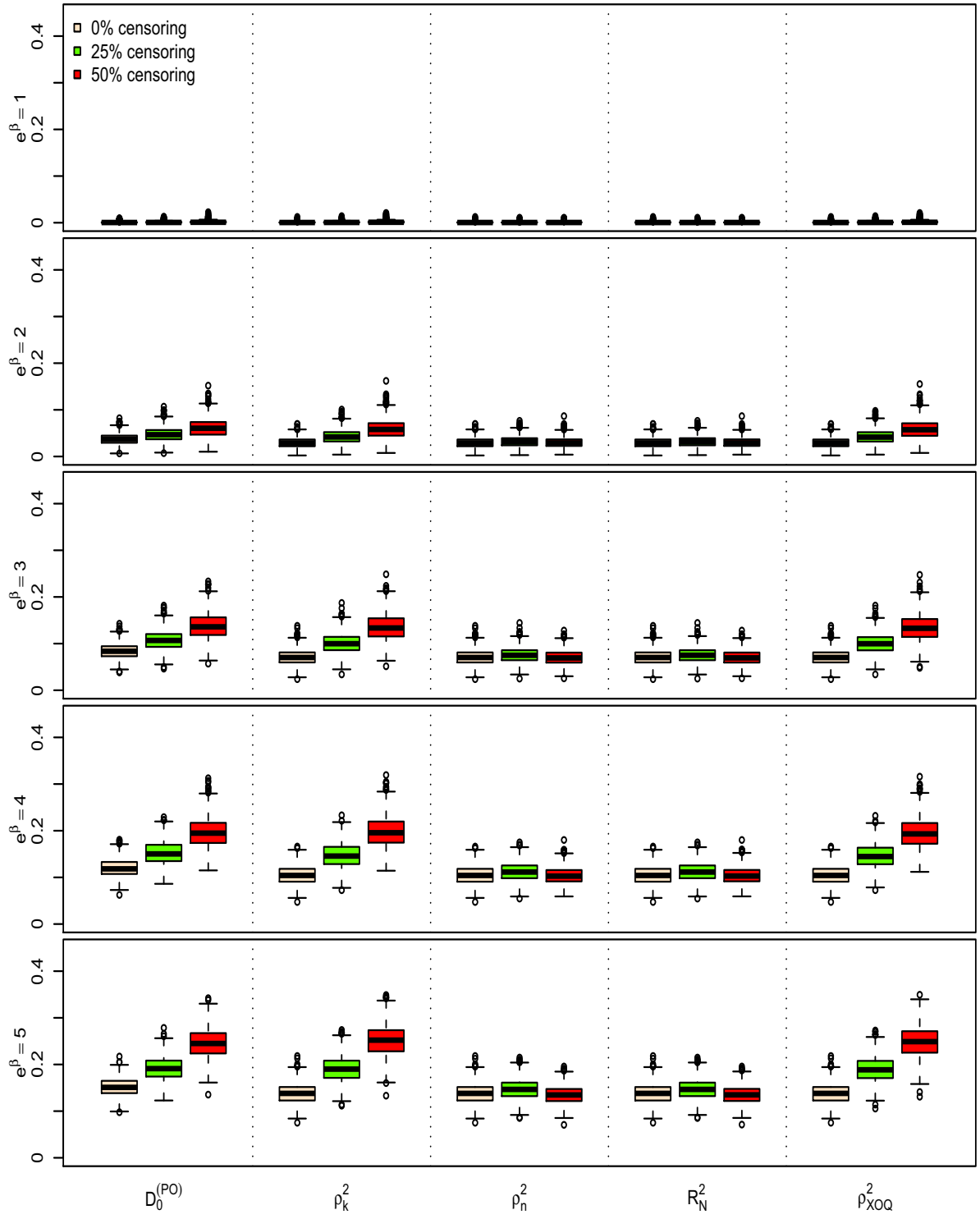


FIGURE B.28 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{PO})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à odds proportionnels, pour différentes valeurs de e^β et différents pourcentages de censure, calculés pour une variable Z de Bernoulli $\mathcal{B}(1/2)$, une censure exponentielle et $n = 50$ (1000 répétitions).

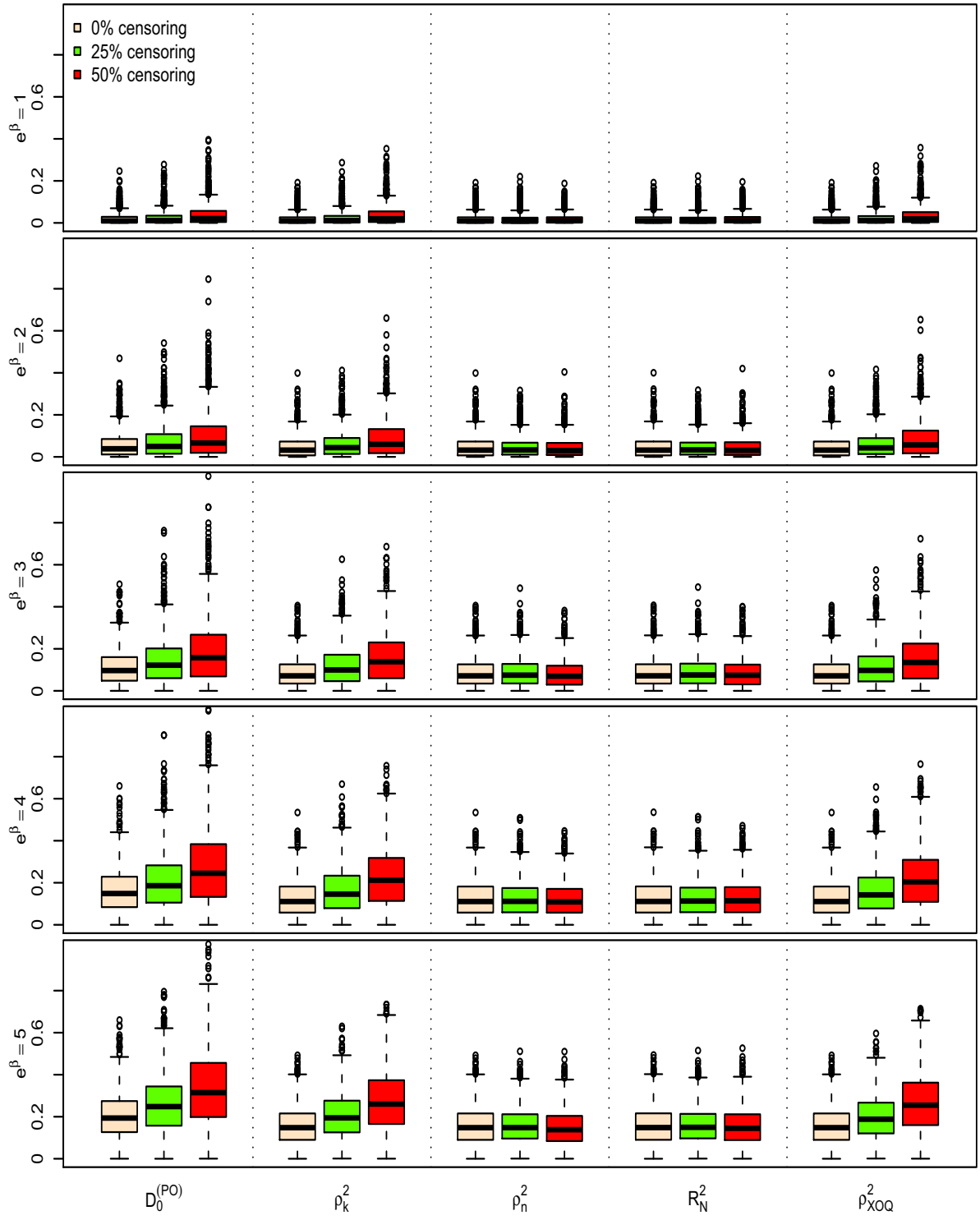


FIGURE B.29 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{PO})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à odds proportionnels, pour différentes valeurs de e^β et différents pourcentages de censure, calculés pour une variable Z uniforme $\mathcal{U}[0, \sqrt{(3)}]$, une censure exponentielle et $n = 50$ (1000 répétitions).

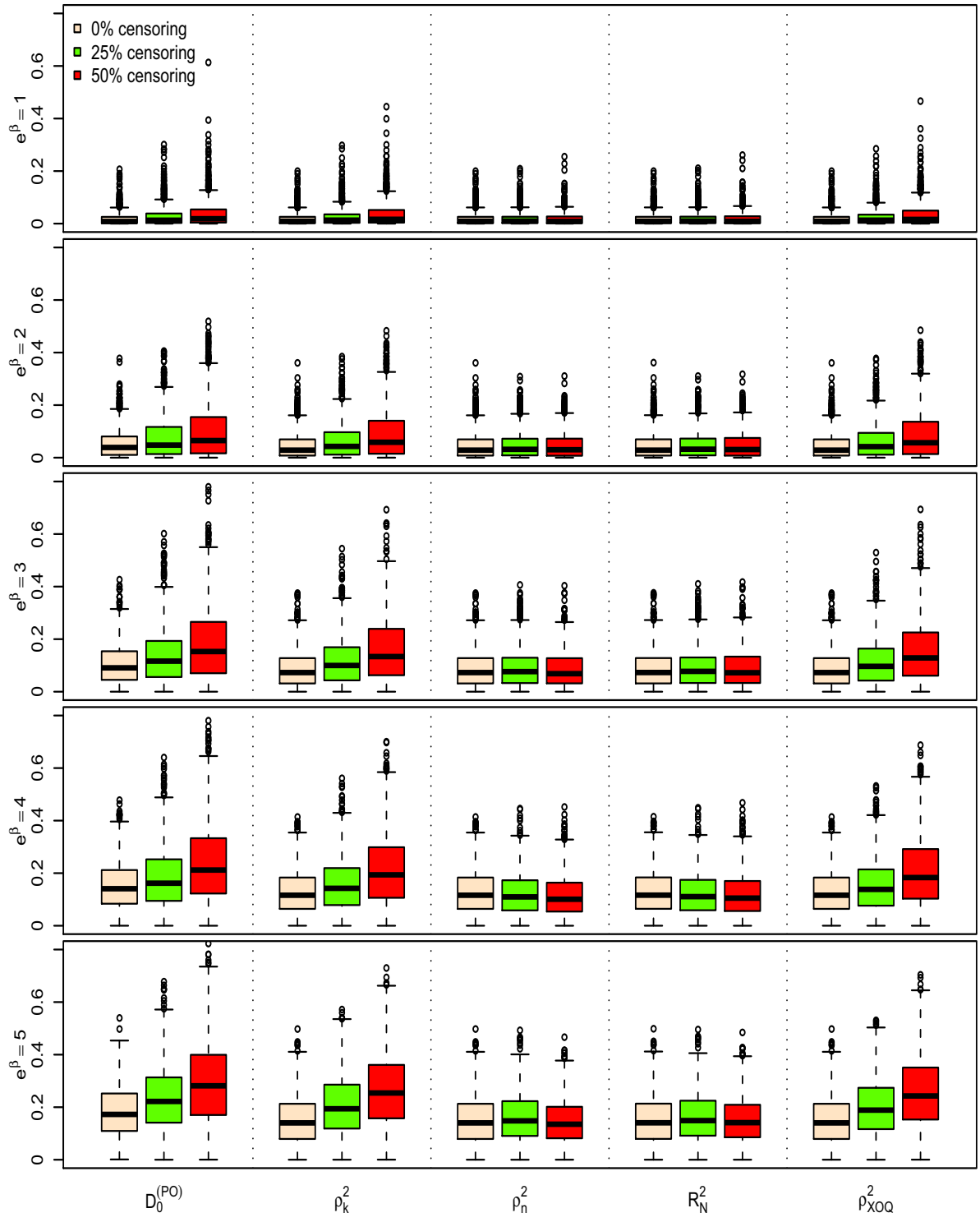


FIGURE B.30 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{PO})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à odds proportionnels, pour différentes valeurs de e^β et différents pourcentages de censure, calculés pour une variable Z normale $\mathcal{N}(0, 1/4)$, une censure exponentielle et $n = 50$ (1000 répétitions).

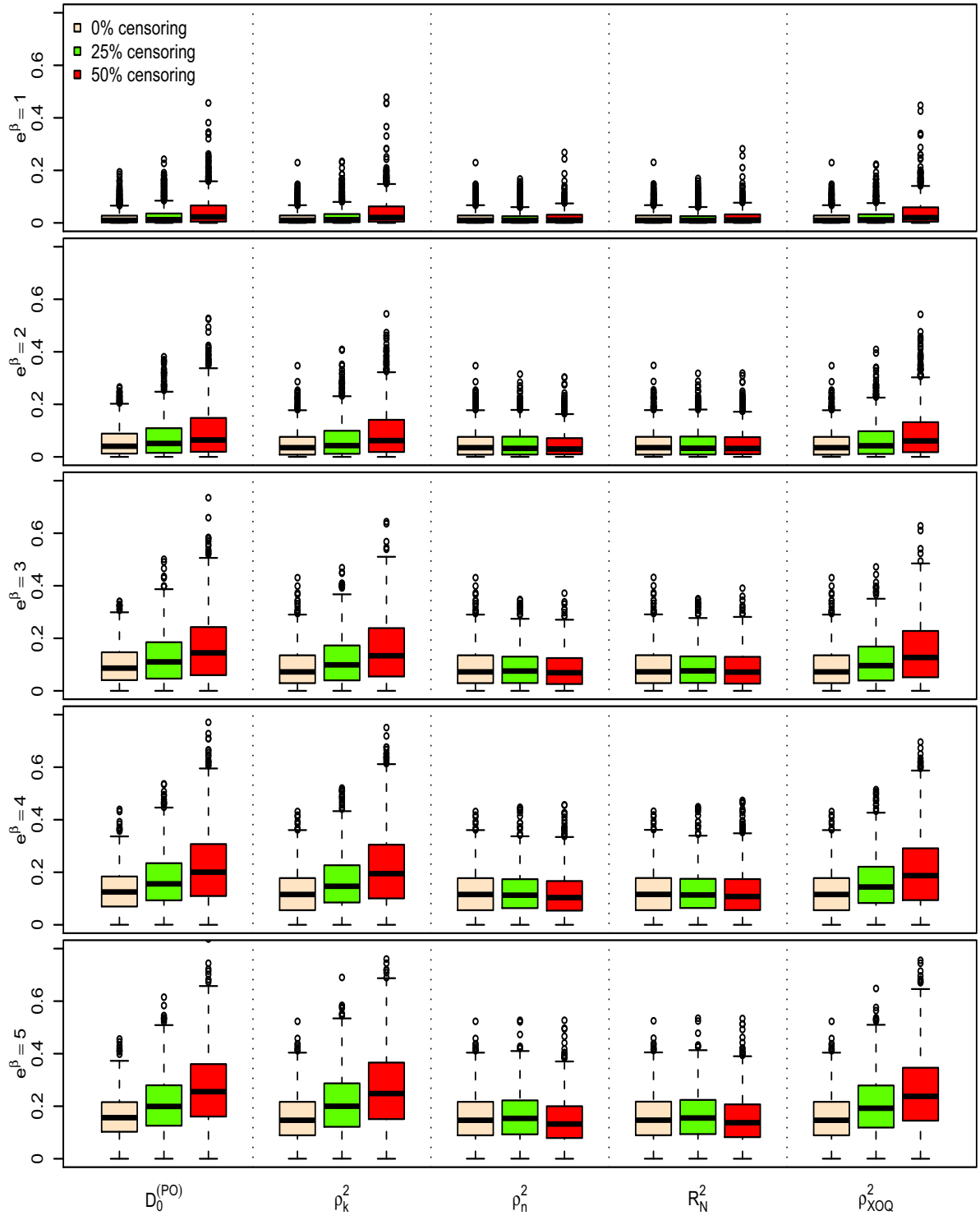


FIGURE B.31 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{PO})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à odds proportionnels, pour différentes valeurs de e^β et différents pourcentages de censure, calculés pour une variable Z de Bernoulli $\mathcal{B}(1/2)$, une censure exponentielle et $n = 100$ (1000 répétitions).

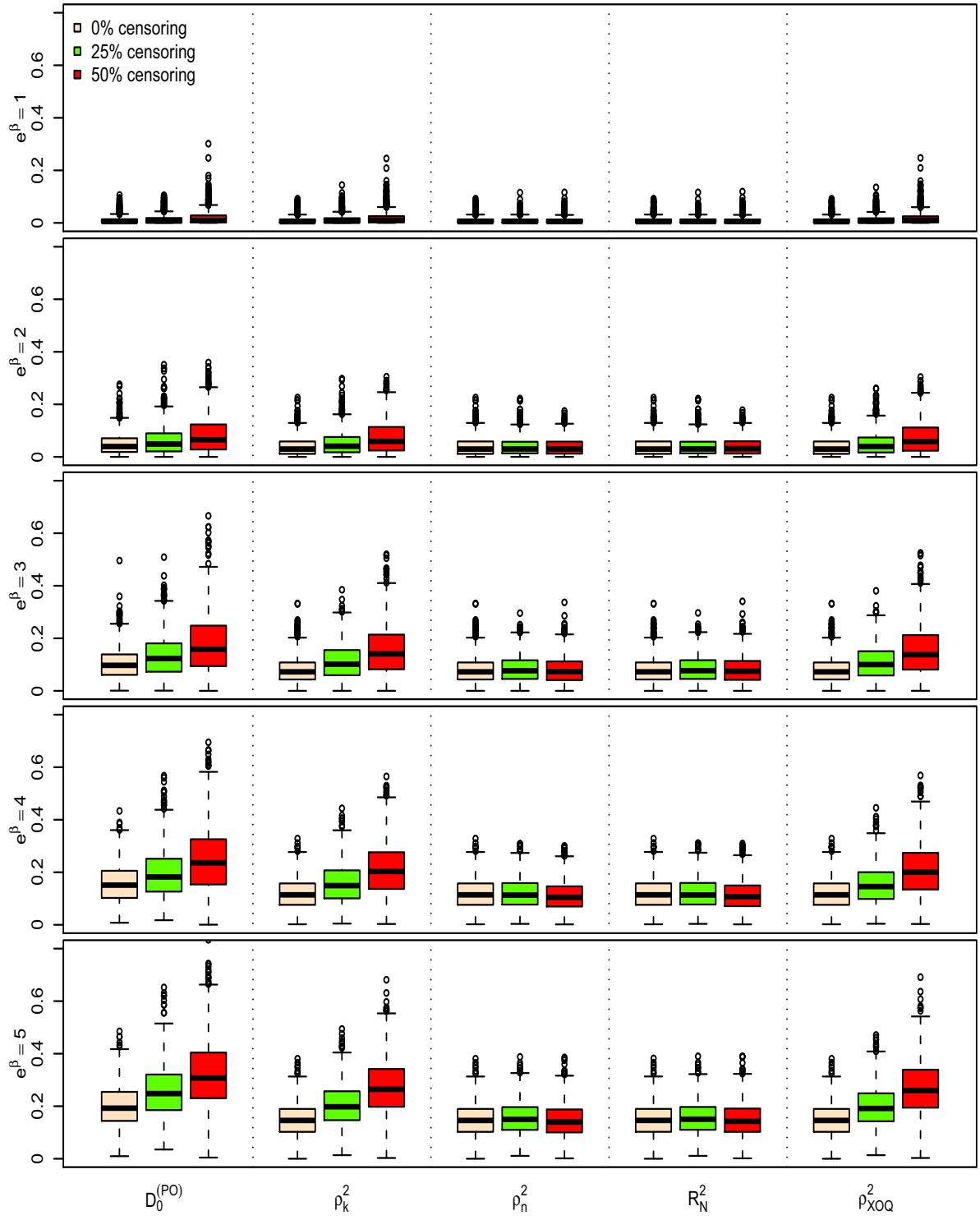


FIGURE B.32 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{PO})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à odds proportionnels, pour différentes valeurs de e^β et différents pourcentages de censure, calculés pour une variable Z uniforme $\mathcal{U}[0, \sqrt{(3)}]$, une censure exponentielle et $n = 100$ (1000 répétitions).

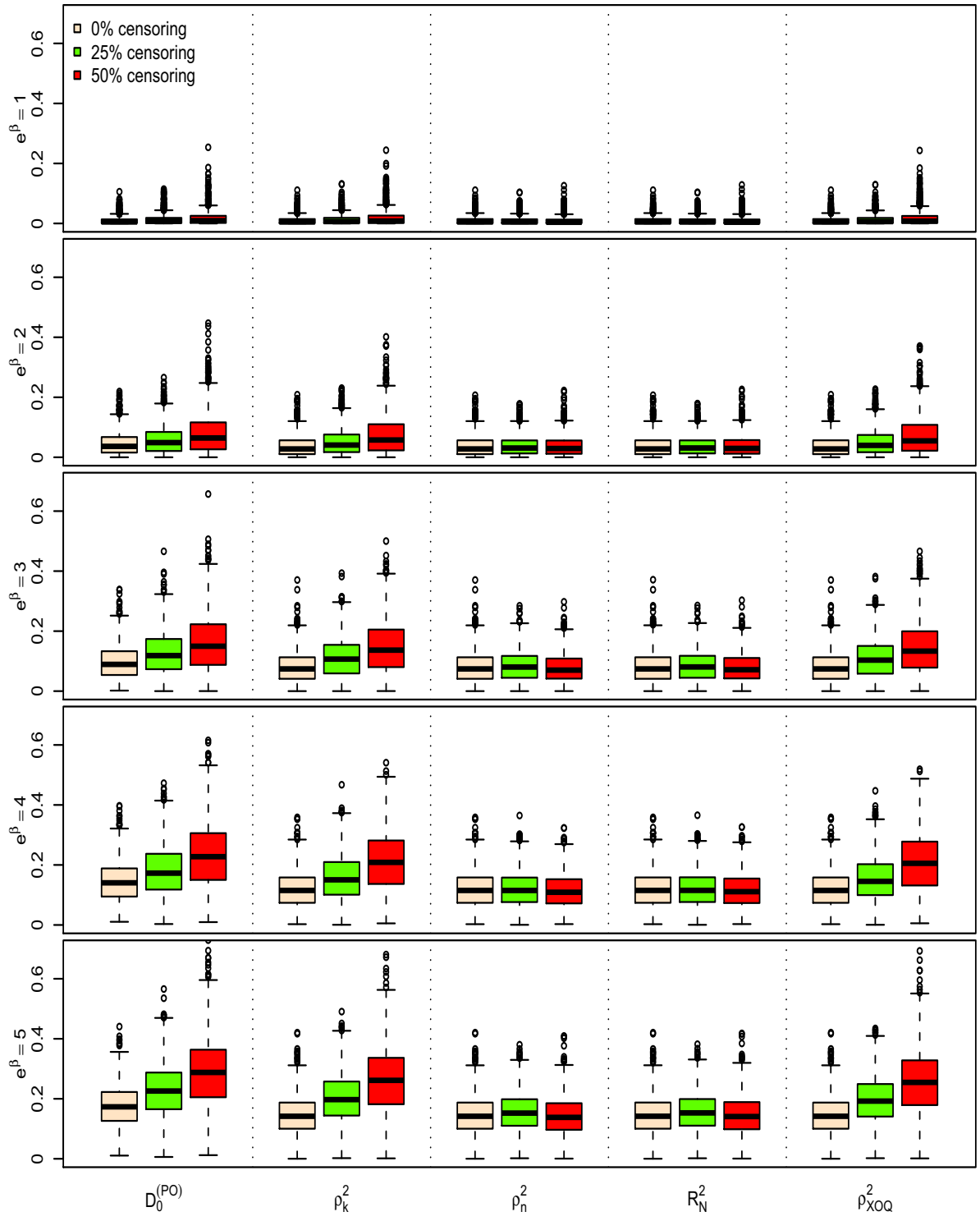


FIGURE B.33 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{PO})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à odds proportionnels, pour différentes valeurs de e^β et différents pourcentages de censure, calculés pour une variable Z normale $\mathcal{N}(0, 1/4)$, une censure exponentielle et $n = 100$ (1000 répétitions).

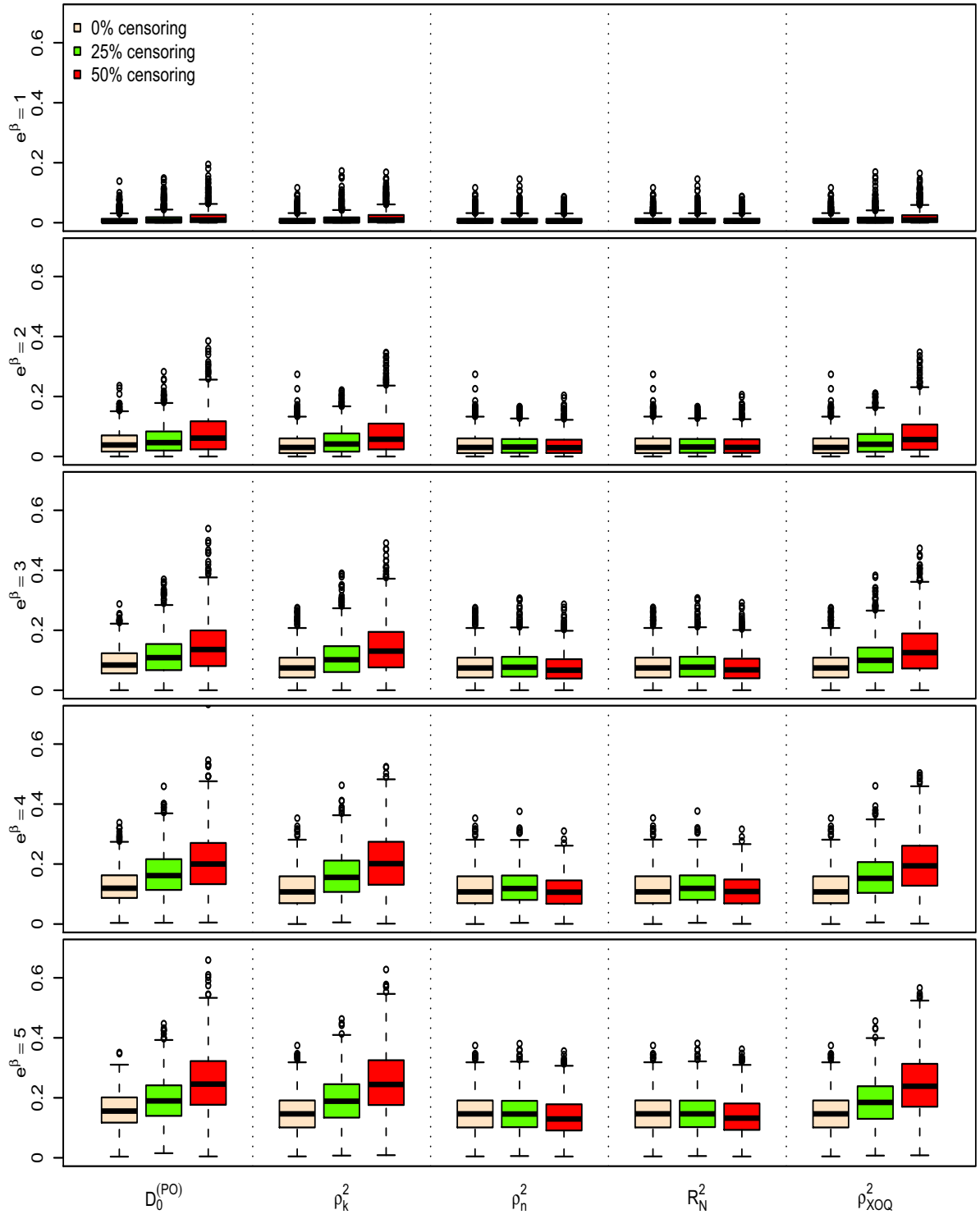


FIGURE B.34 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{PO})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à odds proportionnels, pour différentes valeurs de e^β et différents pourcentages de censure, calculés pour une variable Z de Bernoulli $\mathcal{B}(1/2)$, une censure exponentielle et $n = 1000$ (1000 répétitions).

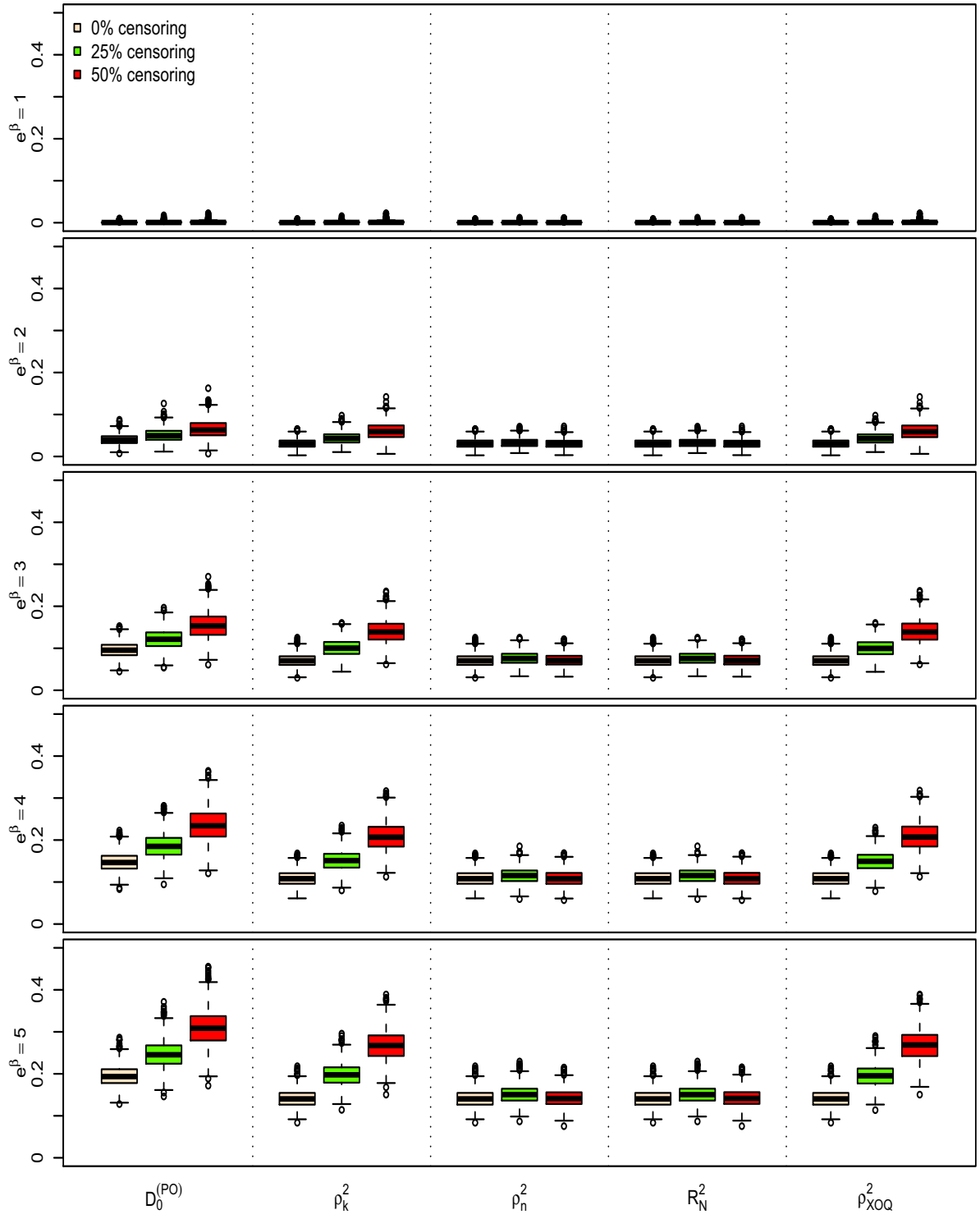


FIGURE B.35 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{PO})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à odds proportionnels, pour différentes valeurs de e^β et différents pourcentages de censure, calculés pour une variable Z uniforme $\mathcal{U}[0, \sqrt(3)]$, une censure exponentielle et $n = 1000$ (1000 répétitions).

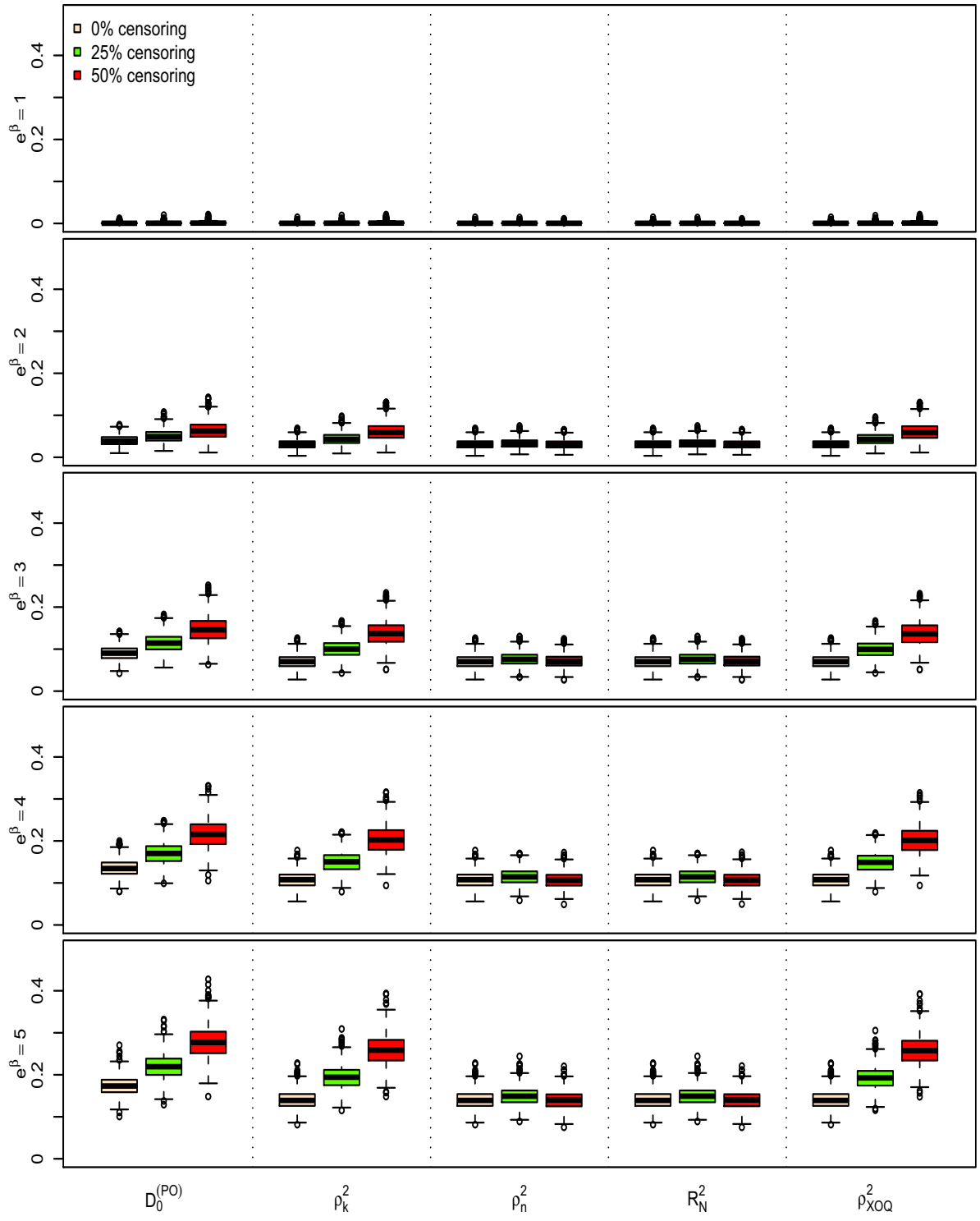


FIGURE B.36 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{PO})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à odds proportionnels, pour différentes valeurs de e^β et différents pourcentages de censure, calculés pour une variable Z normale $\mathcal{N}(0, 1/4)$, une censure exponentielle et $n = 1000$ (1000 répétitions).

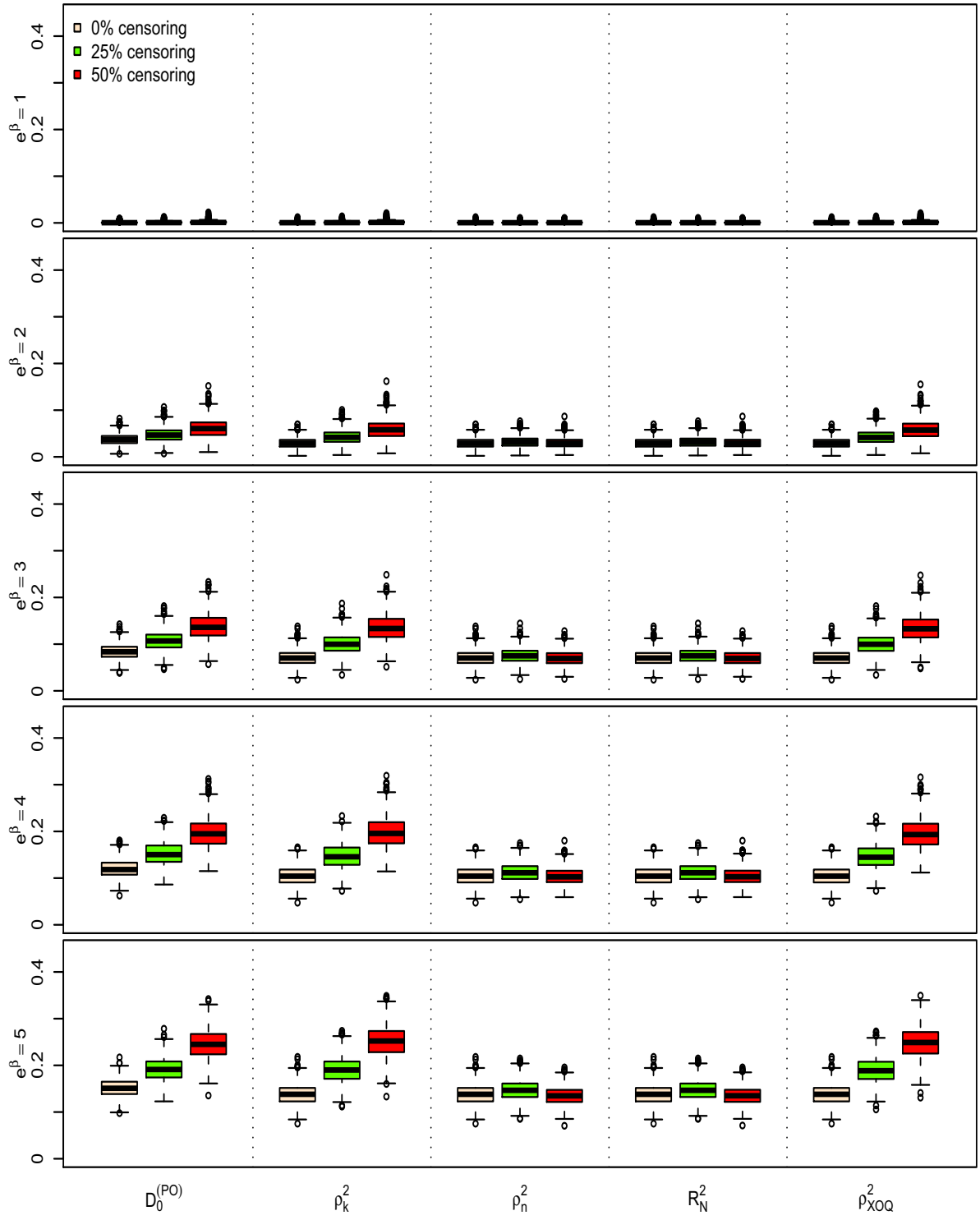


TABLEAU B.7 – Valeurs moyennes de $\mathbf{D}_0^{(\mathrm{NPH})}$ sous un modèle à risques non-proportionnels conduisant à un croisement des risques instantanés, pour différentes valeurs de e^{β}, différents pourcentages de censure p_c, différentes tailles d'échantillon n, différents types de censure, calculées pour une variable de Bernoulli $Z \sim \mathcal{B}(1/2)$ (1000 répétitions). Les écarts-types sont indiqués entre parenthèses.

β	p_c	$C \sim \mathcal{U}[0,r]$				$C \sim \mathcal{E}(\gamma)$			
		$\mathbf{D}_0^{(\mathrm{NPH})}$ ($n=50$)	$\mathbf{D}_0^{(\mathrm{NPH})}$ ($n=100$)	$\mathbf{D}_0^{(\mathrm{NPH})}$ ($n=500$)	$\mathbf{D}_0^{(\mathrm{NPH})}$ ($n=1000$)	$\mathbf{D}_0^{(\mathrm{NPH})}$ ($n=50$)	$\mathbf{D}_0^{(\mathrm{NPH})}$ ($n=100$)	$\mathbf{D}_0^{(\mathrm{NPH})}$ ($n=500$)	$\mathbf{D}_0^{(\mathrm{NPH})}$ ($n=1000$)
1	0	0.0245(0.0350)	0.0114(0.0181)	0.0020(0.0028)	0.0010(0.0014)	0.0231(0.0327)	0.0109(0.0164)	0.0021(0.0029)	0.0011(0.0015)
	0.25	0.0295(0.0403)	0.0140(0.0201)	0.0025(0.0036)	0.0014(0.0019)	0.0294(0.0394)	0.0137(0.0192)	0.0026(0.0036)	0.0013(0.0018)
	0.50	0.0430(0.0595)	0.0211(0.0274)	0.0039(0.0055)	0.0021(0.0029)	0.0430(0.0567)	0.0199(0.0270)	0.0040(0.0054)	0.0019(0.0026)
1.25	0	0.0390(0.0476)	0.0293(0.0320)	0.0242(0.0151)	0.0243(0.0107)	0.0378(0.0486)	0.0302(0.0336)	0.0238(0.0142)	0.0240(0.0102)
	0.25	0.0394(0.0495)	0.0279(0.0334)	0.0210(0.0148)	0.0210(0.0104)	0.0415(0.0532)	0.0299(0.0364)	0.0215(0.0152)	0.0209(0.0103)
	0.50	0.0463(0.0587)	0.0309(0.0387)	0.0201(0.0163)	0.0193(0.0113)	0.0494(0.0617)	0.0343(0.0433)	0.0223(0.0178)	0.0211(0.0119)
1.5	0	0.0709(0.0686)	0.0661(0.0483)	0.0744(0.0247)	0.0763(0.0171)	0.0650(0.0643)	0.0691(0.0506)	0.0730(0.0238)	0.0760(0.0173)
	0.25	0.0646(0.0687)	0.0576(0.0473)	0.0652(0.0251)	0.0664(0.0178)	0.0620(0.0672)	0.0610(0.0516)	0.0635(0.0252)	0.0664(0.0178)
	0.50	0.0641(0.0746)	0.0515(0.0506)	0.0565(0.0259)	0.0570(0.0190)	0.0665(0.0764)	0.0587(0.0543)	0.0624(0.0273)	0.0635(0.0198)
1.75	0	0.1017(0.0764)	0.1182(0.0622)	0.1308(0.0281)	0.1332(0.0206)	0.1017(0.0788)	0.1156(0.0592)	0.1306(0.0288)	0.1323(0.0201)
	0.25	0.0935(0.0786)	0.1066(0.0645)	0.1191(0.0296)	0.1214(0.0213)	0.0920(0.0824)	0.1017(0.0630)	0.1159(0.0308)	0.1186(0.0217)
	0.50	0.0852(0.0817)	0.0934(0.0645)	0.1005(0.0300)	0.1020(0.0211)	0.0909(0.0894)	0.0963(0.0723)	0.1120(0.0346)	0.1133(0.0248)
2	0	0.1404(0.0856)	0.1601(0.0648)	0.1788(0.0285)	0.1826(0.0207)	0.1418(0.0870)	0.1598(0.0662)	0.1800(0.0295)	0.1824(0.0211)
	0.25	0.1303(0.0897)	0.1501(0.0694)	0.1684(0.0313)	0.1725(0.0223)	0.1241(0.0904)	0.1438(0.0714)	0.1662(0.0335)	0.1690(0.0236)
	0.50	0.1124(0.0908)	0.1290(0.0726)	0.1464(0.0333)	0.1497(0.0249)	0.1198(0.0949)	0.1333(0.0777)	0.1598(0.0385)	0.1636(0.0273)
3	0	0.2590(0.0911)	0.2910(0.0648)	0.3123(0.0255)	0.3114(0.0182)	0.2555(0.0901)	0.2866(0.0639)	0.3113(0.0263)	0.3143(0.0178)
	0.25	0.2471(0.0989)	0.2847(0.0714)	0.3096(0.0283)	0.3092(0.0201)	0.2367(0.1044)	0.2724(0.0755)	0.3058(0.0318)	0.3088(0.0222)
	0.50	0.2301(0.1089)	0.2721(0.0847)	0.3027(0.0342)	0.3016(0.0240)	0.2138(0.1129)	0.2615(0.0855)	0.3019(0.0373)	0.3086(0.0260)
4	0	0.3217(0.0872)	0.3612(0.0589)	0.3807(0.0226)	0.3804(0.0157)	0.3196(0.0858)	0.3577(0.0562)	0.3791(0.0226)	0.3808(0.0156)
	0.25	0.3165(0.0919)	0.3592(0.0645)	0.3826(0.0245)	0.3822(0.0174)	0.3056(0.1014)	0.3527(0.0682)	0.3803(0.0274)	0.3825(0.0192)
	0.50	0.3046(0.1021)	0.3581(0.0744)	0.3862(0.0295)	0.3857(0.0212)	0.2826(0.1129)	0.3398(0.0858)	0.3832(0.0347)	0.3867(0.0234)
5	0	0.3601(0.0844)	0.4024(0.0531)	0.4210(0.0199)	0.4217(0.0142)	0.3684(0.0819)	0.3979(0.0541)	0.4210(0.0204)	0.4214(0.0140)
	0.25	0.3580(0.0899)	0.4030(0.0579)	0.4242(0.0216)	0.4251(0.0153)	0.3530(0.0952)	0.3961(0.0662)	0.4266(0.0254)	0.4282(0.0176)
	0.50	0.3524(0.1002)	0.4057(0.0656)	0.4327(0.0256)	0.4337(0.0183)	0.3258(0.1119)	0.3875(0.0770)	0.4333(0.0305)	0.4363(0.0216)

TABLEAU B.8 – Valeurs moyennes de $\mathbf{D}_0^{(\mathrm{NPH})}$ sous un modèle à risques non-proportionnels conduisant à un croisement des risques instantanés, pour différentes valeurs de e^{β}, différents pourcentages de censure p_c, différentes tailles d'échantillon n, différents types de censure, calculées pour une variable uniforme $Z \sim \mathcal{U}[0, \sqrt{(3)}]$ (1000 répétitions). Les écarts-types sont indiqués entre parenthèses.

		$C \sim \mathcal{U}[0,r]$				$C \sim \mathcal{E}(\gamma)$			
β	p_c	$\mathbf{D}_0^{(\mathrm{NPH})}$ ($n=50$)	$\mathbf{D}_0^{(\mathrm{NPH})}$ ($n=100$)	$\mathbf{D}_0^{(\mathrm{NPH})}$ ($n=500$)	$\mathbf{D}_0^{(\mathrm{NPH})}$ ($n=1000$)	$\mathbf{D}_0^{(\mathrm{NPH})}$ ($n=50$)	$\mathbf{D}_0^{(\mathrm{NPH})}$ ($n=100$)	$\mathbf{D}_0^{(\mathrm{NPH})}$ ($n=500$)	$\mathbf{D}_0^{(\mathrm{NPH})}$ ($n=1000$)
1	0	0.0244(0.0323)	0.0105(0.0153)	0.0021(0.0031)	0.0010(0.0014)	0.0240(0.0318)	0.0106(0.0150)	0.0021(0.0030)	0.0011(0.0015)
	0.25	0.0287(0.0370)	0.0136(0.0188)	0.0027(0.0038)	0.0014(0.0019)	0.0304(0.0403)	0.0137(0.0195)	0.0027(0.0038)	0.0014(0.0019)
	0.50	0.0416(0.0522)	0.0203(0.0267)	0.0039(0.0053)	0.0021(0.0029)	0.0435(0.0545)	0.0197(0.0268)	0.0041(0.0055)	0.0019(0.0026)
1.25	0	0.0348(0.0414)	0.0276(0.0294)	0.0226(0.0136)	0.0228(0.0096)	0.0365(0.0454)	0.0283(0.0300)	0.0221(0.0129)	0.0227(0.0096)
	0.25	0.0371(0.0464)	0.0271(0.0310)	0.0207(0.0144)	0.0209(0.0102)	0.0419(0.0509)	0.0294(0.0331)	0.0208(0.0139)	0.0208(0.0103)
	0.50	0.0445(0.0543)	0.0310(0.0370)	0.0193(0.0153)	0.0192(0.0106)	0.0518(0.0616)	0.0338(0.0392)	0.0220(0.0172)	0.0212(0.0122)
1.5	0	0.0638(0.0603)	0.0606(0.0434)	0.0644(0.0215)	0.0658(0.0144)	0.0604(0.0593)	0.0623(0.0441)	0.0642(0.0207)	0.0662(0.0148)
	0.25	0.0613(0.0628)	0.0572(0.0448)	0.0607(0.0225)	0.0617(0.0156)	0.0591(0.0656)	0.0579(0.0474)	0.0592(0.0222)	0.0614(0.0160)
	0.50	0.0590(0.0640)	0.0513(0.0460)	0.0542(0.0236)	0.0547(0.0163)	0.0639(0.0723)	0.0581(0.0521)	0.0591(0.0248)	0.0604(0.0183)
1.75	0	0.0921(0.0676)	0.1005(0.0519)	0.1090(0.0235)	0.1107(0.0171)	0.0896(0.0682)	0.0981(0.0512)	0.1088(0.0240)	0.1101(0.0174)
	0.25	0.0887(0.0708)	0.0955(0.0538)	0.1051(0.0241)	0.1065(0.0179)	0.0843(0.0738)	0.0915(0.0546)	0.1021(0.0255)	0.1037(0.0187)
	0.50	0.0842(0.0765)	0.0893(0.0558)	0.0971(0.0259)	0.0976(0.0183)	0.0858(0.0806)	0.0876(0.0619)	0.1011(0.0290)	0.1018(0.0213)
2	0	0.1208(0.0759)	0.1346(0.0570)	0.1444(0.0244)	0.1473(0.0182)	0.1221(0.0775)	0.1347(0.0566)	0.1463(0.0253)	0.1475(0.0178)
	0.25	0.1163(0.0775)	0.1294(0.0585)	0.1407(0.0256)	0.1439(0.0188)	0.1109(0.0803)	0.1247(0.0609)	0.1402(0.0275)	0.1421(0.0195)
	0.50	0.1079(0.0823)	0.1214(0.0625)	0.1333(0.0269)	0.1363(0.0200)	0.1033(0.0863)	0.1196(0.0679)	0.1387(0.0318)	0.1403(0.0220)
3	0	0.2105(0.0806)	0.2295(0.0584)	0.2437(0.0239)	0.2423(0.0174)	0.2069(0.0793)	0.2290(0.0569)	0.2416(0.0241)	0.2440(0.0176)
	0.25	0.2060(0.0842)	0.2261(0.0609)	0.2423(0.0247)	0.2406(0.0182)	0.1953(0.0902)	0.2194(0.0640)	0.2367(0.0270)	0.2402(0.0196)
	0.50	0.1957(0.0889)	0.2206(0.0648)	0.2393(0.0267)	0.2377(0.0192)	0.1729(0.0999)	0.2088(0.0721)	0.2339(0.0311)	0.2395(0.0218)
4	0	0.2549(0.0774)	0.2812(0.0544)	0.2930(0.0232)	0.2914(0.0162)	0.2543(0.0773)	0.2754(0.0541)	0.2918(0.0239)	0.2911(0.0169)
	0.25	0.2512(0.0776)	0.2791(0.0563)	0.2922(0.0236)	0.2905(0.0168)	0.2409(0.0875)	0.2686(0.0598)	0.2870(0.0267)	0.2871(0.0191)
	0.50	0.2438(0.0809)	0.2747(0.0592)	0.2903(0.0256)	0.2888(0.0175)	0.2144(0.0985)	0.2556(0.0707)	0.2855(0.0308)	0.2859(0.0213)
5	0	0.2815(0.0755)	0.3150(0.0521)	0.3228(0.0228)	0.3213(0.0157)	0.2898(0.0735)	0.3089(0.0520)	0.3216(0.0223)	0.3207(0.0153)
	0.25	0.2791(0.0772)	0.3133(0.0532)	0.3222(0.0233)	0.3204(0.0161)	0.2708(0.0828)	0.2999(0.0571)	0.3168(0.0247)	0.3162(0.0175)
	0.50	0.2734(0.0803)	0.3103(0.0553)	0.3209(0.0245)	0.3192(0.0165)	0.2382(0.0981)	0.2861(0.0666)	0.3149(0.0284)	0.3145(0.0196)

TABLEAU B.9 – Valeurs moyennes de $\mathbf{D}_0^{(\mathrm{NPH})}$ sous un modèle à risques non-proportionnels conduisant à un croisement des risques instantanés, pour différentes valeurs de e^{β}, différents pourcentages de censure p_c, différentes tailles d'échantillon n, différents types de censure, calculées pour une variable normale $Z \sim \mathcal{N}(0, 1/4)$ (1000 répétitions). Les écarts-types sont indiqués entre parenthèses.

		$C \sim \mathcal{U}[0,r]$				$C \sim \mathcal{E}(\gamma)$			
β	p_c	$\mathbf{D}_0^{(\mathrm{NPH})}$ ($n=50$)	$\mathbf{D}_0^{(\mathrm{NPH})}$ ($n=100$)	$\mathbf{D}_0^{(\mathrm{NPH})}$ ($n=500$)	$\mathbf{D}_0^{(\mathrm{NPH})}$ ($n=1000$)	$\mathbf{D}_0^{(\mathrm{NPH})}$ ($n=50$)	$\mathbf{D}_0^{(\mathrm{NPH})}$ ($n=100$)	$\mathbf{D}_0^{(\mathrm{NPH})}$ ($n=500$)	$\mathbf{D}_0^{(\mathrm{NPH})}$ ($n=1000$)
1	0	0.0244(0.0316)	0.0110(0.0156)	0.0021(0.0030)	0.0011(0.0017)	0.0217(0.0279)	0.0104(0.0144)	0.0021(0.0031)	0.0010(0.0014)
	0.25	0.0300(0.0390)	0.0139(0.0185)	0.0026(0.0037)	0.0014(0.0019)	0.0286(0.0358)	0.0137(0.0187)	0.0028(0.0040)	0.0013(0.0018)
	0.50	0.0447(0.0541)	0.0201(0.0275)	0.0040(0.0060)	0.0020(0.0027)	0.0404(0.0484)	0.0200(0.0249)	0.0041(0.0053)	0.0020(0.0027)
1.25	0	0.0345(0.0447)	0.0263(0.0279)	0.0211(0.0121)	0.0209(0.0086)	0.0352(0.0421)	0.0272(0.0292)	0.0213(0.0117)	0.0205(0.0079)
	0.25	0.0360(0.0473)	0.0260(0.0300)	0.0192(0.0130)	0.0185(0.0092)	0.0389(0.0462)	0.0276(0.0311)	0.0200(0.0130)	0.0191(0.0089)
	0.50	0.0477(0.0599)	0.0296(0.0344)	0.0190(0.0148)	0.0180(0.0103)	0.0501(0.0566)	0.0312(0.0356)	0.0209(0.0155)	0.0196(0.0111)
1.5	0	0.0588(0.0535)	0.0538(0.0400)	0.0568(0.0169)	0.0562(0.0126)	0.0576(0.0552)	0.0567(0.0391)	0.0560(0.0175)	0.0556(0.0124)
	0.25	0.0560(0.0558)	0.0491(0.0419)	0.0508(0.0185)	0.0510(0.0137)	0.0570(0.0592)	0.0496(0.0406)	0.0524(0.0187)	0.0526(0.0142)
	0.50	0.0580(0.0606)	0.0486(0.0455)	0.0478(0.0202)	0.0482(0.0151)	0.0635(0.0677)	0.0502(0.0470)	0.0528(0.0217)	0.0530(0.0162)
1.75	0	0.0859(0.0653)	0.0851(0.0417)	0.0874(0.0199)	0.0869(0.0142)	0.0815(0.0601)	0.0843(0.0434)	0.0882(0.0204)	0.0878(0.0143)
	0.25	0.0783(0.0685)	0.0774(0.0461)	0.0809(0.0209)	0.0813(0.0152)	0.0745(0.0660)	0.0775(0.0480)	0.0838(0.0225)	0.0837(0.0154)
	0.50	0.0750(0.0731)	0.0688(0.0491)	0.0750(0.0229)	0.0745(0.0159)	0.0734(0.0710)	0.0747(0.0534)	0.0820(0.0256)	0.0832(0.0175)
2	0	0.1047(0.0680)	0.1141(0.0481)	0.1145(0.0220)	0.1137(0.0155)	0.1063(0.0669)	0.1136(0.0484)	0.1137(0.0212)	0.1129(0.0157)
	0.25	0.0942(0.0714)	0.1060(0.0522)	0.1083(0.0232)	0.1078(0.0159)	0.0975(0.0709)	0.1054(0.0512)	0.1089(0.0230)	0.1082(0.0159)
	0.50	0.0812(0.0685)	0.0943(0.0539)	0.0986(0.0247)	0.0977(0.0167)	0.0896(0.0759)	0.0987(0.0554)	0.1069(0.0253)	0.1063(0.0177)
3	0	0.1687(0.0726)	0.1810(0.0492)	0.1793(0.0242)	0.1755(0.0175)	0.1708(0.0731)	0.1824(0.0496)	0.1796(0.0233)	0.1753(0.0172)
	0.25	0.1569(0.0791)	0.1703(0.0539)	0.1743(0.0251)	0.1705(0.0178)	0.1517(0.0800)	0.1717(0.0545)	0.1749(0.0243)	0.1719(0.0181)
	0.50	0.1283(0.0776)	0.1443(0.0550)	0.1539(0.0248)	0.1512(0.0171)	0.1326(0.0830)	0.1578(0.0587)	0.1691(0.0259)	0.1653(0.0190)
4	0	0.2090(0.0744)	0.2212(0.0535)	0.2117(0.0246)	0.2077(0.0178)	0.2092(0.0723)	0.2181(0.0528)	0.2113(0.0243)	0.2058(0.0177)
	0.25	0.1955(0.0829)	0.2120(0.0577)	0.2085(0.0256)	0.2050(0.0187)	0.1914(0.0775)	0.2096(0.0567)	0.2090(0.0256)	0.2036(0.0180)
	0.50	0.1561(0.0800)	0.1802(0.0543)	0.1839(0.0235)	0.1805(0.0179)	0.1623(0.0792)	0.1924(0.0579)	0.2001(0.0269)	0.1956(0.0192)
5	0	0.2300(0.0718)	0.2402(0.0517)	0.2327(0.0251)	0.2259(0.0177)	0.2318(0.0741)	0.2405(0.0512)	0.2322(0.0251)	0.2246(0.0177)
	0.25	0.2147(0.0761)	0.2324(0.0562)	0.2301(0.0262)	0.2232(0.0181)	0.2167(0.0814)	0.2329(0.0553)	0.2293(0.0267)	0.2226(0.0179)
	0.50	0.1726(0.0747)	0.1954(0.0547)	0.2012(0.0254)	0.1964(0.0173)	0.1818(0.0807)	0.2127(0.0580)	0.2186(0.0269)	0.2127(0.0186)

FIGURE B.37 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{NPH})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à risques non-proportionnels conduisant à un croisement des risques instantanés, pour différentes valeurs de e^β et différents pourcentages de censure, calculés pour une variable Z de Bernoulli $\mathcal{B}(1/2)$ et $n = 50$, une censure uniforme (1000 répétitions).

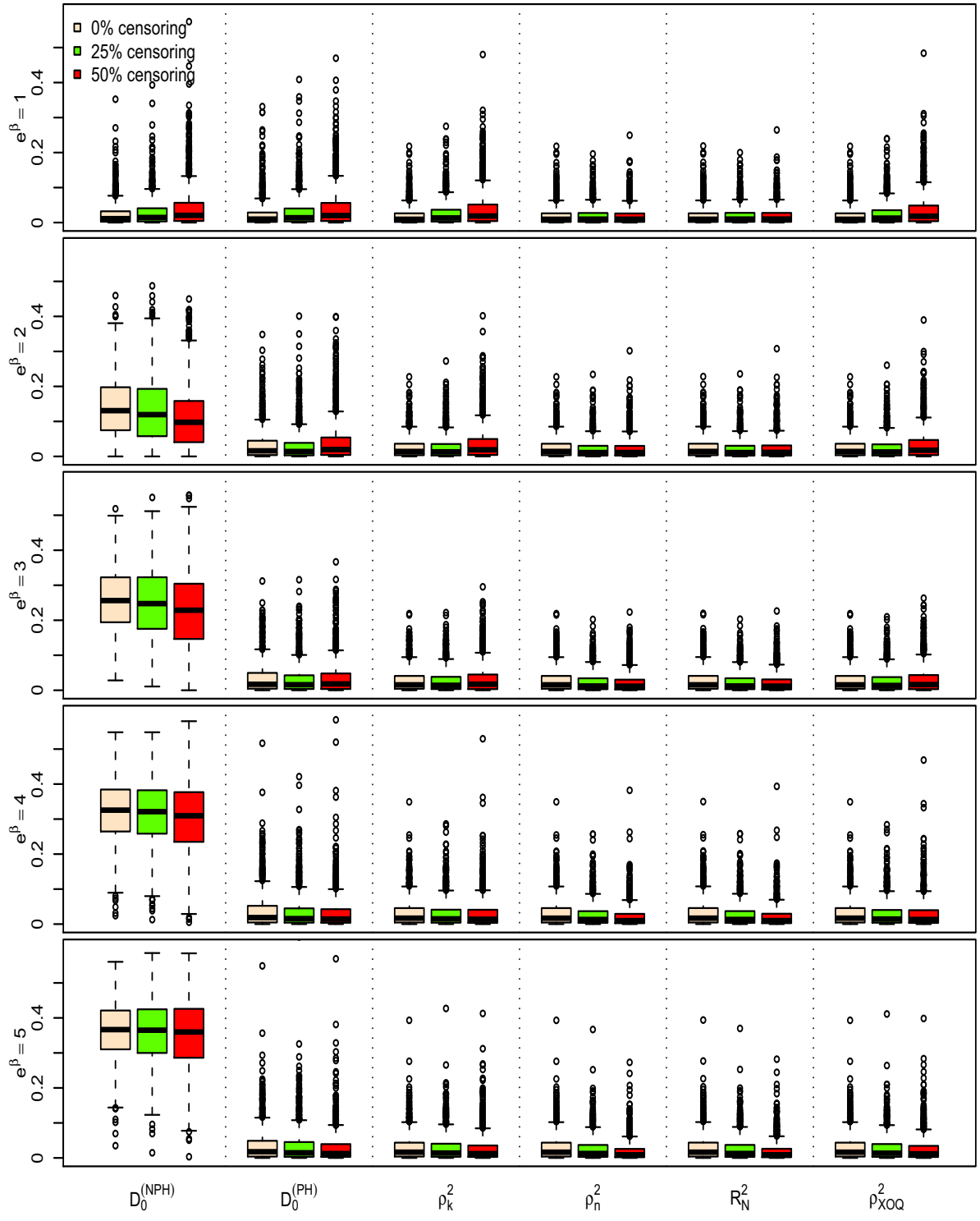


FIGURE B.38 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{NPH})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à risques non-proportionnels conduisant à un croisement des risques instantanés, pour différentes valeurs de e^β et différents pourcentages de censure, calculés pour une variable Z uniforme $\mathcal{U}[0, \sqrt{(3)}]$, une censure uniforme et $n = 50$ (1000 répétitions).

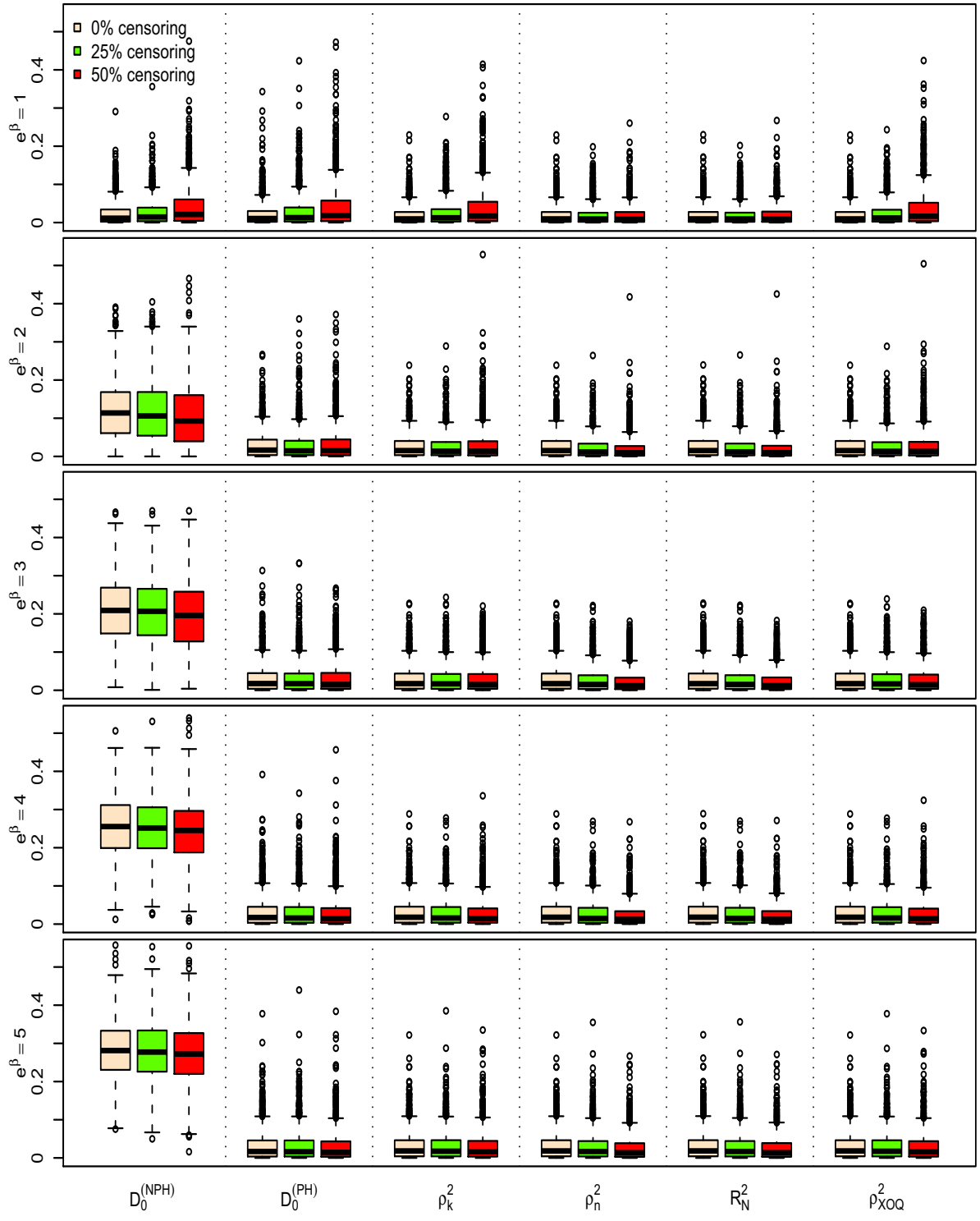


FIGURE B.39 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{NPH})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à risques non-proportionnels conduisant à un croisement des risques instantanés, pour différentes valeurs de e^β et différents pourcentages de censure, calculés pour une variable Z normale $\mathcal{N}(0, 1/4)$, une censure uniforme et $n = 50$ (1000 répétitions).

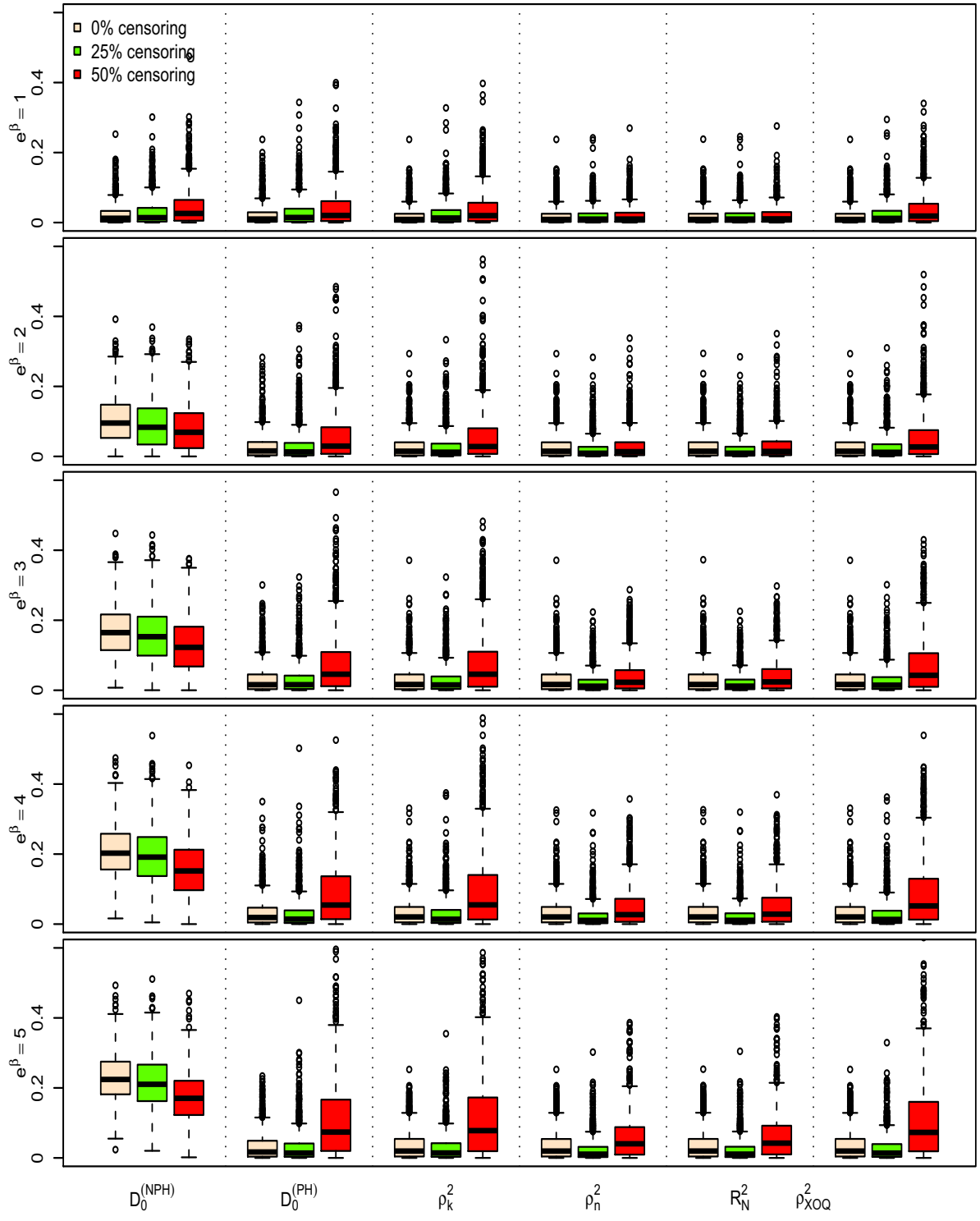


FIGURE B.40 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{NPH})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à risques non-proportionnels conduisant à un croisement des risques instantanés, pour différentes valeurs de e^β et différents pourcentages de censure, calculés pour une variable Z de Bernoulli $\mathcal{B}(1/2)$ et $n = 100$, une censure uniforme (1000 répétitions).

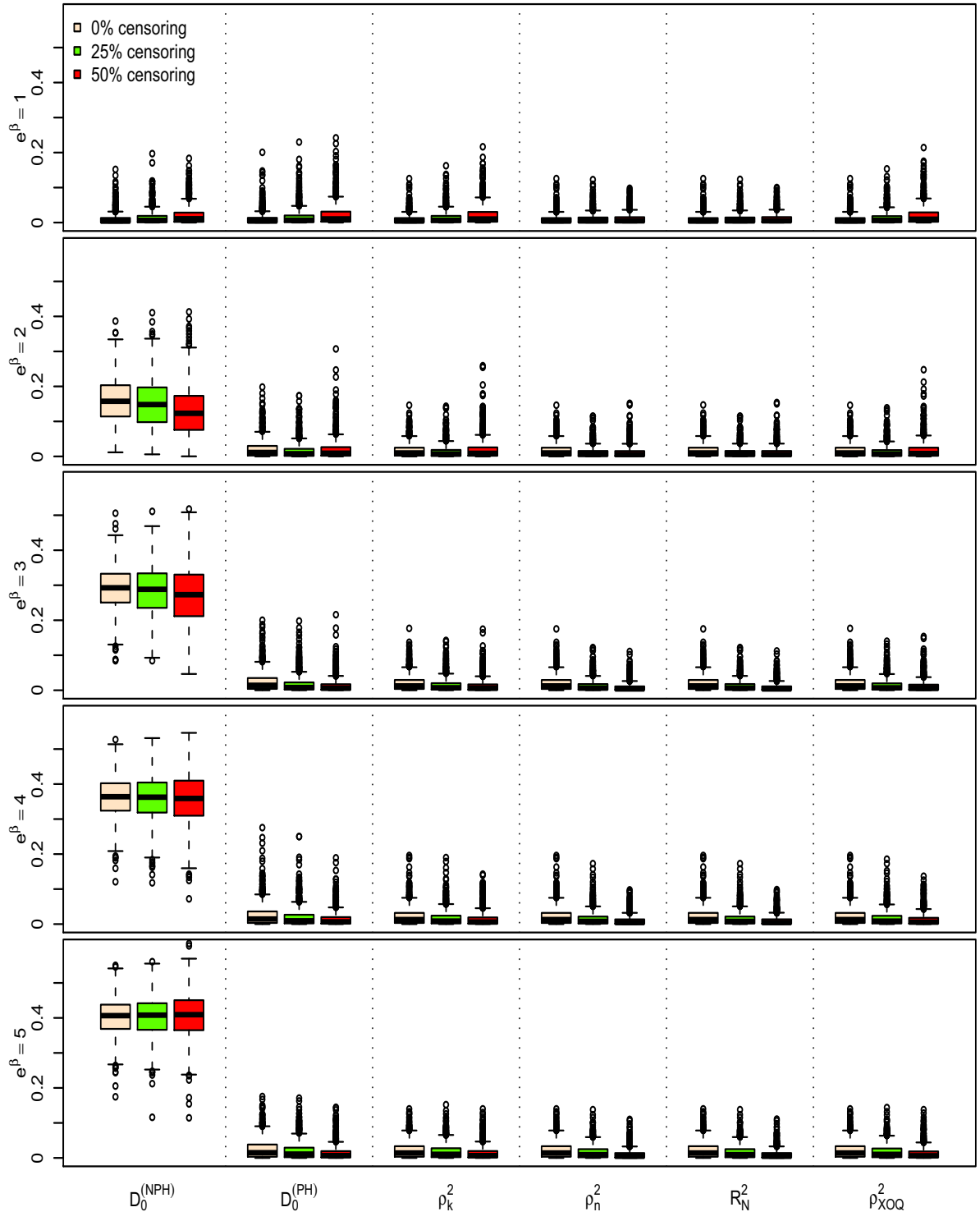


FIGURE B.41 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{NPH})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à risques non-proportionnels conduisant à un croisement des risques instantanés, pour différentes valeurs de e^β et différents pourcentages de censure, calculés pour une variable Z uniforme $\mathcal{U}[0, \sqrt(3)]$, une censure uniforme et $n = 100$ (1000 répétitions).

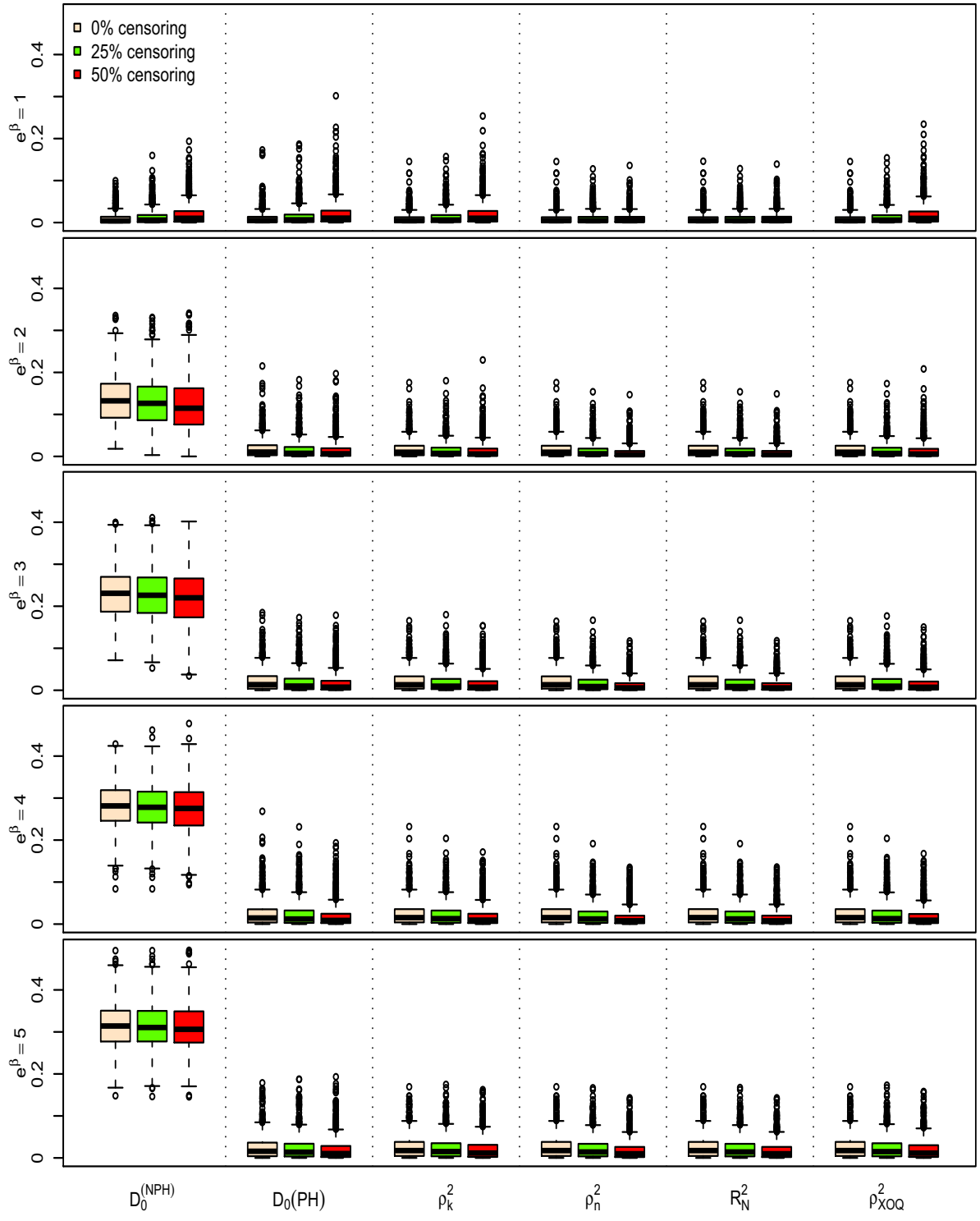


FIGURE B.42 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{NPH})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à risques non-proportionnels conduisant à un croisement des risques instantanés, pour différentes valeurs de e^β et différents pourcentages de censure, calculés pour une variable Z normale $\mathcal{N}(0, 1/4)$, une censure uniforme et $n = 100$ (1000 répétitions).

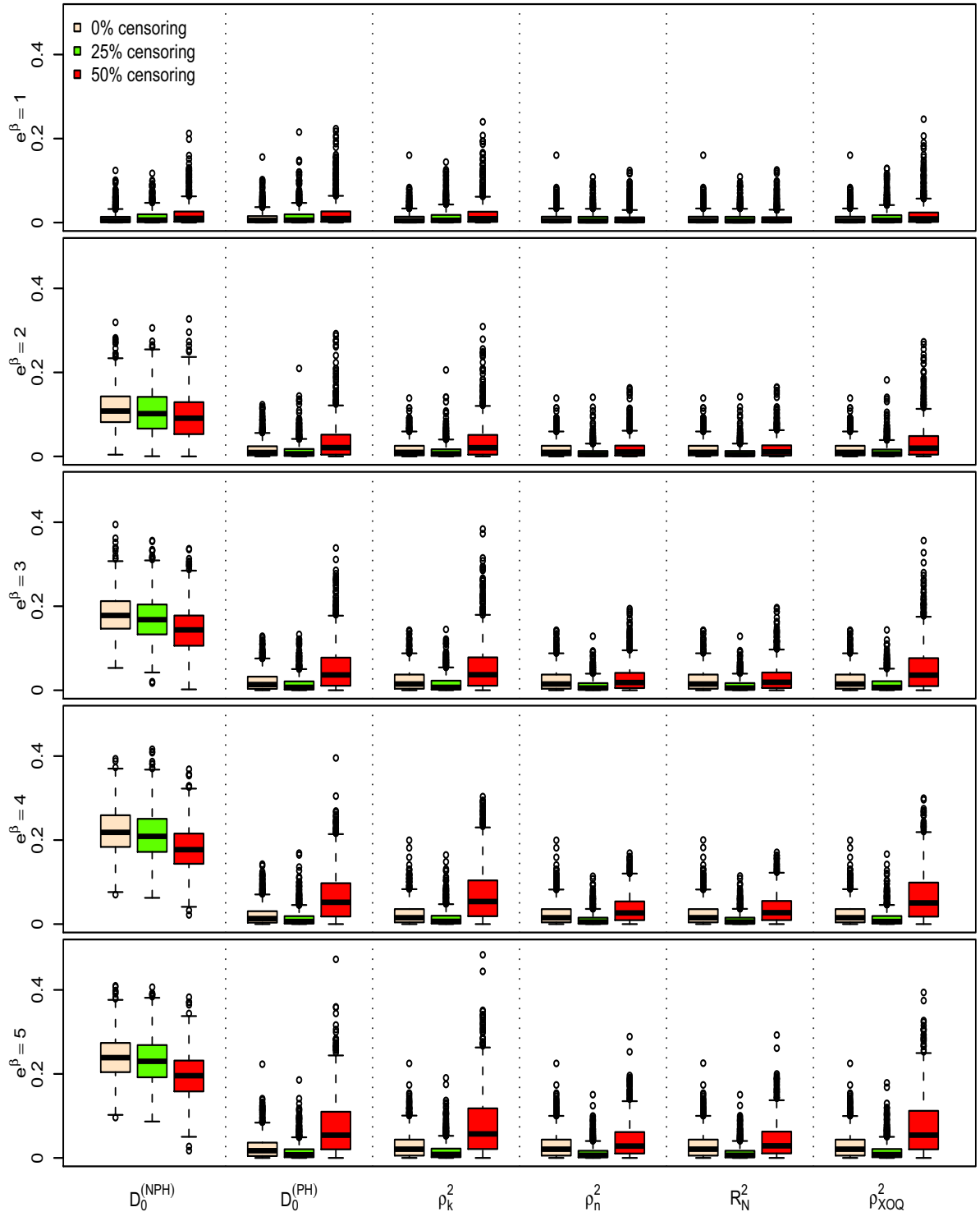


FIGURE B.43 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{NPH})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à risques non-proportionnels conduisant à un croisement des risques instantanés, pour différentes valeurs de e^β et différents pourcentages de censure, calculés pour une variable Z de Bernoulli $\mathcal{B}(1/2)$ et $n = 1000$, une censure uniforme (1000 répétitions).

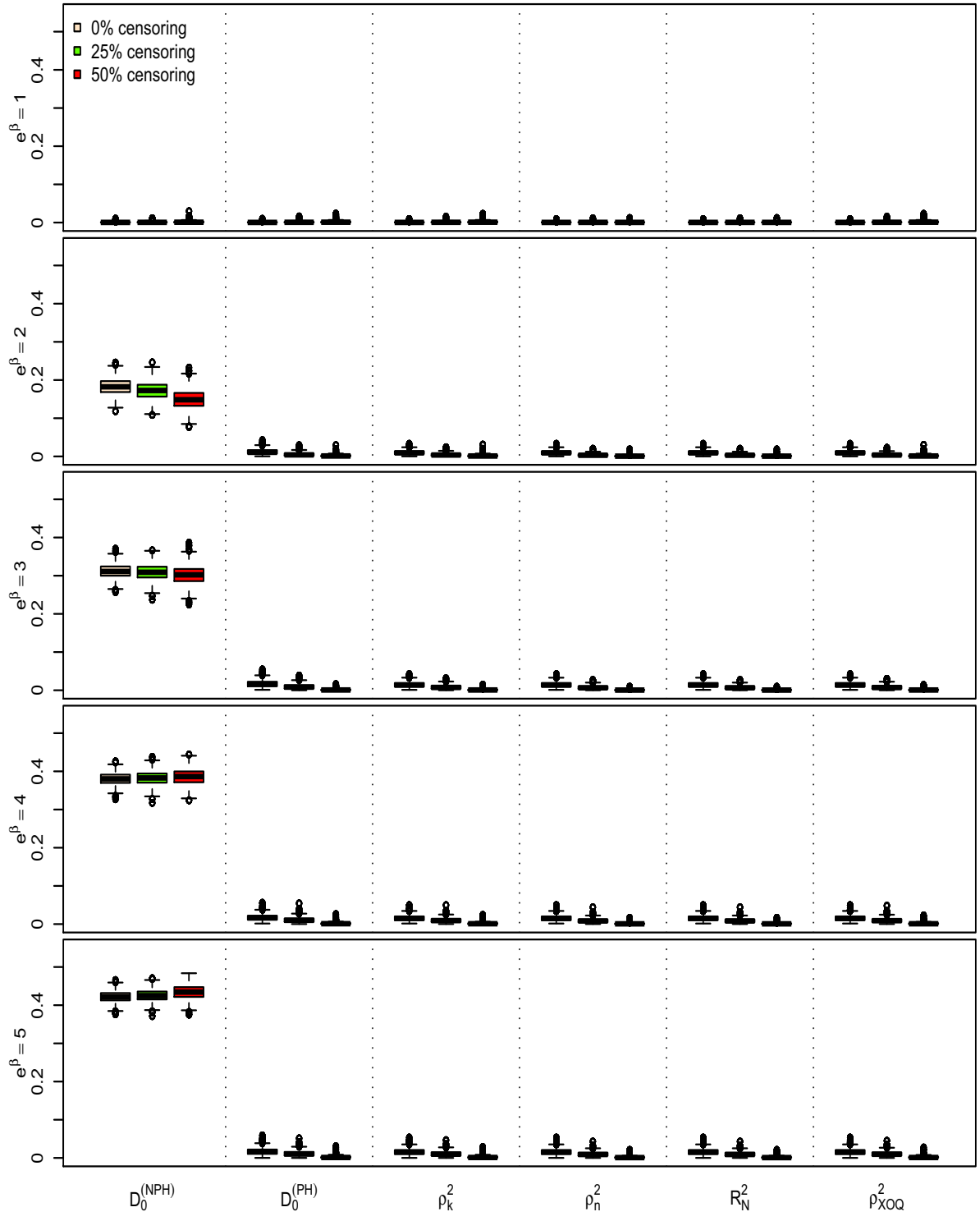


FIGURE B.44 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{NPH})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à risques non-proportionnels conduisant à un croisement des risques instantanés, pour différentes valeurs de e^β et différents pourcentages de censure, calculés pour une variable Z uniforme $\mathcal{U}[0, \sqrt{(3)}]$, une censure uniforme et $n = 1000$ (1000 répétitions).

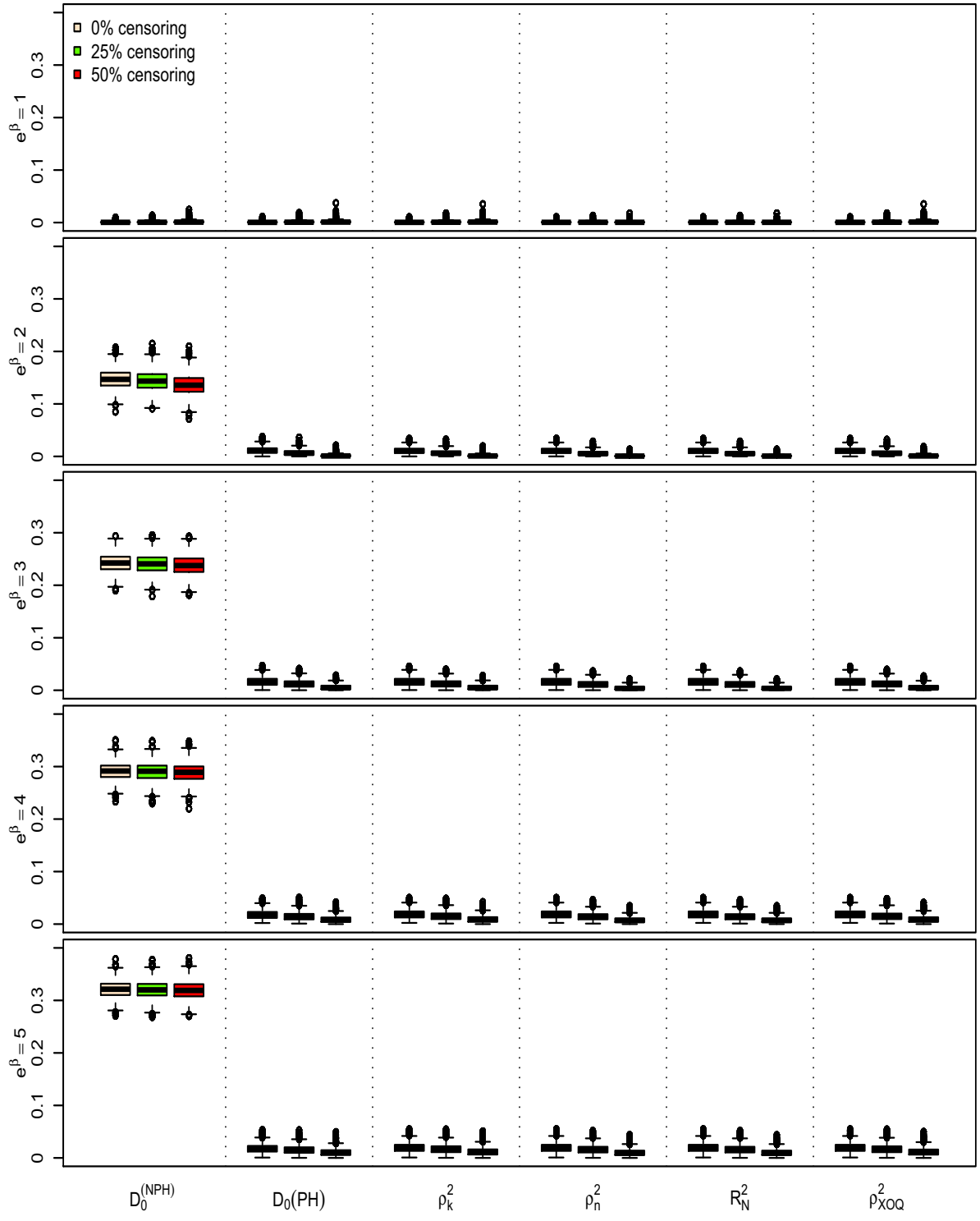


FIGURE B.45 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{NPH})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à risques non-proportionnels conduisant à un croisement des risques instantanés, pour différentes valeurs de e^β et différents pourcentages de censure, calculés pour une variable Z normale $\mathcal{N}(0, 1/4)$, une censure uniforme et $n = 1000$ (1000 répétitions).

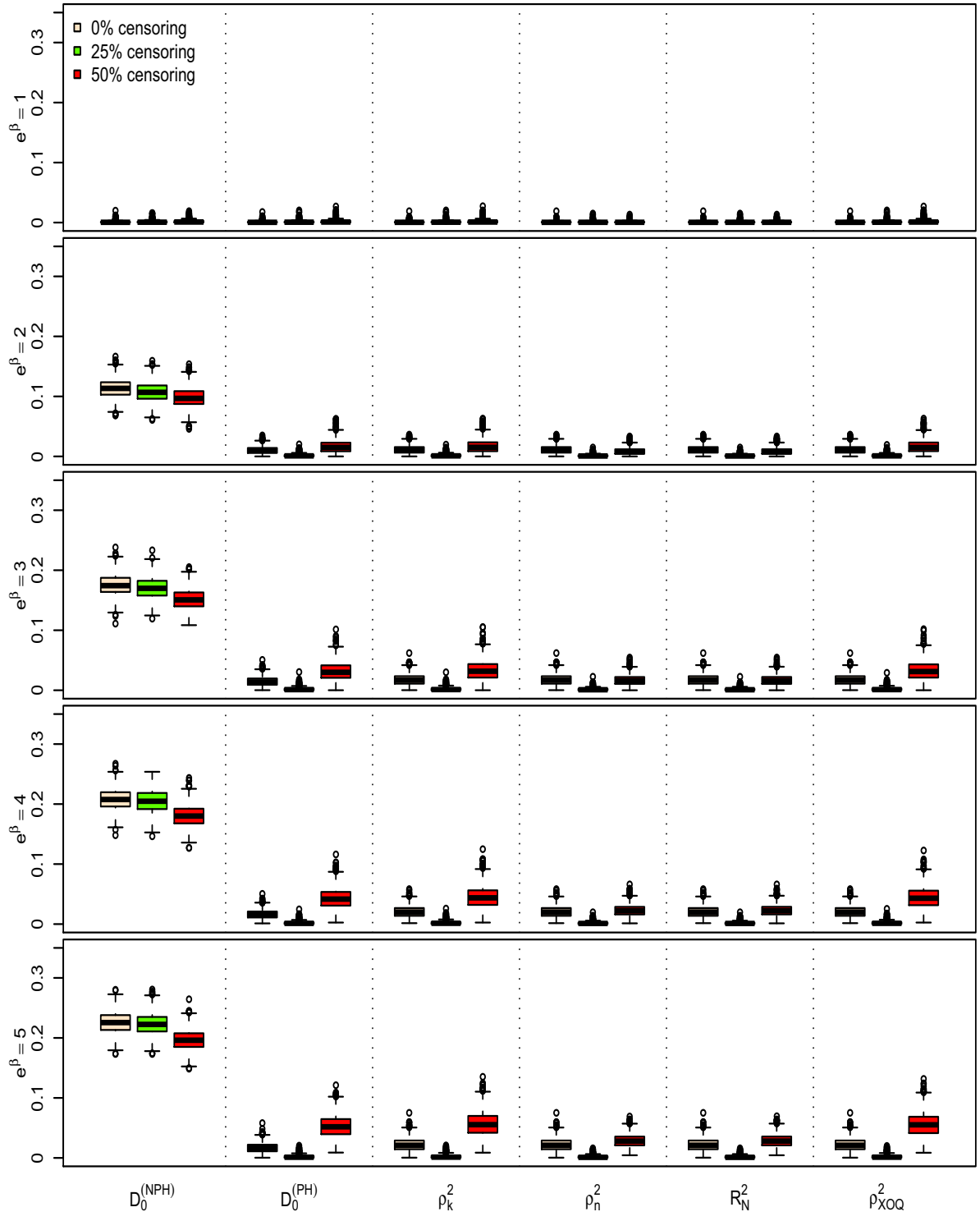


FIGURE B.46 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{NPH})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à risques non-proportionnels conduisant à un croisement des risques instantanés, pour différentes valeurs de e^β et différents pourcentages de censure, calculés pour une variable Z de Bernoulli $\mathcal{B}(1/2)$ et $n = 50$, une censure exponentielle (1000 répétitions).

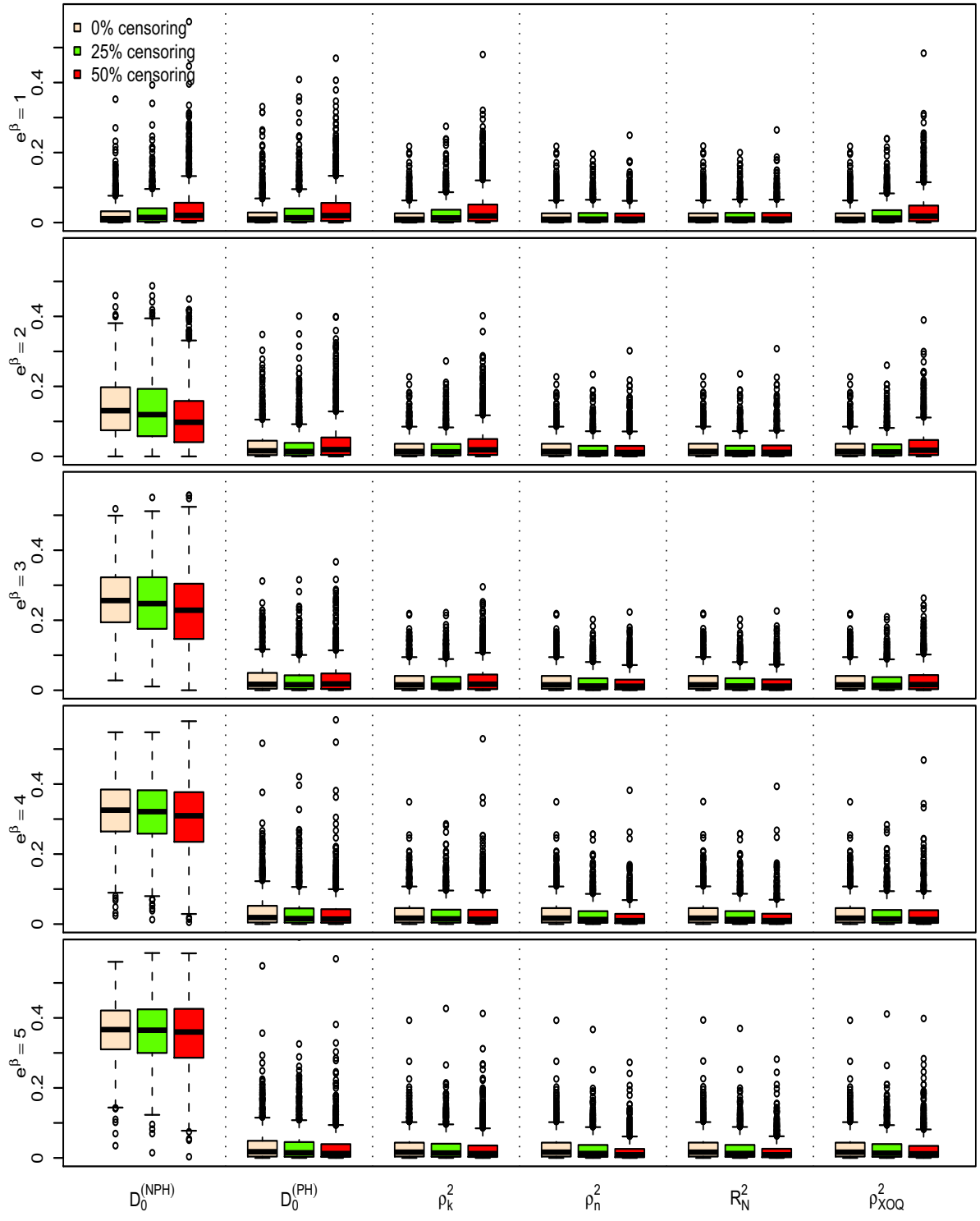


FIGURE B.47 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{NPH})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à risques non-proportionnels conduisant à un croisement des risques instantanés, pour différentes valeurs de e^β et différents pourcentages de censure, calculés pour une variable Z uniforme $\mathcal{U}[0, \sqrt{(3)}]$, une censure exponentielle et $n = 50$ (1000 répétitions).

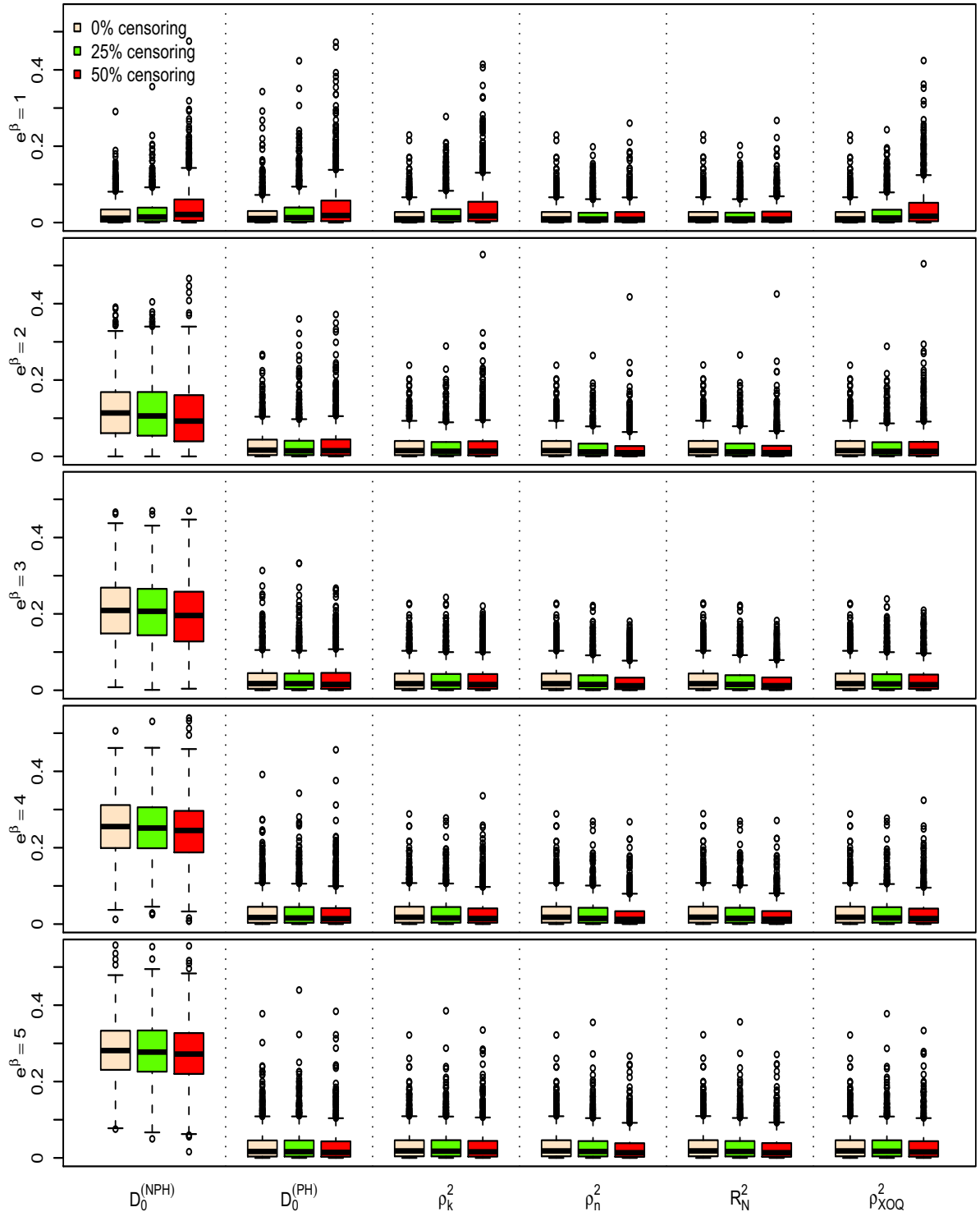


FIGURE B.48 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{NPH})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à risques non-proportionnels conduisant à un croisement des risques instantanés, pour différentes valeurs de e^β et différents pourcentages de censure, calculés pour une variable Z normale $\mathcal{N}(0, 1/4)$, une censure exponentielle et $n = 50$ (1000 répétitions).

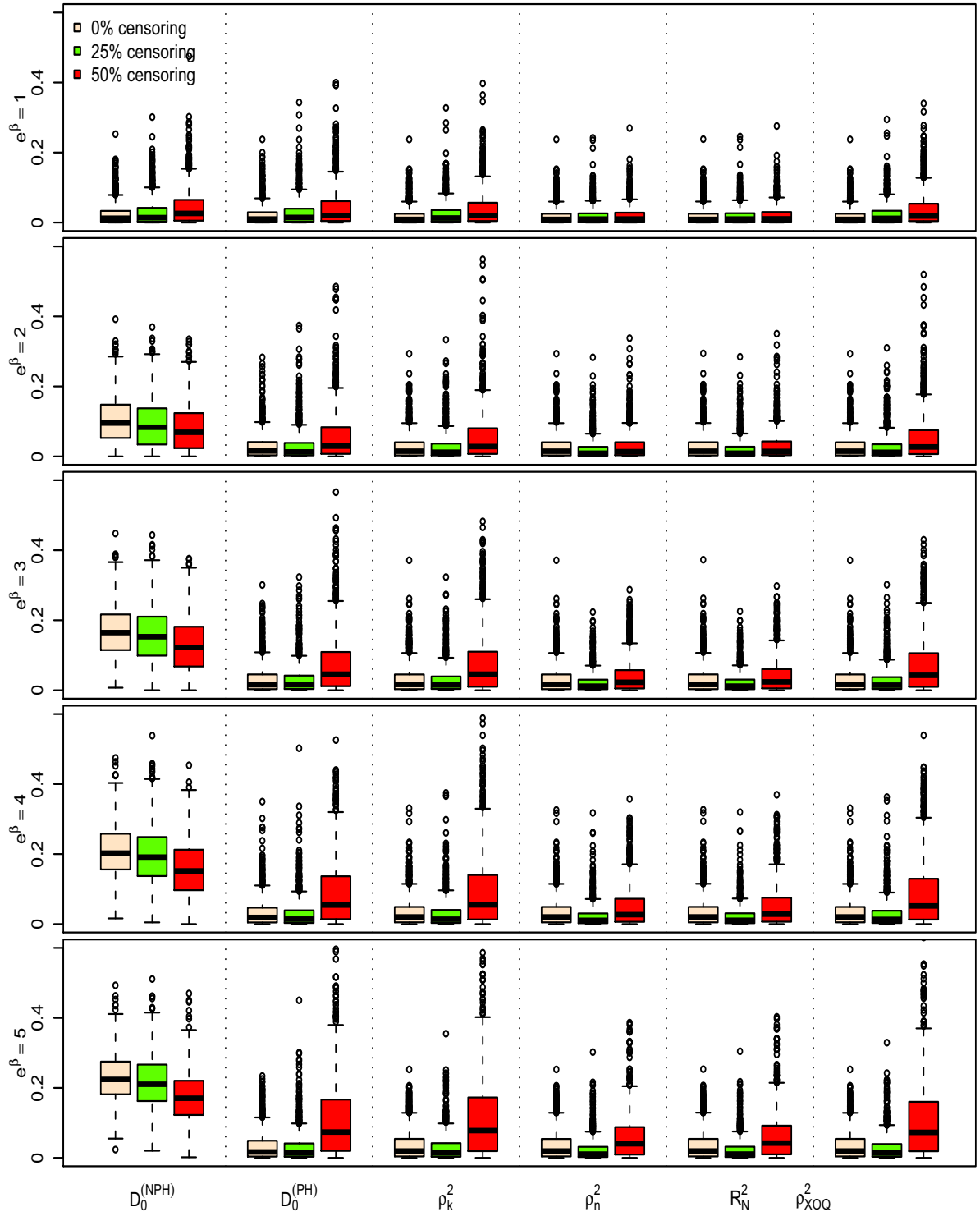


FIGURE B.49 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{NPH})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à risques non-proportionnels conduisant à un croisement des risques instantanés, pour différentes valeurs de e^β et différents pourcentages de censure, calculés pour une variable Z de Bernoulli $\mathcal{B}(1/2)$ et $n = 100$, une censure exponentielle (1000 répétitions).

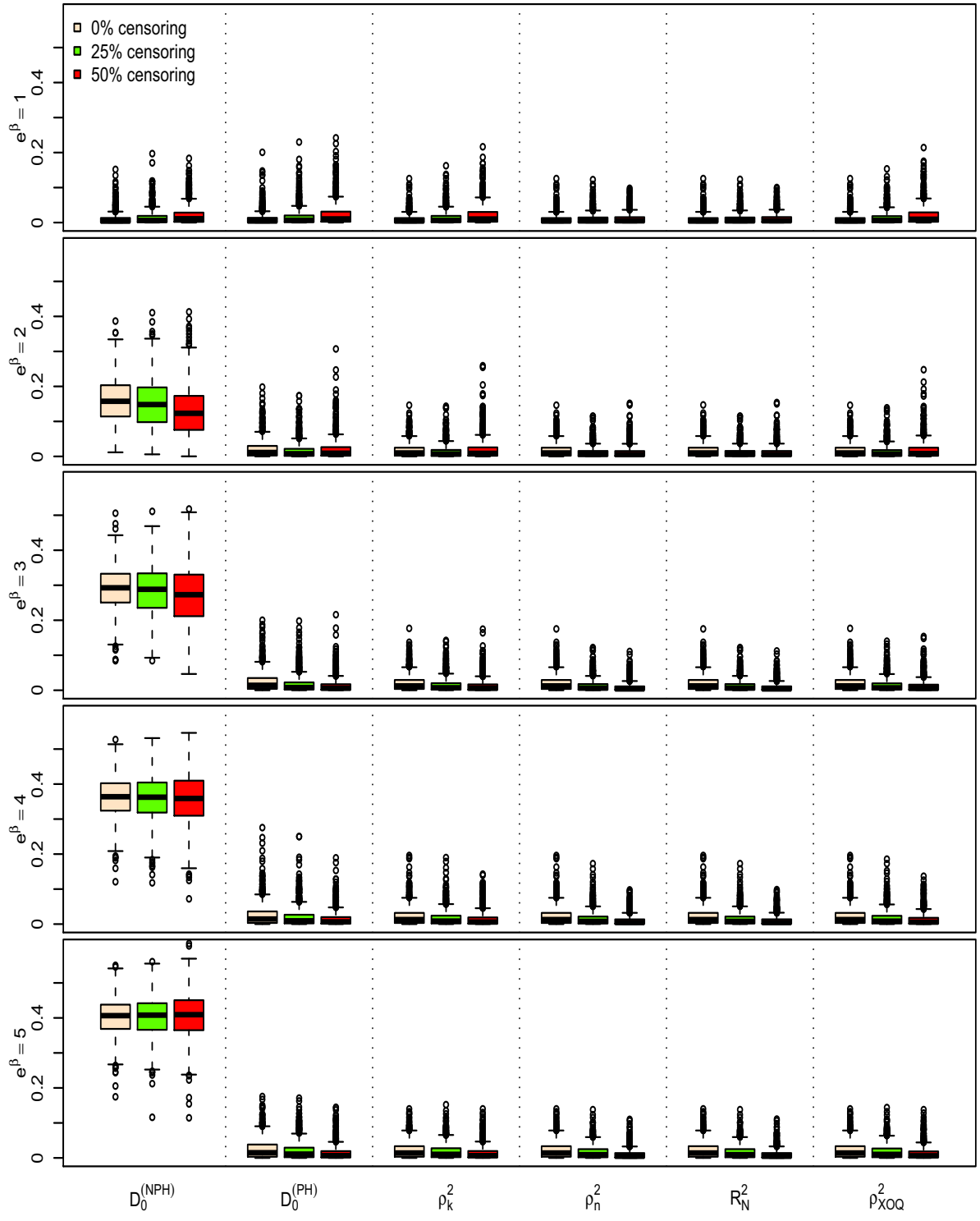


FIGURE B.50 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{NPH})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à risques non-proportionnels conduisant à un croisement des risques instantanés, pour différentes valeurs de e^β et différents pourcentages de censure, calculés pour une variable Z uniforme $\mathcal{U}[0, \sqrt{(3)}]$, une censure exponentielle et $n = 100$ (1000 répétitions).

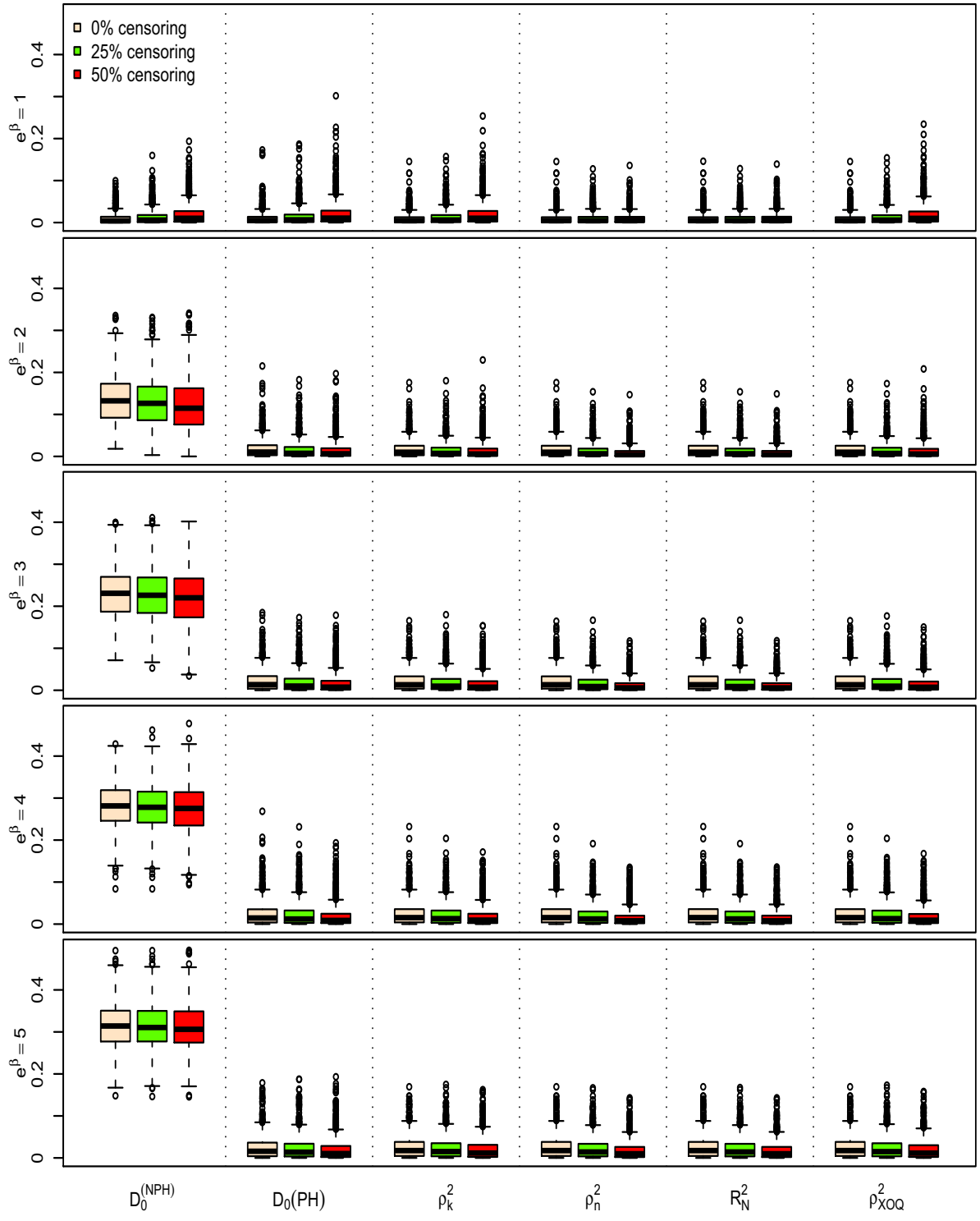


FIGURE B.51 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{NPH})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à risques non-proportionnels conduisant à un croisement des risques instantanés, pour différentes valeurs de e^β et différents pourcentages de censure, calculés pour une variable Z normale $\mathcal{N}(0, 1/4)$, une censure exponentielle et $n = 100$ (1000 répétitions).

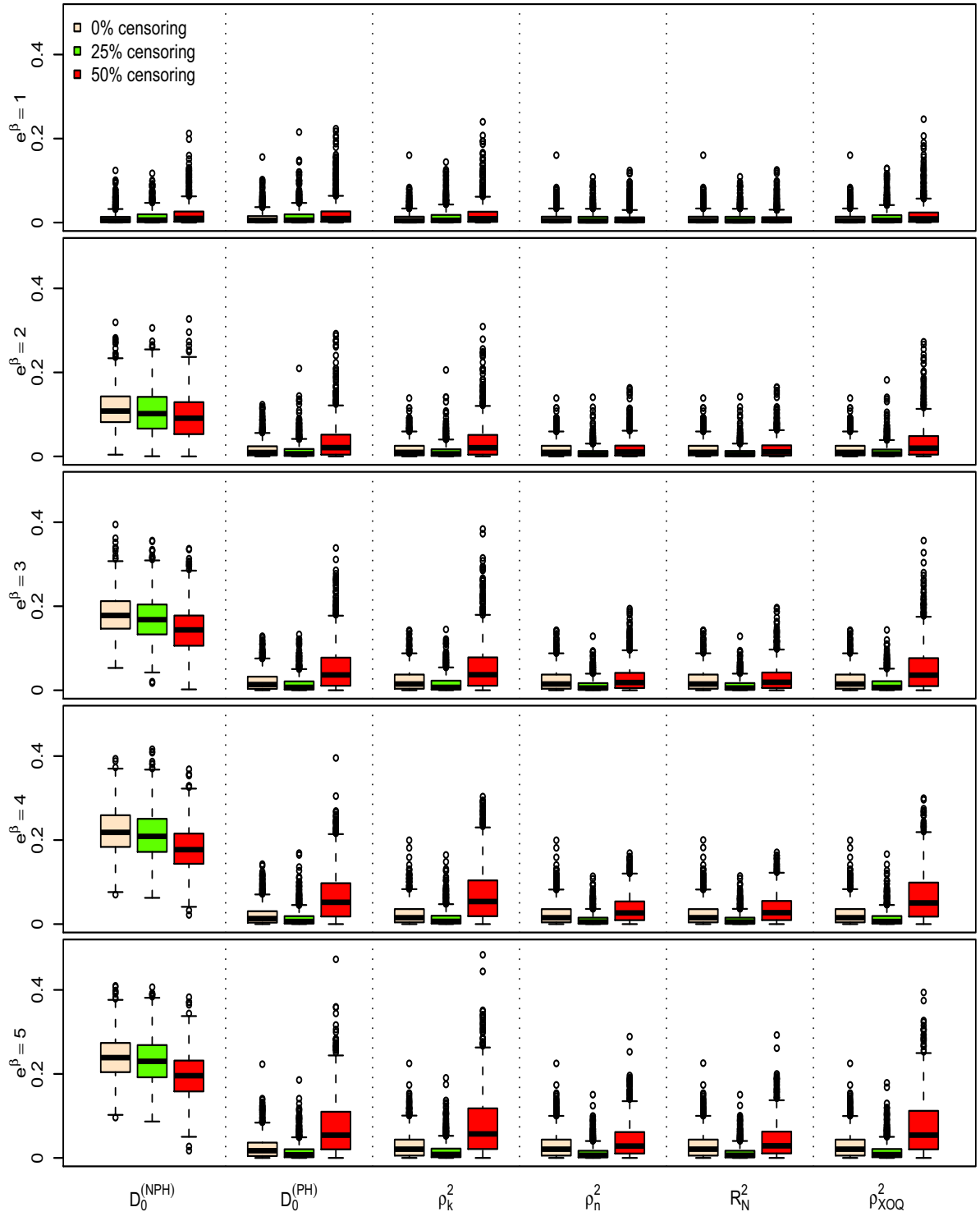


FIGURE B.52 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{NPH})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à risques non-proportionnels conduisant à un croisement des risques instantanés, pour différentes valeurs de e^β et différents pourcentages de censure, calculés pour une variable Z de Bernoulli $\mathcal{B}(1/2)$ et $n = 1000$, une censure exponentielle (1000 répétitions).

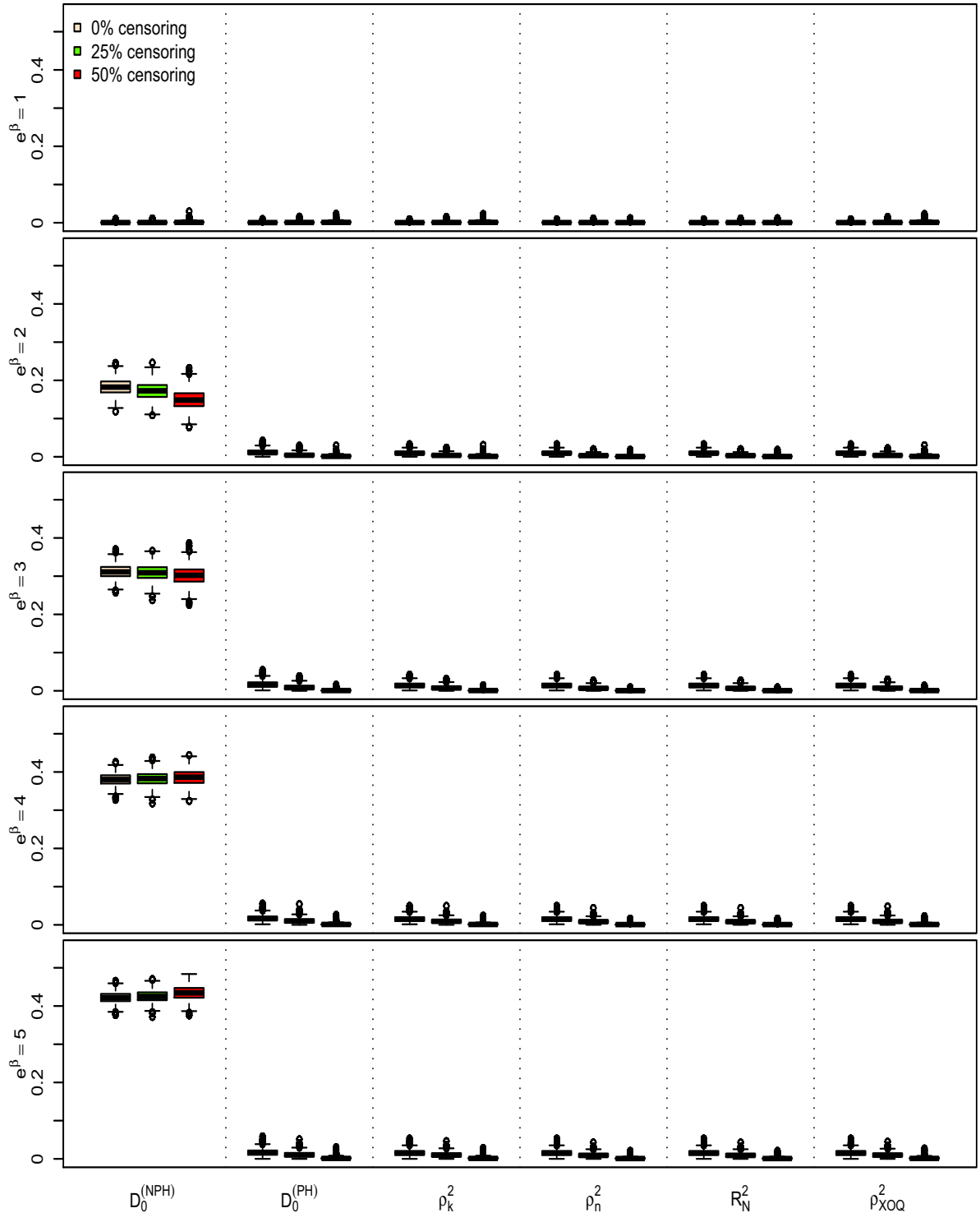


FIGURE B.53 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{NPH})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à risques non-proportionnels conduisant à un croisement des risques instantanés, pour différentes valeurs de e^β et différents pourcentages de censure, calculés pour une variable Z uniforme $\mathcal{U}[0, \sqrt(3)]$, une censure exponentielle et $n = 1000$ (1000 répétitions).

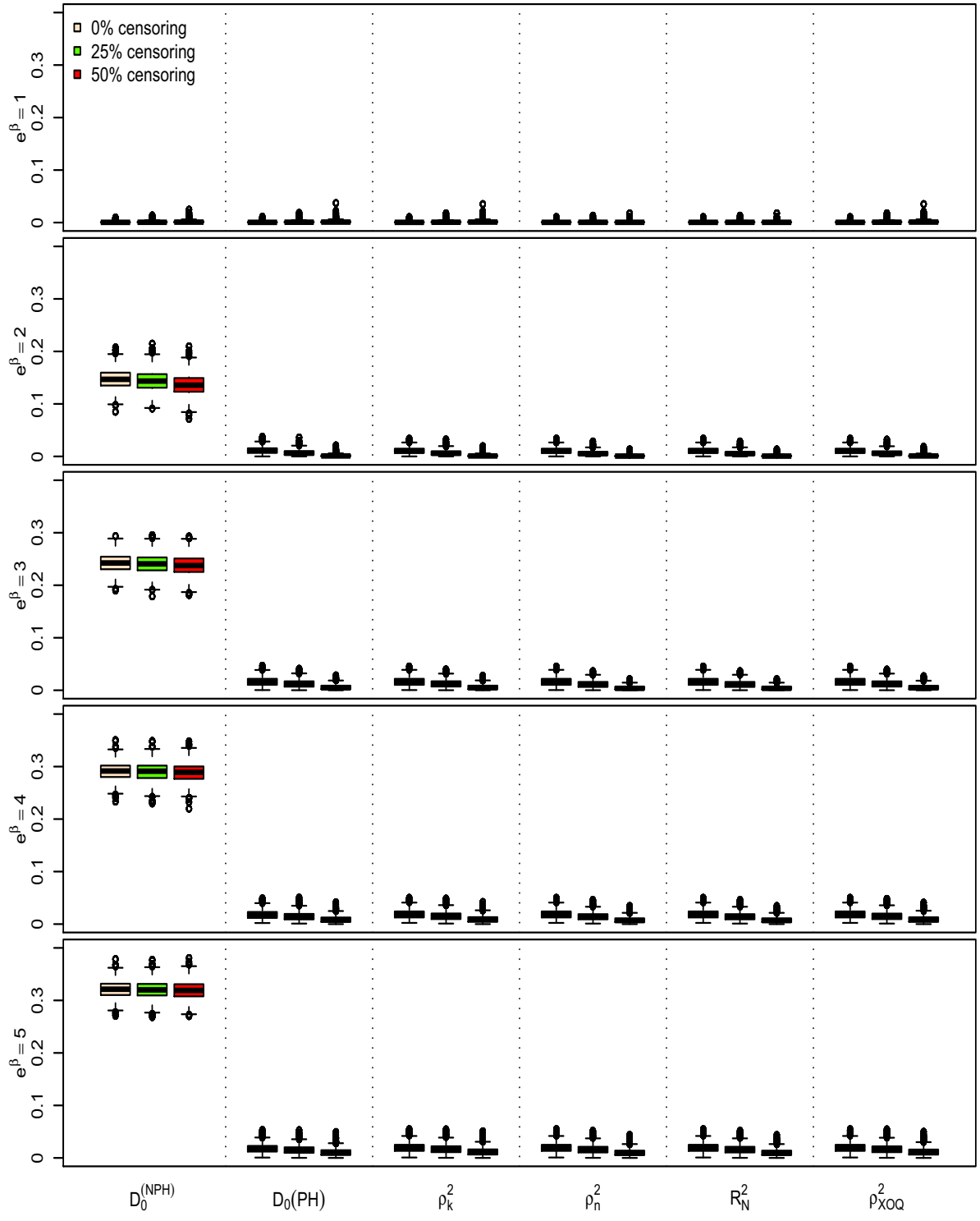


FIGURE B.54 – Boxplot des différents indices $\mathbf{D}_0^{(\mathrm{NPH})}$, ρ_k^2, ρ_k^2, R_N^2, ρ_{XOQ}^2, pour un modèle à risques non-proportionnels conduisant à un croisement des risques instantanés, pour différentes valeurs de e^β et différents pourcentages de censure, calculés pour une variable Z normale $\mathcal{N}(0, 1/4)$, une censure exponentielle et $n = 1000$ (1000 répétitions).

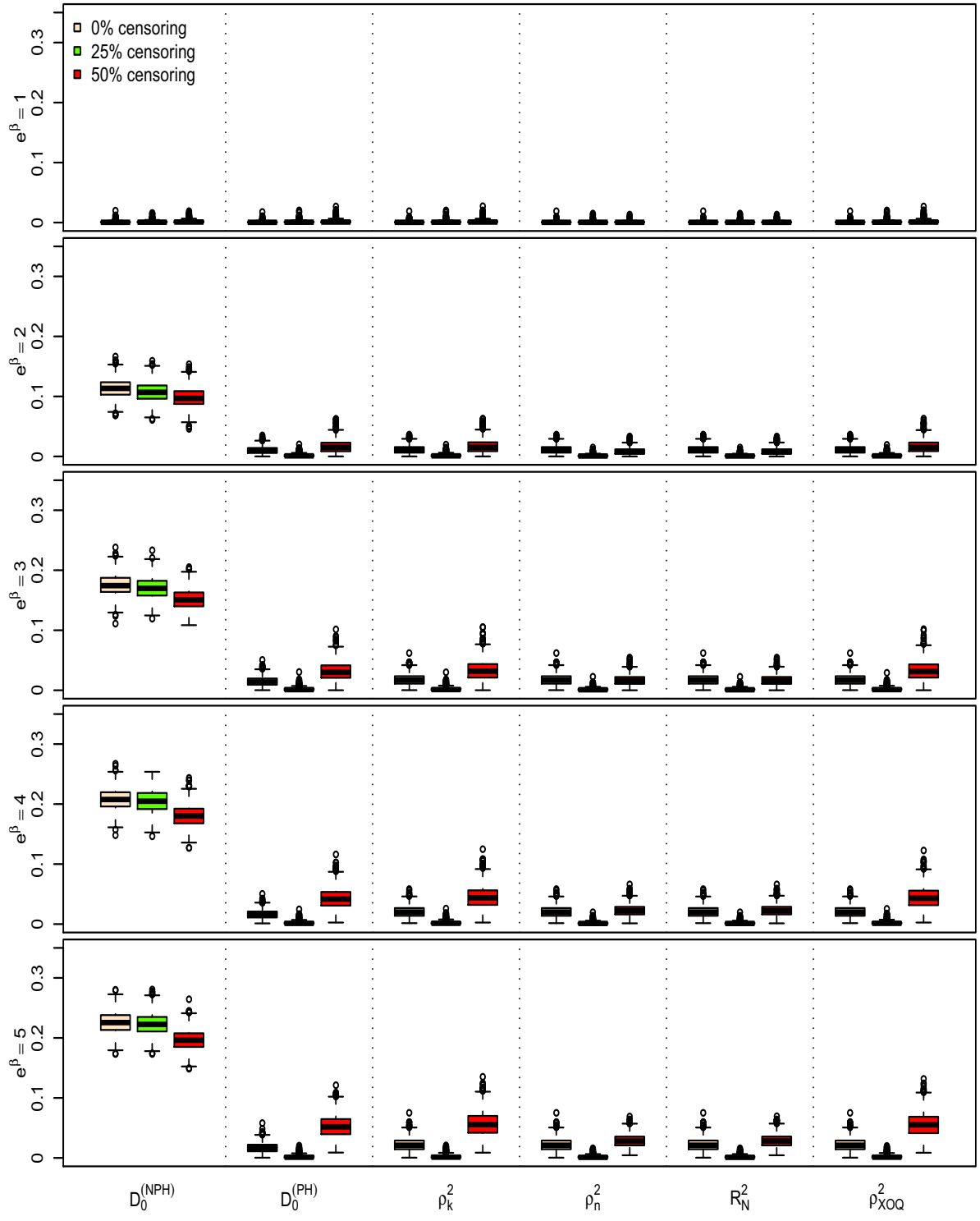


FIGURE B.55 – Courbe du taux de vrais positifs en fonction du taux de faux négatifs de $\mathbf{D}_0^{(PH)}$, ρ_n^2, R_N^2, ρ_k^2, ρ_{XOQ}^2, FDR_{sc} et FDR_L, dans le cadre du modèle de Cox, pour n=100, avec (a) $e^\nu = 2$ et $p_c = 0.25$, (b) $e^\nu = 2$ et $p_c = 0.50$, (c) $e^\nu = 3$ et $p_c = 0.25$ et (d) $e^\nu = 3$ et $p_c = 0.50$.

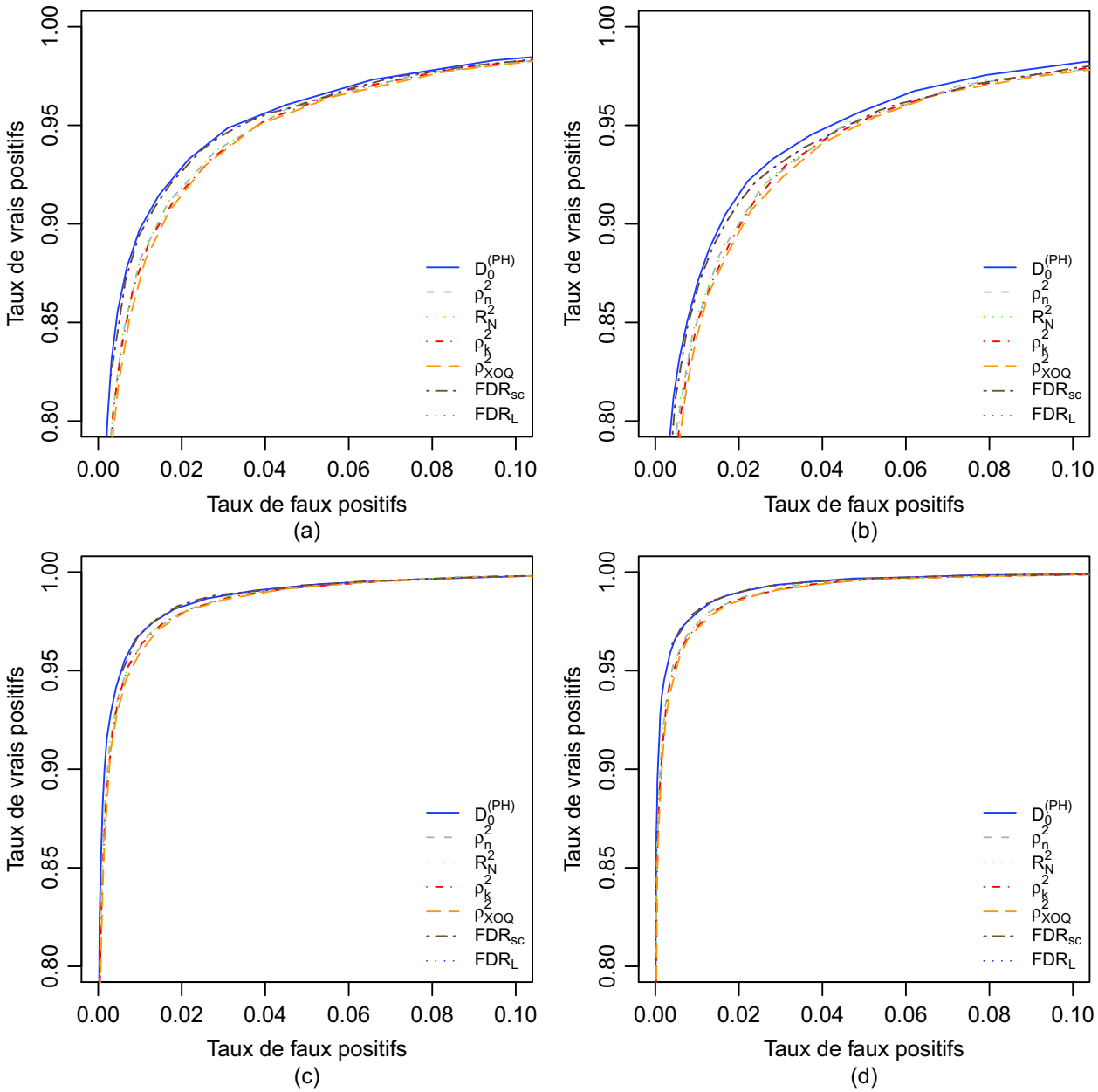


FIGURE B.56 – Courbe du taux de vrais positifs en fonction du taux de faux négatifs de $\mathbf{D}_0^{(PH)}$, ρ_n^2, R_N^2, ρ_k^2, ρ_{XOQ}^2, FDR_{sc} et FDR_L, dans le cadre du modèle de Cox, pour n=200, avec (a) $e^\nu = 2$ et $p_c = 0.25$, (b) $e^\nu = 2$ et $p_c = 0.50$, (c) $e^\nu = 3$ et $p_c = 0.25$ et (d) $e^\nu = 3$ et $p_c = 0.50$.

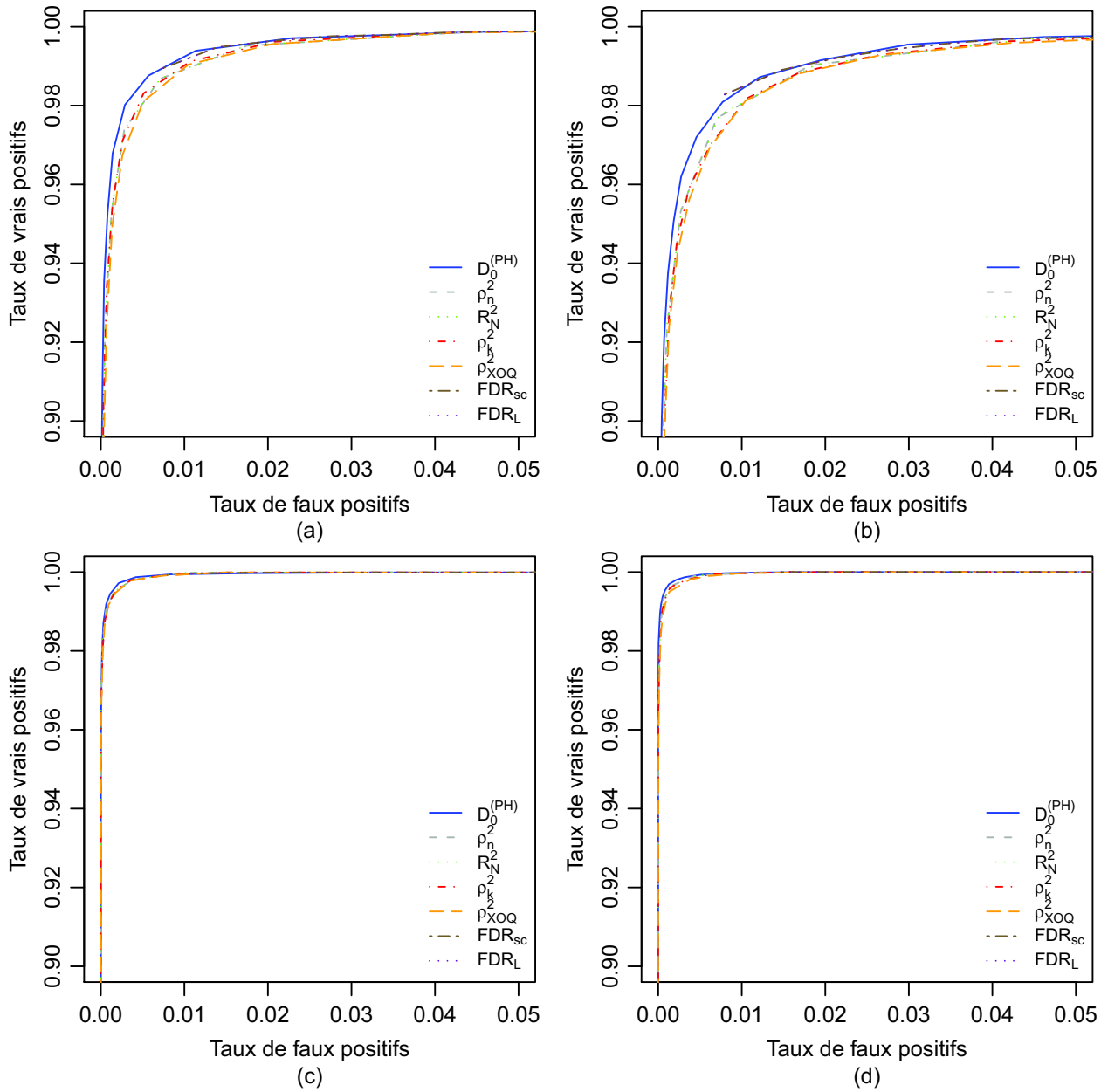


FIGURE B.57 – Courbe du taux de vrais positifs en fonction du taux de faux négatifs de $\mathbf{D}_0^{(\mathrm{POM})}$, ρ_n^2, R_N^2, ρ_k^2, ρ_{XOQ}^2, FDR_{sc} et FDR_L, dans le cadre du modèle à odds proportionnels, pour n=100, avec (a) $e^\nu = 3$ et $p_c = 0.25$, (b) $e^\nu = 3$ et $p_c = 0.50$, (c) $e^\nu = 5$ et $p_c = 0.25$ et (d) $e^\nu = 5$ et $p_c = 0.50$.

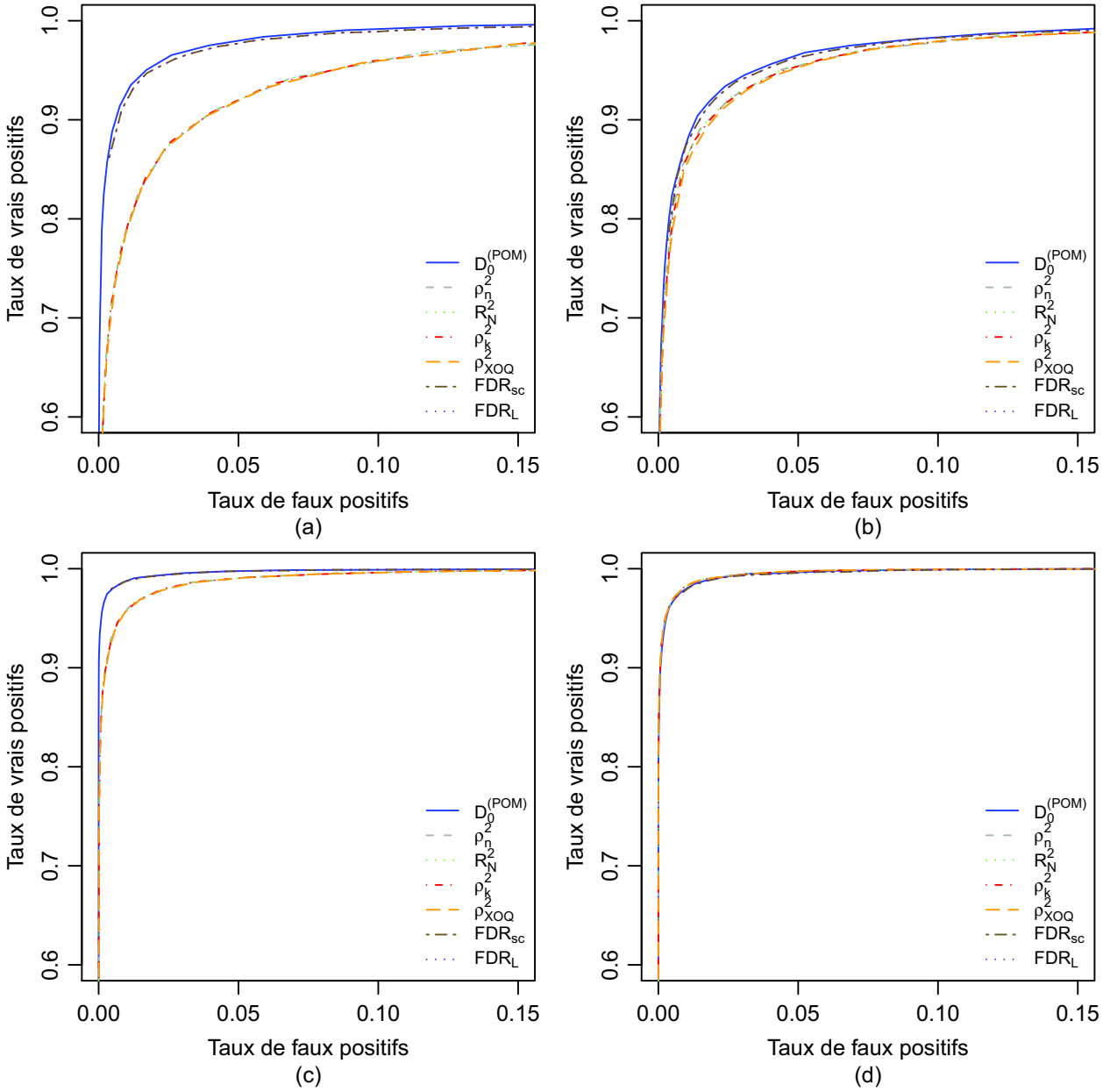


FIGURE B.58 – Courbe du taux de vrais positifs en fonction du taux de faux négatifs de $\mathbf{D}_0^{(\mathrm{POM})}$, ρ_n^2, R_N^2, ρ_k^2, ρ_{XOQ}^2, FDR_{sc} et FDR_L, dans le cadre du modèle à odds proportionnels, pour n=200, avec (a) $e^\nu = 3$ et $p_c = 0.25$, (b) $e^\nu = 3$ et $p_c = 0.50$, (c) $e^\nu = 5$ et $p_c = 0.25$ et (d) $e^\nu = 5$ et $p_c = 0.50$.

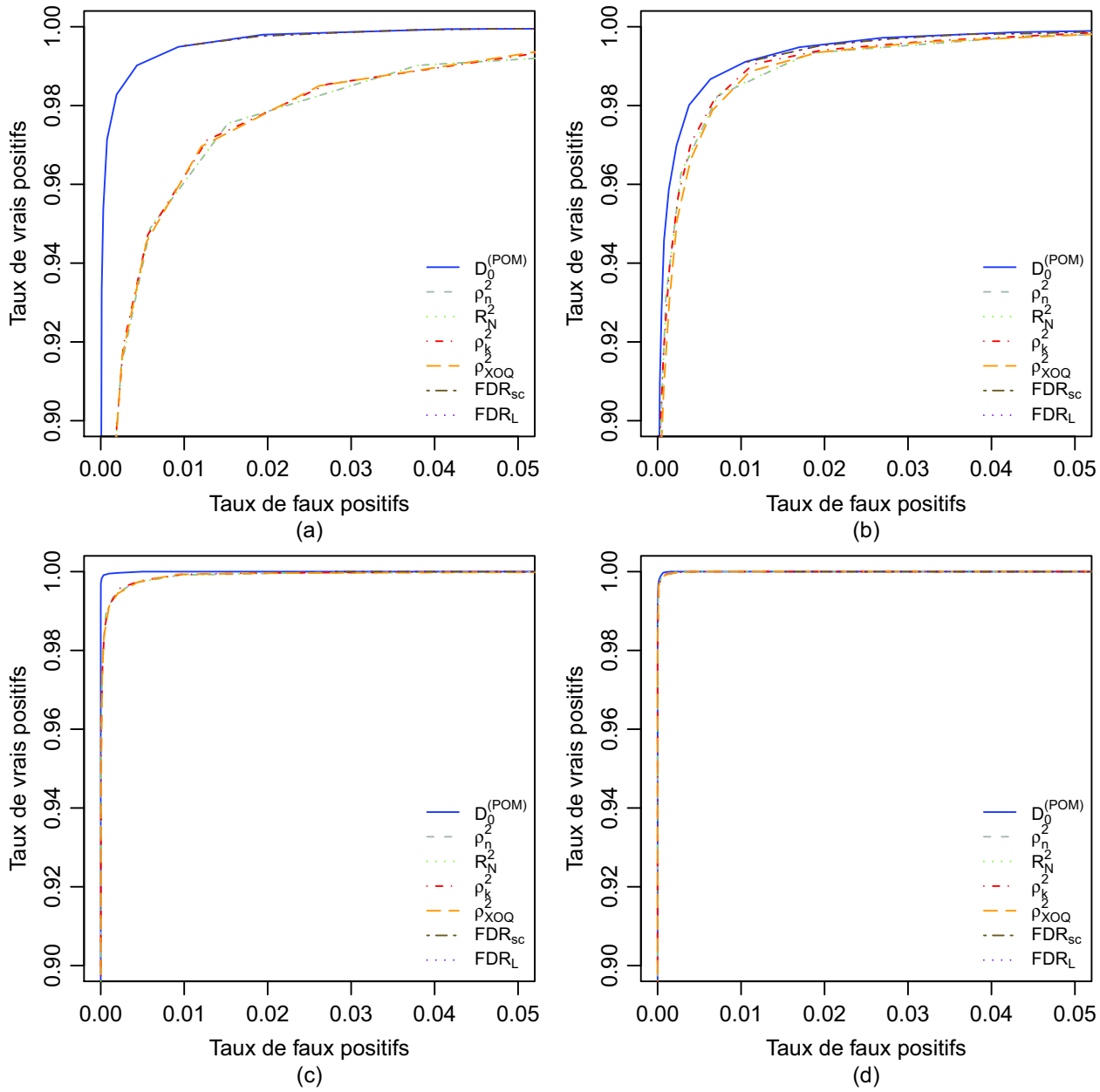


FIGURE B.59 – Courbe du taux de vrais positifs en fonction du taux de faux négatifs de $\mathbf{D}_0^{(\mathrm{NPH})}$, ρ_n^2, R_N^2, ρ_k^2, ρ_{XOQ}^2, FDR_{sc} et FDR_L, dans le cadre du modèle à risques qui se croisent, pour n=100, avec (a) $e^\nu = 2$ et $p_c = 0.25$, (b) $e^\nu = 2$ et $p_c = 0.50$, (c) $e^\nu = 3$ et $p_c = 0.25$ et (d) $e^\nu = 3$ et $p_c = 0.50$.

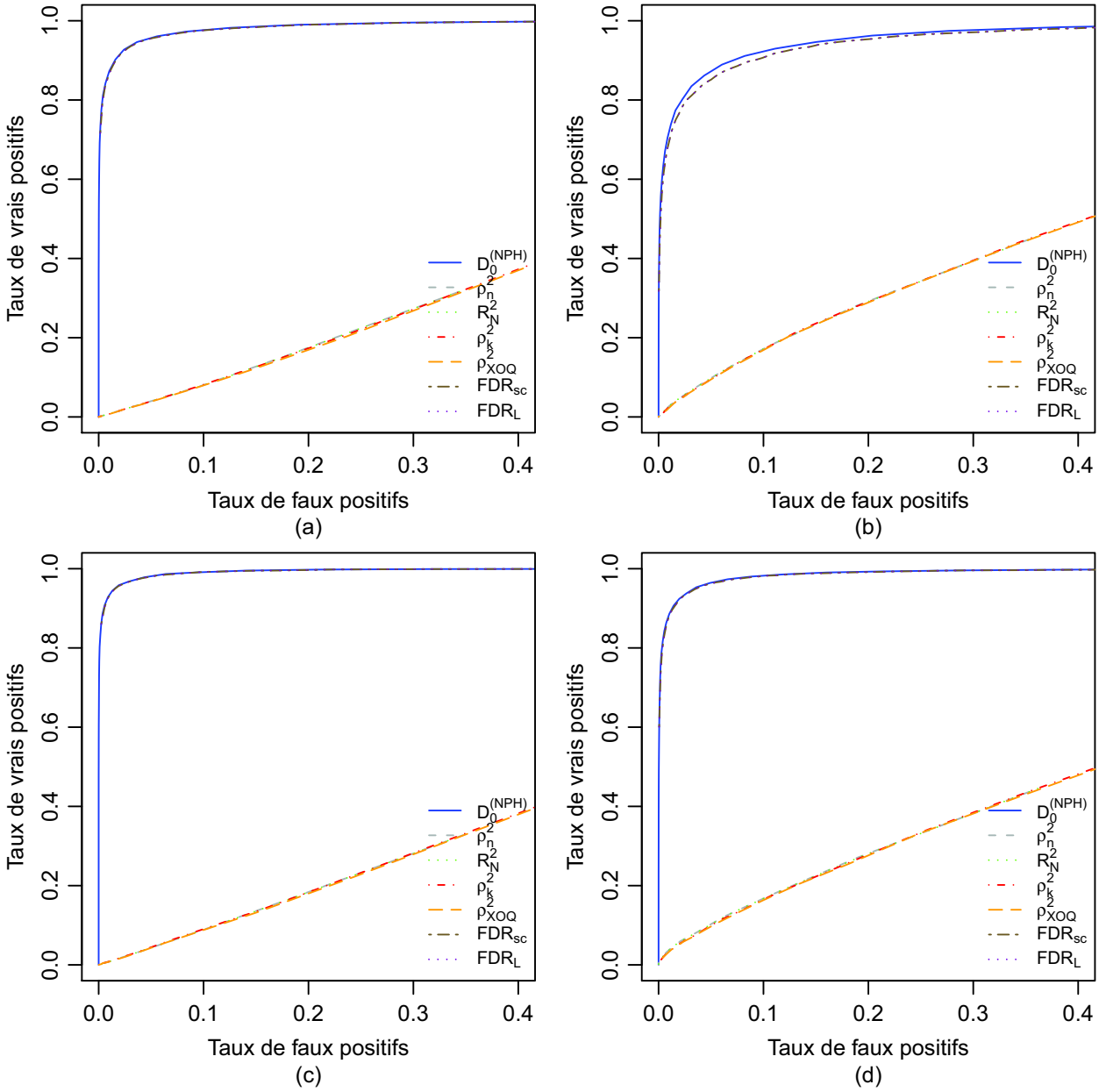
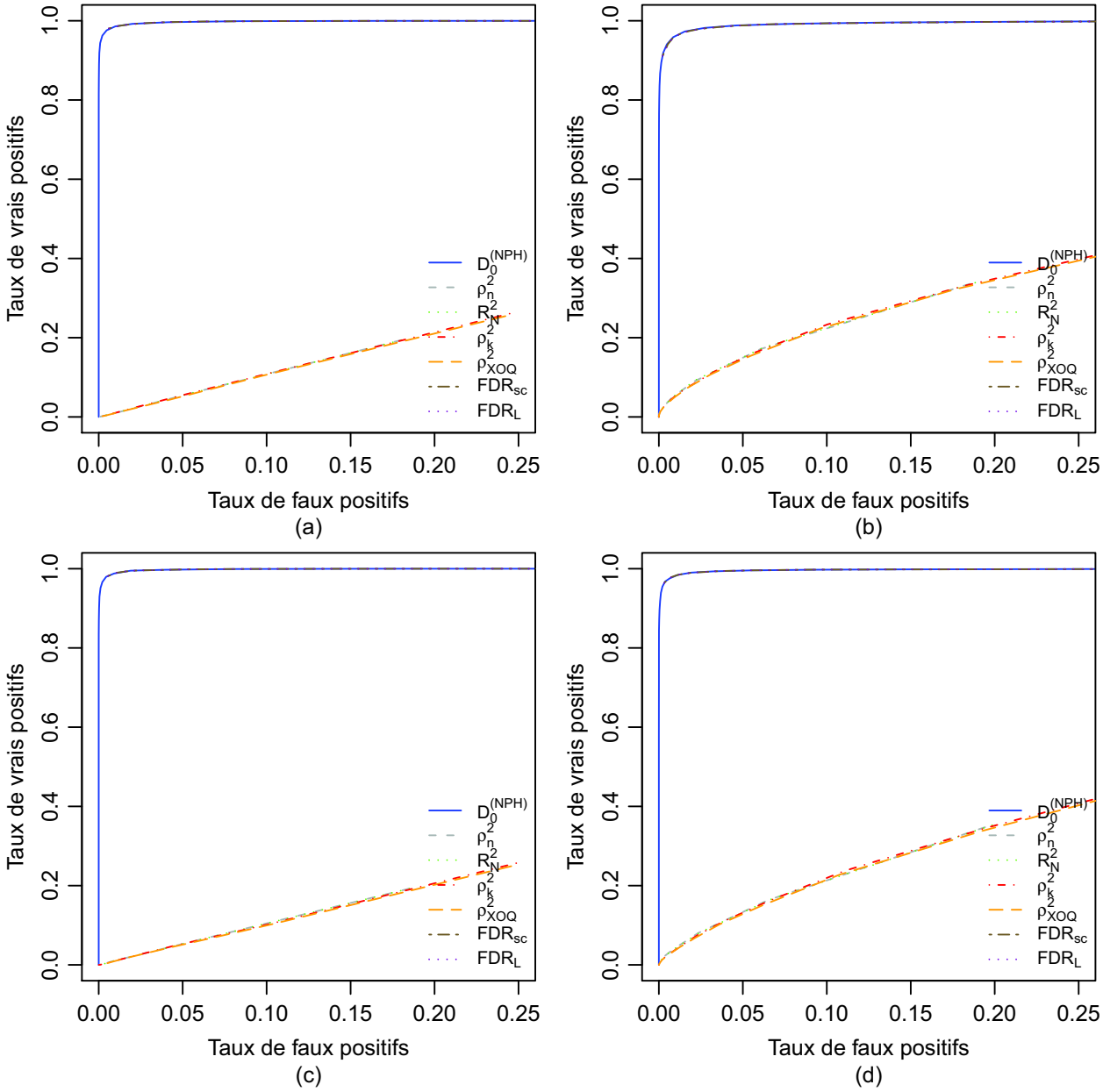


FIGURE B.60 – Courbe du taux de vrais positifs en fonction du taux de faux négatifs de $\mathbf{D}_0^{(\mathrm{NPH})}$, ρ_n^2, R_N^2, ρ_k^2, ρ_{XOQ}^2, FDR_{sc} et FDR_L, dans le cadre du modèle à risques qui se croisent, pour n=200, avec (a) $e^\nu = 2$ et $p_c = 0.25$, (b) $e^\nu = 2$ et $p_c = 0.50$, (c) $e^\nu = 3$ et $p_c = 0.25$ et (d) $e^\nu = 3$ et $p_c = 0.50$.



Annexe C

Résultats complets des exemples

C.1 Courbes de survie complémentaires pour l'exemple 1

FIGURE C.1 – Courbes de survie en fonction du niveau d'expression du gène *LRIG1* dans les huit études. En bleu : faible niveau d'expression, en rouge : haut niveau d'expression

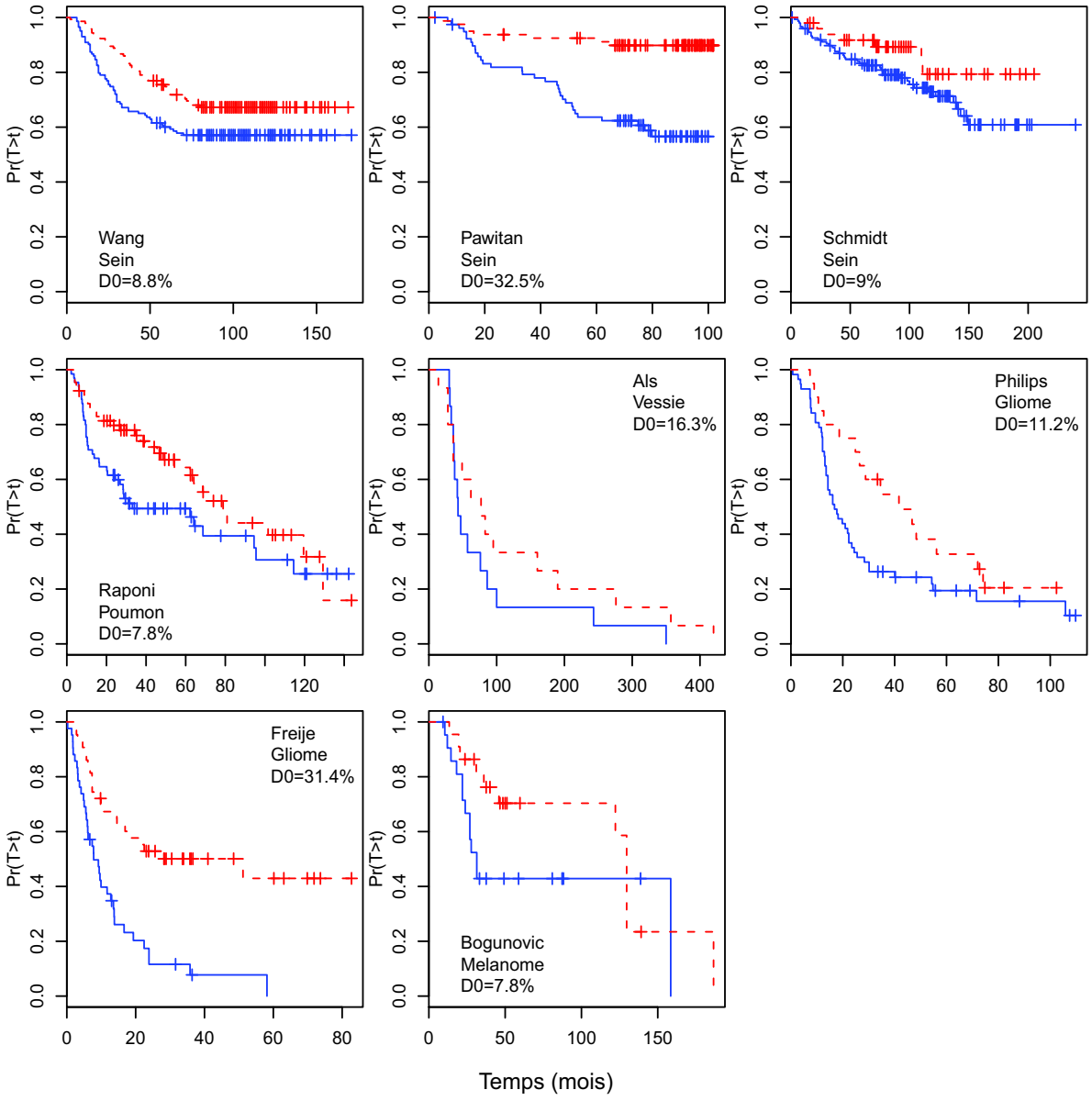


FIGURE C.2 – Courbes de survie en fonction du niveau d'expression du gène *KIF4A* dans les huit études. En bleu : faible niveau d'expression, en rouge : haut niveau d'expression

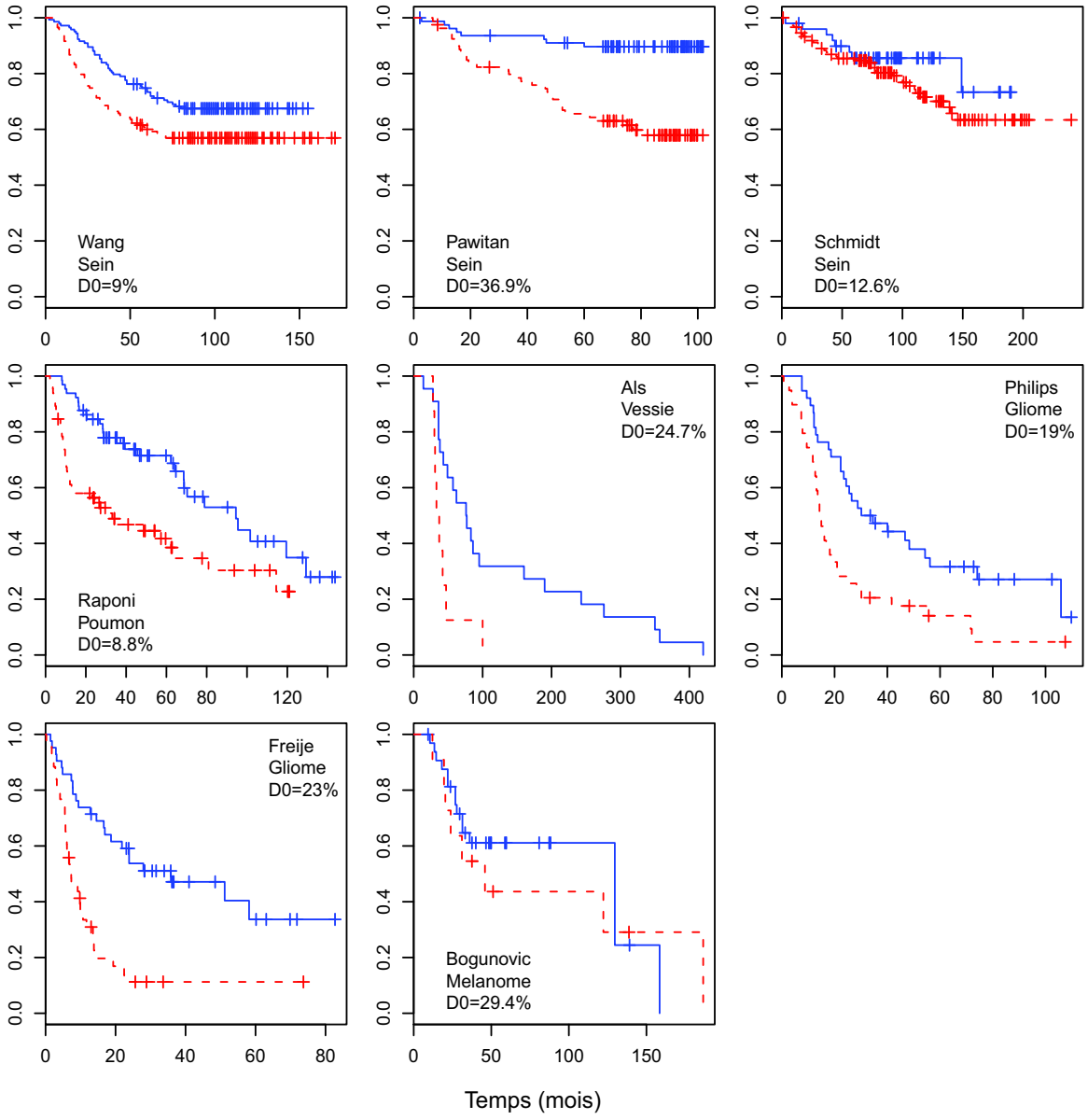


FIGURE C.3 – Courbes de survie en fonction du niveau d'expression du gène *HJURP* dans les huit études. En bleu : faible niveau d'expression, en rouge : haut niveau d'expression

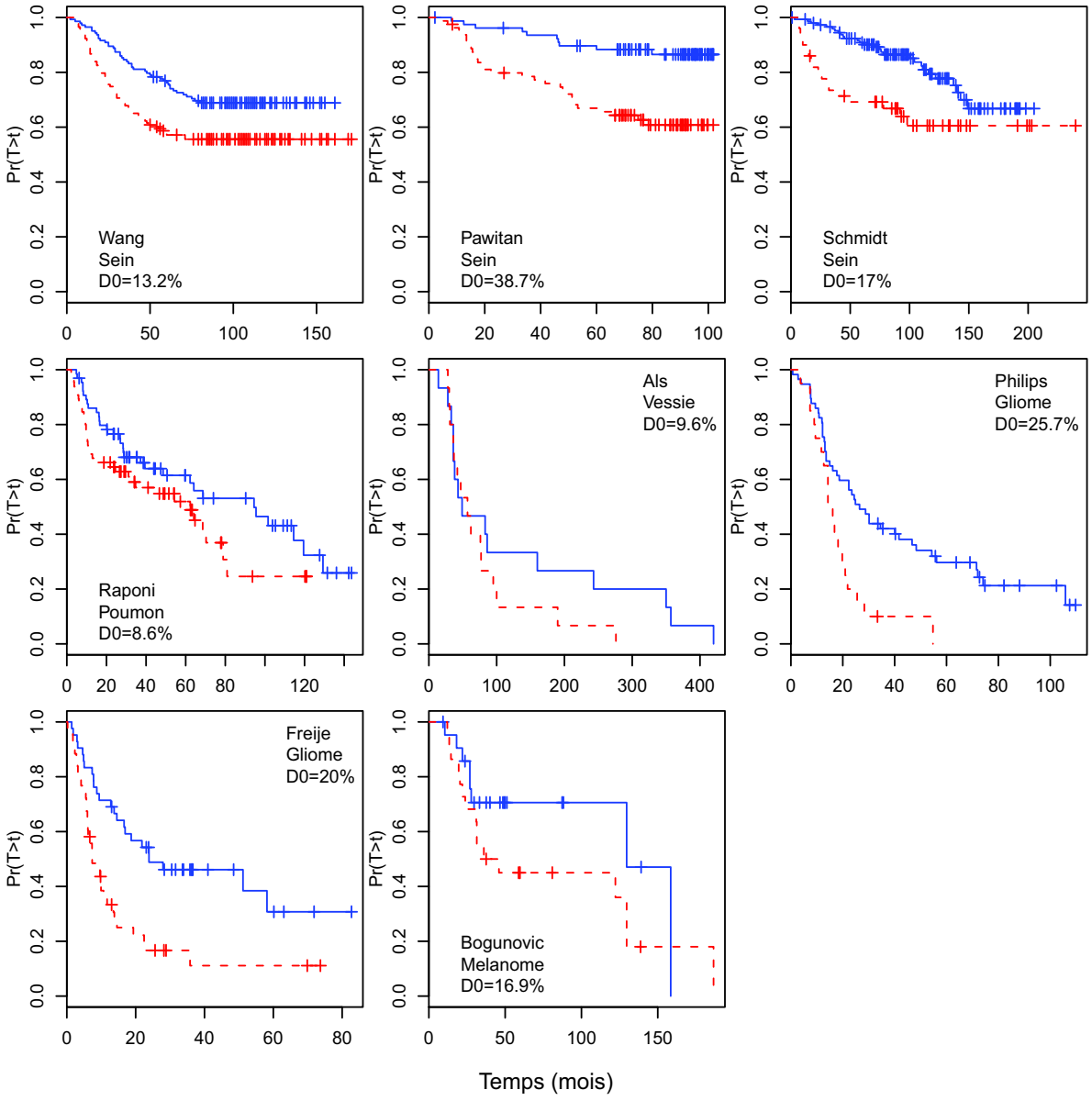


FIGURE C.4 – Courbes de survie en fonction du niveau d'expression du gène *ESPL1* (204817-at) dans les huit études. En bleu : faible niveau d'expression, en rouge : haut niveau d'expression

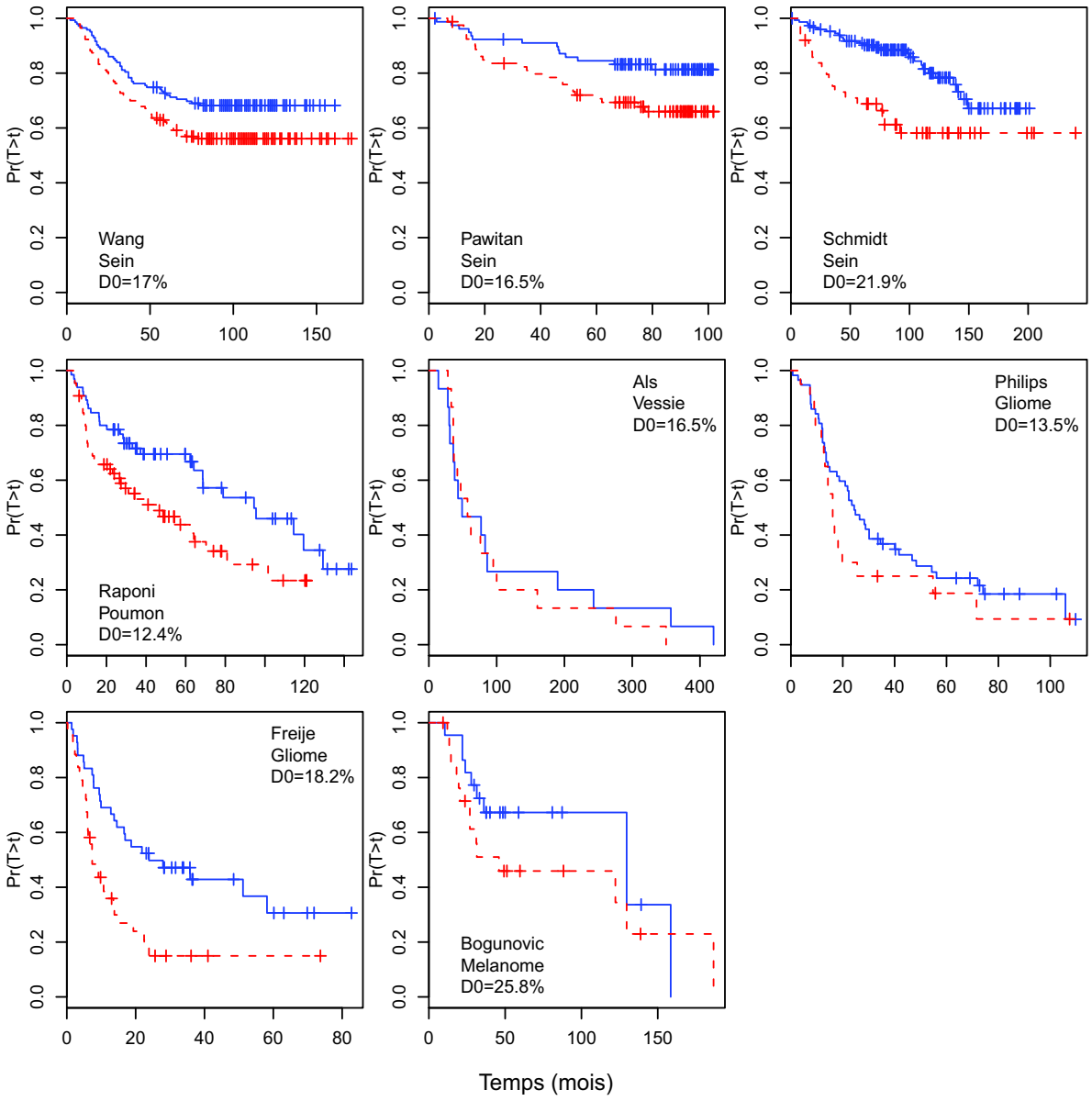
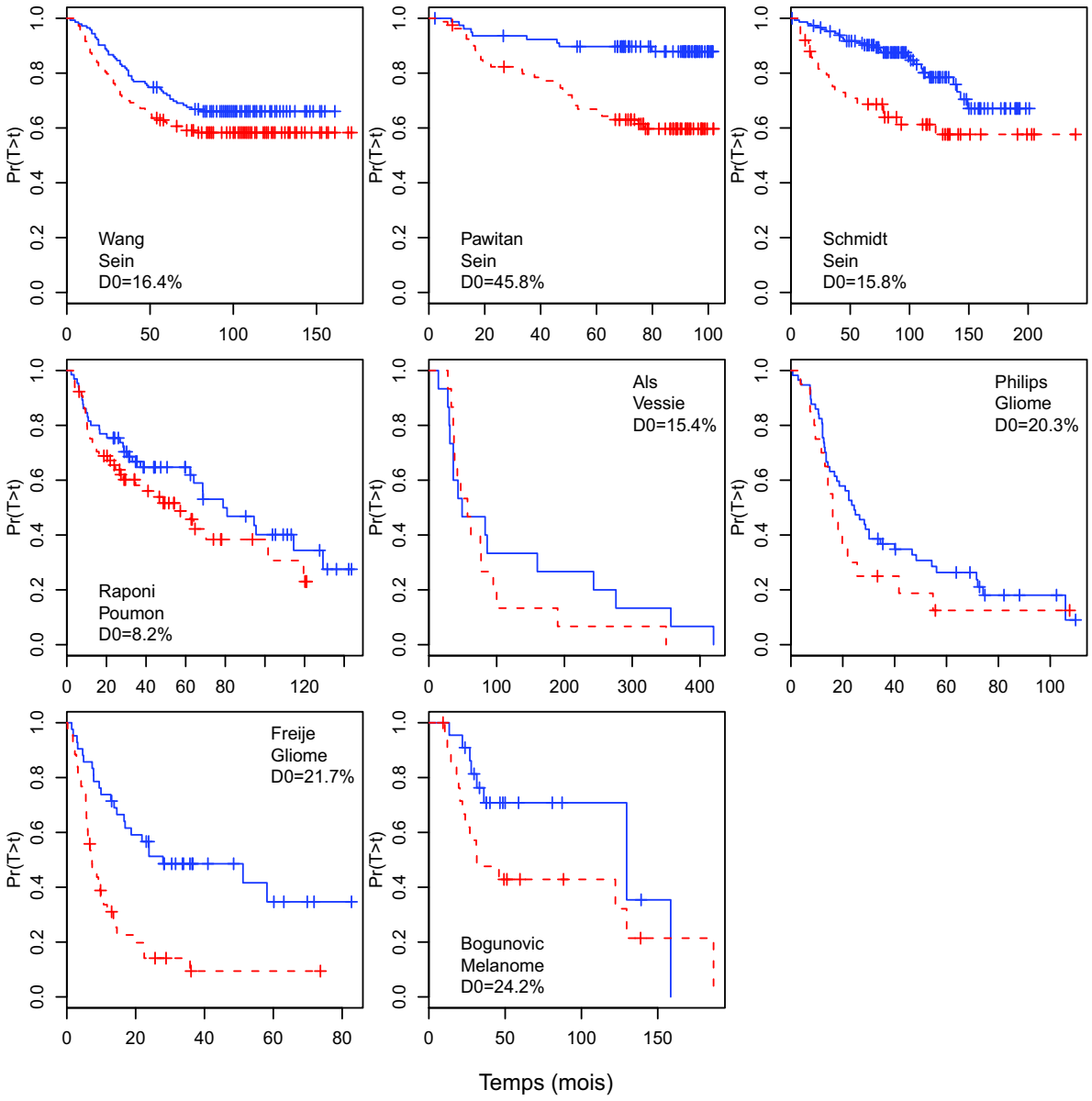


FIGURE C.5 – Courbes de survie en fonction du niveau d'expression du gène *ESPL1* (38158-at) dans les huit études. En bleu : faible niveau d'expression, en rouge : haut niveau d'expression



Annexe D

Codes R pour la programmation de l'indice

D.1 Indice sous le modèle de Cox à risques proportionnels

```
require(survival)

Indice.PH <- function(gene,ydelai,yetat){

  yetat.fake <- rep(1,length(ydelai))
  n <- length(ydelai)
  o <- order(ydelai)

  s.reg <- summary(survfit(Surv(ydelai,yetat)~1))
  s.fake <- summary(survfit(Surv(ydelai,yetat.fake)~1))
  di <- s.reg$n.event
  ni <- s.reg$n.risk[di>0]

  ei <- s.fake$n.event
  tps.tot <- s.fake$time
  tps.deces <- s.reg$time
  delti <- as.numeric(is.element(tps.tot,tps.deces))

  mat<- as.data.frame(matrix(nrow=length(tps.tot),ncol=10))
  names(mat) <- c("tps","n.risk","n.event","n.deces","delta","x.deces","
      x.event","x.risk")
  mat$tps <- tps.tot
  mat$n.risk <- s.fake$n.risk
  mat$n.event <- ei
  mat$delta <- delti
  deb <- c(1,cumsum(ei[-length(ei)])+1)
  finn <- deb+ei-1
  geneo <- gene[o]
  etato <- yetat[o]
  genetat <- geneo*etato
  for (i in 1:length(deb)) {
  mat$n.deces[i] <- sum(etato[deb[i]:finn[i]])
  mat$x.deces[i] <- sum(genetat[deb[i]:finn[i]])
  mat$x.event[i] <- sum(geneo[deb[i]:finn[i]])
  }
```

```
35    mat$x.risk <- rev(cumsum(rev(mat$x.event)))
36
37    Y <- matrix(1,ncol=length(mat$tps),nrow=length(mat$tps))
38    Y[upper.tri(Y)] <- 0
39
40    Ui <- mat$delta*(mat$x.deces-(mat$n.deces/mat$n.risk)*mat$x.risk)
41
42    v1.1 <- mat$delta*mat$n.deces/mat$n.risk
43    m1.1 <- matrix(v1.1,nrow=length(mat$tps),ncol=length(mat$tps),byrow=T)
          *Y
44    n1.1 <- m1.1*mat$x.event
45    v2.1 <- mat$delta*mat$n.deces/(mat$n.risk*mat$n.risk)
46    m2.1 <- matrix(v2.1,nrow=length(mat$tps),ncol=length(mat$tps),byrow=T)
          *Y
47    tmp2.1 <- m2.1*matrix(mat$n.event,ncol=length(mat$tps),nrow=length(mat
          $tps))
48    n2.1 <- matrix(mat$x.risk,ncol=length(mat$tps),nrow=length(mat$tps),
          byrow=T)*tmp2.1
49    EUij <- n1.1-n2.1
50    EUi <- apply(EUij,1,sum)
51    Wi <- Ui - EUi
52
53    Cox.PH <- sum(Wi)^2/sum(Wi^2)
54    Ind.PH <- Cox.PH/length(s.reg$time)
55
56  return(Ind.PH)
57  }
```

D.2 Indice sous le modèle à odds proportionnels

```
1   require(survival)
2   require(MASS)
3
4   Indice.POM <- function(gene,ydelai,yetat){
5
6     n <- length(ydelai)
7     o <- order(ydelai)
8
9     s.reg <- summary(survfit(Surv(ydelai,yetat)~1))
10    ni <- s.reg$n.risk
11
12    tps.deces <- s.reg$time
13    delti <- yetat
14
15    mat<- as.data.frame(matrix(nrow=n,ncol=7))
16    names(mat) <- c("tps","n.risk","delta","x.event","x.deces","x.risk","
          poids")
17    mat$tps <- round(ydelai[o],2)
18    mat$n.risk <- n:1
19    mat$delta <- delti[o]
20    geneo <- gene[o]
21    etato <- yetat[o]
22    mat$x.deces <- as.numeric(geneo*etato)
```

```
23    mat$x.event <- as.numeric(geneo)
24    mat$x.risk <- rev(cumsum(rev(as.numeric(geneo))))
25
26    St0 <- cumprod((mat$n.risk-mat$delta)/mat$n.risk)
27    omega <- c(1,St0[-length(St0)])
28    mat$poids <- omega*mat$delta
29
30    Y <- matrix(1,ncol=n,nrow=n)
31    Y[upper.tri(Y)] <- 0
32
33    Ui.tmp <- mat$delta*(mat$x.deces-(mat$x.risk/mat$n.risk))
34    Ui <- Ui.tmp*mat$poids
35
36    v1 <- mat$delta/mat$n.risk
37    m1 <- matrix(v1,nrow=n,ncol=n,byrow=T)*Y
38    n1 <- m1*mat$x.event
39    v2 <- mat$delta/(mat$n.risk*mat$n.risk)
40    m2 <- matrix(v2,nrow=n,ncol=n,byrow=T)*Y
41    n2 <- matrix(mat$x.risk,ncol=length(mat$tps),nrow=length(mat$tps),
          byrow=T)*m2
42    EUij.tmp <- n1-n2
43    EUij <- EUij.tmp* matrix(mat$poids,ncol=length(mat$tps),nrow=length(
          mat$tps),byrow=T)
44    EUi <- apply(EUij,1,sum)
45    Wi <- Ui - EUi
46    Cox <- sum(Wi)^2/sum(Wi^2)
47    Ind <- Cox /length(s.reg$time)
48  return(Ind)
49  }
```

D.3 Indice sous le modèle conduisant à un croisement des risques instantannés

```
1   Indice.AFT <- function(gene,ydelai,yetat){
2
3     n <- length(ydelai)
4     o <- order(ydelai)
5
6     s.reg <- summary(survfit(Surv(ydelai,yetat)~1))
7     ni <- s.reg$n.risk
8
9     tps.deces <- s.reg$time
10    delti <- yetat
11
12    mat<- as.data.frame(matrix(nrow=n,ncol=7))
13    names(mat) <- c("tps","n.risk","delta","x.event","x.deces","x.risk","
          poids")
14    mat$tps <- round(ydelai[o],2)
15    mat$n.risk <- n:1
16    mat$delta <- delti[o]
17    geneo <- gene[o]
18    etato <- yetat[o]
```

```
19    mat$x.deces <- as.numeric(geneo*etato)
20    mat$x.event <- as.numeric(geneo)
21    mat$x.risk <- rev(cumsum(rev(as.numeric(geneo))))
22
23    Kt0 <- cumsum(mat$delta/mat$n.risk)
24    Kt2 <- c(0,Kt0[-length(Kt0)])
25    omega <- ifelse(Kt2==0,1,1+log(Kt2))
26    mat$poids <- omega*mat$delta
27
28    Y <- matrix(1,ncol=n,nrow=n)
29    Y[upper.tri(Y)] <- 0
30
31    Ui.tmp <- mat$delta*(mat$x.deces-(mat$x.risk/mat$n.risk))
32    Ui <- Ui.tmp*mat$poids
33
34    v1 <- mat$delta/mat$n.risk
35    m1 <- matrix(v1,nrow=n,ncol=n,byrow=T)*Y
36    n1 <- m1*mat$x.event
37    v2 <- mat$delta/(mat$n.risk*mat$n.risk)
38    m2 <- matrix(v2,nrow=n,ncol=n,byrow=T)*Y
39    n2 <- matrix(mat$x.risk,ncol=length(mat$tps),nrow=length(mat$tps),
          byrow=T)*m2
40    EUij.tmp <- n1-n2
41    EUij <- EUij.tmp* matrix(mat$poids,ncol=length(mat$tps),nrow=length(
          mat$tps),byrow=T)
42    EUi <- apply(EUij,1,sum)
43    Wi <- Ui - EUi
44    Cox <- sum(Wi)^2/sum(Wi^2)
45    Ind <- Cox /length(s.reg$time)
46  return(Ind)
47  }
```

Annexe E

Articles

E.1 Identifying commun prognostic factors in genomic cancer studies : a novel index for censored outcomes

http://www.biomedcentral.com/1471-2105/11/150/abstract/

E.2 A pseudo-R^2 measure for selecting genomic markers with crossing hazard functions

http://www.biomedcentral.com/1471-2288/11/28

BIBLIOGRAPHIE

Aalen, O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics*, 6(4):701–726.

Agresti, A. (1986). Applying r2-type measures to ordered categorical data. *Technometrics*, 28(2):133–138.

Allison, P. D. (1995). *Survival Analysis Using SAS : A Practical Guide*. SAS Publishing.

Als, A. B., Dyrskjot, L., von der Maase, H., Koed, K., Mansilla, F., Toldbod, H. E., Jensen, J. L., Ulhoi, B. P., Sengelov, L., Jensen, K. M. E. et Orntoft, T. F. (2007). Emmprin and survivin predict response and survival following cisplatin-containing chemotherapy in patients with advanced bladder cancer. *Clinical Cancer Research*, 13(15):4407–4414.

Antolini, L., Boracchi, P. et Biganzoli, E. (2005). A time-dependent discrimination index for survival data. *Statistics in Medicine*, 24(24):3927–3944.

Bair, E. et Tibshirani, R. (2004). Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biology*, 2:E108.

Barrett, T., Suzek, T. O., Troup, D. B., Wilhite, S. E., Ngau, W.-C., Ledoux, P., Rudnev, D., Lash, A. E., Fujibuchi, W. et Edgar, R. (2005). Ncbi geo : mining millions of expression profiles-database and tools. *Nucleic Acids Research*, 33:D562–D566.

Bennett, S. (1983). Analysis of survival data by the proportional odds model. *Statistics in Medicine*, 2(2):273–277.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*. Springer, New York, USAork, USA, 2nd edition édition.

Bharadwaj, R. et Yu, H. (2004). The spindle checkpoint, aneuploidy, and cancer. *Oncogene*, 23(11):2016–2027.

Blangiardo, M. et Richardson, S. (2007). Statistical tools for synthesizing lists of differentially expressed features in related experiments. *Genome Biology*, 8(4):R54.

Bogunovic, D., O'Neill, D. W., Belitskaya-Levy, I., Vacic, V., Yu, Y.-L., Adams, S., Darvishian, F., Berman, R., Shapiro, R., Pavlick, A. C., Lonardi, S., Zavadil, J., Osman, I. et Bhardwaj, N. (2009). Immune profile and mitotic index of metastatic melanoma lesions enhance clinical staging in predicting patient survival. *Proceedings of the National Academy of Sciences of the United States of America*, 106(48):20429–20434.

Bolstad, B., Irizarry, R., Astrand, M. et Speed, T. (2003). A comparison of normalization methods for high density oligonucleotide array data. *Bioinformatics*, 19(2):185–193.

Breslow, N. E. (1972). Contribution to the discussion on the paper by d.r. cox. *Journal of the Royal Statistical Society Series B*, 34:216–217.

Brier, G. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1–3.

Broët, P., Camilleri-Broët, S., Zhang, S., Alifano, M., Bangarusamy, D., Battistella, M., Wu, Y., Tuefferd, M., Régnard, J.-F., Lim, E., Tan, P. et Miller, L. D. (2009). Prediction of clinical outcome in multiple lung cancer cohorts by integrative genomics : implications for chemotherapy selection. *Cancer Res*, 69(3):1055–62.

Broët, P., Rycke, Y. D., Tubert-Bitter, P., Lellouch, J., Asselain, B. et Moreau, T. (2001). A semiparametric approach for the two-sample comparison of survival times with long-term survivors. *Biometrics*, 57(3):844–852.

Brown, B., Hollander, M. et Korwar, R. (1974). Nonparametric tests of independence for censored data with application to heart transplant studies. *Reliability and Biometry*, 1:327–354.

Cai, T., Wei, L. J. et Wilcox, M. (2000). Semiparametric regression analysis for clustered failure time data. *Biometrika*, 87(4):867–878.

Cameron, A. et Windmeijer, F. (1996). R-Squared Measures for Count Data Regression Models with Applications to Health-Care Utilization. *Journal of Business & Economic Statistics*, 14(2):209–220.

Carter, S. L., Eklund, A. C., Kohane, I. S., Harris, L. N. et Szallasi, Z. (2006). A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. *Nature Genetics*, 38(9):1043–1048.

Cheng, S. C., Wei, L. J. et Ying, Z. (1995). Analysis of transformation models with censored data. *Biometrika*, 82(4):835–845.

Clayton, D. et Cuzick, J. (1986). The semiparametric pareto model for regression analysis of survival times. *In Papers on semiparametric models at the ISI centenary session*, Report MS-R8614, Amsterdam.

Conover, W. et Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician*, 35(3):124–129.

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

Dalmasso, C., Broët, P. et Moreau, T. (2005). A simple procedure for estimating the false discovery rate. *Bioinformatics*, 21(5):660–668.

Devarajan, K. et Ebrahimi, N. (2011). A semi-parametric generalization of the cox proportional hazards regression model : Inference and applications. *Computational Statistics and Data Analysis*, 55(1):667–676.

Efron, B. (1967). The two sample problem with censored data. *Proceedings of the Fifth Berkeley Symposium*, 4.

Efron, B. (1977). The efficiency of cox's likelihood function for censored data. *Journal of the American Statistical Association*, 72(359):557–565.

Efron, B. (1978). Regression and anova with zero-one data : measures of residual variation. *Journal of the American Statistical Association*, 73:113–121.

Fleming, T. et Harrington, D. (2005). *Counting processes and survival analysis*. Wiley, New York.

Freije, W. A., Castro-Vargas, F. E., Fang, Z., Horvath, S., Cloughesy, T., Liau, L. M., Mischel, P. S. et Nelson, S. F. (2004). Gene expression profiling of gliomas strongly predicts survival. *Cancer Research*, 64(18):6503–6510.

Gini, C. (1912). *Variabilita e mutabilita, contributo allo studio delle distribuzioni e delle relazioni statistiche*. Università de Cagliari.

Goodman, L. A. et Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, 49:732–764.

Graf, E., Schmoor, C., Sauerbrei, W. et Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Stat Med*, 18(17-18):2529–2545.

Haberman, S. (1982). Analysis of dispersion of multinomial responses. *Journal of the American Statistical Association*, 77:568–580.

Hacia, J. G., Sun, B., Hunt, N., Edgemon, K., Mosbrook, D., Robbins, C., Fodor, S. P., Tagle, D. A. et Collins, F. S. (1998). Strategies for mutational analysis of the large multiexon atm gene using high-density oligonucleotide arrays. *Genome Research*, 8(12):1245–1258.

Harrell, F., Califf, R., Pryor, D., Lee, K. et Rosati, R. (1982). Evaluating the yield of medical tests. *Journal of the American Medical Association*, 247(18):2543–2546.

Harrington, D. et Fleming, T. (1982). A class of rank test procedures for censored survival data. *Biometrika*, 69:133–143.

Heagerty, P., Lumley, T. et Pepe, M. (2000). Time-dependent roc curves for censored survival data and a diagnostic marker. *Biometrics*, 56(2):337–344.

Heagerty, P. et Zheng, Y. (2005). Survival model predictive accuracy and roc curves. *Biometrics*, 61(1):92–105.

Healy, J. (1984). The use of r2 as a measure of goodness of fit. *Journal of the Royal Statistical Society. Series A*, 147(4):608–609.

Hougaard, P. (1995). *Frailty models for survival data*. Lægeforeningens Forlag.

Kalbfleisch, J. D. et Prentice, R. L. (2002). *The statistical analysis of failure time data*. Wiley series in Probability and Mathematical Statistics. Wiley, New York.

Kaplan, E. et Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481.

Kato, T., Sato, N., Hayama, S., Yamabuki, T., Ito, T., Miyamoto, M., Kondo, S., Nakamura, Y. et Daigo, Y. (2007). Activation of holliday junction recognizing protein involved in the chromosomal stability and immortality of cancer cells. *Cancer Research*, 67(18):8544–8553.

Katoh, M. (2008). Cancer genomics and genetics of fgfr2 (review). *International Journal of Oncology*, 33(2):233–237.

Kendall, M. et Gibbons, J. (1990). *Rank correlation methods.* Edward Arnold, Oxford University Press, 5th edition édition.

Kent, J. T. et O'Quigley, J. (1988). Measures of dependence for censored survival data. *Biometrika*, 75(3):525–534.

Korn, E. et Simon, R. (1990). Measures of explained variation for survival data. *Statistics in Medicine*, 9(5):487–503.

Korn, E. L. et Simon, R. (1991). Explained residual variation, explained risk, and goodness of fit. *The American Statistician*, 45(3):201–206.

Kullback, S. et Leibler, R. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86.

Lai, T. L., Shen, D. B. et Gross, S. (2010). Evaluating probability forecasts. Submitted to the Annals of Statistics.

Lin, D. Y. et Wei, L. J. (1989). The robust inference for the cox proportional hazards model. *Journal of the American Statistical Association*, 84:1074–1078.

Linde, A. V. D. et Tutz, G. (2008). On association in regression : the coefficient of determination revisited. *Statistics*, 42(1):1–24.

Lininger, L., Gail, M. H., Green, S. B. et Byar, D. P. (1979). Comparison of four tests for equality of survival curves in the presence of stratification and censoring. *Biometrika*, 66(3):419–428.

Ma, S. et Huang, J. (2007). Combining multiple markers for classification using roc. *Biometrics*, 63(3):751–757.

Maddala, G. (1983). *Limited-Dependent and Qualitative Variables in Econometrics.* Cambridge University Press.

Magee, L. (1990). R2 measures based on wald and likelihood ratio joint significance tests. *The American Statistician*, 44(3):250–253.

Mann, H. B. et Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60.

Mardia, K., Kent, J. et Bibby, J. (1979). *Multivariate Analysis.* Academic Press.

Martin, B., Paesmans, M., Berghmans, T., Branle, F., Ghisdal, L., Mascaux, C., Meert, A.-P., Steels, E., Vallot, F., Verdebout, J.-M., Lafitte, J.-J. et Sculier, J.-P. (2003). Role of bcl-2 as a prognostic factor for survival in lung cancer : a systematic review of the literature with meta-analysis. *British Journal of Cancer*, 89(1):55–64.

Miller, J. K., Shattuck, D. L., Ingalla, E. Q., Yen, L., Borowsky, A. D., Young, L. J. T., Cardiff, R. D., Carraway, K. L. et Sweeney, C. (2008). Suppression of the negative regulator lrig1 contributes to erbb2 overexpression in breast cancer. *Cancer Research*, 68(20): 8286–8294.

Mittlböck, M. et Schemper, M. (1996). Explained variation for logistic regression. *Statistics In Medicine*, 15(19):1987–1997.

Murphy, S. A., Rossini, A. J. et van der Vaart, A. W. (1997). Maximum likelihood estimation in the proportional odds model. *Journal of the American Statistical Association*, 92(439):968–976.

Nagelkerke, N. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, 78(3):691–692.

Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics*, 42(1):945–965.

O'Quigley, J., Xu, R. et Stare, J. (2005). Explained randomness in proportional hazards models. *Stat Med*, 24(3):479–89.

Parkinson, H., Kapushesky, M., Kolesnikov, N., Rustici, G., Shojatalab, M., Abeygunawardena, N., Berube, H., Dylag, M., Emam, I., Farne, A., Holloway, E., Lukk, M., Malone, J., Mani, R., Pilicheva, E., Rayner, T. F., Rezwan, F., Sharma, A., Williams, E., Bradley, X. Z., Adamusiak, T., Brandizi, M., Burdett, T., Coulson, R., Krestyaninova, M., Kurnosov, P., Maguire, E., Neogi, S. G., Rocca-Serra, P., Sansone, S.-A., Sklyar, N., Zhao, M., Sarkans, U. et Brazma, A. (2009). Arrayexpress update-from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Research*, 37:D868–D872.

Pawitan, Y., Bjöhle, J., Amler, L., Borg, A.-L., Egyhazi, S., Hall, P., Han, X., Holmberg, L., Huang, F., Klaar, S., Liu, E. T., Miller, L., Nordgren, H., Ploner, A., Sandelin, K., Shaw, P. M., Smeds, J., Skoog, L., Wedrén, S. et Bergh, J. (2005). Gene expression profiling spares early breast cancer patients from adjuvant therapy : derived and validated in two population-based cohorts. *Breast Cancer Research*, 7(6):R953–64.

Pencina, M. J. et D'Agostino, R. B. (2004). Overall c as a measure of discrimination in survival analysis : model specic population value and condence overall c as a measure of discrimination in survival analysis : model specic population value and condence interval estimation. *Statistics In Medicine*, 23(13):2109–2123.

Peto, R. (1972). Contribution to the discussion of the paper by dr cox. *Journal of the Royal Statistical Society Series B*, 34:205–207.

Phillips, H. S., Kharbanda, S., Chen, R., Forrest, W. F., Soriano, R. H., Wu, T. D., Misra, A., Nigro, J. M., Colman, H., Soroceanu, L., Williams, P. M., Modrusan, Z., Feuerstein, B. G. et Aldape, K. (2006). Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. *Cancer Cell*, 9(3):157–173.

Pierce, D. A. (1982). The asymptotic effect of substituting estimators for parameters in certain types of statistics. *The Annals of Statistics*, 10(2):475–478.

Qiu, X., Klebanov, L. et Yakovlev, A. (2005). Correlation between gene expression levels and limitations of the empirical bayes methodology for finding differentially expressed genes. *Statistical pplications in Genetics and Molecular Biology*, 4(1):Article 34.

Quantin, C., Moreau, T., Asselain, B., Maccario, J. et Lellouch, J. (1996). A regression survival model for testing the proportional hazards hypothesis. *Biometrics*, 52(3):874–885.

Raponi, M., Zhang, Y., Yu, J., Chen, G., Lee, G., Taylor, J. M. G., Macdonald, J., Thomas, D., Moskaluk, C., Wang, Y. et Beer, D. G. (2006). Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung. *Cancer Research*, 66(15):7466–7472.

Read, C. B. (2006). *Encyclopedia of Statistical Sciences*, chapitre Mean Deviation, pages 4655–4656. Wiley-Interscience.

Rhodes, D. R., Kalyana-Sundaram, S., Mahavisno, V., Varambally, R., Yu, J., Briggs, B. B., Barrette, T. R., Anstet, M. J., Kincead-Beal, C., Kulkarni, P., Varambally, S., Ghosh, D. et Chinnaiyan, A. M. (2007). Oncomine 3.0 : genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia*, 9:166–80.

Royston, P. et Sauerbrei, W. (2004). A new measure of prognostic separation in survival data. *Stat Med*, 23(5):723–748.

Rubin, D. (1987). *Multiple imputation for nonresponse in surveys*. Wiley.

Rubin, D. (1991). Em and beyond. *Psychometrika*, 56(2):241–254.

Schemper, M. (1990). The explained variation in proportional hazards regression. *Biometrika*, 77(1):216–218.

Schemper, M. (1994). Correction : the explained variation in proportional hazards regression. *Biometrika*, 81(3):361.

Schemper, M. (2003). Predictive accuracy and explained variation. *Stat Med*, 22(14):2299–2308.

Schemper, M. et Henderson, R. (2000). Predictive accuracy and explained variation in cox regression. *Biometrics*, 56(1):249–255.

Schemper, M. et Kaider, A. (1997). A new approach to estimate correlation coefficients in the presence of censoring and proportional hazards. *Computational Statistics and Data Analysis*, 23(4):467–476.

Schemper, M. et Smith, T. L. (1996). A note on quantifying follow-up in studies of failure time. *Controlled Clinical Trials*, 17(4):343–346.

Schemper, M. et Stare, J. (1996). Explained variation in survival analysis. *Statistics in Medicine*, 15(19):1999–2012.

Schmidt, M., Böhm, D., von Törne, C., Steiner, E., Puhl, A., Pilch, H., Lehr, H.-A., Hengstler, J. G., Kölbl, H. et Gehrmann, M. (2008). The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer Res*, 68(13):5405–13.

Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27.

Snedecor, G. et Cochran, W. (1989). *Statistical methods*. Iowa State University Press, 8tn edition édition.

Somers, R. (1962). A new asymmetric measure of association for ordinal variables. *American Sociological Review*, 27(6):799–811.

Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A. et Narechania., A. (2003). Panther : a library of protein families and subfamilies indexed by function. *Genome Research*, 13:2191–2141.

Vaupel, J. W., Manton, K. G. et Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16(3):439–454.

Wang, Y., Klijn, J. G. M., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., Talantov, D., Timmermans, M., van Gelder, M. E. M., Yu, J., Jatkoe, T., Berns, E. M. J. J., Atkins, D. et Foekens, J. A. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, 365:671–679.

Weibull, W. (1951). A statistical distribution function of wide applicability. *Journal of Applied Mechanics*, 18:292–297.

Winkler, R. (1967). The quantification of judgement : some methodological suggestions. *Journal of the American Statistical Association*, 62:1105–1120.

Winkler, R. et Murphy, A. (1968). "good" probability assessors. *Journal of Applied Meteorology*, 7:751–758.

Wu, G., Zhou, L., Khidr, L., Guo, X. E., Kim, W., Lee, Y. M., Krasieva, T. et Chen, P.-L. (2008). A novel role of the chromokinesin kif4a in dna damage response. *Cell Cycle*, 7(13):2013–2020.

Wu, H.-D. I. (2007). A partial score test for difference among heterogeneous populations. *Journal of Statistical Planning and Inference*, 137(2):527–537.

Xu, R. et O'Quigley, J. (1999). A r2 type measure of dependence for proportional hazards models. *Journal of Nonparametric Statistics*, 12(1):83–107.

Yang, S. et Prentice, R. L. (1999). Semiparametric inference in the proportional odds regression model. *Journal of the American Statistical Association*, 94(445):125–136.

Zhang, N., Ge, G., Meyer, R., Sethi, S., Basu, D., Pradhan, S., Zhao, Y.-J., Li, X.-N., Cai, W.-W., El-Naggar, A. K., Baladandayuthapani, V., Kittrell, F. S., Rao, P. H., Medina, D. et Pati, D. (2008). Overexpression of separase induces aneuploidy and mammary tumorigenesis. *Proceedings of the National Academy of Sciences of the United States of America*, 105(35):13033–13038.

Zinkel, S., Gross, A. et Yang, E. (2006). Bcl2 family in dna damage and cell cycle control. *Cell Death Differ*, 13(8):1351–1359.