UCLouvain

FACULTE DES SCIENCES

Faculté des sciences

Analyse de données de survie en présence d'une fraction d'individus "cure" à l'aide de modèles de "frailty"

Auteur: **Catherine BERTRAND**
Promoteur: **Catherine LEGRAND**
Lecteur: **Aurélie BERTRAND**
Année académique 2023–2024
Master [120] en statistique, orientation biostatistiques

Résumé

Dans ce mémoire, nous nous intéressons à des données comprenant une proportion d'individus cure. Plutôt que d'utiliser les modèles de cure habituels, nous optons pour les modèles de frailty. Nous utilisons des modèles de frailty univariés avec une distribution de frailty discrète afin de donner un poids à la frailty en 0, représentant ainsi les individus cure. Nous explorons d'abord trois distributions discrètes simples : la distribution de Poisson, la géométrique et la binomiale négative. Nous construisons des modèles avec ces trois distributions, puis procédons à des simulations pour évaluer leurs performances sur différents échantillons de données. Les simulations révèlent que la proportion de cure est systématiquement sous-estimée par les modèles. Pour remédier à cela, nous complexifions nos trois distributions pour augmenter le nombre de zéros. Nous utilisons pour cela des distributions de la famille ZMPS (Zero Modified Power Series). Ces distributions ajoutent un paramètre p pour augmenter ou diminuer le nombre de zéros des distributions de la famille PS (power series). Nous construisons trois modèles avec des distributions de la famille ZMPS : la ZMP (Zero Modified Poisson), la ZMG (Zero Modified géométrique) et la ZMBN (Zero Modified binomial négative). Nous constatons à travers des simulations que cela améliore l'estimation du taux de cure. Ensuite, nous explorons la distribution compound Poisson comme distribution de la frailty. C'est une distribution continue à laquelle nous ajoutons une composante discrète pour inclure un poids en 0 et pouvoir estimer une proportion de cure. La distribution compound Poisson est encore une fois évaluée via des simulations pour évaluer son comportement sur différentes données. Enfin, nous appliquons tous ces modèles à la base de données "retinopathy" du package "survival" en R, pour discuter les résultats obtenus avec ces différents modèles.

Remerciements

Je tiens tout d'abord à exprimer ma plus profonde gratitude à ma promotrice, Catherine Legrand, pour m'avoir accompagnée avec autant de bienveillance et d'exigence tout au long de ce mémoire, ainsi que pendant mes trois années de master. Elle m'a aidée et poussée à donner le meilleur de moi-même.

Je remercie également mes lecteurs de consacrer de leur temps pour lire mon travail.

Je souhaite aussi exprimer ma reconnaissance à tous mes professeurs de mathématiques et de biostatistiques qui m'ont suivie et formée durant mon parcours.

Un merci particulier à mes parents : à ma maman, qui est plutôt littéraire, pour ses nombreuses relectures de mon orthographe et pour avoir appris à utiliser LaTeX et ce qu'est une "frailty" spécialement pour m'aider. Et à mon papa, de m'avoir transmis le goût des mathématiques et des statistiques. Merci à eux de m'avoir soutenue dans mes échecs et mes réussites, et de m'avoir encouragée jusqu'au bout.

Merci à tous mes proches pour leur soutien et leurs encouragements.

Table des matières

Chapitre 1

Introduction

L'objectif de ce mémoire est de modéliser des données de survie dans lesquelles une partie de la population est considérée comme "guérie" (cured), mais au lieu d'utiliser des modèles de cure traditionnels, nous allons opter pour des modèles de frailty. Nous allons utiliser un modèle de frailty univarié avec une distribution de frailty discrète et non négative. Cette approche nous permettra de mettre un certain poids à la valeur de la frailty 0. Les individus avec une frailty nulle correspondent alors aux individus "cured". Nous allons passer en revue une série de distributions discrètes pour la frailty afin de déterminer les meilleurs modèles à utiliser selon les données de survie à analyser.

Nous commencerons dans le chapitre 2 par rappeler les concepts fondamentaux de l'analyse de survie. Dans la section 2.1, nous aborderons les principes de base tels que la fonction de survie, la fonction de hazard, ainsi que les différents types de modèles de survie paramétriques et non paramétriques, et enfin les méthodes d'estimation des paramètres. Ensuite, dans la section 2.2, nous réviserons les modèles de frailty en définissant ce qu'est une frailty et en discutant des formes de ces modèles, des frailty univariés/multivariés et des méthodes d'estimation des paramètres. Enfin, dans la dernière section 2.3, nous ferons un rappel sur ce qu'est un individu cure, sur les modèles de cure, et plus particulièrement sur les "mixture cure model" (MCM).

Le chapitre 3 introduira les modèles de survie avec une frailty discrète, en construisant le modèle et en montrant ses particularités. Nous construirons le modèle avec trois distributions discrètes différentes, à savoir dans la section 3.1 une distribution de Poisson, dans la section 3.2 une distribution géométrique, et dans la section 3.3 une distribution binomiale négative.

Dans le chapitre 4, nous expliquerons l'importance des simulations. Nous réali-

serons deux types de simulations : d'abord en générant des échantillons de données selon un modèle de frailty avec une distribution discrète, puis en modélisant les échantillons de données avec le même modèle (section 4.1) ; et ensuite en générant les échantillons avec un MCM, puis en les modélisant avec les 3 modèles de frailty avec une distribution discrète construits précédemment (section 4.2).

Nous verrons grâce à ces simulations que les trois distributions discrètes (Poisson, géométrique et binomiale négative) sont trop simples. Dans ces distributions, la proportion d'individus avec une frailty égale à 0 par rapport aux proportions des autres valeurs de la frailty est trop contraint. Nous voudrons donc tester d'autres distributions discrètes pour la frailty afin de pouvoir avoir suffisamment de 0 et une diversité dans la proportion de valeurs non-nulles de la frailty. Tel sera l'objet du chapitre 5. Nous commencerons par expliquer ce que sont les distributions PS (section 5.1), puis les distributions ZMPS (section 5.2). Nous construirons ensuite un modèle de frailty avec une distribution ZMPS (section 5.3). Enfin, nous réaliserons des simulations pour voir comment se comporte ce modèle sur différents échantillons de données (chapitre 6).

Dans le chapitre 7, nous allons introduire un modèle de frailty où la distribution de la frailty utilisée sera une compound Poisson. C'est une distribution continue à laquelle on ajoute une partie discrète pour donner un poids à la valeur de la frailty 0. On va commencer par présenter la distribution compound Poisson (section 7.1), ensuite construire le modèle de survie (section 7.2), et enfin passer à des simulations (chapitre 8).

Pour finir, dans le chapitre 9, on va appliquer nos différents modèles à la base de données "retinopathy" du package "survival" en R. On regardera comment nos modèles estiment la proportion de cure de cette base de données et on discutera les résultats obtenus avec ces différents modèles.

Chapitre 2

Analyse de survie

Dans ce chapitre, on rappelle les concepts fondamentaux de l'analyse de survie classique tirés du cours "Analysis of survival and duration data" d'I.Van keilegom [16]. On va ensuite faire des rappels des modèles de "frailty" et des modèles de "cure" tirés du cours "Advanced Survival Models" de C.Legrand [8] et du livre "Advanced survival models" [7].

2.1 Analyse de survie classique

En analyse de survie classique, l'objectif principal est d'examiner le temps écoulé depuis une certaine origine jusqu'à la survenue d'un événement spécifique. Cet événement peut être la mort d'un patient, l'atteinte d'une certaine taille par une tumeur, l'arrivée à un stade particulier d'une maladie, etc.

Nous allons d'abord définir la variable aléatoire T, continue et non négative, qui représente le temps écoulé depuis une origine choisie jusqu'à l'occurrence d'un événement d'intérêt.

Voici plusieurs fonctions de base que l'on va utiliser tout au long de ce mémoire :

La fonction de survie qui représente la probabilité qu'un individu expérimente l'événement après un temps donné t :

$$S(t) = \mathbb{P}(T > t),$$

c'est une fonction décroissante non-négative, de plus au temps 0, $S(0) = 1$ et à un temps infini $\lim_{t\to\infty} S(t) = 0$.

La fonction de hazard qui représente la probabilité conditionnelle instantanée que l'événement se produise au temps t, étant donné que l'événement ne s'est pas encore produit :

$$h(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t < T < t + \Delta t | T > t)}{\Delta t}.$$

La fonction de hazard cumulatif,

$$H(t) = \int_0^\infty h(u)du.$$

Ces trois fonctions sont liées par l'équation ci-dessous :

$$S(t) = \exp\left(-\int_0^t h(u)du\right) = \exp(-H(t)). \tag{2.1}$$

Nous avons également la fonction de répartition de T qui représente la probabilité que l'événement survienne avant le temps t :

$$F(t) = P(T \leq t) = 1 - S(t)$$

c'est une fonction croissante. De plus, au temps 0, $F(t) = 0$ et à un temps infini $\lim_{t\to\infty} F(t) = 1$. C'est à dire que au temps 0, personne n'a expérimenté l'événement et à un temps ∞, tout le monde a expérimenté l'événement.

La fonction de densité de T :

$$f(t) = \frac{d}{dt}F(t) = -\frac{d}{dt}S(t). \tag{2.2}$$

Une caractéristique importante que l'on retrouve dans l'analyse de survie est la présence de censure. La censure se produit lorsque le moment survenue de l'événement n'est pas connu. La censure peut se produire pour différentes raisons : la fin de l'étude sans qu'un patient n'ait eu l'événement, un patient qui quitte l'étude pour différentes raisons, ... Le temps de censure est une variable aléatoire positive C qui suit une certaine fonction de distribution. On va se concentrer dans ce mémoire sur la censure à droite. On dira qu'une personne est censurée à droite quand $min(T, C) = C$, et qu'une personne est non-censurée quand $min(C, T) = T$. On associe à cette censure un indicateur de censure $\delta = I(T \leq C)$. Une personne censurée aura donc $\delta = 0$ et une personne qui expérimente l'événement aura $\delta = 1$. On va faire l'hypothèse que le temps de censure est indépendant du temps de l'événement ($C \perp T$).

Si notre étude concerne n individus, les données de survie de l'individu $i = 1, \ldots, n$ sont représentées par le vecteur (y_i, δ_i), où $y_i = \min(c_i, t_i)$. Dans cette

équation, c_i représente le temps de censure et t_i représente le temps d'événement de l'individu i, tandis que δ_i est défini comme indiqué précédemment.

Le but principal de l'analyse de survie est d'estimer les paramètres de la fonction de survie. Si l'on considère n observations indépendantes, si l'on suppose que les temps de censure c_i sont indépendants des temps de d'événement t_i, et que l'on considère seulement la censure à droite, la fonction de vraisemblance sera la suivante :

$$L = \prod_{i=1}^{n} h(y_i)^{\delta_i} S(y_i)$$

Par la fonction (2.1) et la formule de la dérivée du logarithme, on peut réécrire la fonction de vraisemblance en termes de la fonction de densité $f(y_i)$ (2.2) :

$$\begin{aligned} L &= \prod_{i=1}^{n} \left[-\frac{d}{dy_i} \log(S(y_i)) \right]^{\delta_i} S(y_i) \\ &= \prod_{i=1}^{n} \left[-\frac{d}{dy_i} S(y_i) \right]^{\delta_i} \frac{1}{S(y_i)^{\delta_i}} S(y_i) \\ &= \prod_{i=1}^{n} f(y_i)^{\delta_i} (S(y_i))^{1-\delta_i}. \end{aligned} \tag{2.3}$$

On peut estimer paramétriquement les fonctions de survie et de hazard en supposant que T suive une distribution paramétrique non négative par exemple une distribution exponentielle $h(t) = \lambda$ avec $\lambda > 0$ et donc $S(t) = \exp(-\lambda t)$, une distribution de Weibull $h(t) = \rho\lambda t^{\rho-1}$ avec $\rho, \lambda > 0$ et donc $S(t) = \exp(-\lambda t^{\rho})$,... Pour estimer les paramètres, on va remplacer dans la formule de la vraisemblance (2.3) les fonctions de survie et de hazard par la distribution paramétrique choisie. Puis, on va maximiser cette fonction de vraisemblance pour obtenir les estimateurs du maximum de vraisemblance des paramètres de la distribution choisie. C'est sur cette estimation paramétrique que l'on va se concentrer dans ce mémoire.

Il existe également des estimateurs non-paramétriques, tel que l'estimateur de Kaplan-Meier [7], pour la fonction de survie. On considère ici plus l'ordre dans lequel les événements et les censures se passent que leur temps d'événement/de censure. Cet estimateur a la forme suivante :

$$\hat{S}(t) = \prod_{j:t_{(j)} \leq t} \left(1 - \frac{d_{(j)}}{n_{(j)}} \right) \tag{2.4}$$

avec $t_{(j)}$ le j ème temps où un événement s'est produit, $d_{(j)}$ le nombre d'événements survenus au temps $t_{(j)}$ et $n_{(j)}$ le nombre de sujets encore à risque juste avant le temps $t_{(j)}$.

En analyse de survie, il est important de pouvoir prendre en compte l'effet de covariables dans les modèles. Par exemple l'effet du sexe, de l'âge, d'un traitement, etc sur la survie des individus. Il existe plusieurs modèles paramétriques qui prennent en considération ces covariables, et parmi eux, un modèle largement utilisé est le modèle des hazards proportionnels.

Celui-ci s'exprime comme suit :

$$h_i(t) = h_b(t) \exp(\beta' X_i) \tag{2.5}$$

où $i = 1, \ldots, n$ avec n le nombre d'individus, $X = (X_{i1}, X_{i2}, ..., X_{ip})$ est le vecteur des covariables avec p le nombre de covariables, $\beta' = (\beta_1, \beta_2, ..., \beta_p)'$ est la transposée du vecteur des coefficients mesurant l'impact des covariables, et $h_b(t)$ est la fonction de hazard de base (elle est supposée identique pour tous les individus).

On peut également grâce à la fonction (2.1) le réécrire en termes de la fonction de survie :

$$S_i(t) = S_b(t)^{\exp(\beta' X_i)} \tag{2.6}$$

avec $S_b(t)$ la fonction de survie de base associée à la fonction de hazard de base $h_b(t)$ grâce à (2.1).

L'hypothèse de base que l'on pose pour ce modèle est celle des hazards proportionnels : le ratio des fonctions de hazard de deux individus avec des covariables x_i et x_j ne dépend pas du temps,

$$\frac{h_i(t)}{h_j(t)} = \frac{h_b(t) \exp(\beta' x_i)}{h_b(t) \exp(\beta' x_j)} = \frac{\exp(\beta' x_i)}{\exp(\beta' x_j)}.$$

La fonction de hazard de base $h_b(t)$ peut être choisie de façon à ce que T suive une distribution paramétrique comme une distribution exponentielle, une distribution de Weibull, une distribution log-normal...

Dans le cas paramétrique, l'estimation des paramètres se fera par l'estimation du maximum de vraisemblance, avec comme formule pour la vraisemblance :

$$
\begin{aligned}
L &= \prod_{i=1}^{n} h_i(y)^{\delta_i} S_i(y) \\
&= \prod_{i=1}^{n} \Big(h_b(y) \exp(\beta' X_i) \Big)^{\delta_i} S_b(y)^{\exp(\beta' X_i)} \\
&= \prod_{i=1}^{n} \Big(f_b(y) \exp(\beta' X_i) \Big)^{\delta_i} \Big(S_b(y)^{\exp(\beta' X_i)} \Big)^{1-\delta_i}
\end{aligned} \tag{2.7}
$$

avec $y_i = min(c_i, t_i)$, δ_i qui est l'indicateur de censure, n le nombre d'individus, $S_i(.)$ la fonction de survie (2.6), $h_i(.)$ la fonction de hazard (2.5) et $f_b(.)$ la fonction de densité de base associée à $S_b(.)$. Pour passer à la fonction de vraisemblance en terme de $f_b(.)$, on procède comme dans (2.3).

On peut également laisser la fonction de hazard de base du modèle des hazards proportionnels non spécifiée, ce qui donnera le "modèle de Cox" [7]. Par ailleurs, il existe d'autres modèles paramétriques qui ne font pas l'hypothèse des hazards proportionnels tel que le "accelerated failure time model" (AFT) [7].

Dans les sections suivantes, on va introduire des modèles plus complexes tels que les modèles de "frailty" et les modèles de "cure". Ces modèles offrent une plus grande flexibilité pour analyser les données de survie dans des contextes réels.

2.2 Modèle de frailty

Dans le modèle des hazards proportionnels (2.5), les individus avec les mêmes valeurs des covariables x_i ont les mêmes valeurs de hazard. On suppose dans le modèle des hazards proportionnels que les covariables capturent toute la variabilité entre les individus, et donc que la population est homogène.

Mais dans la vie réelle, une variabilité supplémentaire entre les individus qui ont les mêmes valeurs de covariables peut exister ; ils peuvent avoir un hazard différent malgré leurs valeurs de covariables identiques. En effet, le modèle n'est jamais entièrement spécifié. Une certaine partie de l'information ne sera pas prise en compte par les covariables du modèle car on ne sait jamais tout mesurer, tout observer dans des situations réelles. Il y aura entre les individus une certaine hétérogénéité non observée.

Les modèles de frailty que l'on va introduire dans cette section sont des modèles développés pour tenir compte de cette hétérogénéité non observée. Pour la prendre

en compte, ce modèle va introduire une variable aléatoire non négative Z appelée "frailty" (fragilité en français) dans la fonction de hazard.

La fonction de hazard conditionnel à la frailty a la forme suivante :

$$h(t|Z = z, X) = zh_b(t) \exp(\beta' X). \tag{2.8}$$

Elle exprime le hazard étant donné la valeur des covariables X et la valeur spécifique de la frailty Z.

La fonction de survie conditionnelle à la frailty est la suivante :

$$S(t|Z = z, X) = \mathbb{P}(T > t|Z = z, X) = \exp\left(-z \exp(\beta' X) \int_0^t h_b(u)du\right) \tag{2.9}$$

Elle représente la probabilité que l'événement survienne après un temps donné t étant donné les covariables X et la valeur spécifique de la frailty Z.

Comme la frailty Z multiplie la fonction de hazard, elle doit être non-négative. La distribution de la frailty pourra être n'importe quelle distribution positive et on notera sa fonction de densité $f_Z(z)$. Pour éviter les problèmes d'identifiabilité, cette distribution sera souvent standardisée en $E(Z) = 1$. La variance $V(Z)$ peut être vue comme le paramètre d'hétérogénéité des hazards dans la population. Parmi les distributions souvent utilisées, il y a la gamma, la log-normal, la inverse-Gaussian, ... qui sont toutes des distributions continues. Comme expliqué dans l'introduction, nous allons dans ce mémoire considérer des distributions de frailty discrètes plus au moins compliquées.

Il existe deux types de modèles de frailty :

Le premier est le modèle de frailty multivarié qui s'applique au cas où la population est divisée en différents clusters. Les individus d'un même cluster auront la même valeur de la frailty. Pour l'individu i avec $i = 1, ...n_j$, du cluster j avec $j = 1, ..., J$ avec les covariables X_{ij}, la fonction de hazard conditionnelle est la suivante :

$$h_{ij}(t|Z, X) = h(t|Z = z_i, X_{ij}) = z_i h_b(t) \exp(\beta' X_{ij}).$$

Les individus qui viennent d'un cluster avec une frailty élevée ont un hazard plus grand de subir l'événement, tandis que ceux qui viennent d'un cluster avec une frailty plus faible ont moins de chance de le subir.

Le deuxième modèle est celui avec une frailty univariée. Ici chaque individu de la population aura sa propre frailty. Pour la suite de ce mémoire, on va parler uniquement de ce modèle de frailty univarié, dans lequel la fonction de hazard conditionnelle à la frailty Z pour un individu i $(i = 1, ..., n)$ avec les covariables X_i est la suivante :

$$h_i(t|Z, X) = h(t|Z = z_i, X_i) = z_i h_b(t) \exp(\beta' X_i). \tag{2.10}$$

Le hazard global, aussi appellé hazard marginal, a la forme suivante :

$$h(t) = E(h(t|Z, X)), \tag{2.11}$$

avec l'espérance qui est prise par rapport à la distribution de la frailty.

Les individus avec une frailty élevée ont un hazard plus grand de subir l'événement que ceux avec une frailty plus faible. Ainsi, au sein de la population étudiée, les individus avec une frailty élevée quitteront plus vite l'étude que les individus avec une petite frailty car ils expérimenteront l'événement plus tôt. Au plus on avance dans le temps, au plus il ne restera dans l'étude que des individus avec une faible frailty, et donc le hazard global aura tendance à diminuer au cours du temps.

La fonction de survie globale sur la population, ou autrement dit la fonction de survie non-conditionnelle, a la forme suivante :

$$S(t) = E(S(t|Z, X)), \tag{2.12}$$

avec l'espérance également prise par rapport à la distribution de la frailty.

Le modèle de frailty peut être paramétrique ou semi-paramétrique. Si la fonction de hazard de base $h_b(t)$ est laissée non-spécifiée, le modèle de frailty sera semi-paramétrique. Si la fonction de base du hazard est spécifiée de manière paramétrique avec par exemple une exponentielle, ou une Weibull, l'estimation des paramètres du modèle paramétrique se fait par la maximisation de la log-vraisemblance marginale.

La fonction de vraisemblance conditionnelle pour l'individu i de la population de n individus est la suivante :

$$\begin{aligned} L_i(\zeta, \beta|z_i) &= h_i(y|Z, X)^{\delta_i} S_i(y|Z, X) \\ &= (z_i h_b(y_i) \exp(\beta' X_i))^{\delta_i} \exp\left(-z_i \exp(\beta' X_i) \int_0^t h_b(u)du\right) \end{aligned} \tag{2.13}$$

avec ζ les paramètres de la fonction de hazard de base h_b, $h_i(y|Z,X)$ spécifiée comme dans (2.10) et $S_i(y|Z,X)$ la fonction de survie conditionnelle associée, et avec β, X_i, y_i et δ_i spécifiés comme dans la section (2.1).

La fonction de vraisemblance marginale de l'individu i est la suivante :

$$L_{marg,i}(\xi) = \int_0^\infty h_i(y|Z,X)^{\delta_i} S_i(y|Z,X) f_Z(z) dz \tag{2.14}$$

avec $f_Z(z)$ la fonction de distribution de la frailty, et ξ qui contient les coefficients β, les paramètres de la fonction de hazard de base ζ et les paramètres de la fonction de distribution de la frailty θ.

Et pour finir, la fonction de vraisemblance marginale est donnée par :

$$L_{marg}(\xi) = \prod_{i=1}^{n} L_{marg,j}(\xi) \tag{2.15}$$

Il suffit alors d'estimer les paramètres du modèle en maximisant le log de cette fonction.

On va maintenant introduire la transformée de Laplace qui va nous aider à construire des fonctions de survie.

Pour une variable aléatoire non négative W avec une fonction de densité $f_W(\cdot)$, la transformée de Laplace est définie par :

$$\begin{aligned} \mathcal{L}(s) &= E[\exp(-Ws)] \\ &= \int_0^\infty \exp(-sw) f_W(w) dw \end{aligned} \tag{2.16}$$

Nous pouvons ainsi réécrire la fonction de survie globale du modèle de frailty en termes de la transformée de Laplace :

$$\begin{aligned} S(t) &= E[S(t|Z)] \\ &= E\left[\exp\left(-Z\int_0^\infty h_b(u)du\right)\right] \\ &= \mathcal{L}_Z\left(\int_0^\infty h_b(u)du\right) \\ &= \int_0^\infty \exp\left(-\left(\int_0^\infty h_b(u)du\right)z\right) f_Z(z)dz \end{aligned} \tag{2.17}$$

avec Z la frailty, h_b la fonction de hazard de base, f_Z la fonction de densité de Z et $\mathcal{L}_Z$ la transformée de Laplace de Z.

Notons que l'intégrale de l'équation (2.14) ne peut pas toujours être résolue avec une forme analytique exacte. Dans certaines situations, il faut approximer numériquement cette intégrale. Comme on va se concentrer ici sur des distributions de frailty discrètes, nous n'aurons pas ce problème.

2.3 Modèle de cure

Dans la section 2.1, on a fait l'hypothèse que tous les individus auront l'événement si le temps est assez long ($\lim_{t\to\infty} S(t) = 0$). Mais dans la vie réelle, cette hypothèse n'est pas toujours respectée. Dans certaines situations, les individus peuvent être "cured", immunisés, non-susceptibles d'avoir l'événement. Ces individus qui n'expérimenteront jamais l'événement sont aussi souvent appelés des "long term survivors" ; on écrira alors que leur temps d'événement est $T = \infty$.

Le fait d'avoir une partie de la population qui n'expérimentera jamais l'événement d'intérêt a un impact sur la fonction de survie. Lorsque t tend vers l'infini, la fonction de survie ne tend plus vers 0, mais elle se stabilisera à un niveau correspondant au taux de guérison. Ce sera alors une fonction de survie impropre :

$$\lim_{t\to\infty} S(t) = \omega$$

où $\omega > 0$ représente la proportion d'individus cure. Le fonction $F(t) = 1 - S(T)$ ne sera donc plus une fonction de répartition propre.

Dans un échantillon, on peut facilement voir si il y a une fraction d'individus cured en regardant la courbe de survie estimée par l'estimateur de Kaplan-Meier (2.4) introduit précédemment. Si la courbe de survie finit par un long plateau plus haut que zéro, cela indique que l'on a probablement une partie de la population qui est cure. Sur la figure (2.1), on a modélisé la fonction de survie de Kaplan-Meier sur une base de données qui a été générée de façon à ce que 20% de la population soit censurée. On y voit clairement le long plateau à la fin.

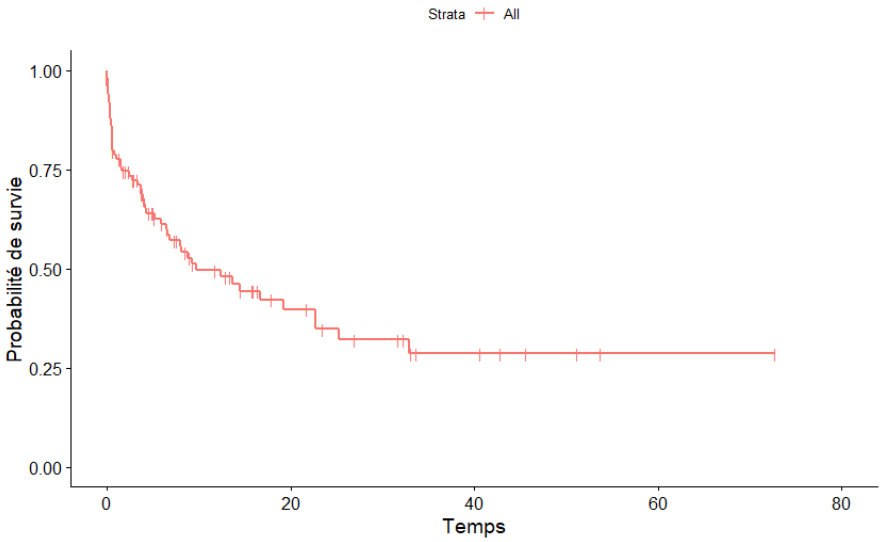


FIGURE 2.1 – Estimation de Kaplan-Meier pour une base de données avec une proportion de cure

Dans les données de survie comprenant une fraction d'individus cure, tous les individus cure sont censurés. Cependant, un individu censuré n'est pas nécessairement cure. Dans la figure (2.1), le long plateau inclut à la fois les individus cure et les individus censurés mais non cure. C'est là que réside la difficulté des données de survie en présence d'une fraction d'individus cure : il faut pouvoir distinguer les individus censurés en ceux qui sont cure et ceux qui auraient pu avoir l'événement mais qui ont simplement quitté l'étude.

En général, en présence d'une fraction d'individus cure, la proportion d'individus cure va devenir un paramètre d'intérêt. Si on étudie un nouveau traitement, on voudra savoir si la proportion de cure est plus grande avec ce nouveau traitement que avec une ancien traitement. On va donc vouloir estimer ce paramètre ω de proportion d'individus cure. On va également vouloir mesurer l'impact des covariables sur la proportion de cure et sur le temps d'événement des individus non cure.

Quand on a une fraction d'individus cure, on peut considérer que la population peut être divisée en deux parties : une partie cure et une partie non-cure. Nous avons deux groupes ayant chacun des hazards différents. Pour les individus cure, la fonction de hazard est égale à zéro. Un modèle souvent utilisé pour gérer ces différences de hazard dans les deux groupes est le "Mixture Cure Modèle" (MCM).

Le MCM se construit en deux parties :

Une appelée la partie d'incidence, qui modélise la probabilité d'être susceptible d'avoir l'événement étant donné la covariable X :

$$\psi(x) = P(B = 1|X), \tag{2.18}$$

avec $B = 0$ si l'individu est cure et $B = 1$ si l'individu est susceptible d'avoir l'événement. Cette partie d'incidence est souvent modélisée par une régression logistique, $\psi(x) = \frac{\exp(\gamma' x)}{1+\exp(\gamma' x)}$.

Et une partie appelée partie de latence, qui modélise la fonction de survie conditionnelle des individus susceptibles d'avoir l'événement :

$$S_u(t|W) = P(T > t|B = 1, W) \tag{2.19}$$

avec W un vecteur de covariable, il peut être le même que X ou non. Cette fonction de survie conditionnelle est une fonction de survie propre ($\lim_{t\to\infty} S_u(t) = 0$). On peut modéliser cette partie de latence en utilisant un modèle paramétrique, semi-paramétrique ou non-paramétrique.

La fonction de survie du MCM construite à partir des deux modèles (2.18) et (2.19) est la suivante :

$$S^*(t|X, W) = (1 - \psi(X)) + \psi(X)S_u(t|W) \tag{2.20}$$

Si on prend un modèle paramétrique pour la partie de latence (2.19), l'estimation du modèle MCM se fait par maximisation de la fonction de vraisemblance suivante :

$$\begin{aligned} L^{MCM}(\zeta) &= \prod_{i=1}^{n} f^*(y_i)^{\delta_i}(S^*(y_i))^{1-\delta_i} \\ &= \prod_{i=1}^{n} (\psi(x_i)f_u(y_i|w_i))^{\delta_i} \prod_{i=1}^{n}(1 - \psi(x_i) + \psi(x_i)S_u(y_i|w_i))^{1-\delta_i} \end{aligned} \tag{2.21}$$

avec ζ le vecteur des paramètres du modèle, f^* la fonction de densité associée à S^*, f_u la fonction de densité associée à S_u et $y_i = min(c_i, t_i)$.

Chapitre 3

Frailty avec une distribution discrète

Comme expliqué dans l'introduction, l'objectif de ce mémoire est d'analyser des données de survie dans lesquelles une partie de la population est cure en utilisant des modèles de frailty univariés avec une frailty discrète non-négative.

Avec une distribution continue pour la frailty, la probabilité associée à un point précis est nulle, donc la probabilité qu'un individu ait une certaine valeur de la frailty est nulle. Alors que avec une distribution discrète de la frailty, nous pouvons attribuer directement une probabilité pour chaque valeur de la frailty, ce qui va nous permettre d'estimer une proportion d'individus cure comme nous le verrons dans la suite ce ce chapitre.

Dans ce mémoire, différentes distributions discrètes non-négatives vont être utilisées, en commençant par des distributions très simples, pour aller vers des distributions plus compliquées.

Comme on a vu dans la section 2.2 sur les modèles de frailty, la fonction de hazard conditionnelle se note $h(t|Z) = Zh_b(t)$ (lorsqu'on ne considère pas de covaraibles), avec $h_b(t)$ la fonction de hazard de base.

La fonction de survie conditionnelle à la frailty lorsqu'on ne considère pas de covariable est donnée par :

$$S(t|Z) = \exp\left(-Z\int_0^t h_b(u)du\right) = S_b(t)^Z \tag{3.1}$$

avec $S_b(t) = \exp(-\int_0^t h_b(u)du)$ qui va être le fonction de survie associée à la fonction de hazard de base $h_b(t)$. Notons que $S_b(t)$ est une fonction de survie propre.

Comme la distribution de la frailty Z est discrète, il existe une infinité dénombrable de valeurs non-négatives entières possibles pour cette frailty ($k \in 0, 1, 2, 3, ...$). Ainsi, certains individus peuvent se retrouver avec la même valeur de frailty. Malgré cela, nous parlons quand même ici d'un modèle de frailty univarié. Chaque individu possède en effet une frailty indépendante des autres même s'ils partagent la même valeur de frailty. Cela diffère du modèle multivarié où la population est divisée en clusters, et où chaque cluster se voit attribuer une valeur de la frailty spécifique, comme expliqué dans le chapitre 2.2.

On notera la distribution de probabilité de la frailty comme suit :

$$q_k = \mathbb{P}(Z = k)$$

cela représente le taux d'individus partageant la même valeur de la frailty k.

La fonction génératrice de probabilité de la frailty est notée :

$$G_Z(s) = \sum_{k=0}^{\infty} \mathbb{P}(Z = k)s^k = \sum_{k=0}^{\infty} q_k s^k.$$

L'article de C.Caroni [4] nous donne la fonction de survie non conditionnelle quand on suppose une distribution discrète pour la frailty,

$$\begin{aligned} S(t) &= E_Z(S_b(t)^Z) \\ &= \sum_{k=0}^{\infty} S_b(t)^k q_k \\ &= G_z(S_b(t)) \end{aligned} \tag{3.2}$$

avec $G_Z(.)$, la fonction génératrice de probabilité de Z, q_k la distribution de probabilité de Z et avec $S_b(t)$ et $S_b(t)^Z$ comme dans (3.1).

Les individus ayant une frailty nulle représentent des individus ayant une fonction de hazard conditionnelle nulle ; ils correspondent donc à des individus cure.

$$h(t|Z = 0, X) = 0h_b(t) \exp(\beta' X) = 0.$$

Nous allons maintenant chercher la proportion d'individus cure. Regardons la

fonction de survie (3.2) quand le temps t tend vers l'infini :

$$\begin{aligned}\lim_{t\to\infty} S(t) &= \lim_{t\to\infty} \sum_{k=0}^{\infty} S_b(t)^k q_k \\ &= \lim_{t\to\infty} S_b(t)^0 q_0 + \sum_{k=1}^{\infty} S_b(t)^k q_k \\ &= q_0 + \sum_{k=1}^{\infty} \lim_{t\to\infty} S_b(t)^k q_k\end{aligned}$$

Comme la fonction de survie $S_b(t)$ est une fonction de survie propre (décroissante et $\lim_{t\to\infty} S_b(t) = 0$), on a que $\lim_{t\to\infty} S_b(t)^k = 0$ car k est un entier.

On trouve alors la proportion de cure qui est égale à la probabilité que la frailty soit égale à 0 :

$$\lim_{t\to\infty} S(t) = q_0 = \mathbb{P}(Z = 0). \tag{3.3}$$

On va maintenant s'intéresser à estimer les paramètres du modèle de frailty avec distribution discrète. La fonction de hazard de base $h_b(t)$ sera choisie en supposant que T suive une distribution paramétrique. On estimera les paramètres du modèle par la méthode du maximum de vraisemblance.

Nous avons rappelé la formule de la fonction de vraisemblance dans le chapitre 2.1 à l'équation (2.3). Nous allons dans cette formule remplacer les fonctions de survie $S(t)$ par la fonction de survie trouvée en (3.2), avec $h(t)$ et $f(t) = -\frac{d}{dt}S(t)$ respectivement la fonction de hazard et la fonction de densité de survie associée à $S(t)$. :

$$\begin{aligned}L(\theta) &= \prod_{i=1}^{n} h(t_i)^{\delta_i} S(t_i) \\ &= \prod_{i=1}^{n} \left[-\frac{d}{dt_i} \log(G_z(S_b(t)))\right]^{\delta_i} G_z(S_b(t_i)) \\ &= \prod_{i=1}^{n} \left[-\frac{d}{dy_i} G_z(S_b(t_i))\right]^{\delta_i} \frac{1}{G_z(S_b(t_i))^{\delta_i}} G_z(S_b(t_i)) \\ &= \prod_{i=1}^{n} \left[-\frac{d}{dy_i} G_z(S_b(t_i))\right]^{\delta_i} G_z(S_b(t_i))^{1-\delta_i}.\end{aligned} \tag{3.4}$$

Avec θ le vecteur des paramètres du modèle. Ce dernier comprend le ou les paramètre(s) de la fonction de base du hazard et le ou les paramètre(s) de la distribution

de la frailty.

De la formule précédente, on obtient facilement la fonction de log-vraisemblance

$$\begin{aligned} l(\theta) &= \log(L(\theta)) \\ &= \sum_{i=1}^{n} \delta_i \log\left(\left[-\frac{d}{dy_i} G_z(S_b(t_i))\right]\right) + (1-\delta_i)\log(G_z(S_b(t_i))) \end{aligned} \tag{3.5}$$

L'article de C.Caroni [4] utilise les distributions discrètes suivantes pour la frailty : la distribution de Poisson, la distribution géométrique et la binomiale négative. Dans la suite, on va reprendre ces 3 distributions pour construire 3 modèles différents. Pour chacun de ces modèles, nous allons utiliser comme fonction de hazard de base une Weibull, $h_b(t) = \rho\lambda t^{\rho-1}$ avec $\lambda > 0$ et $\rho > 0$ et donc $S_b(t) = \exp(-\lambda t^\rho)$.

3.1 Frailty avec une distribution de Poisson

Le premier modèle est un modèle avec comme distribution de la frailty une Poisson, $Z \sim \text{Pois}(\eta)$ avec $\eta \in (0, \infty)$. La distribution de probabilité de la frailty est donc la suivante :

$$q_k = \frac{\exp(-\eta)\eta^k}{k!}.$$

La moyenne et la variance de la distribution de Poisson sont données par $E(Z) = Var(Z) = \eta$, et la fonction génératrice de probabilité est $G_Z(s) = \exp(\eta(s-1))$.

En substituant ceci dans la fonction de survie (3.2), on obtient comme fonction de survie non conditionnelle pour ce modèle :

$$\begin{aligned} S(t) &= G_Z(S_b(t)) \\ &= \exp(\eta(S_b(t)-1)) \\ &= \exp(\eta(\exp(-\lambda t^\rho)-1)). \end{aligned} \tag{3.6}$$

La fonction de densité de survie est donnée par

$$f(t) = -\frac{d}{dt}S(t) = \eta\lambda\rho t^{\rho-1}\exp[-\lambda t^\rho + \eta(\exp(-\lambda t^\rho)-1)].$$

En réinjectant ces deux équations dans la formule de la vraisemblance (3.4), on trouve :

$$L(\eta,\lambda,\rho) = \prod_{i=1}^{n}\left[\eta\lambda\rho t^{\rho-1}\exp(-\lambda t^\rho + \eta(\exp(-\lambda t^\rho)-1))\right]^{\delta_i}\left[\exp(\eta(\exp(-\lambda t^\rho)-1))\right]^{1-\delta_i}$$

Il s'en suit que la log-likelihood de ce modèle est la suivante :

$$l(\eta, \lambda, \rho) = \sum_{i=1}^{n} \delta_i log(f(t)) + (1 - \delta_i) log(S(t)) \tag{3.7}$$

La proportion de cure de ce modèle est donnée par :

$$\begin{aligned} q_0 &= \lim_{t \to \infty} \exp(\eta(\exp(-\lambda t^\rho) - 1)) \\ &= \exp(\eta(0 - 1)) \\ &= \exp(-\eta) \end{aligned} \tag{3.8}$$

ce qui est bien égal à $P(Z = 0) = \frac{\exp(-\eta)\eta^0}{0!} = \exp(-\eta)$.

3.2 Frailty avec une distribution géométrique

Le deuxième modèle est un modèle avec comme distribution de la frailty une distribution géométrique, $Z \sim \text{Geom}\,(\pi)$ avec $\pi \in (0, 1)$. La distribution de probabilité de la frailty est donc la suivante :

$$q_k = \pi^k (1 - \pi).$$

La moyenne et la variance de la distribution géométrique sont respectivement données par $E(Z) = \frac{\pi}{1-\pi}$ et $Var(Z) = \frac{\pi}{(1-\pi)^2}$. La fonction génératrice de probabilité est $G_Z(s) = \frac{1-\pi}{1-\pi s}$.

En substituant ceci dans la fonction de survie (3.2), on obtient comme fonction de survie non conditionnelle pour ce modèle :

$$\begin{aligned} S(t) &= G_Z(S_b(t)) \\ &= \frac{1 - \pi}{1 - \pi S_b(t)} \\ &= \frac{1 - \pi}{1 - \pi \exp(-\lambda t^\rho)}. \end{aligned} \tag{3.9}$$

La fonction de densité de survie est donnée par :

$$f(t) = \frac{\lambda \rho \pi (1 - \pi) t^{\rho - 1} \exp(-\lambda t^\rho)}{[1 - \pi \exp(-\lambda t^\rho)]^2}.$$

En réinjectant ces deux équations dans la formule de la vraisemblance (3.4), on trouve :

$$L(\pi, \lambda, \rho) = \prod_{i=1}^{n} \left[\frac{\lambda \rho \pi (1 - \pi) t^{\rho - 1} \exp(-\lambda t^\rho)}{[1 - \pi \exp(-\lambda t^\rho)]^2} \right]^{\delta_i} \left[\frac{1 - \pi}{1 - \pi \exp(-\lambda t^\rho)} \right]^{1 - \delta_i}$$

Il s'en suit que la log-likelihood de ce modèle est la suivante :

$$l(\pi, \lambda, \rho) = \sum_{i=1}^{n} \delta_i \log(f(t)) + (1 - \delta_i) log(S(t)) \tag{3.10}$$

La proportion de cure de ce modèle est donnée par :

$$\begin{aligned} q_0 &= \lim_{t \to \infty} \frac{1 - \pi}{1 - \pi \exp(-\lambda t^\rho)} \\ &= \frac{1 - \pi}{1} \\ &= 1 - \pi \end{aligned} \tag{3.11}$$

ce qui est bien égal à $P(Z = 0) = \pi^0(1 - \pi) = (1 - \pi)$.

3.3 Frailty avec une distribution binomiale négative

Pour le troisième modèle, la distribution de la frailty sera une binomiale négative, $Z \sim \text{BinNeg}(\pi, ; \nu)$ avec $\pi \in (0, 1)$ et $\nu > 0$. La distribution de probabilité de la frailty est donc la suivante :

$$q_k = \binom{k + \nu - 1}{k} \pi^k (1 - \pi)^\nu$$

La moyenne et la variance de la distribution binomiale négative sont respectivement données par $E(Z) = \frac{\nu\pi}{1-\pi}$ et $Var(Z) = \frac{\nu\pi}{(1-\pi)^2}$. La fonction génératrice de probabilité est $G_Z(s) = \left[\frac{1-\pi}{1-\pi s}\right]^\nu$.

En substituant ceci dans la fonction de survie (3.2), on obtient comme fonction de survie non conditionnelle pour ce modèle :

$$\begin{aligned} S(t) &= G_Z(S_b(t)) \\ &= \left[\frac{1 - \pi}{1 - \pi S_b(t)}\right]^\nu \\ &= \left[\frac{1 - \pi}{1 - \pi \exp(-\lambda t^\rho)}\right]^\nu . \end{aligned} \tag{3.12}$$

La fonction de densité de survie est donnée par :

$$f(t) = \frac{\lambda\rho\nu\pi(1 - \pi)^\nu t^{\rho-1} \exp(-\lambda t^\rho)}{[1 - \pi \exp(-\lambda t^\rho)]^{\nu+1}}.$$

En réinjectant ces deux équations dans la formule de la vraisemblance (3.4), on trouve :

$$L(\pi,\nu,\lambda,\rho)=\prod_{i=1}^{n}\left[\frac{\pi\lambda\nu(1-\pi)^{\nu}\exp(-\lambda t)}{[1-\pi\exp(-\lambda t)]^{\nu+1}}\right]^{\delta_i}\left[\left[\frac{1-\pi}{1-\pi\exp(-\lambda t)}\right]^{\nu}\right]^{1-\delta_i}$$

Il s'en suit que la log-likelihood de ce modèle est la suivante :

$$l(\pi,\nu,\lambda,\rho)=\sum_{i=1}^{n}\delta_i\log(f(t))+(1-\delta_i)log(S(t)) \tag{3.13}$$

La proportion de cure de ce modèle est donnée par :

$$\begin{aligned} q_0 &= \lim_{t\to\infty}\left[\frac{1-\pi}{1-\pi\exp(-\lambda t^{\rho})}\right]^{\nu} \\ &= \left[\frac{1-\pi}{1}\right]^{\nu} \\ &= (1-\pi)^{\nu} \end{aligned} \tag{3.14}$$

ce qui est bien égal à $P(Z=0)=\binom{0+\nu-1}{0}\pi^0\,(1-\pi)^{\nu}=(1-\pi)^{\nu}$.

Chapitre 4

Simulations

Dans le chapitre 3, nous avons construit 3 modèles d'après l'article de C.Caroni [4], le premier avec une frailty qui suit une Poisson (3.6), le deuxième une géométrique (3.9), et le dernier une binomiale négative (3.12). Dans cet article, C.Caroni soulève le fait que en utilisant ces 3 distributions discrètes assez simples, la proportion d'individus cure (q_0) par rapport aux proportions des autres valeurs de la frailty (q_k) avec $k > 0$ est assez contraint (nous verrons cela plus en détails dans la section 4.3 de ce chapitre). Par conséquent, ces trois modèles basés sur ces distributions pourraient ne pas être adaptés pour modéliser efficacement des ensembles de données plus complexes.

Nous allons générer des ensembles d'échantillons de données à l'aide de différentes fonctions de survie, et ensuite leur appliquer nos 3 modèles afin de voir dans quelles situations ils sont les plus performants. Pour cela, nous allons faire des simulations. Nous commencerons par définir les objectifs des simulations et expliquerons de manière générale la démarche suivie. Ensuite, nous procéderons à plusieurs types de simulations inspirées de celles faites dans les articles [2], [12], [3] et [11]. Le langage de programmation utilisé pour ce faire est R.

Lors de la modélisation des données dans une étude réelle, il est impossible de déterminer avec certitude si les paramètres estimés correspondent aux véritables paramètres, étant donné que ces derniers sont inconnus. C'est pourquoi nous utilisons des simulations. Nous générons nous-mêmes un certain nombre d'échantillons de données de survie, tous d'après une même fonction de survie avec des paramètres fixés. Ensuite, nous estimons les paramètres de la fonction de survie pour chaque échantillon généré à l'aide du modèle que nous souhaitons étudier. Grâce au grand nombre d'échantillons créés et donc au grand nombre de fois que chaque paramètre est estimé, on va pouvoir vérifier si le modèle choisi reflète correctement les paramètres et les caractéristiques que nous avons définis lors de la génération

de nos échantillons.

4.1 Génération des échantillons et modélisation avec le même modèle

Dans un premier temps, nous allons générer des échantillons de données avec la même fonction de survie que celle du modèle que l'on veut étudier pour valider notre méthodologie. On va donc générer des échantillons avec les modèles de frailty avec une distribution discrète introduits dans le chapitre 3. Pour rappel, la fonction de survie conditionnelle à la frailty est donnée (3.1) par :

$$S_b(t)^Z = \exp\left(-Z\int_0^t h_b(u)du\right)$$

Pour commencer, on va générer une réalisation de la frailty Z_i pour chaque individu i. Les individus cure seront ceux avec une frailty nulle ($Z_i = 0$), on notera leur temps d'événement $t_i = \infty$. La proportion d'individus cure sera donnée par $q_0 = \mathbb{P}(Z = 0)$.

Pour générer des temps d'événement pour les individus non-cure ($Z \neq 0$), revenons à une notion plus générale :
Soit Y une variable aléatoire avec une fonction de répartition continue $F(Y)$. Alors $U = F(Y)$ suit une distribution uniforme sur $[0, 1]$. De plus, nous avons également que $1 - U$ suit une distribution uniforme.

Appliquons maintenant ce même principe à la génération de nos données. Soient T la variable aléatoire et $F(t|Z) = 1 - S_b(t)^Z$ la fonction de répartition conditionnelle à la frailty, où $S_b(t)^Z$ représente la fonction de survie conditionnelle à la frailty Z. Nous savons que $F(t|Z) \sim \text{Unif}[0, 1]$. Ainsi, $1 - F(t|Z) = S_b(t)^Z \sim \text{Unif}[0, 1]$.

Pour obtenir nos temps d'événement t_i, nous recherchons alors les racines de l'équation suivante :

$$S_b(t_i)^{z_i} - u_i = 0,$$

où u_i est une réalisation de la distribution Unif$[0, 1]$, et z_i est la réalisation de la frailty. Nous trouvons ces racines à l'aide de la fonction "uniroot" du package "stats" en R, permettant ainsi d'obtenir un temps d'événement t_i pour chaque individu non-cure.

On a maintenant un temps d'événement fini t_i pour chaque individu non cure et ∞ pour les individus cure. Il faut encore rajouter la censure. On va générer avec une distribution exponentielle un temps de censure c_i pour chaque individu. Selon le paramètre de la distribution exponentielle, on aura plus ou moins de censure dans l'échantillon. Pour chaque individu, si le temps de censure est plus petit que le temps d'événement ($c_i < t_i$), il sera censuré ($\delta = 0$) et $y_i = c_i$; si le temps d'événement est plus petit que le temps de censure ($t_i < c_i$), il sera non censuré ($\delta = 1$) et $y_i = t_i$. Comme les personnes cure ont un temps d'événement égal à $t_i = \infty > c_i$, les personnes cure seront toutes censurées et $y_i = \infty$.

On va générer 1000 échantillons de données sur base de chacun des 3 modèles (3.6), (3.9) et (3.12), et ensuite leur appliquer le même modèle pour estimer les paramètres. On va faire varier dans un premier temps le nombre d'individus dans les échantillons de données pour voir si cela a une influence sur l'estimation des paramètres, et ensuite modifier le taux d'individus censurés ajouté à la censure déjà présente avec les individus cure.

4.1.1 Frailty de Poisson

Nous commençons par générer 1000 échantillons de données de survie selon le modèle de surive avec une frailty de Poisson (3.6), rappelé ci-dessous :

$$S(t) = \exp(\eta(\exp(-\lambda t^{\rho}) - 1)). \tag{4.1}$$

Nous fixons les paramètres de la Weibull à $\lambda = 0.2$, $\rho = 0.5$, et le paramètre de la frailty de Poisson à $\eta = 1.5$, ce qui nous donne un taux de cure

$$q_0 = \frac{\exp(-\eta)\eta^0}{0!} = 0.22.$$

La censure suit une distribution exponentielle de paramètre 0.005, ajoutant ainsi en moyenne 7% de censure supplémentaire à celle déjà présente avec les individus cures.

On va d'abord générer 1000 échantillons de données avec $n = 200$ individus par échantillon, ensuite $n = 500$, puis $n = 1000$.

Dans le tableau 4.1, nous présentons la moyenne, le biais, la variance et la MSE des 1000 estimations pour chaque paramètre et pour les différents nombres d'individus. Nous remarquons que plus le nombre d'individus augmente, plus l'estimation des paramètres est précise ; le biais, la variance et donc la MSE diminuent avec

l'augmentation du nombre d'individus.

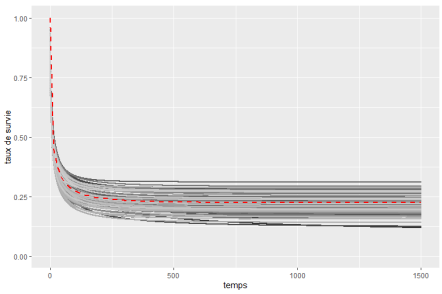
On peut également observer les boxplots pour chaque paramètre dans l'annexe A à la figure A.1, où il est clair que l'estimation est meilleure lorsque le nombre d'individus est élevé.

Sur le graphique 4.1, nous avons sélectionné de façons aléatoire 100 jeux de paramètres parmi les 1000 estimés et reconstruit les courbes de survie à l'aide de l'équation de survie rappelée au début de la section. Nous avons réintégré les paramètres estimés dans la fonction de survie pour les tracer. Bien que l'estimation soit meilleure avec plus d'individus, nous constatons que nos paramètres avec 200 individus sont tout de même bien estimés et que les courbes de survie se rapprochent de celles ayant généré les échantillons de données (courbe rouge). Ainsi, nous pouvons conclure que notre méthode de génération de données et d'estimation des paramètres pour ce modèle est bien réalisée.

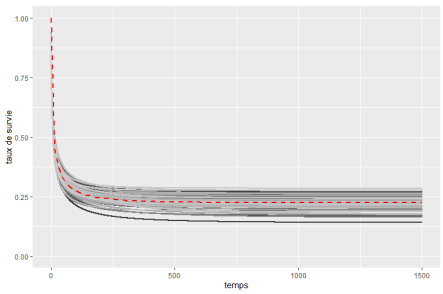
En faisant varier le paramètre de censure, on peut voir que la précision des estimations des paramètres change aussi. Au plus le paramètre de la distribution exponentielle augmente, au plus les temps de censure diminuent ; au plus la censure ajoutée augmente, au moins l'estimation est précise. Les résultats pour la Poisson sont disponibles dans la table A.1 de l'annexe A. On garde pour ces simulations ci les mêmes λ, ρ et η que juste au dessus, on fixe $n = 500$. On commence par fixer le paramètre de censure à 0.005 pour avoir une censure ajoutée de 7%, ensuite à 0.05 ce qui donne une censure ajoutée de 25%. On verra cela plus en détails pour le modèle qui a une frailty binomiale négative.

Paramètre	**Moyenne**	**Biais**	**Variance**	**MSE**
		$n = 200$		
$\lambda = 0.2$	0.1994	-0.0006	0.0011	0.0011
$\rho = 0.5$	0.5047	0.0047	0.0017	0.0017
$\eta = 1.5$	1.5248	0.0248	0.0336	0.0342
$q_0 = 0.22$	0.2212	-0.0019	0.0015	0.0015
		$n = 500$		
λ	0.1995	-0.0005	0.0004	0.0004
ρ	0.5024	0.0024	0.0006	0.0006
η	1.5157	0.0157	0.0139	0.0142
q_0	0.2212	-0.0020	0.0006	0.0006
		$n = 1000$		
λ	0.2000	-0.0001	0.0002	0.0002
ρ	0.5012	0.0012	0.0003	0.0003
η	1.5037	0.0037	0.0059	0.0059
q_0	0.2230	-0.0002	0.0003	0.0003

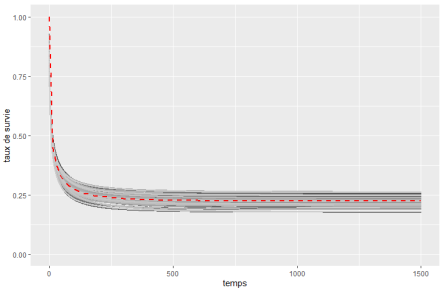
TABLE 4.1 – Résultats des simulations pour le modèle de Poisson (3.6) avec différentes tailles d'échantillons



(a) n=200



(b) n=500



(c) n=1000

FIGURE 4.1 – Courbes de survie du modèle Poisson construites à partir des paramètres estimés avec différentes tailles d'échantillons.

4.1.2 Frailty géométrique

On génère 1000 échantillons de données de survie d'après le modèle de survie avec une frailty géométrique (3.9) qui pour rappel est donné par :

$$S(t) = \frac{1-\pi}{1-\pi\exp(-\lambda t^{\rho})}. \tag{4.2}$$

en faisant également varier le nombre d'individus.

On va fixer les paramètres de la distribution Weibull : $\lambda = 0.1$, $\rho = 0.5$, et le paramètre de la frailty géométrique : $\pi = 0.75$, ce qui nous donne un taux de guérison

$$q_0 = \pi^0(1-\pi) = 0.25.$$

La censure suit une distribution exponentielle de paramètre 0.005, ajoutant ainsi en moyenne 10% de censure supplémentaire à celle déjà présente avec les individus guéris.

Dans le tableau 4.2, nous présentons la moyenne, le biais, la variance et la MSE des 1000 estimations pour chaque paramètre et pour les différents nombres d'individus. Comme pour le modèle de Poisson, on observe que plus le nombre d'individus augmente, plus les estimations sont précises.

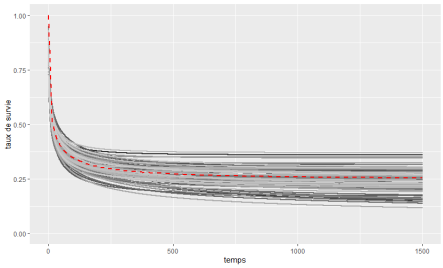
Dans l'annexe A à la figure A.2, les boxplots de chaque paramètre montrent clairement que l'estimation est meilleure lorsque le nombre d'individus est élevé.

Sur le graphique 4.2, nous retraçons 100 courbes de survie sélectionnées aléatoirement dans les 1000 estimées. On remarque encore une fois que au plus le nombre d'individus augmente, au plus les paramètres sont bien estimés.

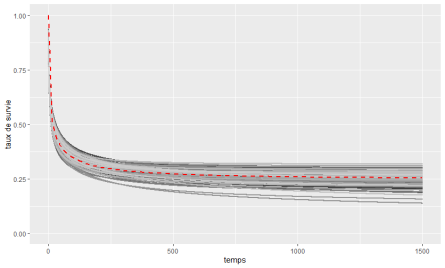
Comme pour le modèle de Poisson, au plus la censure ajoutée augmente, au moins l'estimation des paramètres est précise. On peut voir ces résultats dans la table A.2 de l'annexe A. On garde les mêmes λ, ρ et η que juste au dessus, on fixe $n = 500$. On commence par fixer le paramètre de censure à 0.005 pour avoir une censure ajoutée de 10%, ensuite à 0.05 ce qui donne une censure ajoutée de 29%.

Paramètre	**Moyenne**	**Biais**	**Variance**	**MSE**
		$n = 200$		
$\lambda = 0.1$	0.0962	-0.0038	0.0011	0.0012
$\rho = 0.5$	0.5039	0.0039	0.0021	0.0022
$\pi = 0.75$	0.7615	0.0115	0.0040	0.0041
$q_0 = 0.25$	0.2385	-0.0115	0.0040	0.0041
		$n = 500$		
λ	0.0984	-0.0016	0.0004	0.0005
ρ	0.5022	0.0022	0.0009	0.0009
π	0.7534	0.0034	0.0015	0.0015
q_0	0.2466	-0.0034	0.0015	0.0015
		$n = 1000$		
λ	0.0987	-0.0013	0.0002	0.0002
ρ	0.5015	0.0015	0.0004	0.0004
π	0.7514	0.0014	0.0007	0.0007
q_0	0.2486	-0.0014	0.0007	0.0007

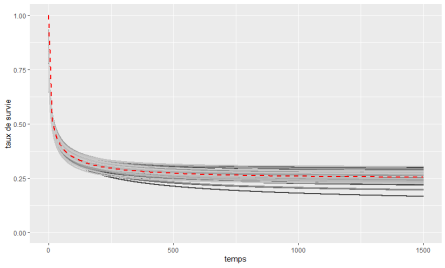
TABLE 4.2 – Résultats des simulations pour le modèle géométrique (3.9) avec différentes tailles d'échantillons



(a) n=200



(b) n=500



(c) n=1000

FIGURE 4.2 – Courbes de survie du modèle géométrique construites à partir des paramètres estimés avec différentes tailles d'échantillons.

4.1.3 Frailty binomiale négative

Pour le modèle avec la frailty binomiale négative, on va commencer par faire varier le taux de censure pour voir si cela à une influence sur l'estimation des paramètres. On génère 1000 échantillons de données de survie d'après le modèle (3.12) qui pour rappel est donné par :

$$S(t) = \left[\frac{1-\pi}{1-\pi\exp(-\lambda t^{\rho})}\right]^{\nu}.$$

On va fixer les paramètres de la distribution Weibull : $\lambda = 0.3$, $\rho = 0.6$, et les paramètres de la frailty binomiale négative : $\pi = 0.2$ et $\nu = 5$, ce qui nous donne un taux de guérison

$$q_0 = \binom{0+\nu-1}{0}\pi^0(1-\pi)^{\nu} = 0.32.$$

On va fixer le nombre d'individus par échantillon à 500 et faire varier le paramètre de la censure qui suit une distribution exponentielle. On va commencer par le fixer à 0.005, ajoutant une moyenne de 2% de censure supplémentaire, et ensuite

le fixer à 0.05 ajoutant en moyenne 13% de censure.

Dans le tableau 4.3, nous présentons la moyenne, le biais, la variance et la MSE des 1000 estimations pour chaque paramètre et pour les différents taux de censure ajoutés. Plus le taux de censure ajoutée augmente, moins les estimations des paramètres sont précises (augmentation du biais, de la variance et de la MSE).

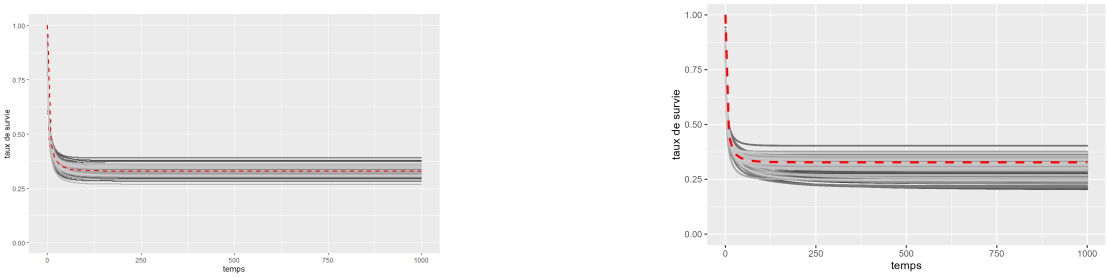
Dans l'annexe A à la figure A.3, on voit également sur les boxplots que plus la censure augmente, moins les estimations sont précises.

Sur le graphique 4.3, nous avons sélectionné 100 ensembles de paramètres parmi les 1000 estimés pour retracer les courbes de survie. On remarque encore une fois que au plus la censure augmente, au moins l'estimation des paramètres est précise.

Comme pour les deux autres modèles, on a également fait varier le nombre d'individus dans les échantillons en fixant le paramètre de censure à 0.005. Les résultats sont dans l'annexe A au tableau A.3. Ces résultats confirment encore que au plus on a d'individus, au mieux les paramètres sont estimés.

Paramètre	**Moyenne**	**Biais**	**Variance**	**MSE**
Censure ajoutée 2%				
$\lambda = 0.3$	0.2534	-0.0466	0.0055	0.0076
$\rho = 0.6$	0.6367	0.0367	0.0035	0.0076
$\pi = 0.2$	0.3869	0.1869	0.0962	0.1311
$\nu = 5$	7.5995	2.5995	67.0042	73.6945
$q_0 = 0.32$	0.3269	-0.0008	0.0005	0.0005
Censure ajoutée 13%				
λ	0.2392	-0.0608	0.0098	0.0135
ρ	0.6382	0.0382	0.0040	0.0135
π	0.4163	0.2163	0.1177	0.1644
ν	7.0810	2.0810	55.1605	59.4358
q_0	0.3173	-0.0104	0.0017	0.0018

TABLE 4.3 – Résultats des simulations pour le modèle binomial négatif (3.12) avec différentes tailles d'échantillons



(a) censure ajoutée 2%

(b) censure ajoutée 13%

FIGURE 4.3 – Courbes de survie du modèle binomial négatif construites à partir des paramètres estimés avec différents taux de censure ajoutés.

4.2 Génération des échantillons avec un MCM

Dans cette section, on va générer les échantillons de données à l'aide d'un MCM (mixture cure model introduit dans la section 2.3). Pour rappel, la fonction de survie du MCM est donnée (2.20) par :

$$S^*(t) = (1 - \psi) + \psi S_u(t)$$

avec $\psi = P(B = 1)$ la probabilité qu'un individu soit susceptible d'avoir l'événement et $S_u(t) = P(T > t|B = 1)$ la fonction de survie conditionnelle des individus susceptibles d'avoir l'événement, et avec $B = 1$ si l'individu est susceptible et $B = 0$ si il est cure.

Pour commencer, on va générer une réalisation de B_i pour chaque individu i à l'aide d'une Bernoulli de paramètre ψ. Les individus cure seront ceux avec un $B_i = 0$. On notera leur temps d'événement $t_i = \infty$. La proportion d'individus cure sera donnée par $1 - \psi$.

Pour les individus non-cure ($B = 1$), on va procéder comme dans la section précédente pour générer leur temps de survie mais en utilisant l'équation $S_u(t_i) - u_i = 0$ pour générer les temps.

Dans ces simulations, la censure suit une distribution de Weibull plutôt qu'une distribution exponentielle.

On génère 1000 échantillons de données avec 500 individus chacun. La baseline suit une Weibull de paramètres $\lambda = 0.06$ et $\rho = 1.5$. On va faire varier dans la

génération des données la proportion de cure $1 - \psi = P(B = 0)$ pour voir comment se comportent nos 3 modèles. Nous allons utiliser comme proportion de cure $1-\psi = 0.6$, puis $1-\psi = 0.4$ et pour finir $1-\psi = 0.2$. Comme nous faisons varier le taux de cure, différents paramètres sont nécessaires pour la distribution de Weibull de la censure afin de maintenir une censure supplémentaire moyenne d'environ 10%. Pour les simulations avec $1 - \psi = 0.6$, les paramètres nécessaires sont 4.5 et 9 pour obtenir une censure supplémentaire de 10% ; pour $1 - \psi = 0.4$, les paramètres nécessaires sont 5 et 11 ; et pour $1-\psi = 0.2$, les paramètres nécessaires sont 6 et 12.

Nous allons appliquer à ces échantillons de données générés le modèle MCM (2.20), le modèle de Poisson (3.6), le modèle géométrique (3.9) et le modèle binomial négatif (3.12). Nous allons nous concentrer sur la comparaison des taux de cure estimés, car les autres paramètres estimés ne sont pas comparables aux paramètres qui ont généré nos données. Nous examinerons la moyenne, le biais, la variance et la MSE pour les différents $1 - \psi$ fixés. Ensuite, comme pour les simulations précédentes, nous avons sélectionné 100 ensembles de paramètres de façon aléatoire parmi les 1000 estimés afin de tracer les courbes de survie.

4.2.1 Taux de cure fixé à $1 - \psi = 0.6$

Commençons par appliquer nos modèles aux échantillons de données générés avec une proportion de cure de $1 - \psi = 0.6$. Le tableau 4.4 indique que dans le MCM, les estimations des paramètres convergent à 100%, pour le modèle de Poisson à 99%. En revanche, pour les distributions géométrique et binomiale négative, les taux de convergence sont légèrement plus bas, à 92% et 93% respectivement. Cette différence peut être due au fait que la fonction de vraisemblance pour ces deux distributions contient une division qui, pour certains échantillons de données, peut entraîner une division par zéro et donc une vraisemblance non finie.

On peut également observer dans le tableau 4.4 que l'estimation du taux de cure a une meilleur MSE pour le modèle de Poisson que pour le MCM. Cependant, si l'on considère le biais, le MCM est nettement plus précis que le modèle de Poisson. Par ailleurs, le modèle de Poisson présente une variance plus faible que le MCM. Les modèles géométrique et binomial négatif estiment moins bien le taux de cure, ayant tous deux une MSE assez grande par rapport aux deux autres modèles.

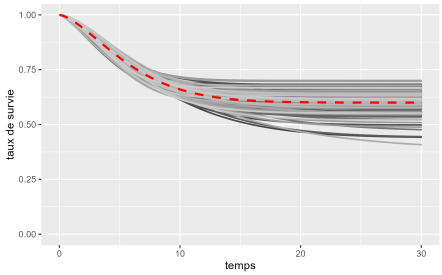
La figure 4.4 met en évidence que le modèle MCM ajuste le mieux les données. Les courbes grises sont plus proches de la courbe rouge, qui a généré les données. Pour le modèle de Poisson, on constate que les courbes sont assez proches de celles observées dans le MCM, mais quelques courbes se détachent légèrement en dessous

de la courbe rouge. Ensuite, on remarque que les modèles géométrique et binomial négatif ajustent nettement moins bien les données, avec une grande partie des courbes bien en dessous de la courbe rouge.

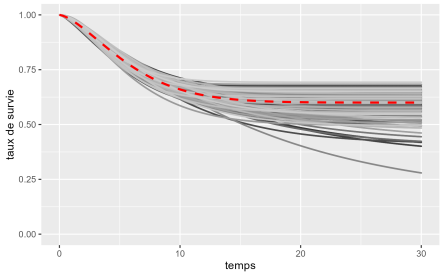
Comme nous l'avons vu dans le tableau, le biais est toujours négatif. De plus, pour les modèles géométrique et binomial négatif, un grand nombre de courbes de survie sont largement en dessous de la courbe de génération des données. Cela suggère que l'estimation de la proportion de cure n'est pas assez élevée. Pour améliorer nos modèles, il faudra donc augmenter le nombre d'individus cure et donc de zéros dans la frailty.

$1-\psi=0.6$	**Moyenne**	**Biais**	**Variance**	**MSE**	**Taux de convergence**
MCM	0.5893	-0.0107	0.0063	0.0064	100%
Poisson	0.5804	-0.0196	0.0047	0.0051	99%
Géométrique	0.5436	-0.0564	0.0132	0.0164	92%
Binomial Négatif	0.5393	-0.0607	0.0157	0.0193	93%

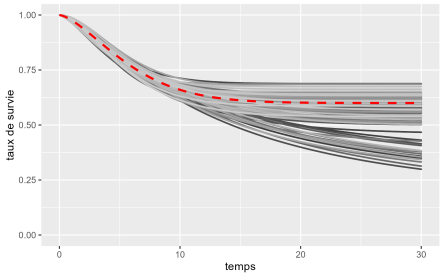
TABLE 4.4 – Résultats des estimations de la proportion de cure pour $1-\psi=0.6$ avec les différents modèles



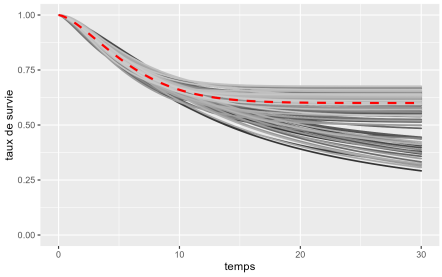
(a) MCM



(b) Modèles de Poisson



(c) Modèle géométrique



(d) Modèles binomial négatif

FIGURE 4.4 – Courbes de survie pour $1-\psi = 0.6$ construites à partir des paramètres estimés avec les différents modèles.

4.2.2 Taux de cure fixé à $1 - \psi = 0.4$

Appliquons maintenant nos modèles aux échantillons de données générés avec une proportion de cure $1-\psi = 0.4$. Le tableau 4.5 montre que dans le MCM, toutes les estimations des paramètres convergent à 100%, tandis que dans le modèle de Poisson, ce taux est de 99%. Dans les deux derniers modèles, il est respectivement de 71% et 78%, ce qui est nettement inférieur, encore une fois à cause de la division dans la fonction de vraisemblance. On devra faire attention à l'interprétation des résultats du tableau, les deux derniers taux de convergence étant assez bas.

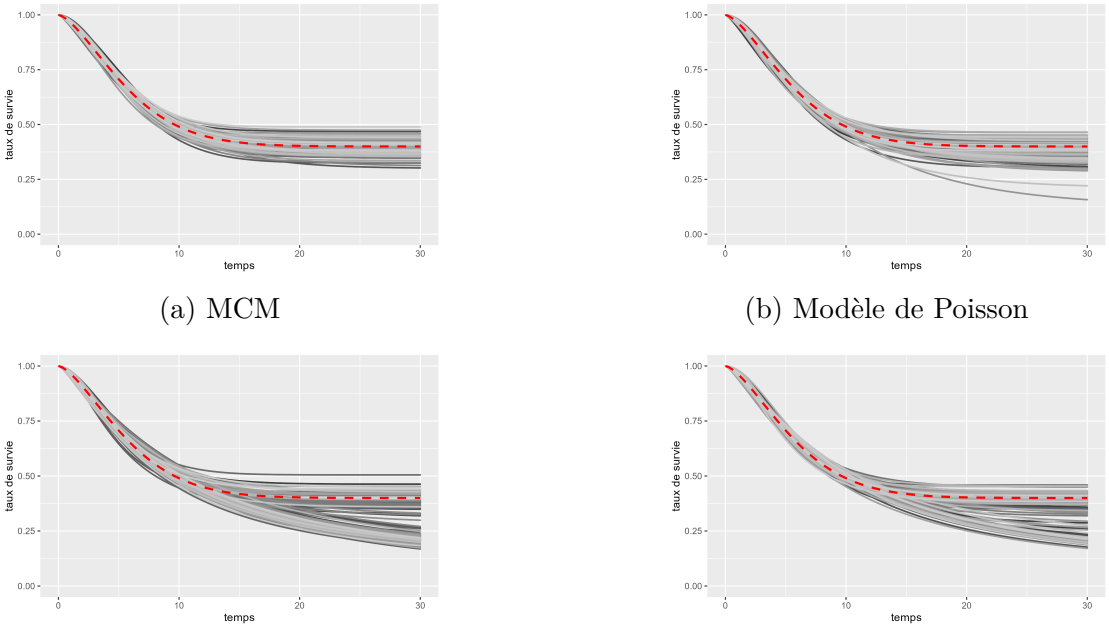
Le modèle MCM présente la meilleure précision d'estimation du taux de cure, suivi du modèle de Poisson ; tandis que les modèles géométrique et binomial négatif estiment vraiment moins bien le taux de cure. Encore une fois, tous les biais sont négatifs, le taux de personnes cure estimé n'est pas assez élevé.

Sur les graphes de la figure 4.5, on voit que le modèle MCM ajuste le mieux les données. Le modèle de Poisson se rapproche fort des résultats pour le MCM mais avec quelques courbes de survie qui sont nettement en dessous des autres. Pour les modèles géométrique et binomial négatif, on observe que de nombreuses courbes

de survie se terminent bien en dessous de la courbe rouge qui a généré les données. Il semble une fois de plus que, dans de nombreux cas, le taux d'individus cure a été sous-estimé.

$\psi = 0.4$	**Moyenne**	**Biais**	**Variance**	**MSE**	**Taux de convergence**
MCM	0.3990	−0.0010	0.0016	0.0016	100%
Poisson	0.3792	−0.0208	0.0025	0.0029	99%
Géométrique	0.2634	−0.1366	0.0184	0.0370	71%
Binomial Négatif	0.2852	−0.1148	0.0185	0.0317	78%

TABLE 4.5 – Résultats des estimations de la proportion de cure pour $1 - \psi = 0.4$ avec les différents modèles



(a) MCM

(b) Modèle de Poisson

(c) Modèles géométrique

(d) Modèles binomial négatif

FIGURE 4.5 – Courbes de survie pour $1-\psi = 0.4$ construites à partir des paramètres estimés avec les différentes modèles.

4.2.3 Taux de cure fixé à $1 - \psi = 0.2$

Pour finir, passons aux échantillons de données générés avec une proportion de cure de $1 - \psi = 0.2$. Dans le tableau 4.6, on constate que dans le modèle MCM,

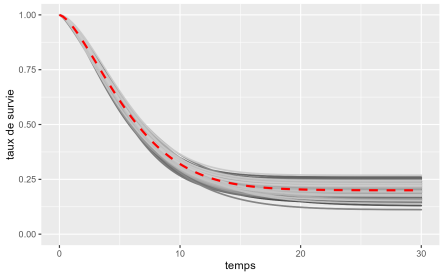
toutes les estimations des paramètres convergent à 100%. Dans le modèle de Poisson, ce taux descend à 97%. Pour le modèle géométrique, le taux de convergence est de 51%, tandis que pour le modèle binomial négatif, il est de 61%. C'est nettement moins bien que pour les taux de cure qui avaient été fixés à 0.4 et 0.6. La baisse du taux de cure semble entraîner une augmentation des vraisemblances non finies, et donc une baisse du taux de convergence des estimations.

Le modèle MCM affiche la meilleure précision, suivi du modèle de Poisson, alors que les modèles géométrique et binomial négatif estiment à nouveau nettement moins bien le taux de cure. Encore une fois, tout les biais sont négatifs.

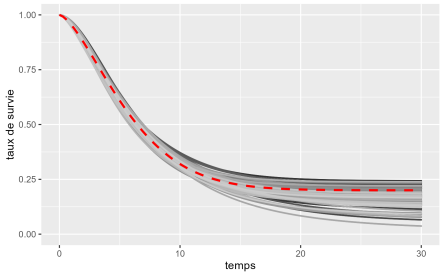
Sur les graphiques de la figure 4.6, on remarque que le modèle de Poisson ajuste les données moins bien que le MCM. Pour les deux autres modèles, bien qu'ils semblent légèrement moins performants que le MCM, cette différence n'est pas aussi évidente que les résultats obtenus dans le tableau. Toutefois, il est important de rappeler que nous avons un taux de convergence des estimations très bas, ce qui pourrait expliquer la difficulté à interpréter les graphiques.

$1-\psi=0.2$	**Moyenne**	**Biais**	**Variance**	**MSE**	**Taux de convergence**
MCM	0.2000	0.0000	0.0011	0.0011	100%
Poisson	0.1626	−0.0374	0.0019	0.0033	97%
Géométrique	0.0877	−0.1123	0.0011	0.0137	51%
Binomial Négatif	0.0890	−0.1110	0.0024	0.0148	61%

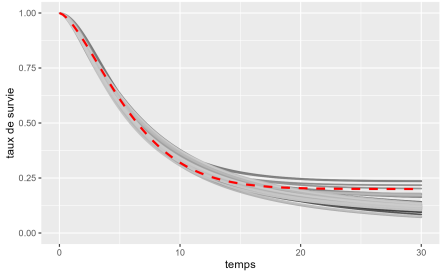
TABLE 4.6 – Résultats des estimations de la proportion de cure pour $1-\psi=0.2$ avec les différents modèles



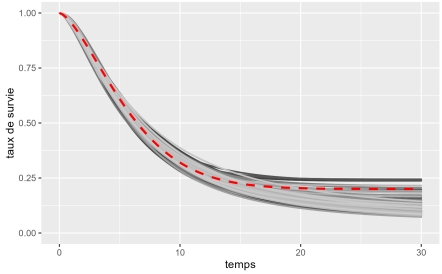
(a) MCM



(b) Modèle de Poisson



(c) Modèle géometrie



(d) Modèle binomial négatif

FIGURE 4.6 – Courbes de survie pour $1-\psi = 0.2$ construites à partir des paramètres estimés avec les différents modèles.

4.3 Conclusion des simulations

Dans les premières simulations (4.1), nous avons constaté que le nombre d'individus inclus dans notre base de données affecte la précision de nos résultats. Plus le nombre d'individus est grand, plus nos estimations sont performantes. De plus, nous avons observé que plus le taux de censure est élevé, moins le modèle est performant.

Dans les deuxièmes simulations (4.2), le modèle MCM semble mieux ajuster les données, suivi du modèle de Poisson. En revanche, les modèles géométrique et binomial négatif estimaient systématiquement beaucoup moins bien les données, bien que le modèle binomial négatif se comporte légèrement mieux. De plus, nous avons observé une convergence de l'estimation à 100% pour le MCM, ainsi qu'un taux proche de 100% pour le modèle de Poisson. En revanche, pour les modèles géométrique et binomial négatif, le taux de convergence est bien plus bas en raison de la division par zéro dans la fonction de vraisemblance. De plus, lorsque le taux de cure diminue, le nombre d'estimations convergentes diminue également. Nous avons également constaté sur les graphiques que les données semblaient être bien ajustées

au début par tous les modèles, mais que dès que l'on atteignait la proportion de cure restante, c'est là que l'estimation devenait moins précise. Comme nous l'avons observé à chaque fois, le taux de cure était beaucoup trop bas sur les graphiques, ce qui se reflète également dans le biais qui était systématiquement négatif.

Nos observations rejoignent celle de l'article de C. Caroni [4]. Il suggère que l'utilisation des distributions de Poisson, géométrique et binomiale négative comme distributions de frailty ne modélise pas correctement les données. Il remarque qu'avec ces distributions, la proportion d'individus cure (q_0) par rapport aux proportions des autres valeurs de frailty (q_k) est trop contrainte, et que pour certains types de données, cela posera des problèmes dans l'estimation des paramètres.

Dans ces 3 distributions, lorsque le nombre de zéros est élevé, cela signifie que les autres valeurs de frailty restent relativement faibles. En effet, dans les distributions de Poisson, géométrique et binomiale négative, les valeurs de frailty ont tendance à rester petites lorsque le nombre de zéros dans les données est important. Par conséquent, si l'on souhaite obtenir des valeurs de frailty plus élevées, cela entraînera une diminution du nombre de zéros dans les données. À un certain point, cela pourrait même conduire à la disparition totale des zéros dans les données, ce qui rendrait impossible la modélisation de données de survie contenant une fraction d'individus cure.

Initialement, une simple distribution exponentielle a été utilisée pour modéliser la censure dans la génération des données. Avec cette approche "simple", les trois modèles ont montré un très bon ajustement aux données, comme en témoignent les résultats de ces simulations disponibles dans les tableaux B.1, B.2 et B.3 de l'annexe B. Dans ces tableaux, nous observons que les trois modèles, ainsi que le MCM, ont une MSE de maximum 0.0007 pour chaque paramètre et chaque valeur de $1 - \psi$. De plus, pour les modèles binomial négatif et géométrique, la convergence des estimations est légèrement supérieure à celle des données générées avec une distribution de Weibull, tandis que le taux de convergence du modèle de Poisson est légèrement inférieur.

Comme les limites des performances de ces modèles sur des données plus complexes ont été soulignées dans l'article de C. Caroni, nous avons décidé d'utiliser une distribution de Weibull pour générer la censure. Dans ces simulations, les modèles binomial négatif et géométrique n'ajustaient pas bien les données, tandis que le modèle de Poisson montrait des résultats encore acceptables. Comme la moyenne est égale à la variance dans la distribution de Poisson, si on a des données avec une grande variance, on peut supposer que même le modèle de Poisson n'estimera

pas bien les données.

Ainsi, pour des ensembles de données moins complexes où la censure n'est pas aussi problématique, les trois modèles peuvent fonctionner efficacement. Toutefois, dans des cas plus complexes, il faudra trouver des moyens d'augmenter le nombre d'individus cure, qui était systématiquement sous-estimé dans les simulations avec une censure Weibull.

Dans son article, C. Caroni propose de modifier les distributions de Poisson, géométrique et binomiale négative pour que le rapport entre q_0 et q_k soit moins contraignant. Il suggère de traiter q_0 comme un paramètre distinct, en posant que la proportion de guérison est donnée par $P(Z = 0) = p$ et que pour le reste des individus (les non guéris), $P(Z = k) = (1 - p)r_k$, où $k = 1, 2, \ldots$ et r_k suit l'une des trois distributions utilisées jusqu'à présent.

Chapitre 5

Zéro modified power series

K.Molina propose dans son article [11] une solution qui rejoint celle de C. Caroni [4]. Il propose d'utiliser des distributions de la famille ZMPS (Zero modified power series) qui sont construites en modifiant les zéros des distributions de la famille PS (Power series). On espère de cette manière pouvoir introduire des valeurs de la frailty assez variées tout en gardant une certaine proportion de 0 pour pouvoir ajuster correctement le modèle aux données.

On va commencer par introduire ce que sont les distributions de la famille PS (Power series distribution) d'après les articles [9] et [10]. Ensuite, on va définir les distributions de la famille ZMPS (zero modified power series distribution) d'après les articles [5] et [11]. Pour finir, à l'aide des même articles, on va construire des modèles de survie avec différentes distributions de la famille ZMPS comme distributions de la frailty.

5.1 Distributions de la famille power series (PS)

Les articles [9] et [10] font une bonne introduction aux distributions de la famille PS. Si la variable aléatoire Z suit une distribution de la famille PS, sa distribution de probabilité peut s'écrire sous la forme suivante :

$$\xi_{PS}(z;\mu,\phi) = \mathbb{P}(Z=z) = \frac{\alpha(z,\phi)g(\mu,\phi)^z}{b(\mu,\phi)} \tag{5.1}$$

avec $\mu > 0$ le paramètre de moyenne et $\phi \geq 0$ le paramètre de dispersion. Avec $\alpha(k,\phi) > 0$, $g(\mu,\phi) > 0$ et avec

$$b(\mu, \phi) = \sum_{k=0}^{\infty} \alpha(k, \phi) g(\mu, \phi)^k$$

.

La variance des distributions de la famille PS est donnée par

$$\sigma^2_{PS} = \frac{g(\mu, \phi)}{\frac{d}{d\mu} g(\mu, \phi)}$$

et la fonction génératrice de probabilité est donnée par

$$G_{PS}(s) = \frac{b(s\mu, \phi)}{b(\mu, \phi)}.$$

.

Beaucoup de distributions couramment utilisées sont des distributions de la famille PS. Les 3 distributions utilisées dans le chapitre précédent (Poisson, géométrique et binomiale négative) sont des distributions PS.

Distribution de Poisson

Commençons par la Poisson, on veut montrer que sa distribution de probabilité $\frac{\exp(-\eta)\eta^z}{z!}$ avec $\eta \in (0, \infty)$ peut se réécrire sous la forme (5.1).

Si on pose $\alpha(z) = \frac{1}{z!}$ et $g(\mu) = \mu = \eta$ on a alors,

$$b(\mu) = \sum_{k=0}^{\infty} \frac{1}{k!} \mu^k = \exp(\mu) = \exp(\eta)$$

ce qui donnera bien la distribution de probabilité

$$\begin{aligned} \xi_{PS}(z; \mu) &= \frac{\frac{1}{z!}\mu^z}{\exp(\mu)} \\ &= \frac{\exp(-\mu)\mu^z}{z!} \\ &= \frac{\exp(-\eta)\eta^z}{z!}. \end{aligned} \tag{5.2}$$

Sa variance est donnée par

$$\sigma^2_{PS} = \frac{\mu}{\frac{d}{d\mu}\mu} = \mu = \eta$$

et sa fonction génératrice de probabilité sera

$$G_{PS}(s) = \frac{\exp(s\mu)}{\exp(\mu)} = \exp(\mu(s-1)) = \exp(\eta(s-1)).$$

Nous retombons bien sur toutes les caractéristiques d'une distribution définie dans la section 3.1. Donc, la distribution de Poisson est bien une distribution de la famille PS.

Distribution géométrique

Ensuite pour la géométrique, on veut montrer que sa distribution de probabilité $\pi^z(1-\pi)$ avec $\pi \in (0,1)$ peut se réécrire sous la forme (5.1).

Si on pose $\alpha(z) = 1$ et $g(\mu) = \frac{\mu}{1+\mu} = \pi$ on a alors,

$$b(\mu) = \sum_{k=0}^{\infty} 1(\frac{\mu}{1+\mu})^k = \frac{1}{1-\frac{\mu}{1+\mu}} = 1+\mu = 1+\frac{\pi}{1-\pi} = \frac{1}{1-\pi}$$

ce qui donnera bien la distribution de probabilité

$$\begin{aligned}\xi_{PS}(z;\mu) &= \frac{1(\frac{\mu}{1+\mu})^z}{1+\mu} \\ &= \frac{\pi^z}{\frac{1}{1-\pi}} \\ &= \pi^z(1-\pi). \end{aligned} \tag{5.3}$$

Sa variance est donnée par

$$\sigma^2_{PS} = \frac{\frac{\mu}{1+\mu}}{\frac{d}{d\mu}\frac{\mu}{1+\mu}} = \frac{\frac{\mu}{1+\mu}}{\frac{1}{(1+\mu)^2}} = \mu(1+\mu) = \frac{\pi}{1-\pi}\left(1+\frac{\pi}{1-\pi}\right) = \frac{\pi}{(1-\pi)^2}$$

et sa fonction génératrice de probabilité sera

$$G_{PS}(s) = \frac{1+s\mu}{1+\mu} = \frac{1-\pi}{1-s\pi}$$

Nous retombons bien sur toutes les caractéristiques de la distribution géométrique définie dans la section 3.2. Donc, la distribution géométrique est bien une distribution de la famille PS.

Distribution binomiale négative

Enfin, pour la binomiale négative, montrons que sa distribution de probabilité $\binom{z+\nu-1}{z}\pi^z\,(1-\pi)^\nu$ avec $\pi \in (0,1)$ et $\nu > 0$ peut se réécrire sous la forme (5.1).

Si on pose $\phi = \nu$, $\alpha(z,\phi) = \binom{z+\phi-1}{z} = \binom{z+\nu-1}{z}$, et $g(\mu,\phi) = \frac{\mu}{\mu+\phi} = \frac{\mu}{\mu+\nu} = \pi$ on a alors,

$$
\begin{aligned}
b(\mu,\phi) &= \sum_{k=0}^{\infty} \binom{k+\phi-1}{k} \left(\frac{\mu}{\mu+\phi}\right)^k \\
&= \left(1-\frac{\mu}{\mu+\phi}\right)^{-\phi} \\
&= (1-\pi)^{-\nu} \\
&= \frac{1}{(1-\pi)^{\nu}}
\end{aligned}
$$

ce qui donnera la distribution de probabilité

$$
\begin{aligned}
\xi_{PS}(z;\mu,\phi) &= \binom{z+\phi-1}{z} \frac{(\frac{\mu}{\mu+\phi})^z}{(\frac{\phi}{\phi+\mu})^{-\phi}} \\
&= \binom{z+\phi-1}{z} \frac{(\pi)^z}{(1-\pi)^{-\nu}} \\
&= \binom{z+\nu-1}{z} \pi^z (1-\pi)^{\nu}. \qquad (5.4)
\end{aligned}
$$

Sa variance est donnée par

$$
\sigma^2_{PS} = \frac{\frac{\mu}{\phi+\mu}}{\frac{d}{d\mu}\frac{\mu}{\phi+\mu}} = \frac{\frac{\mu}{\phi+\mu}}{\frac{\phi}{(\phi+\mu)^2}} = \frac{\mu(\phi+\mu)}{\phi} = \frac{\frac{\pi\nu}{(1-\pi)}}{(1-\pi)} = \frac{\pi\nu}{(1-\pi)^2}
$$

et sa fonction génératrice de probabilité sera

$$
G_{PS}(s) = \frac{(\frac{\phi}{\phi+s\mu})^{-\phi}}{(\frac{\phi}{\phi+\mu})^{-\phi}} = \left(\frac{\frac{\phi}{\phi+\mu}}{\frac{\phi}{\phi+s\mu}}\right)^{\phi} = \left(\frac{1-\pi}{1-s\pi}\right)^{\nu}
$$

Nous retombons bien sur toutes les caractéristiques de la distribution binomiale négative définie dans la section 3.3. Donc, la distribution binomiale négative est bien une distribution de la famille PS.

5.2 Distributions de la famille zero-modified power series (ZMPS)

On va maintenant construire des distributions de la famille ZMPS en modifiant la probabilité d'avoir des zéros dans les distributions de la famille PS introduites

dans la section 5.1. On va modifier le nombre de zéros de la distribution en ajoutant un paramètre p à la fonction de distribution (5.1). On verra plus tard qu'en fonction de la valeur de ce paramètre p, le nombre de zéros sera modifié de manières différentes.

Si Z suit une distribution de la famille ZMPS, sa fonction de distribution est donnée par :

$$\xi_{ZMPS}(z;\mu,\phi,p) = (1-p)I(z) + p(\xi_{PS}(z;\mu,\phi)) \tag{5.5}$$

avec $z \in \{0,1,...\}$, avec le paramètre de moyenne $\mu > 0$, avec le paramètre de dispersion $\phi \geq 0$, avec la fonction indicatrice $I(z)$, qui est telle que $I(z) = 1$ si $z = 0$ et $I(z) = 0$ si $z > 0$, et avec p le paramètre responsable de modifier la probabilité d'avoir un zéro de la distribution PS.

Le paramètre p prend ses valeurs dans l'intervalle $\left[0, \frac{1}{1-\xi_{PS}(0;\mu,\phi)}\right]$. En effet, si $p < 0$, alors on aura une probabilité négative d'obtenir une valeur plus grande que 0 et $\xi_{ZMPS}(z;\mu,\phi,p)$ ne sera plus une fonction de distribution. Si p est plus grand que $\frac{1}{1-\xi_{PS}(0;\mu,\phi)}$, cela signifie que la probabilité d'obtenir un zéro de la distribution de la famille ZMPS est négative, ce qui ne respecte pas non plus la définition de distribution de probabilité.

La moyenne de Z est $E(Z) = p\mu$, sa variance est donnée par $Var(Z) = p(\sigma^2_{PS} + \mu^2(1-p))$ et sa fonction génératrice de probabilité par

$$G_{ZMPS}(s) = 1 - p(1 - G_{PS}(s))$$

avec σ^2_{PS} et $G_{PS}(s)$ la variance et la fonction génératrice de probabilité de la distribution PS.

Comme expliqué précédemment, notre but est de modifier le nombre de zéros des distributions de la famille PS. Pour voir comment la distribution de la famille ZMPS change cette proportion de zéros, regardons la différence entre la distribution ZMPS (5.5) au point zéro et la distribution PS (5.1) au point zéro :

$$\begin{aligned}
&\xi_{ZMPS}(0;\mu,\phi,p) - \xi_{PS}(0;\mu,\phi) \\
&= [(1-p)I(0) + p(\xi_{PS}(0;\mu,\phi))] - [\xi_{PS}(0;\mu,\phi)] \\
&= [(1-p)1 + p(\xi_{PS}(0;\mu,\phi))] - [\xi_{PS}(0;\mu,\phi)] \\
&= (1-p) + (p-1)\xi_{PS}(0;\mu,\phi) \\
&= (1-p)(1 - \xi_{PS}(0;\mu,\phi)).
\end{aligned} \tag{5.6}$$

Cette différence nous montre qu'en fonction de la valeur du paramètre p, la distribution de la famille ZMPS modifiera les zéros de la distribution de la famille PS de manière différente :

Si $p = 0$, la différence (5.6) devient, $\xi_{ZMPS}(0;\mu,\phi,p) - \xi_{PS}(0;\mu,\phi) = 1 - \xi_{PS}(0;\mu,\phi)$ et donc

$$\xi_{ZMPS}(0;\mu,\phi,p) = 1$$

Toutes les valeurs de la distribution de la famille PS seront alors transformées en 0 dans la distribution de la famille ZMPS.

Si $p \in (0,1)$, la différence (5.6) sera positive et donc

$$\xi_{ZMPS}(0;\mu,\phi,p) > \xi_{PS}(0;\mu,\phi)$$

On aura alors plus de zéros dans la distribution de la famille ZMPS que dans la famille PS. On dira que cette distribution appartient à la famille zero inflated power series (ZIPS), car le nombre de zéros de la distribution de la famille PS est augmenté.

Si $p = 1$, la différence (5.6) sera égale à 0 donc

$$\xi_{ZMPS}(0;\mu,\phi,p) = \xi_{PS}(0;\mu,\phi)$$

La probabilité d'avoir un zéro est la même dans la distribution de la famille PS que dans la distribution de la famille ZMPS.

Si $p \in (1, \frac{1}{1-\xi_{PS}(0;\mu,\phi)})$, la différence (5.6) sera négative et donc

$$\xi_{ZMPS}(0;\mu,\phi,p) < \xi_{PS}(0;\mu,\phi).$$

On aura donc moins de zéros dans la distribution de la famille ZMPS que dans la famille PS. On dira que cette distribution appartient à la famille zero deflated power series (ZDPS), car le nombre de zéros de la distribution de la famille PS est diminué.

Si $p = \frac{1}{1-\xi_{PS}(0;\mu,\phi)}$, la différence (5.6) sera égale à $-\xi_{PS}(0;\mu,\phi)$ et donc

$$\xi_{ZMPS}(0;\mu,\phi,p) = 0.$$

Tous les 0 de la distribution de la famille PS seront alors tronqués dans la distribution de la famille ZMPS. On aura donc plus aucun zéro. On dira que cette distribution appartient à la famille des zero truncated power series(ZTPS), car on a plus aucun zéro dans la distribution de la famille PS.

5.3 Modèle de survie avec une frailty de la famille ZMPS

Maintenant que l'on a défini les distributions de la famille ZMPS, on va pouvoir utiliser celles-ci comme distributions de la frailty dans les modèles de survie. Repartons du modèle de survie avec une frailty discrète (3.1) pour définir la fonction de survie de ce nouveau modèle avec une frailty de la famille ZMPS :

$$\begin{aligned} S_{ZMPS}(t) &= G_{ZMPS}(S_b(t)) \\ &= 1 - p(1 - G_{PS}(S_b(t))) \\ &= 1 - p\left(1 - \frac{b(S_b(t)\mu, \phi)}{b(\mu, \phi)}\right) \end{aligned} \tag{5.7}$$

avec $S_b(t)$ la fonction de survie de base, $p \in \left[0, \frac{1}{1-\xi_{PS}(0;\mu,\phi)}\right]$ le paramètre qui modifie la probabilité d'avoir des zéros de la distribution PS. Avec G_{PS} et G_{ZMPS} respectivement les fonctions génératrices de probabilité de la distribution de la famille PS et de la distribution de la famille ZMPS. Avec $b(), \mu, \phi$ définis comme dans l'équation (5.1).

La proportion d'individus ayant une frailty nulle est donnée par la formule suivante avec ξ_{PS} et ξ_{ZMPS} les fonctions de distribution de la famille PS et de la famille ZMPS,

$$\begin{aligned} q_0 &= \mathbb{P}(Z = 0) \\ &= \xi_{ZMPS}(0; \mu, \phi, p) \\ &= (1-p)I(0) + p\xi_{PS}(0; \mu, \phi) \\ &= (1-p) + p\xi_{PS}(0; \mu, \phi). \end{aligned} \tag{5.8}$$

Cette proportion d'individus avec une frailty égale à zéro représente les individus cure, comme nous l'avons vu dans les chapitres précédents. Cependant, quand nous somme dans le cas d'un ZIPS pour rappelle quand $p \in (0, 1)$, K.Molina [11] et G.Cancho [5] considerent que cette proportion comporte deux composantes distinctes.

La première composante, $(1-p)$, représente les individus qui n'ont jamais été exposés au risque d'avoir l'événement, ce sont les individus immunisés.

La deuxième composante, notée

$$q_0^* = p\xi_{PS}(0; \mu, \phi)$$

représente les individus qui étaient initialement à risque mais qui ont été guéris grâce à un traitement ou une intervention depuis leur inclusion dans l'étude.

Le fait que la fraction d'individus guéris soit séparée en deux catégories est très intéressant dans les études où l'on s'intéresse à l'effet d'un traitement. Cela permet de différencier les personnes immunisées de base de celles qui ont été guéries grâce au traitement.

Comme nous l'avons vu précédemment, les distributions de Poisson, géométrique et binomiale négative sont des distributions PS. Nous allons modifier les zéros de ces distributions pour construire une Zero Modified Poisson (ZMP) à partir de la Poisson, une Zero Modified Géométrique (ZMG) à partir de la géométrique, et une Zero Modified Négative Binomiale (ZMNB) à partir de la binomiale négative. Ensuite, nous allons construire des modèles de survie dans lesquels ces distributions ZMPS serviront de frailty.

5.3.1 Zero modified Poisson (ZMP)

Commençons par construire la distribution Zero modified Poisson (ZMP) qui est une distribution de la famille ZMPS construite à partie de la distribution de Poisson qui est une distribution de la famille PS.

La fonction de survie de la ZMP est alors donnée par :

$$\begin{aligned} S_{ZMP}(t) &= G_{ZMP}(S_b(t)) \\ &= 1 - p(1 - G_{Poiss}(S(t)) \\ &= 1 - p(1 - \exp(\eta(S_b(t) - 1))) \end{aligned} \tag{5.9}$$

avec $\eta > 0$ le paramètre de la frailty de Poisson.

Le paramètre p sera compris entre 0 et

$$\frac{1}{1 - \xi_{Poiss}(0; \mu, \phi)} = \frac{1}{1 - \exp(-\eta)}.$$

On pose que la survie de base suit une Weibull de paramètre $\lambda, \rho > 0$, la fonction de survie est alors donnée par :

$$S_{ZMP}(t) = 1 - p + p \exp(\eta(\exp(-\lambda t^{\rho}) - 1)).$$

La fonction de densité de survie est :

$$f(t) = -\frac{d}{dt} S_{ZMP}(t) = p\eta\lambda\rho t^{\rho-1} \exp[-\lambda t^{\rho} + \eta(\exp(-\lambda t^{\rho}) - 1)].$$

On peut alors construire la fonction de vraisemblance :

$$L_{ZMP}(\eta, \lambda, \rho, p) =$$
$$\prod_{i=1}^{n} \left[p\eta\lambda\rho t^{\rho-1} \exp[-\lambda t^{\rho} + \eta(\exp(-\lambda t^{\rho}) - 1)]\right]^{\delta_i} [1 - p + p \exp(\eta(\exp(-\lambda t^{\rho}) - 1))]^{1-\delta_i}$$

La proportion d'individus cure est donnée par :

$$q_0 = (1 - p) + p \exp(-\eta).$$

5.3.2 Zero modified géométrique (ZMG)

Passons ensuite à la distribution Zero modified géométrique (ZMG) qui est une distribution de la famille ZMPS construite à partie de la distribution géométrique qui est une distribution de la famille PS.

La fonction de survie de la ZMG est donnée par :

$$\begin{aligned} S_{ZMG}(t) &= G_{ZMG}(S_b(t)) \\ &= 1 - p(1 - G_{geom}(S(t)) \\ &= 1 - p\left(1 - \frac{1 - \pi}{1 - \pi S(t)}\right) \end{aligned} \tag{5.10}$$

avec $\pi \in (0, 1)$ le paramètre de la frailty géométrique.

Le paramètre p sera compris entre 0 et

$$\frac{1}{1 - \xi_{geom}(0; \mu, \phi)} = \frac{1}{1 - (1 - \pi)} = \frac{1}{\pi}.$$

On pose que la fonction de survie de base va suivre une Weibull de paramètre $\lambda, \rho > 0$, la fonction de survie est alors donnée par :

$$S_{ZMG}(t) = 1 - p + p\frac{1 - \pi}{1 - \pi \exp(-\lambda t^{\rho})}$$

La fonction de densité de survie est :

$$f(t) = -\frac{d}{dt}S_{ZMG}(t) = \frac{p\lambda\rho\pi(1-\pi)t^{\rho-1}\exp(-\lambda t^{\rho})}{(1-\pi\exp(-\lambda t^{\rho}))^2}.$$

On peut alors construire la fonction de vraisemblance :

$$L_{ZMG}(\eta,\lambda,\rho,p) = \prod_{i=1}^{n}\left[\frac{p\lambda\rho\pi(1-\pi)t^{\rho-1}\exp(-\lambda t^{\rho})}{(1-\pi\exp(-\lambda t^{\rho}))^2}\right]^{\delta_i}\left[1-p+p\frac{1-\pi}{1-\pi\exp(-\lambda t^{\rho})}\right]^{1-\delta_i}$$

La proportion d'individus cure est donnée par :

$$q_0 = (1-p) + p(1-\pi).$$

5.3.3 Zero modified binomial négative (ZMBN)

Pour finir, construisons la distribution Zero modified binomiale négative (ZMBN) qui est une distribution de la famille ZMPS construite à partie de la distribution binomiale négative qui est une distribution de la famille PS.

La fonction de survie de la ZMBN est alors donnée par :

$$\begin{aligned} S_{ZMBN}(t) &= G_{ZMBN}(S_b(t)) \\ &= 1-p(1-G_{binneg}(S(t)) \\ &= 1-p\left(1-\left(\frac{1-\pi}{1-\pi S(t)}\right)^{\nu}\right) \end{aligned} \tag{5.11}$$

avec $\pi \in (0,1)$ et $\nu > 0$ les paramètres de la frailty binomiale négative.

Le paramètre p sera compris entre 0 et

$$\frac{1}{1-\xi_{binneg}(0;\pi,\nu)} = \frac{1}{1-(1-\pi)^{\nu}}.$$

On pose que la fonction de survie de base va suivre une Weibull de paramètre $\lambda, \rho > 0$, la fonction de survie est alors donnée par :

$$S_{ZMBN} = 1-p+p\left(\frac{1-\pi}{1-\pi S(t)}\right)^{\nu}$$

La fonction de densité de survie est :

$$f(t) = -\frac{d}{dt} S_{ZMBN}(t) = \frac{p\lambda\rho\nu\pi(1-\pi)^{\nu} t^{\rho-1} \exp(-\lambda t^{\rho})}{(1-\pi\exp(-\lambda t^{\rho}))^{\nu+1}}.$$

On peut alors construire la fonction de vraisemblance :

$$L_{ZMG}(\eta, \lambda, \rho, p) =$$
$$\prod_{i=1}^{n} \left[\frac{p\lambda\rho\nu\pi(1-\pi)^{\nu} t^{\rho-1} \exp(-\lambda t^{\rho})}{(1-\pi\exp(-\lambda t^{\rho}))^{\nu+1}} \right]^{\delta_i} \left[1 - p + p\left(\frac{1-\pi}{1-\pi S(t)} \right)^{\nu} \right]^{1-\delta_i}$$

La proportion d'individus cure est donnée par :

$$q_0 = (1-p) + p(1-\pi)^{\nu}.$$

Chapitre 6

Simulation ZMPS

Dans ce chapitre, on va générer différents échantillons de données et leur appliquer les trois modèles construits dans la chapitre précédant : le modèle ZMPS (5.9), le modèle ZMG (5.10) et le modèle ZMBN (5.11).

Nous avons généré les données de la même manière que dans la section 4.2. Des échantillons de données de 500 individus ont été générés, en fixant les paramètres λ et ρ de la fonction de hazard de base qui suit une distribution de Weibull à 0.06 et 1.5. Ensuite, nous avons varié le taux d'individus cure en commençant par $1 - \psi = 0.2$, puis $1 - \psi = 0.4$, et enfin $1 - \psi = 0.6$. Les données ont été générées avec une censure suivant une distribution de Weibull, les paramètres de cette distribution ayant été choisis de manière à ce que la moyenne de la censure ajoutée soit de 10%.

Dans le chapitre 5, nous avons défini que le paramètre p devait être compris dans l'intervalle :

$$\left[0, \frac{1}{1 - \xi_{PS}(0; \mu, \phi)}\right]$$

où $\xi_{PS}(0; \mu, \phi)$ est la proportion d'individus cure dans la distribution PS, homologue à la distribution ZMPS choisie. Ainsi, dans ces simulations, nous allons borner le paramètre p pour chaque échantillon entre 0 et $\frac{1}{1-\xi_{PS}(0;\hat{\mu},\hat{\phi})}$, qui est le taux de cure estimé par le modèle PS homologue sur le même échantillon de données. Cependant, comme pour certains échantillons de données l'estimation ne converge pas en raison d'une vraisemblance infinie, nous allons borner le paramètre p pour ces échantillons entre 0 et 1. Nous avons choisi cet intervalle car, comme observé dans les simulations précédentes, le taux de cure était systématiquement sous-estimé. Cela nous permet de rester directement dans un cadre ZIPS.

Dans l'article [11], l'auteur soulève le fait que lorsque l'indice p est proche de 1, le modèle a du mal à déterminer s'il doit passer du modèle PS à un modèle ZIPS (qui augmente le nombre de zéros) ou à un modèle ZDPS (qui diminue le nombre de zéros). Comme nous l'avons vu dans la section 4, le nombre de zéros était systématiquement trop faible, nécessitant une augmentation pour mieux estimer le taux de cure. Nous allons donc également réaliser des simulations avec le paramètre p contraint à rester entre 0 et 1 pour les 1000 échantillons de données, comme le fait G. Cancho dans son article [5]. Cela nous donnera un modèle ZIPS qui augmente la proportion de cure estimée par le modèle PS de base. De plus, grâce au modèle ZIPS, nous pourrons identifier si, parmi la proportion d'individus cure, il y a une proportion d'individus immunisés selon l'interprétation de K. Molina [11] et G. Cancho [5].

Nous noterons les modèles avec $p \in \left[0, \frac{1}{1-\xi_{PS}(0;\hat{\mu},\hat{\phi})}\right]$ par ZMP, ZMG et ZMBN, et les modèles avec $p \in [0,1]$ par ZMP*, ZMG* et ZMBN* (ZMPS*) afin d'éviter toute confusion.

On présentera pour chaque modèle la moyenne des estimations du taux de cure, la variance, la MSE ainsi que le taux de convergence des estimations. Pour les modèles ZMPS, nous afficherons également la moyenne du paramètre p et sa variance. Pour les modèles ZMPS*, nous présenterons le taux d'individus immunisés et le taux d'individus guéris selon l'interprétation de K. Molina [11] et G. Cancho [5].

6.1 Taux de cure fixé à $1 - \psi = 0.2$

Dans cette section, on va appliquer nos différents modèles aux données générées avec un taux de cure fixé à $1 - \psi = 0.2$.

Dans la table 6.1, on constate que pour le modèle de Poisson, ainsi que pour les modèles ZMP et ZMP*, l'estimation de la proportion de cure reste pratiquement identique. Les moyennes et les variances sont très proches. Le taux de convergence est également le même pour les trois modèles, et il est très élevé.

En revanche, on observe une meilleure estimation de la proportion de cure dans les modèles ZMG* et ZMBN* par rapport aux modèles ZMG et ZMBN, qui sont eux-mêmes meilleurs que les modèles géométrique et binomial négatif. Le biais est plus faible et la variance est également réduite. De plus, le taux de convergence a augmenté dans les modèles ZMG et ZMBN, et encore plus dans les modèles ZMG* et ZMBN*.

Dans la table 6.2, on remarque que les moyennes des paramètres p, qui modifie la proportion de zéros de la distribution de Poisson en ZMP et ZMP*, sont très proches de 1 (1.0001 et 0.9993), ce qui explique la similarité des résultats d'estimation de proportion de cure entre les deux modèles. La proportion de zéro estimée par le modèle de Poisson est déjà assez élevée, les modèles ZMP et ZMP* ne rajoutent pas de zéro supplémentaire.

Pour le modèle ZMG*, la moyenne du paramètre p vaut 0.9864, tandis que pour le modèle ZMG, elle est de 1.0120. Cependant, ces deux modèles augmentent le nombre de zéros estimé par le modèle géométrique, alors que le modèle ZMG devrait le réduire puisque son p est supérieur à 1. Mais rappelons que 1.0120 est une moyenne et que sa variance est de 0.0028. On a donc que 68% des paramètres p estimés se trouvent dans l'intervalle $[0.95, 1.06]$. Donc, pour une partie des échantillons, le nombre de zéros estimé de la distribution PS est augmenté et pour l'autre diminué. Pour les modèles ZMBN et ZMBN*, les deux moyennes des paramètres p sont inférieures à 1.

On observe également dans la table 6.2, selon l'interprétation de K.Molina [11] et G.Cancho [5], que la proportion moyenne d'individus immunisés dans le modèle ZMP* est de $1 - p = 0.0007$, ce qui est assez faible en raison de la valeur élevée du paramètre p. C'est une information que nous n'avions pas avec le modèle de Poisson ou avec les modèles MCM qui mélangeaient tous les individus cure dans une même proportion sans distinguer les individus immunisés des individus guéris de la maladie. On voit pour le modèle ZMG* que la proportion moyenne d'individus immunisés est de 0.0136 et la proportion moyenne d'individus guéris de la maladie est de 0.1438, donnant ainsi un taux de cure total moyen de 0.1574. Nous avons donc une proportion notable d'individus immunisés parmi les individus guéris, ce qui n'aurait pas été perceptible avec le modèle géométrique ou le MCM qui nous donnent seulement le taux de guérison totale. Pour le modèle ZMBN*, le taux moyen d'individus immunisés est de 0.0193, celui des guéris de 0.1111, et le taux moyen de guérison totale de 0.1304.

Dans la figure 6.1, on peut observer que l'allure générale des courbes de survie estimées s'ajustent mieux aux données dans le modèle ZMP* que dans le modèle de Poisson, même si cette amélioration n'était pas visible dans l'estimation du taux de cure. Pour les modèles ZMG* et ZMBN*, on observe sur les graphes une amélioration de l'estimation de la proportion de cure par rapport à leurs modèles de la famille PS respectifs. Pour les six modèles, on constate sur les graphiques qu'ils s'ajustent assez bien aux données jusqu'à arriver à la proportion de cure, où

l'on observe une différence entre les modèles avec une frailty de la famille PS et ceux avec une frailty de la famille ZMPS.

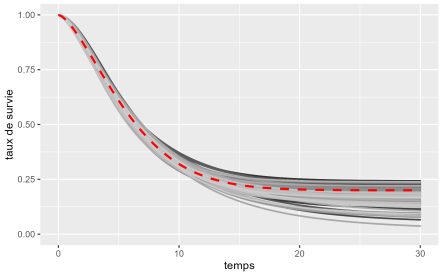
Les comparaisons des graphes pour les distributions ZMP, ZMG et ZMBN par rapport aux distributions ZMP*, ZMG* et ZMBN* sont disponibles dans la figure C.1 de l'annexe C. Ces graphes ont sensiblement la même allure, mais on observe néanmoins que la variance du taux de cure des modèles ZMG et ZMBN est plus grande que celle des modèles ZMG* et ZMBN*.

$1-\psi=0.2$	**Moyenne**	**Biais**	**Variance**	**MSE**	**Taux de convergence**
Poisson	0.1626	−0.0374	0.0019	0.0033	97%
ZMP	0.1625	−0.0375	0.0019	0.0033	97%
ZMP*	0.1625	−0.0375	0.0019	0.0033	97%
Géométrique	0.0877	−0.1123	0.0011	0.0137	51%
ZMG	0.1469	−0.0531	0.0016	0.0044	61%
ZMG*	0.1574	−0.0426	0.0007	0.0025	66%
Binom Négatif	0.0890	−0.1110	0.0024	0.0148	61%
ZMBN	0.1268	−0.0732	0.0015	0.0069	68%
ZMBN*	0.1304	−0.0696	0.0009	0.0057	74%

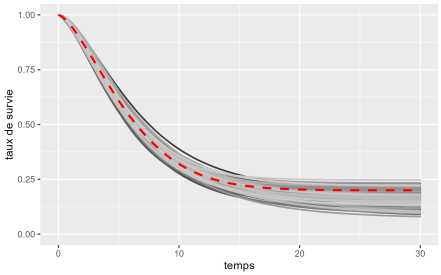
TABLE 6.1 – Résultats des estimations de la proportion de cure pour $1-\psi=0.2$ avec les 7 différents modèles

$1-\psi=0.2$	**Cure**	**Moyenne p**	**Variance p**	**Imunne (1-p)**	**Guerri**
ZMP	0.1625	1.0001	0.0051	-	-
ZMP*	0.1625	0.9993	0.0001	0.0007	0.1617
ZMG	0.1469	1.0120	0.0028	-	-
ZMG*	0.1574	0.9864	0.0004	0.0136	0.1438
ZMBN	0.1268	0.9975	0.0027	-	-
ZMBN*	0.1304	0.9807	0.0007	0.0193	0.1111

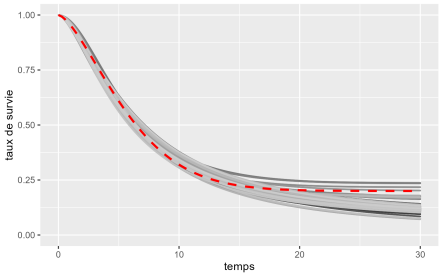
TABLE 6.2 – Résultats des estimations de la proportion de cure pour $1-\psi=0.2$ avec les 6 différents modèles



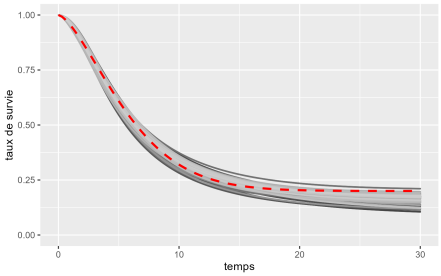
(a) Poisson



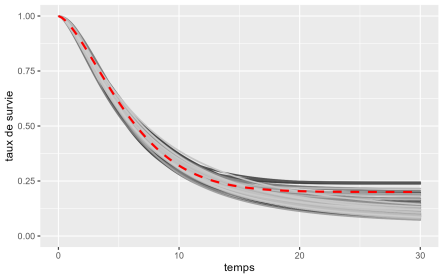
(b) ZMP*



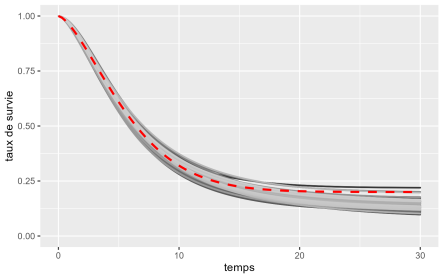
(c) Géométrique



(d) ZMG*



(e) Binomial négatif



(f) ZMBN*

FIGURE 6.1 – Courbes de survie des modèles de Poisson, géométrique, binomial négatif, ZMP*, ZMG* et ZMBN* avec la proportion de cure fixée à $1 - \psi = 0.2$

6.2 Taux de cure fixé à $1 - \psi = 0.4$

Dans cette section, nous appliquons nos différents modèles aux données générées avec un taux de guérison fixé à $1 - \psi = 0.4$.

Comme indiqué dans le tableau 6.3, pour le modèle de Poisson et le ZMP*, les résultats pour l'estimation du taux de cure sont très similaires. Le modèle ZMP présente une estimation légèrement plus élevée du taux de cure. Comme dans la section précédente, presque toutes les estimations de ces modèles convergent.

Les modèles ZMG* et ZMBN* estiment nettement mieux la proportion de cure que les modèles géométrique et binomial négatif. De plus, les modèles ZMG et ZMBN améliorent encore le biais de l'estimation du taux de cure par rapport aux ZMG* et ZMBN*, mais leur variance est légèrement plus grande. Cela a pour conséquence que les modèles ZMG et ZMG* ont la même MSE, de même que les modèles ZMBN et ZMBN*.

Le taux d'estimation qui converge augmente en passant du modèle géométrique au modèle ZMG, puis augmente encore en passant au ZMG*, de même qu'il augmente en passant du binomial négatif au ZMBN, et encore plus avec les modèles ZMBN*.

La table 6.4 nous montre que les moyennes des paramètres p des trois modèles ZMPS dépassent 1. On voit également que les moyennes des paramètres p des 3 modèles ZMPS* sont plus petits que 1. Les taux moyens d'individus immunisés selon l'interprétation de K.Molina [11] et G.Cancho [5] des modèles ZMP* et ZMG* sont assez proches (0.0351 et 0.0386), mais pour le modèle ZMBN*, ce taux moyen d'individus immunisés est de 0.0569, ce qui est plus élevé.

La figure 6.2 montre clairement que les modèles de la famille ZMPS* ajustent mieux les données que les modèles de la famille PS comme on l'a vu dans la section précédente. La différence est plus forte entre le modèle de Poisson et le modèle ZMP* qu'entre les modèles géométrique et ZMG* et binomial négatif et ZMBN*, ce qui reflète bien les paramètres estimés dans les tableaux.

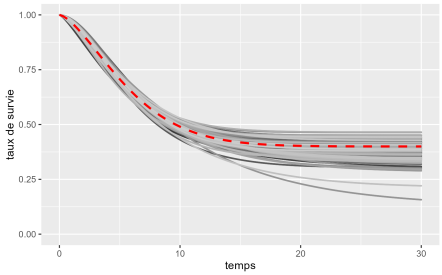
Les comparaisons des graphes pour les distributions ZMPS par rapport aux distributions ZMPS* sont disponibles dans la figure C.2 de l'annexe C.

$\psi = 0.4$	**Moyenne**	**Biais**	**Variance**	**MSE**	**Taux de convergence**
Poisson	0.3792	−0.0208	0.0025	0.0029	99%
ZMP	0.3859	−0.0141	0.0021	0.0023	99%
ZMP*	0.3779	−0.0221	0.0025	0.0030	99%
Géométrique	0.2634	−0.1366	0.0184	0.0370	71%
ZMG	0.3459	−0.0541	0.0048	0.0077	83%
ZMG*	0.3388	−0.0612	0.0040	0.0077	86%
Binom Négatif	0.2852	−0.1148	0.0185	0.0317	78%
ZMBN	0.3513	−0.0477	0.0050	0.0072	86%
ZMBN*	0.3479	−0.0520	0.0045	0.0072	91%

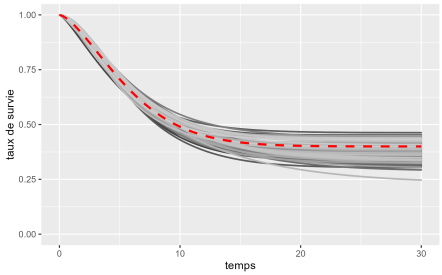
TABLE 6.3 – Résultats des estimations de la proportion de cure pour $1 - \psi = 0.4$ avec les différents modèles

$1 - \psi = 0.4$	**Cure**	**Moyenne p**	**Variance p**	**Imunne (1-p)**	**Guerri**
ZMP	0.3859	1.3813	0.1021	-	-
ZMP*	0.3779	0.9649	0.0068	0.0351	0.3428
ZMG	0.3459	1.1140	0.0622	-	-
ZMG*	0.3388	0.9614	0.0033	0.0386	0.3002
ZMBN	0.3513	1.0777	0.0767	-	-
ZMBN*	0.3479	0.9431	0.0035	0.0569	0.2910

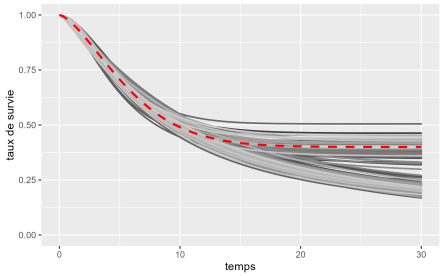
TABLE 6.4 – Résultats des estimations de la proportion de cure pour $1 - \psi = 0.4$ avec les différents modèles



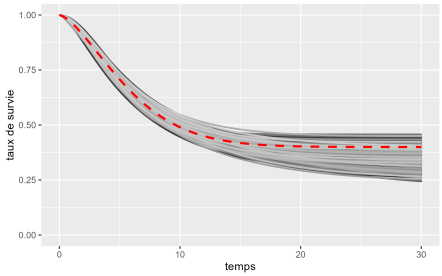
(a) Poisson



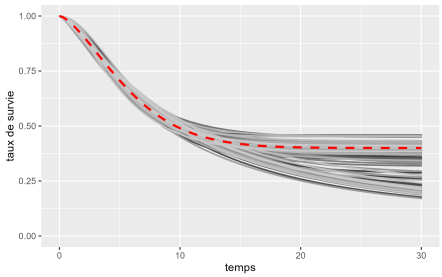
(b) ZMP*



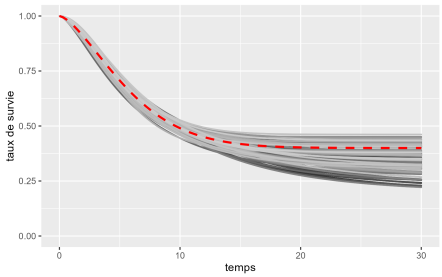
(c) Géométrique



(d) ZMG*



(e) Binomial négatif



(f) ZMBN*

FIGURE 6.2 – Courbes de survie des modèles de Poisson, géométrique, binomial négatif, ZMP*, ZMG* et ZMBN* avec la proportion de cure fixée à $1 - \psi = 0.4$

6.3 Taux de cure fixé à $1 - \psi = 0.6$

Dans cette section, nous appliquons nos différents modèles aux données générées avec un taux de cure fixé à $1 - \psi = 0.6$.

La table 6.5 montre encore une fois que pour les modèle ZMP, ZMP* et le modèle de Poisson, les résultats pour l'estimation du taux de cure sont très similaires.

Contrairement aux deux autres sections (où le taux de cure était plus faible), on

constate qu'ici le modèle géométrique estime mieux le taux de cure que les modèles ZMG et ZMG*, et le modèle binomial négatif estime mieux le taux de cure que les modèles ZMBN et ZMBN*.

La table 6.5 nous montre que les moyennes des estimations des paramètres p pour les 3 modèles ZMPS sont plus grandes que 1. Pour les 3 modèles ZMPS*, les moyennes des estimations des paramètres p sont plus petites que pour les taux de cure de 0.2 et 0.4, ce qui donne une grande proportion moyenne d'individus immunisés selon l'interprétation de K.Molina [11] et G.Cancho [5] pour les 3 modèles ZMPS*.

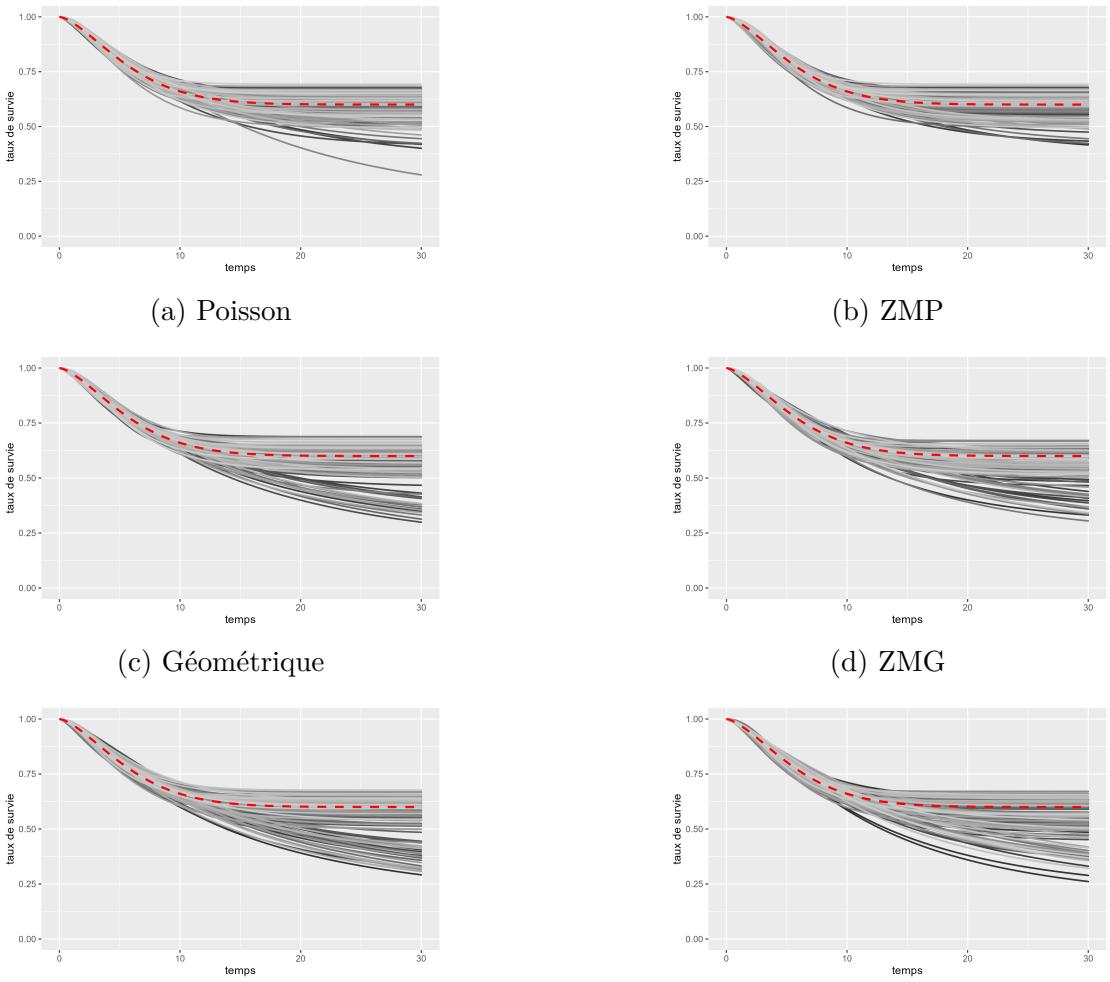
La figure 6.3 montre qu'il n'y a pas vraiment de différence entre l'ajustement des données des modèles avec des frailty de la famille ZMPS* et des modèles avec des frailty de la famille PS. On peut observer la même chose sur le graphe C.3 de l'annexe C qui compare les modèles ZMPS au modèles ZMPS* .

$1-\psi=0.6$	**Moyenne**	**Biais**	**Variance**	**MSE**	**Taux de convergence**
Poisson	0.5804	−0.0196	0.0047	0.0051	99%
ZMP	0.5793	−0.0207	0.0047	0.0051	99%
ZMP*	0.5771	−0.0229	0.0046	0.0051	99%
Géométrique	0.5436	−0.0564	0.0132	0.0164	92%
ZMG	0.5400	−0.0599	0.0140	0.0175	90%
ZMG*	0.5321	−0.0679	0.0133	0.0179	91%
Binom Négatif	0.5393	−0.0607	0.0157	0.0193	93%
ZMBN	0.5375	−0.0625	0.0160	0.0199	90%
ZMBN*	0.5239	−0.0761	0.0177	0.0235	91%

TABLE 6.5 – Résultats des estimations de la proportion de cure pour $1-\psi=0.6$ avec les différents modèles

$1-\psi=0.6$	**Cure**	**Moyenne p**	**Variance p**	**Imunne (1-p)**	**Guerri**
ZMP	0.5793	1.2594	0.5062	-	-
ZMP*	0.5771	0.8427	0.0351	0.1573	0.4198
ZMG	0.5400	1.3023	0.5797	-	-
ZMG*	0.5321	0.8952	0.0203	0.1048	0.4273
ZMBN	0.5375	1.2978	0.5948	-	-
ZMBN*	0.5239	0.8848	0.0208	0.1152	0.4087

TABLE 6.6 – Résultats des estimations de la proportion de cure pour $1-\psi=0.6$ avec les différents modèles



(a) Poisson

(b) ZMP

(c) Géométrique

(d) ZMG

(e) Binomial négatif

(f) ZMBN

FIGURE 6.3 – Courbes de survie des modèles de Poisson, géométrique, binomial négatif, ZMP*, ZMG* et ZMBN* avec la proportion de cure fixée à $1 - \psi = 0.6$

6.4 Conclusion des simulations ZMPS

Pour les taux de cure fixés à 0.2 et 0.4, les simulations ont montré que les modèles ZMPS* estiment mieux le taux de cure et ajustent mieux les données que les modèles PS, en augmentant leur proportion de cure estimé. Les modèles ZMPS ajustent également mieux les données que les modèles PS, ils sont souvent à égalité avec les modèles ZMPS* (même MSE). Les modèles avec une distribution de la frailty PS ont donné les moins bons résultats.

En revanche, pour un taux de cure de 0.6, nous avons observé de meilleurs résultats avec les frailty ayant une distribution de la famille PS, comparé aux modèles ZMPS et ZMPS*. Cela s'explique probablement par la proportion élevée d'individus cure, ce qui rend difficile pour le modèle ZMPS d'ajouter encore de tels individus.

Selon K.Molina [11] et G.Cancho [5], un avantage des modèles ZMPS* par rapport aux modèles MCM et PS est qu'ils permettent de séparer la proportion de cure en deux parties, celle avec les individus immunisés (qui n'ont jamais été à risque) et celle comprenant ceux qui ont été guéris (qui étaient à risque au départ mais ne le sont plus). Il serait intéressant de mener des simulations où l'on impose une proportion d'individus immunisés pour voir si les modèles arrivent à bien l'estimer. Il serait aussi intéressant d'ajouter une covariable traitement qui influencerait la partie des individus qui ont été guéris mais pas la partie immunisée.

Nous avons principalement exploré le cas des modèles où nous devions augmenter le nombre de zéros de la distribution PS. Il serait également intéressant d'examiner le cas où nous avons trop de zéros et où l'on doit donc diminuer le nombre de zéros en utilisant un modèle ZDPS.

Chapitre 7

Compound Poisson

Dans les chapitres précédents, nous avons utilisé des modèles de frailty avec différentes distributions discrètes pour modéliser des données de survie présentant une fraction de cure. Ici, nous allons utiliser comme distribution de la frailty une compound Poisson (CP). C'est une distribution continue à laquelle on ajoute une partie discrète en 0. Cela va permettre de donner un poids à la valeur de la frailty $z = 0$ pour prendre en compte les individus cure.

Nous allons commencer par introduire ce qu'est la distribution compound Poisson d'après les articles [13] et [1], puis construire le modèle de survie avec la compound Poisson comme distribution de la frailty d'après les articles [15], [14], [13]et [17].

7.1 Distribution compound Poisson

La distribution compound Poisson est la distribution d'une variable Z définie par :

$$Z = \begin{cases} X_1 + X_2 + \ldots + X_N & \text{si } N > 0 \\ 0 & \text{si } N = 0 \end{cases} \tag{7.1}$$

avec N qui suit une distribution de Poisson (discrète) de paramètre $\Phi > 0$, sa distribution de probabilité est pour rappel donnée par :

$$P(N = k) = \frac{\Phi^k \exp(-\Phi)}{k!}.$$

($N \sim \text{Poiss}(\Phi)$).

avec $X_1, X_2, \ldots$ indépendants et identiquement distribués selon une certaine distribution continue. Ici, on va se restreindre à la distribution Gamma de paramètre de dispersion $\gamma > 0$ et de forme $\theta > 0$, la densité de probabilité de la Gamma étant donnée par :

$$f_X(x;\theta,\gamma) = \frac{\gamma^\theta x^{\theta-1}\exp(-\gamma x)}{\Gamma(\theta)},$$

$(X \sim \text{Gamma}(\theta,\gamma))$ avec $\Gamma(.)$ la fonction Gamma.

La fonction de distribution de la compound Poisson est une fonction de distribution en deux parties, une partie discrète et une partie continue :

La partie discrète correspond à la probabilité que Z soit égale à 0 :

$$P(Z=0) = P(N=0) = \exp(-\Phi).$$

qui est donc donnée par la probabilité d'avoir un zéro de la distribution de Poisson.

La partie continue est trouvée en prenant en compte l'effet de la variable aléatoire N sur la distribution Gamma des variables X_i :

Pour chaque valeur de $N > 0$, Z est la somme de N variables aléatoires X_i. Donc, pour chaque N, la densité de probabilité de Z est la convolution des fonctions de densité de N variables aléatoires indépendantes X_i, qui sont toutes distribuées selon une distribution Gamma.

La convolution des fonctions de densité de N variables aléatoires identiquement distribuées avec une distribution Gamma est donnée par :

$$\begin{aligned} f_Z(z|N=k) &= (f_X * f_X * \ldots * f_X)(z) \\ &= \frac{\gamma^{k\theta} z^{k\theta-1}\exp(-\gamma z)}{\Gamma(k\theta)}, \end{aligned} \tag{7.2}$$

qui est aussi une distribution Gamma de paramètre de forme $k\theta$ et de dispersion γ. On a donc $(Z|N \sim \text{Gamma}(k\theta,\gamma))$.

Enfin, la densité de probabilité de Z pour $N > 0$ est obtenue en pondérant chaque densité conditionnelle par la probabilité correspondante de N, donnée par la distribution de Poisson :

$$
\begin{aligned}
f(z,\gamma,\theta,\Phi) &= \sum_{k=1}^{\infty} P(N=k) f(z|N=k) \\
&= \sum_{k=1}^{\infty} \frac{\Phi^k \exp(-\Phi)}{k!} \frac{\gamma^{k\theta} z^{k\theta-1} \exp(-\gamma z)}{\Gamma(k\theta)} \\
&= \exp(-(\Phi+\gamma z)) \frac{1}{z} \sum_{k=1}^{\infty} \frac{\Phi^k (\gamma z)^{k\theta}}{k!\Gamma(k\theta)}
\end{aligned}
$$

On va maintenant réécrire la densité de probabilité générale ($N \geq 0$) pour la compound Poisson :

$$
f_Z(z,\gamma,\theta,\Phi) = \begin{cases} \exp(-(\Phi+\gamma z)) \frac{1}{z} \sum_{k=1}^{\infty} \frac{\Phi^k (\gamma z)^{k\theta}}{k!\Gamma(k\theta)} & \text{si } z > 0 \\ \exp(-\Phi) & \text{si } z = 0 \end{cases} \tag{7.3}
$$

On pose que l'espérance de Z doit être égale à 1 pour assurer l'identifiabilité du modèle, ce qui implique

$$
E(Z) = \frac{\Phi\theta}{\gamma} = 1.
$$

On obtient ainsi $\Phi = \gamma/\theta$. On a alors un lien directement entre la distribution de Poisson et la distribution Gamma.

7.2 Modèle de survie avec une frailty compound Poisson

Pour construire notre modèle de survie avec la frailty compound Poisson, nous allons utiliser la transformée de Laplace introduite plus tôt (2.17).

La transformée de Laplace de la distribution de Poisson de paramètre Φ est donnée par :

$$
\mathcal{L}_N(s) = \exp(-\Phi + \Phi \exp(-s))
$$

La transformée de Laplace de la distribution Gamma de paramètre θ et γ est donnée par :

$$
\mathcal{L}_X(s) = \left(\frac{\gamma}{\gamma+s}\right)^{\theta}
$$

L'article [17] nous explique comment trouver la transformée de Laplace de la compound Poisson :

$$\begin{aligned}
\mathcal{L}_Z(s) &= E[\exp(-Zs)] \\
&= E[\exp(-s(X_1 + \ldots + X_N)] \\
&= E[\mathcal{L}_X(s)^N] \qquad (7.4) \\
&= \mathcal{L}_N[-\ln(\mathcal{L}_X(s))] \qquad (7.5)
\end{aligned}$$

en replaçant la transformée de Laplace de la Poisson et de la Gamma dans l'équation, on obtient :

$$\begin{aligned}
\mathcal{L}_Z(s) &= \mathcal{L}_N\left[-\ln\left(\left(\frac{\gamma}{\gamma+s}\right)^\theta\right)\right] \\
&= \exp\left[-\Phi + \Phi\exp\left(\ln\left(\left(\frac{\gamma}{\gamma+s}\right)^\theta\right)\right)\right] \\
&= \exp\left[-\Phi + \Phi\left(\frac{\gamma}{\gamma+s}\right)^\theta\right] \\
&= \exp\left[-\Phi\left(1 - \left(\frac{\gamma}{\gamma+s}\right)^\theta\right)\right]. \qquad (7.6)
\end{aligned}$$

On a alors la fonction de survie du modèle de survie avec une frailty compound Poisson qui est donnée par :

$$\begin{aligned}
S(t) &= \mathcal{L}_Z\left(\int_0^\infty h_b(u)du\right) \\
&= \exp\left(-\Phi\left[1 - \left(\frac{\gamma}{\gamma + \int_0^\infty h_b(u)du}\right)^\theta\right]\right) \qquad (7.7)
\end{aligned}$$

avec $h_b(.)$ le hazard de base.

On pose que le hazard de base $h_b(.)$ suit une Weibull de paramètre $\lambda, \rho > 0$, la fonction de survie est alors donnée par :

$$S(t) = \exp\left(-\Phi\left[1 - \left(\frac{\gamma}{\gamma + \lambda t^\rho}\right)^\theta\right]\right).$$

La fonction de densité de survie est donnée par :

$$f(t) = \frac{\Phi\theta\gamma^\theta\lambda\rho t^{\rho-1}}{(\gamma+\lambda t^\rho)^{\theta+1}}\exp\left(-\Phi\left[1 - \left(\frac{\gamma}{\gamma+\lambda t^\rho}\right)^\theta\right]\right).$$

On peut alors trouver la fonction de vraisemblance qui est donnée par

$$L(\theta,\gamma,\lambda,\rho) = \prod_{i=1}^{n}\left[\frac{\Phi\theta\gamma^{\theta}\lambda\rho t^{\rho-1}}{(\gamma+\lambda t^{\rho})^{\theta+1}}\exp\left(-\Phi\left[1-\left(\frac{\gamma}{\gamma+\lambda t^{\rho}}\right)^{\theta}\right]\right)\right]^{\delta_i}\left[\exp\left(-\Phi\left[1-\left(\frac{\gamma}{\gamma+\lambda t^{\rho}}\right)^{\theta}\right]\right)\right]^{1-\delta_i}.$$

La proportion d'individus cure est donnée par :

$$\begin{aligned} q_0 &= \lim_{t\to\infty} S(t) \\ &= \lim_{t\to\infty}\exp\left(-\Phi\left[1-\left(\frac{\gamma}{\gamma+\lambda t^{\rho}}\right)^{\theta}\right]\right) \\ &= \exp\left(-\Phi\left[1-0^{\theta}\right]\right) \\ &= \exp(-\Phi) \\ &= \exp(-\frac{\gamma}{\theta}). \end{aligned} \tag{7.8}$$

Chapitre 8

Simulations compound Poisson

Pour les simulations des distributions Compound Poisson, nous avons généré les données de la même manière que dans la section 4.2. Des échantillons de données de 500 individus ont été générés en fixant les paramètres λ et ρ de la fonction de hazard de base qui suit une distribution de Weibull à 0.06 et 1.5. Ensuite, nous avons varié le taux d'individus cure en commençant par $1 - \psi = 0.2$, puis $1 - \psi = 0.4$, et enfin $1 - \psi = 0.6$.

8.1 Distribution des temps de censure : Weibull

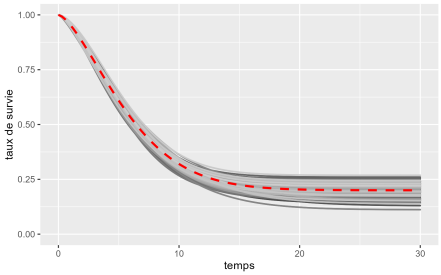
Nous avons d'abord utilisé une distribution de Weibull pour générer les temps de censure avec des paramètres différents pour chaque taux de cure fixé afin d'atteindre en moyenne un ajout de 10% de censure à la censure déjà présente avec les individus cure.

Dans le tableau 8.1, on observe que les estimations du taux de cure sont moins précises pour le modèle compound Poisson que pour le modèle MCM. Pour chaque taux fixé, la moyenne de nos estimations est inférieure à la valeur réelle du taux de cure (tous les biais sont négatifs) et la variance est plus grande pour le modèle compound Poisson que pour le MCM. Le taux de convergence des estimations est assez élevé pour tous les taux fixés (plus de 95%).

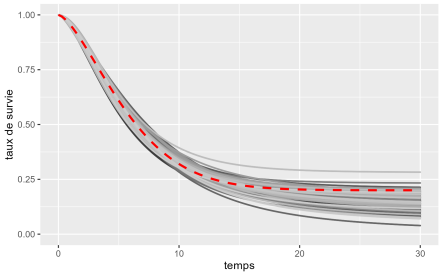
À la figure 8.1, on voit qu'au début le modèle compound Poisson ajuste bien les données, mais dès que l'on atteint la proportion de cure, les courbes de survie s'éparpillent en dessous de la ligne rouge qui a généré les données. On remarque également sur les graphes que la variance est plus faible pour le modèle MCM que pour le modèle Compound Poisson.

$1-\psi$	**Modèle**	**Moyenne**	**Biais**	**Variance**	**MSE**	**Taux de convergence**
0.2	MCM	0.2000	0.0000	0.0011	0.0011	100%
0.2	CP	0.1232	-0.0768	0.0024	0.0083	96%
0.4	MCM	0.3990	-0.0010	0.0016	0.0016	100%
0.4	CP	0.3481	-0.0519	0.0044	0.0071	95%
0.6	MCM	0.5893	-0.0107	0.0063	0.0064	100%
0.6	CP	0.5303	-0.0700	0.0112	0.0161	98%

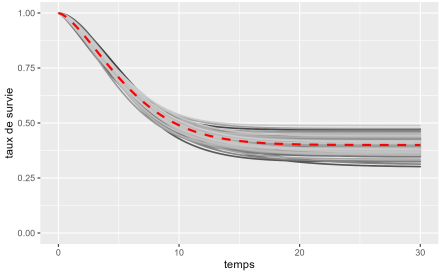
TABLE 8.1 – Résultats des estimations de la proportion de cure avec le modèle compound Poisson et le MCM pour différents taux de cure fixés et avec une censure qui suit une distribution de Weibull



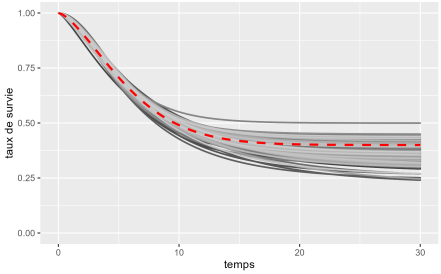
(a) MCM (taux de cure de 0.2)



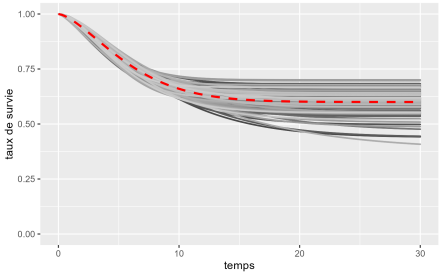
(b) CP (taux de cure de 0.2)



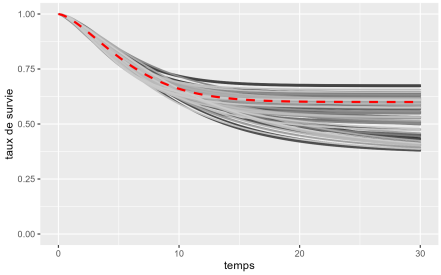
(c) MCM (taux de cure de 0.2)



(d) CP (taux de cure de 0.4)



(e) MCM (taux de cure de 0.2)



(f) CP (taux de cure de 0.6)

FIGURE 8.1 – Courbes de survie du modèle compound Poisson et du MCM pour différents taux de cure fixés avec une censure qui suit une Weibull

8.2 Distribution des temps de censure : exponentielle

Nous allons dans cette section utiliser une distribution exponentielle pour générer les temps de censure. Pour ce faire, nous allons fixer les paramètres de la distribution exponentielle de manière à ce que le taux de censure ajouté soit de 10%.

Les résultats de l'estimation du taux de cure des données générées avec une

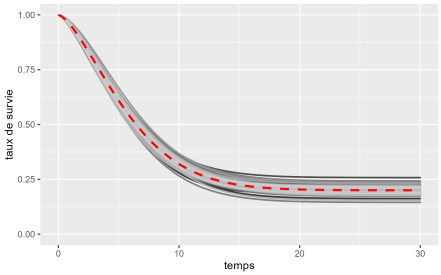
censure exponentielle sont présentés dans le tableau 8.2. Nous constatons que, sur ces échantillons de données, le modèle compound Poisson estime nettement mieux le taux de cure. La MSE est plus faible que lorsque la censure était distribuée selon une distribution de Weibull. Sur ces données, le modèle compound Poisson est aussi performant que le MCM, ils ont la même MSE pour chaque taux de cure fixé. Cela est probablement dû au fait que la censure générée avec une distribution exponentielle donne des temps de censure bien plus grands que la censure générée par la distribution de Weibull. La censure arrive plus tard pour la distribution exponentielle, ce que le modèle semble mieux gérer.

Le taux de convergence des estimations du modèle compound Poisson est un peu moins bon que celui du MCM, mais reste assez satisfaisant.

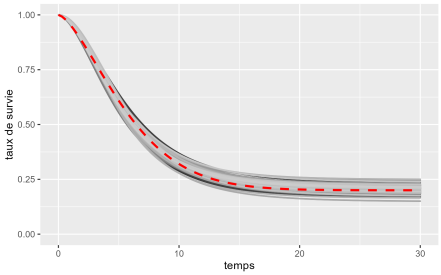
La figure 8.2 montre également que le modèle s'ajuste mieux aux données lorsque la censure est distribuée selon une distribution exponentielle. Nous remarquons que les courbes grises sont toutes assez proches de la courbe rouge qui a généré les données.

$1-\psi$	**Modèle**	**Moyenne**	**Biais**	**Variance**	**MSE**	**Taux de convergence**
0.2	MCM	0.1993	-0.0007	0.0004	0.0004	99%
0.2	CP	0.2037	0.0037	0.0004	0.0004	81%
0.4	MCM	0.4007	-0.0007	0.0006	0.0006	100%
0.4	CP	0.4018	0.0018	0.0006	0.0006	80%
0.6	MCM	0.6007	0.0007	0.0007	0.0007	100%
0.6	CP	0.5989	-0.0010	0.0007	0.0007	98%

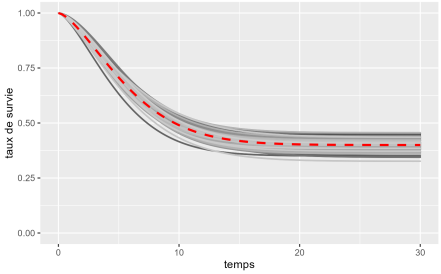
TABLE 8.2 – Résultats des estimations de la proportion de cure avec le modèle compound Poisson et le MCM pour différents taux de cure fixés et avec une censure qui suit une distribution exponentielle



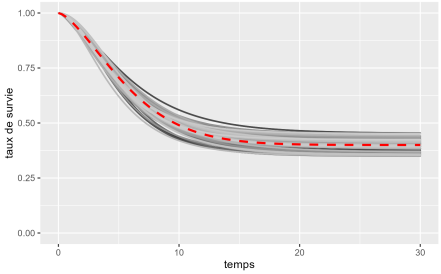
(a) MCM (taux de cure de 0.2)



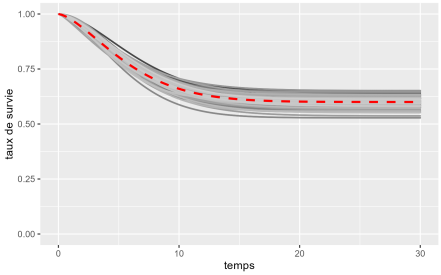
(b) CP (taux de cure de 0.2)



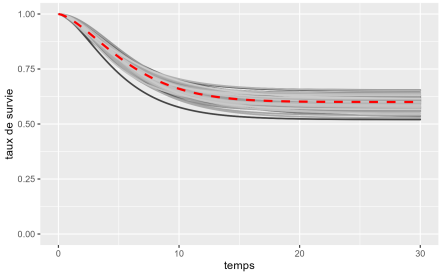
(c) MCM (taux de cure de 0.4)



(d) CP (taux de cure de 0.4)



(e) MCM (taux de cure de 0.6)



(f) CP (taux de cure de 0.6)

FIGURE 8.2 – Courbes de survie du modèle compound Poisson et du MCM pour différents taux de cure fixés avec une censure qui suit une exponentielle

8.3 Conclusion des simulations compound Poisson

Ces simulations ont montré que quand les temps de censure arrivent plus tard avec la distribution exponentielle, le modèle compound Poisson estime aussi bien le taux de cure que le MCM. Quand la censure suit une distribution de Weibull où la censure arrive plus tôt, le modèle ajuste moins bien les données.

L'article [13] propose une extension du modèle compound Poisson où le paramètre Φ qui est le paramètre de la distribution de Poisson sera une variable aléatoire qui suit une certaine distribution. Dans l'article, les auteurs proposent des distributions telles que la Gamma, l'inverse Gaussian et la positive stable pour la distribution de la variable aléatoire Φ. Mais ces approches concernent les modèles de frailty multivariés, où les individus d'un même cluster auront les même valeurs de Φ générées par la distribution choisie. La variance de la variable aléatoire Φ indiquera le variation entre les différents cluster. Notre étude se limitant aux modèles de survie avec une frailty univariée, nous n'avons pas exploré plus loin cette piste.

Chapitre 9

Application

Dans ce mémoire, nous avons examiné théoriquement plusieurs modèles, notamment le MCM, le modèle de Poisson, le modèle géométrique, le modèle binomial, ainsi que les modèles de frailty ZMP, ZMG et ZMBN, puis le modèle compound Poisson. Nous avons ensuite appliqué ces modèles à des échantillons de données que l'ont a générés avec différents taux de cure pour observer comment les modèles estimaient ce taux de cure. Nous allons à présent appliquer tous les modèles étudiés à la base de données "retinopathy" du package "survival" en R. Contrairement aux simulations que nous avons menées précédemment, nous ne connaissons pas le taux de cure "réel" de cette base de données ; nous allons l'estimer avec nos différents modèles.

Nous commencerons par expliquer comment on va calculer nos intervalles de confiance, puis présenterons brièvement la base de données. Nous allons ensuite ajuster nos différents modèles, et enfin les comparer à l'aide de l'indice AIC afin de vérifier si l'un d'eux ajuste mieux les données que les autres.

9.1 Intervalle de confiance

On va pour chaque taux de cure calculer son intervalle de confiance de 95%. Pour ce faire, nous utilisons la propriété asymptotique de normalité de l'estimateur du maximum de vraisemblance. L'intervalle de confiance de l'estimateur du taux de cure $\hat{q_0}$ sera donné par :

$$[\hat{q_0} - Z_\alpha \sqrt{V(\hat{q_0})}, \hat{q_0} + Z_\alpha \sqrt{V(\hat{q_0})}]$$

avec Z_α le quantile d'ordre $1 - \frac{\alpha}{2}$ de la distribution normale réduite centrée, et $V(q_0)$ la variance de l'estimation de la proportion de cure.

Quand la proportion de cure dépend de un paramètre θ, la variance de son estimation est donnée par :

$$V(\hat{q_0}) = \frac{\partial \hat{q_0}(\theta)}{\partial \theta} V(\theta)$$

avec $V(\theta)$ trouvée à l'aide de l'inverse de la matrice Hessienne. La fonction "optim", qui a servi à maximiser notre fonction de vraisemblance, nous donne la Matrice Hessienne.

Quand la proportion de cure dépend de n paramètre $\theta_1, \ldots, \theta_n$, la variance de son estimation est donnée par :

$$V(\hat{q_0}) = \left(\frac{\partial \hat{q_0}(\theta_1)}{\partial \theta_1}, \ldots, \frac{\partial \hat{q_0}(\theta_n)}{\partial \theta_n} \right)' * V(\theta_1, \ldots, \theta_n) * \left(\frac{\partial \hat{q_0}(\theta_1)}{\partial \theta_1}, \ldots, \frac{\partial \hat{q_0}(\theta_n)}{\partial \theta_n} \right)$$

avec $V(\theta_1, \ldots, V(\theta_n))$ la matrice de variance covariance trouvée à l'aide de l'inverse de la matrice Hessienne. La fonction "optim", qui a servi à maximiser notre fonction de vraisemblance, nous donne la Matrice Hessienne.

9.2 Présentation de la base de données

Pour commencer, la base de données "retinopathy" contient deux mesures pour 197 patients, ce qui nous donne 394 observations au total. Les patients souffrent de rétinopathie diabétique, une complication du diabète qui affecte les yeux et peut entraîner une perte de vision si elle n'est pas traitée. Chaque patient de cette base de données a un oeil soumis à un traitement au laser, tandis que l'autre oeil ne reçoit aucun traitement. L'événement d'intérêt est la perte de vision. Les durées de survie dans cet échantillon représentent le temps réel jusqu'à la perte de la vision en mois. La censure était causée par le décès, l'abandon ou la fin de l'étude.

Comme les modèles étudiés dans ce travail utilisent une frailty univariée et pas multivariée, nous ne savons pas prendre en compte le fait que les informations pour les deux yeux d'un patient sont corrélées. Nous allons donc restreindre la base de données à un seul oeil par patient. Nous allons garder l'oeil de contrôle, c'est-à-dire l'oeil non traité. Nous aurons donc dans notre base de données 197 patients avec un oeil non traité.

Dans cette base de données, nous avons 96 mesures censurées et 101 mesures qui ont eu l'événement, donc 101 yeux qui ont perdu la vision. Le temps moyen

de censure est de 46 mois et le temps moyen d'événement est de 18 mois. Le plus grand temps d'événement est 61 mois et le plus petit 0.3. Nous allons regarder la courbe de l'estimation non paramétrique de Kaplan-Meier que nous avions introduite dans le chapitre 2.1 pour voir si la courbe suggère la présence d'individus cure.

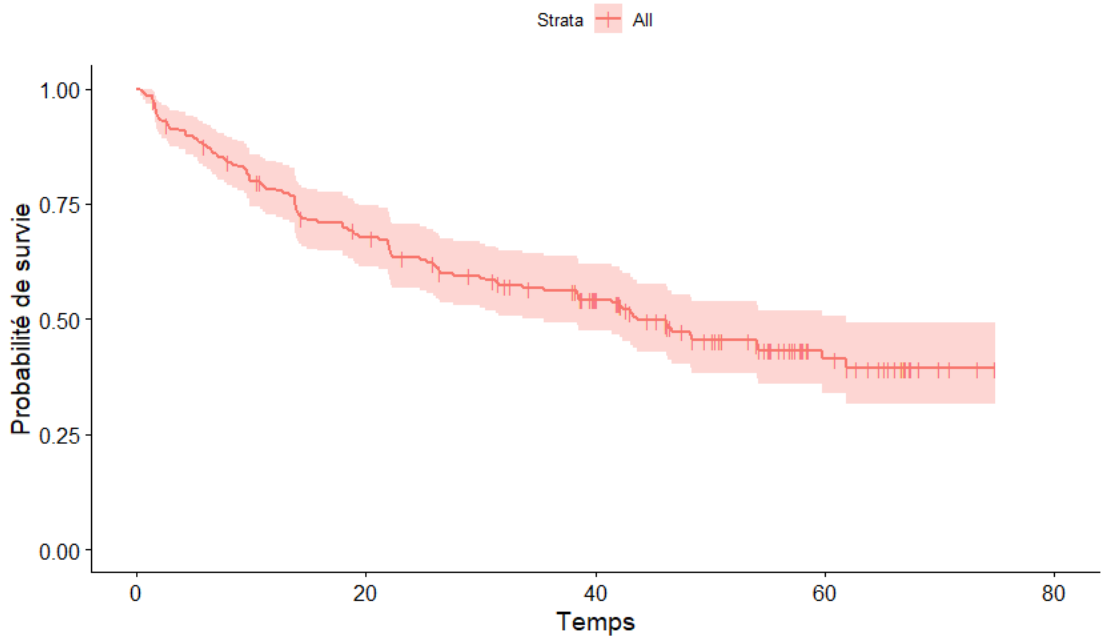


FIGURE 9.1 – Estimateur de Kaplan-Meier de la fonction de survie de la base de données rétinopathie

On peut voir à la figure 9.1 que la courbe de Kaplan-Meier suggère qu'il y a une fraction d'individus cure dans la base de données car la courbe se finit en un long plateau bien au-dessus du taux de survie de 0. Avec ce graphe, on peut penser que l'on a un taux de cure d'environ 30% dans notre base de données. Ce graphe nous confirme que c'est approprié d'appliquer nos modèles, qui prennent en compte la proportion de cure, à cette base de donnée.

9.3 Application des modèles PS

Nous commençons par appliquer le modèle MCM et les modèles de la famille de distribution PS (Poisson (3.6), géométrique (3.9) et binomial négatif (3.12)) à la

base de données. On peut voir dans la table 9.1 que les taux de cure estimés sont assez différents. Le MCM estime le plus haut taux de cure avec 29%, ce qui reflète assez bien ce que la courbe de Kaplan-Meier nous montre. Le taux de cure des modèles Poisson, gométrique et binomial négatif sont un peu plus faibles. Au vu de nos résultats de simulation, il semble logique de se demander s'il ne faudrait pas utiliser les modèles avec une frailty de la famille ZMPS pour augmenter le nombre de zéros dans la mesure où il y a une certaine différence entre le taux de cure du MCM et celui des 3 autres modèles.

Les intervalles de confiance pour les estimations des taux cure sont assez larges pour les 4 modèles.

Sur la figure 9.2, on observe que les trois courbes de survie estimées par les modèles PS sont légèrement plus basses que celle estimée par le modèle MCM (noir). Comme l'indique le tableau, la courbe du modèle binomial négatif (vert) est la plus proche de celle du MCM. Les courbes du modèle de Poisson (rouge) et du modèle géométrique (bleu) sont un peu plus basses. Toutes les courbes ont une allure similaire, avec seulement une légère différence observable au niveau du taux de cure.

Modèle	**Taux de cure**	**IC**
MCM	0.2890	[0.06, 0.52]
Poisson	0.2453	[-0.05, 0.54]
Géométrique	0.2345	[0.04, 0.43]
Binom négatif	0.2585	[0.04, 0.48]

TABLE 9.1 – Taux de guérison estimés par le MCM et les différents modèles PS.

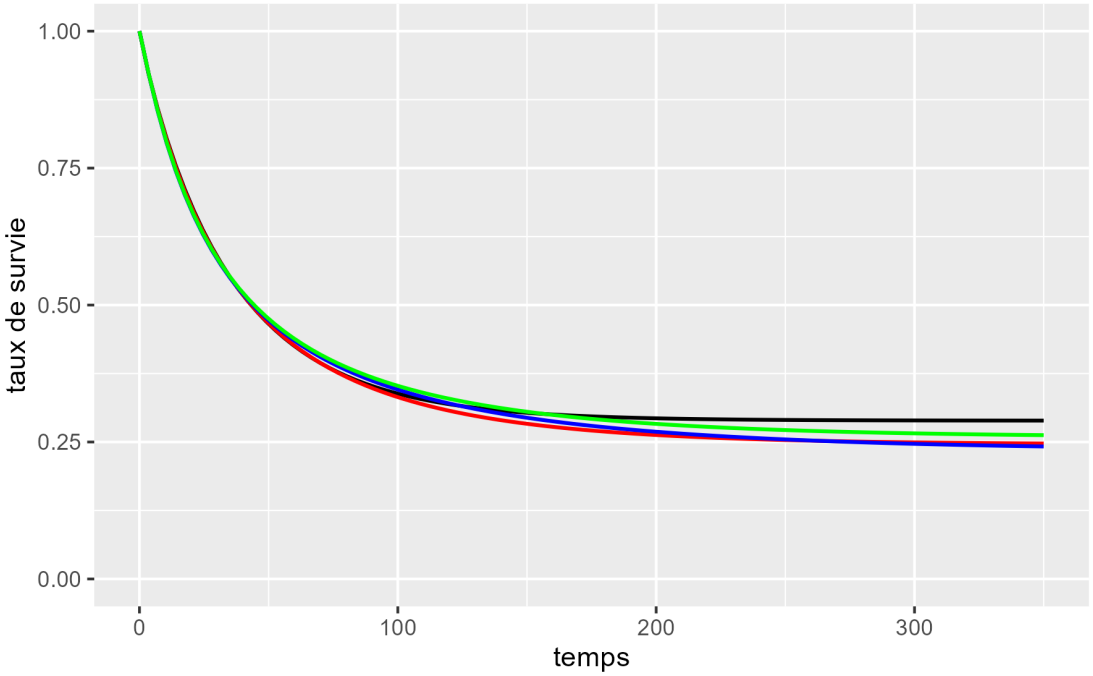


FIGURE 9.2 – Courbes de survie estimées par les modèles MCM (noir), Poisson (rouge), geométrique (bleu) et binomial négatif(vert).

9.4 Application des modèles ZMPS

On va maintenant appliquer les modèles de survie avec une frailty de la famille de distribution ZMPS : ZMP (5.9), ZMG (5.10) et ZMBN (5.11). On va borner le paramètre p des distributions ZMPS entre 0 et $\frac{1}{1-\xi_{PS}(0;\hat{\mu},\hat{\phi})}$.

Donc, pour le modèle ZMP on a

$$p \in \left[0, \frac{1}{1-0.2453} = 1.3248\right],$$

pour le ZMG on a

$$p \in \left[0, \frac{1}{1-0.2345} = 1.3060\right],$$

et pour le ZMBN on a

$$p \in \left[0, \frac{1}{1-0.2585} = 1.3484\right].$$

On voit dans la table 9.2 que les taux de cure estimés des modèles ZMPS sont plus petits que les taux de cure estimés des modèles PS. On constate que les modèles ZMG et ZMB ont fortement diminué le nombre de zéros estimé de leur distribution PS homologue. Pour le modèle ZMP, la proportion de zéros estimée n'a pas fortement diminué. En outre, le tableau montre que les intervalles de confiance des distributions PS sont toujours inclus dans les intervalles de confiance des distributions ZMPS, qui sont plus larges.

Modèle	**Taux de cure**	**IC**
Poisson	0.2453	[-0.05, 0.54]
ZMP	0.2434	[-0.09, 0.55]
Géométrique	0.2345	[0.04, 0.43]
ZMG	0.1984	[-0.15, 0.55]
Binom négatif	0.2585	[0.04, 0.48]
ZMBN	0.2154	[-0.12, 0.55]

TABLE 9.2 – Taux de cure estimés par les différents modèles

On peut voir dans le tableau 9.3 que tous les paramètres p sont supérieurs à 1. Comme nous l'avons observé dans l'autre tableau, cela réduit le nombre de zéros estimés dans la distribution PS. En ce qui concerne les intervalles de confiance des paramètres p, nous remarquons qu'ils sont assez larges. De plus, ils incluent systématiquement 1, ce qui, rappelons-le, signifie qu'il n'est pas nécessaire d'ajouter de zéro au zéros estimés par la distribution PS. Par conséquent, on peut se demander si passer à une distribution ZMPS est vraiment nécessaire et si une distribution PS ne serait pas suffisante.

Modèle	**Cure**	**Paramètre p**	**IC p**
ZMP	0.2434	1.1073	[-2.48, 4.69]
ZMG	0.1984	1.2009	[0.20, 2.20]
ZMBN	0.2154	1.2264	[-1.33, 3.79]

TABLE 9.3 – Taux de cure et paramètres p estimés par les modèles ZMPS

Sur le graphique 9.3, on observe que, comme le montrent les valeurs du tableau, la courbe orange du modèle ZMP se confond avec celle du modèle de Poisson (rouge), et qu'elles sont toutes deux légèrement en dessous de celle du MCM (noir).

Le graphique 9.4 confirme ce que les données du tableau montrent : la courbe du modèle ZMG (bleu clair) est plus basse que celle du modèle géométrique (bleu foncé), ce qui reflète le fait que le taux de cure du ZMG est plus faible que celui du modèle géométrique.

Le graphique 9.5 confirme ce que le tableau de valeurs indiquait : la courbe du modèle binomial négatif (vert clair) est très proche de la courbe du MCM (noir), tandis que celle du ZMBN (vert foncé) se trouve en dessous de la courbe binomiale négative.

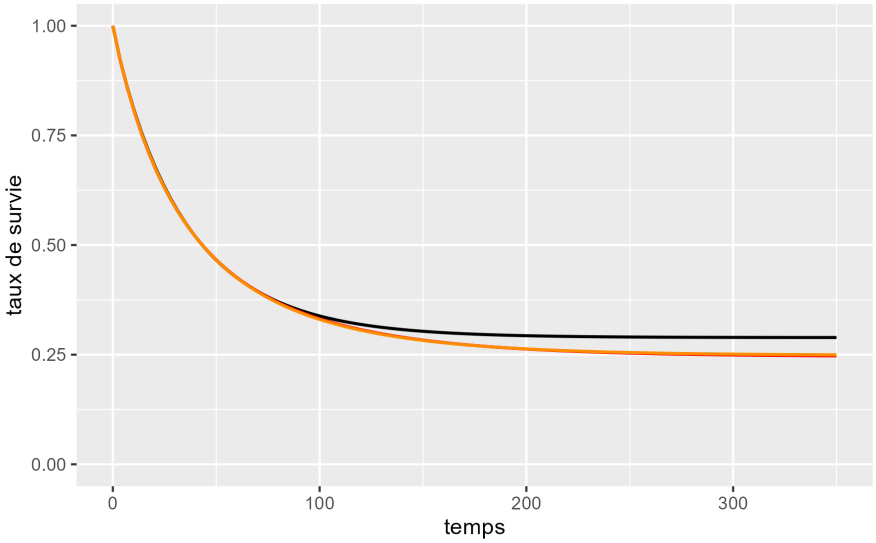


FIGURE 9.3 – Courbes de survie estimées par les modèles MCM (noir), Poisson (rouge) et ZMP (orange).

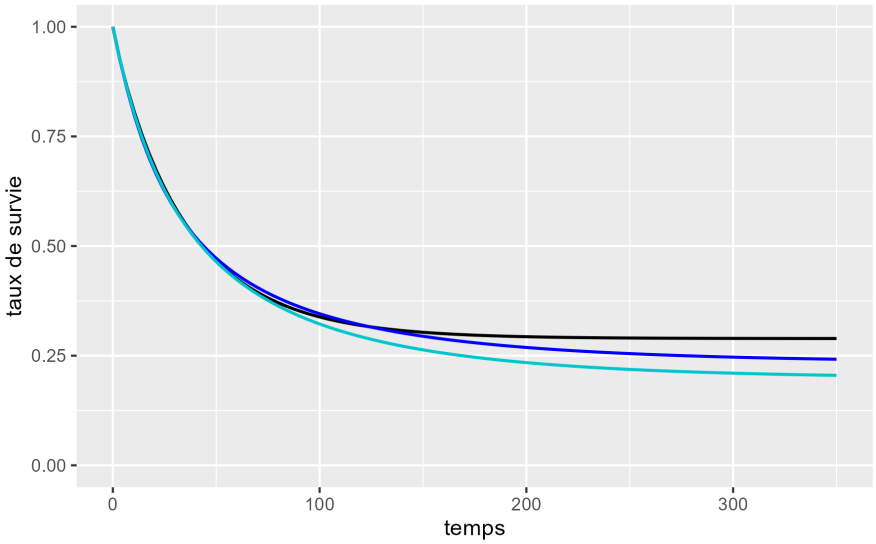


FIGURE 9.4 – Courbes de survie estimées par les modèles MCM (noir), géométrique (bleu) et ZMG (bleu clair).

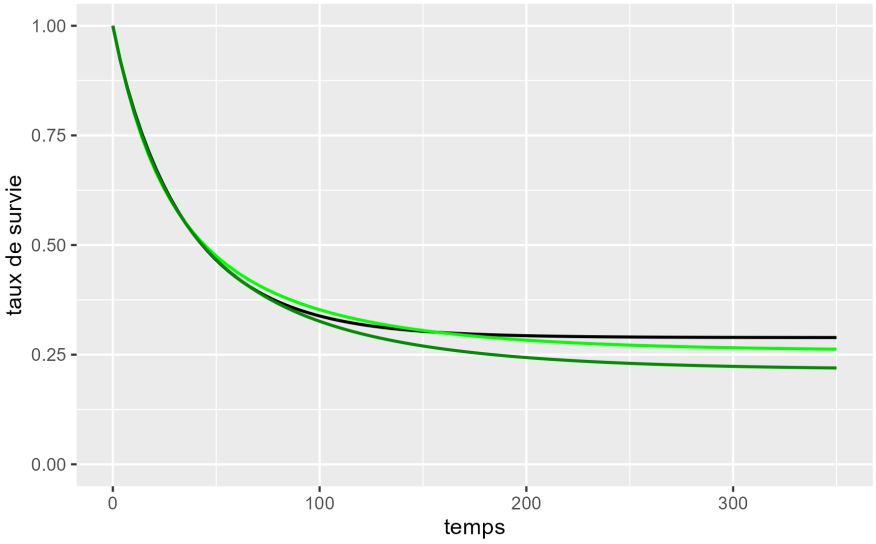


FIGURE 9.5 – Courbes de survie estimées par les modèles MCM (noir),binomial négatif (vert) et ZMBN (vert foncé).

9.5 Application du modèle compound Poisson

On applique maintenant le modèle compound Poisson. On obtient pour ce modèle une estimation de taux de cure très basse par rapport au taux de cure que les autres modèles estiment et que la courbe de Kaplan-Meier indique. De plus, son intervalle de confiance est très large. Ces résultats sont disponibles à la table 9.4.

Modèle	**Taux de cure**	**IC**
CP	0.1735	[-0.10, 0.45]

TABLE 9.4 – Taux de cure estimés par le modèle compound Poisson

Sur le graph 9.6, on voit que la courbe du compound Poisson (mauve) est bien plus basse que la courbe du MCM (noir).

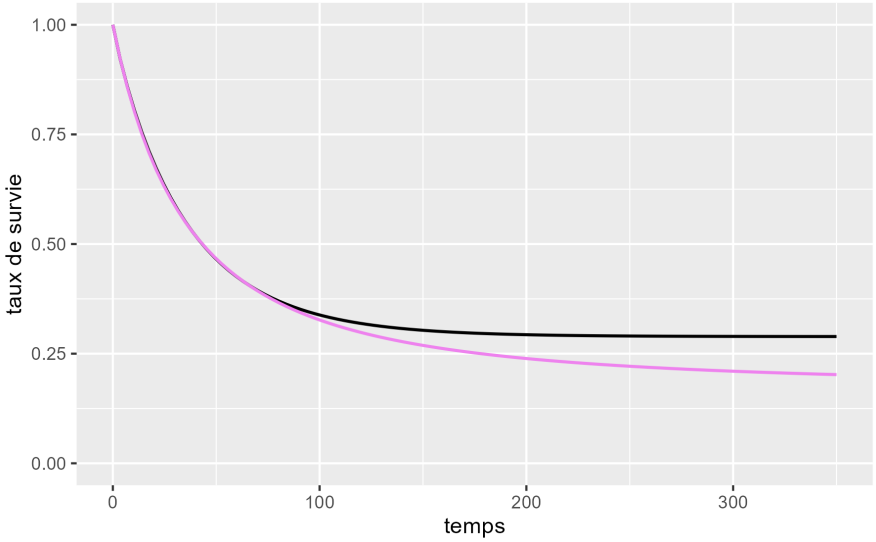


FIGURE 9.6 – Courbes de survie estimées par les modèles MCM (noir), compound Poisson (mauve).

9.6 Comparaison de tous les modèles

Pour comparer les différents modèles, on va comparer leurs critères d'information d'Akaike (AIC) respectifs, donnés par

$$AIC = -2 * log(L) + 2 * P$$

où L est le maximum de la vraisemblance et P le nombre de paramètres du modèle.

Les résultats confirment ce que nous avons observé dans les sections précédentes grâce aux tables et aux graphiques : le modèle MCM semble mieux ajuster les données, avec le plus petit AIC et un intervalle de confiance plus restreint pour son taux de cure estimé.

En ce qui concerne les modèles de Poisson, géométrique et binomial négatif, nous constatons que dans chaque cas, leurs équivalents ZMPS ont un AIC plus élevé. Ainsi, en termes de préférence, nous pourrions favoriser les modèles PS plutôt que les modèles ZMPS.

Les modèles ZMPS comportent un paramètre supplémentaire, ce qui entraîne une pénalisation plus importante de l'AIC, mais dans le cas où $p \in [0, 1]$ (ZIPS), ils offrent l'avantage d'estimer le taux d'individus immunisés. Dans notre étude, nous avons appliqué les modèles ZMP*, ZMG*, ZMBN* lorsque $p \in [0, 1]$, mais pour les trois distributions, le paramètre p estimé était de 1. Ainsi, la proportion de cure estimée des modèles PS n'a pas été modifiée et les taux d'individus immunisés sont égaux à zéro. Cependant, dans d'autres ensembles de données, il pourrait être intéressant d'utiliser les modèles ZIPS, même si l'AIC est moins favorable, surtout lorsque l'estimation de la proportion d'individus immunisés est un objectif.

On constate que pour les modèles PS, le meilleur AIC est celui du modèle de Poisson, suivi du modèle géométrique et ensuite du modèle binomial négatif. On observe le même ordre pour les AIC des modèles ZMPS. Donc, s'il faut faire un choix parmi les trois types de distributions discrètes, la distribution de Poisson semble préférable, avec un AIC très proche de celui du MCM.

Quant au modèle compound Poisson, il présente le plus mauvais AIC (après le ZMBN que nous avons déjà écarté), et il estime une proportion de cure nettement inférieure à celle proposée par les autres modèles et par la courbe de Kaplan-Meier. Nous ne recommandons donc pas son utilisation pour cette base de données.

Modèle	**Taux de cure**	**IC**	**AIC**	**Nbr paramètres**
MCM	0.2890	[0.06, 0.52]	-1025.961	3
Poisson	0.2453	[-0.05, 0.54]	-1025.654	3
ZMP	0.2484	[-0.09, 0.50]	-1023.661	4
Géométrique	0.2345	[0.04, 0.43]	-1025.556	3
ZMG	0.1984	[-0.15, 0.55]	-1023.396	4
Binom négatif	0.2585	[0.04, 0.48]	-1023.668	4
ZMBN	0.2154	[-0.12, 0.55]	-1021.457	5
CP	0.1735	[-0.10, 0.45]	-1023.54	4

TABLE 9.5 – Table reprenant les taux de cure, les IC de taux de cure, les AIC et le nombre de paramètres de chacun des modèles

Chapitre 10

Conclusion

Dans ce mémoire, nous avons exploré l'analyse de données de survie où une partie de la population est considérée comme cure, en utilisant des modèles de frailty univariés avec une frailty discrète non négative. Une distribution discrète pour la frailty permet d'attribuer directement une probabilité à chaque valeur de la frailty. La proportion d'individus cure est la probabilité que la frailty soit égale à zéro. Cette approche nous a permis de mettre en évidence la relation entre la distribution de la frailty et la proportion d'individus cure.

Nous avons construit 3 modèles de frailty avec une distribution discrète, en utilisant les distributions de Poisson, géométrique et binomiale négative que C.Caroni [4] suggérait.

Ensuite, nous avons générer des échantillons de données avec les 3 modèles discrets, pour dans un second temps appliquer à ces échantillons les mêmes modèles. Cette démarche nous a permis d'observer que plus la taille de l'échantillon est grande, meilleur est l'ajustement des données ; tandis que plus le taux de censure est élevé, plus l'ajustement des données est mauvais.

Ensuite, nous avons généré des échantillons de données avec un mixture cure modèle dont la censure suit une distribution exponentielle. Dans ce cas, les trois modèles estiment correctement le taux de cure et s'ajustent bien aux données. En revanche, lorsque la censure suit une distribution de Weibull, le modèle de Poisson semble bien fonctionner, mais les modèles géométrique et binomial négatif estiment un taux de guérison beaucoup trop bas. De plus, l'article de C. Caroni [4] souligne les limitations de ces trois distributions dans le contexte de la distribution discrète de la frailty, car le rapport entre la proportion des valeurs de la frailty égale à zéro et les proportions des autres valeurs de la frailty est trop contraint. Cela conduit dans le cas des simulations réalisées à une estimation trop faible de la

proportion de cure pour les modèles géométrique et binomial négatif. Il est donc nécessaire de complexifier ces distributions afin d'obtenir une meilleure estimation de la proportion de cure. Il serait intéressant de trouver un modèle capable de générer des échantillons de données qui mettent à mal le modèle de Poisson. C. Caroni [4] a mentionné que sur certaines bases de données, ce modèle pourrait rencontrer des difficultés. Mais ce n'est pas ce que nous observons ici. Il serait donc utile de déterminer dans quelles situations ce modèle peine à estimer le taux de cure.

Pour modifier le nombre de zéros de la distribution de la frailty, nous nous sommes tournés vers les distributions de la famille ZMPS, comme suggéré par K. Molina [11]. Ces distributions sont à la base des distributions de la famille PS, mais elles intègrent un paramètre supplémentaire, p, qui permet d'ajuster la proportion de zéros, en augmentant celle-ci (ZIPS) ou en diminuant celle-ci (ZDPS). Nous avons ainsi construit des modèles ZMPS en utilisant nos trois modèles de Poisson, géométrique et binomial négatif, puisque nous avons montré qu'ils appartenaient à la famille PS. Nous les avons appelé ZMP, ZMG, et ZMBN. Nous espérions ainsi mieux estimer le taux de cure et ajuster les données car ces distributions offrent une plus grande flexibilité entre la proportion de valeurs de la frailty égale à zéro et les proportions des autres valeurs de la frailty grâce au paramètre p ajouté. De plus, selon l'interprétation de K.Molina [11] et G.Cancho [**?**], les modèles ZIPS ($p \in [0, 1]$) offrent l'avantage de séparer la proportion de cure en deux composantes : une pour les individus immunisés et une pour ceux qui ont été guéris de la maladie.

Nous avons réalisé de nouvelles simulations sur les échantillons de données générés avec le MCM, utilisant une censure suivant une distribution de Weibull, afin de comparer les modèles avec une frailty de la famille de distributions ZMPS (avec $p \in \left[0, \frac{1}{1-\xi_{PS}(0;\hat{\mu},\hat{\phi})}\right]$) et les modèles ZMPS* (avec $p \in [0, 1]$) à leurs homologues ayant une distribution de la frailty appartenant à la famille PS. Les résultats des simulations suggèrent que pour les taux de cure fixés à 0.2 et 0.4, les modèles de survie avec la distribution de frailty ZMPS* (suivi de près par le modèles ZMPS) ajustent mieux les données que les modèles PS. Pour le taux de cure fixé à 0.6, on voit que les estimations sont meilleures avec les modèles PS. Comme le taux de cure estimé est déjà très élevé, les modèles ont du mal a encore augmenter ce taux.

Il serait intéressant de trouver des modèles de survie capables de générer des données où le nombre de zéros est trop grand, conduisant ainsi à une estimation trop élevée de la proportion de zéros. Dans de tels cas, l'utilisation de distributions ZDPS avec un paramètre $p > 1$ serait nécessaire pour mieux ajuster les données. Il serait tout aussi intéressant de générer des données avec un certain taux d'individus immunisés pour voir si les 3 modèles savent estimer correctement cette proportion

d'individus immunisés.

Ensuite, nous avons utilisé comme distribution de la frailty la distribution "compound Poisson". C'est une distribution continue où on ajoute une partie discrète pour mettre un poids en 0 et savoir estimer la proportion de cure.

Des simulations ont été effectuées en générant encore une fois les échantillons de données avec un MCM avec une censure qui suit une distribution de Weibull. Le modèle ajuste moins bien les données que le MCM, mais on a un bon taux de convergence. Nous avons utilisé ensuite une distribution exponentielle pour la censure, ce qui fait que la censure arrive plus tard. Cela a donné de meilleurs résultats pour l'estimation du taux de cure mais un moins bon taux de convergence. L'article [13] propose une extension du modèle compound Poisson, où le paramètre Φ de la distribution de Poisson est une variable aléatoire. Mais cela dans le contexte des modèles de frailty multivariés où la population est divisée en clusters et chaque individu d'un même cluster à la même valeur du paramètre Φ. On s'est concentré dans ce mémoire sur des modèles de frailty univariés, mais il serait intéressant d'étudier plus en profondeur ce modèle de compound Poisson dans la contexte des shared frailty model.

Nous avons appliqué nos différents modèles à la base de données "retinopathy" du package survival en R. En utilisant l'indice AIC et les intervalles de confiance du taux de guérison, nous avons constaté que le meilleur modèle était le MCM, suivi des modèles avec une distribution de la frailty de la famille PS, en particulier le modèle de Poisson. Les modèles ZMPS ont également donné de bons résultats mais ils sont pénalisés en raison du nombre supplémentaire de paramètres, bien qu'ils soient intéressants si l'on souhaite estimer le taux d'immunisés. Le modèle compound Poisson a présenté les pires performances. Il serait intéressant d'appliquer tous ces modèles à d'autres bases de données pour déterminer dans quelles situations les modèles avec une distribution discrète de la frailty ajustent mieux les données et estiment plus précisément le taux de cure par rapport au MCM.

Pour finir, ce mémoire m'a permis d'en apprendre plus sur les modèles de frailty utilisant une distribution discrète pour modéliser des données de survie en présence d'une fraction d'individus cure. Il a permis de synthétiser, de regrouper et de comparer les travaux de divers auteurs sur ce sujet, en appliquant leurs modèles sur des données générées de la même façon.

Dans la continuité de cette étude, il serait pertinent d'intégrer des covariables afin d'analyser comment l'estimation du taux cure peut être influencée par la

présence de covariables. Enfin, ce mémoire abordant uniquement les modèles de frailty univariés, il serait intéressant aussi de voir ce que l'on pourrait faire avec ces frailty discrètes dans le cadre de modèles de frailty multivariés.

Bibliographie

[1] Aalen O. (1992). *Modelling Heterogeneity in Survival Analysis by the Compound Poisson Distribution.* The Annals of Applied Probability, 2(4) :951-972.

[2] Bender R, Augustin T, Blettner M. (2005). *Generating survival times to simulate Cox proportional hazards models.* Statistics in Medicine, 24 :1713-1723.

[3] Boulesteix A, Groenwold R, Abrahamowicz M. (2020). *Introduction to statistical simulations in health research.* BMJ Open, 10 :e039921.

[4] Caroni C, Crowder M, Kimber A. (2010). *Proportional hazards models with discrete frailty.* Lifetime Data Analysis, 16 :374–384.

[5] Cancho V, Macera M, Suzuki A, Louzada F, Zavaleta K. (2020). *A new long-term survival model with dispersion induced by discrete frailty.* Lifetime Data Analysis, 26(2) :221-244.

[6] de Souza D, Cancho VG, Rodrigues J, Balakrishnan N. (2017). *Bayesian cure rate models induced by frailty in survival analysis.* Statistical Methods in Medical Research, 26(5) :2011-2028.

[7] Legrand C. (2021). *Advanced survival models.* Chapman Hall.

[8] Legrand C. (2023). *Advanced survival models.* Cours dispensé à l'Université UCLouvain.

[9] Luyede BO, Mashabe B, Fagbamigbe A, Makubate B, Wanduku D. (2020). *The exponentiated generalized power series : Family of distributions : theory, properties and applications.* Heliyon, 6 :e04653.

[10] Kemp A. (2010). *Families of power series distributions, with particular reference to the Lerch family.* Journal of Statistical Planning and Inference, 140 :2255-2259.

[11] Molina K, Calsavara V, Tomazella V, Milani E. (2021). *Survival models induced by zero-modified power series discrete frailty : Application with a melanoma data set.* Statistical Methods in Medical Research, 30(8) :1874-1889.

[12] Morris T, White I, Crowther M. (2019). *Using simulation studies to evaluate statistical methods.* Statistics in Medicine, 38 :2074-2102.

[13] Moger T, Aalen O. (2005). *Distribution for multivariate frailty based on the compound Poisson distribution with random scale*. Lifetime Data Analysis, 11(1) :41-59.

[14] Rahmati M, Rezanejad Asl P, Mikaeli J, Zeraati H, Rasekhi A. (2021). *Compound Poisson frailty model with a gamma process prior for the baseline hazard : accounting for a cured fraction*. Journal of Applied Statistics, 49(13) :3377-3391.

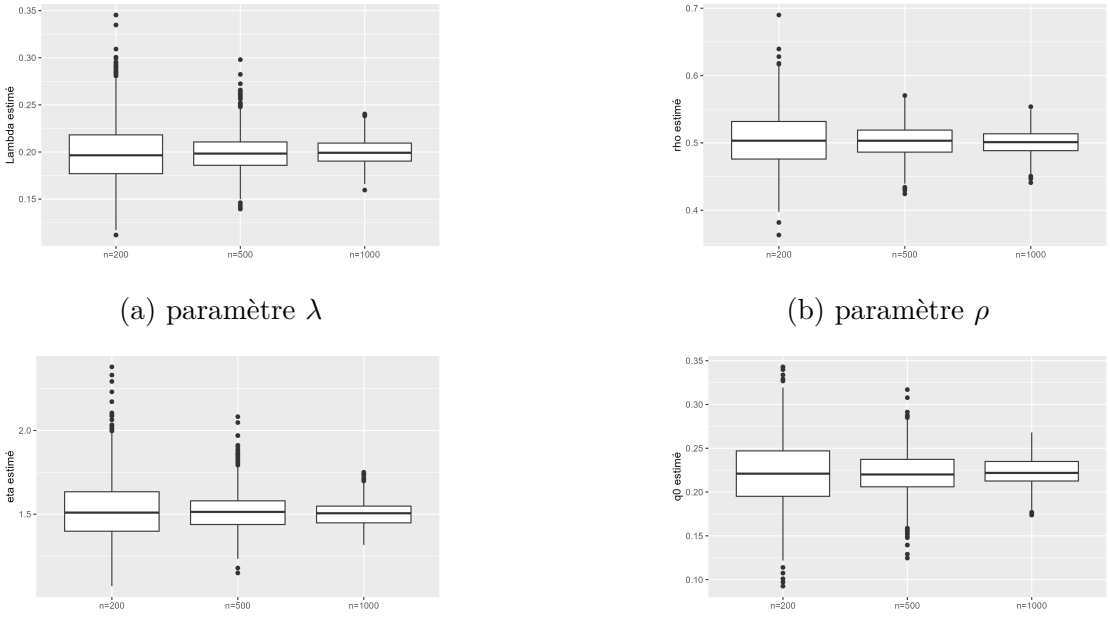
[15] Price D, Manatunga A. (2001). *Modelling survival data with a cured fraction using frailty models*. Statistics in Medicine, 20(9-10) :1515-1527.

[16] Van Keilegom I. (2021). *Analysis of survival and duration data*. Cours dispensé à l'Université UCLouvain.

[17] Wienke A, Ripatti S, Palmgren J, Yashin A. (2010). *A bivariate survival model with compound Poisson frailty*. Statistics in Medicine, 29(2) :275–283.

Annexe A

Frailty de Poisson



(a) paramètre λ

(b) paramètre ρ

(c) paramètre η

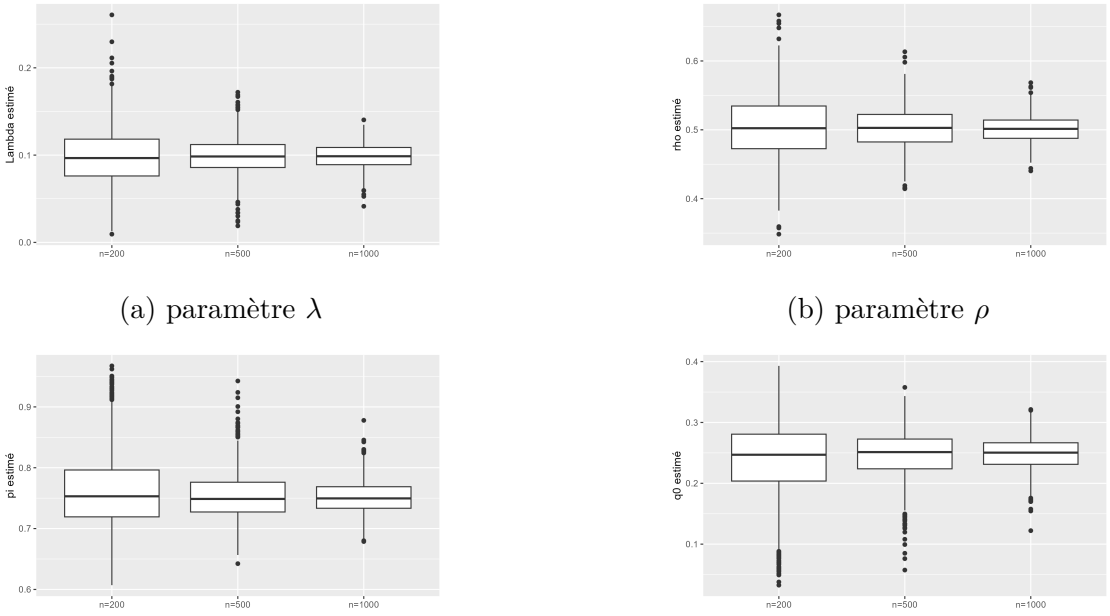
(d) paramètre q_0

FIGURE A.1 – Boxplot des estimations des différents paramètres du modèle de Poisson pour différentes tailles d'échantillons.

Paramètre	Moyenne	Biais	Variance	MSE
censure ajoutée 10%				
$\lambda = 0.2$	0.1995	-0.0005	0.0004	0.0004
$\rho = 0.5$	0.5024	0.0024	0.0006	0.0006
$\eta = 1.5$	1.5157	0.0157	0.0139	0.0142
$q_0 = 0.22$	0.2212	-0.0020	0.0006	0.0006
censure ajoutée 25%				
λ	0.1489	-0.0511	0.0044	0.0070
ρ	0.4869	-0.0131	0.0014	0.0016
η	2.7292	1.2292	5.4360	6.9415
q_0	0.1447	-0.0784	0.0085	0.01464

TABLE A.1 – Résultats des simulations pour le modèle de Poisson (3.6) avec différents taux de censure ajoutés

Frailty geométrique



(a) paramètre λ

(b) paramètre ρ

(c) paramètre π

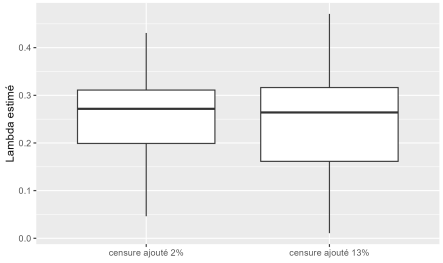
(d) paramètre q_0

FIGURE A.2 – Boxplot des estimations des différents paramètres du modèle géométrique pour différentes tailles d'échantillons.

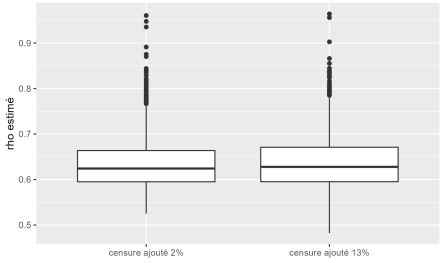
Paramètre	**Moyenne**	**Biais**	**Variance**	**MSE**
censure ajoutée 10%				
$\lambda = 0.1$	0.0984	-0.0016	0.0004	0.0005
$\rho = 0.5$	0.5022	0.0022	0.0009	0.0009
$\pi = 0.75$	0.7534	0.0034	0.0015	0.0015
$q_0 = 0.25$	0.2466	-0.0034	0.0015	0.0015
censure ajoutée 29%				
λ	0.0533	-0.04665	0.0021	0.0043
ρ	0.4950	$0. - 0.0050$	0.0011	0.0012
π	0.8583	0.1083	0.0085	0.0203
q_0	0.1416	-0.1083	0.0085	0.0203

TABLE A.2 – Résultats des simulations pour le modèle géométrique (3.9) avec différents taux de censure ajoutés

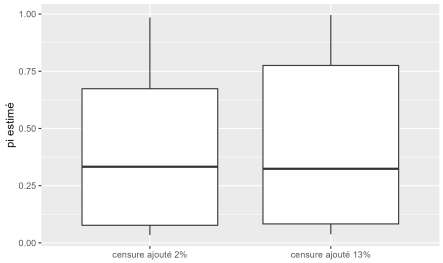
Frailty binomiale négative



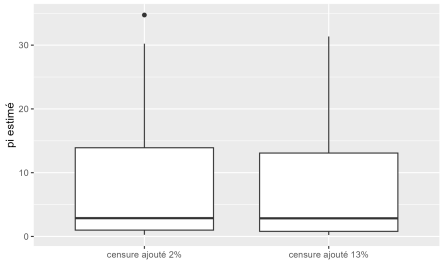
(a) paramètre λ



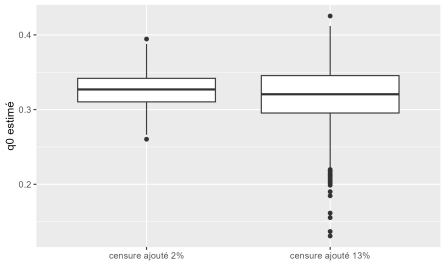
(b) paramètre ρ



(c) paramètre π



(d) paramètre ν



(e) paramètre q_0

FIGURE A.3 – Boxplot des estimations des différents paramètres du modèle géométrique pour différents taux de censure ajoutés.

Paramètre	**Moyenne**	**Biais**	**Variance**	**MSE**
		$n = 200$		
λ	0.2347	-0.0653	0.0102	0.0145
ρ	0.6681	0.0681	0.0099	0.0145
π	0.4323	0.2323	0.1345	0.1883
ν	9.3373	4.3373	111.4096	130.1102
q_0	0.3255	-0.0022	0.0012	0.0012
		$n = 500$		
λ	0.2534	-0.0466	0.0055	0.0076
ρ	0.6367	0.0367	0.0035	0.0076
π	0.3869	0.1869	0.0962	0.1311
ν	7.5995	1.5995	67.0042	73.6945
q_0	0.3269	-0.0008	0.0005	0.0005
		$n = 1000$		
λ	0.2632	-0.0368	0.0032	0.0046
ρ	0.6242	0.0242	0.0014	0.0046
π	0.3637	0.1637	0.0689	0.0956
ν	6.3150	1.3150	42.8202	44.5064
q_0	0.3280	0.0003	0.0002	0.0002

TABLE A.3 – Résultats des simulations pour le modèle binomial négatif avec différentes tailles d'échantillons

Annexe B

$1 - \psi = 0.6$

$1 - \psi = 0.6$	**Moyenne**	**Biais**	**Variance**	**MSE**	**Taux de convergence**
MCM	0.6007	0.0007	0.0007	0.0007	100%
Poisson	0.5989	−0.0011	0.0007	0.0007	97%
Curegeo	0.6000	0.0000	0.0007	0.0007	97%
Binomial Négatif	0.5994	−0.0006	0.0007	0.0007	97%

TABLE B.1 – Résultats des estimations de la proportion de cure pour $1 - \psi = 0.6$ avec une censure exponentielle pour les différents modèles

$1 - \psi = 0.4$

$1 - \psi = 0.4$	**Moyenne**	**Biais**	**Variance**	**MSE**	**Taux de convergence**
MCM	0.4007	0.0007	0.0006	0.0006	100%
Poisson	0.4019	0.0019	0.0007	0.0007	91%
Curegeo	0.4049	0.0049	0.0006	0.0007	81%
Binomial Négatif	0.4030	0.0030	0.0007	0.0007	84%

TABLE B.2 – Résultats des estimations de la proportion de cure pour $1 - \psi = 0.4$ avec une censure exponentielle pour les différents modèles

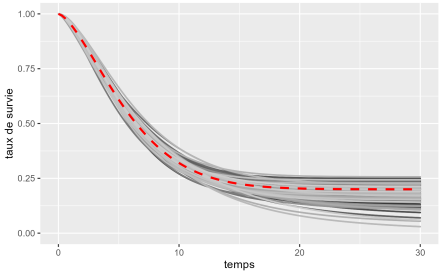
$1 - \psi = 0.2$

$1 - \psi = 0.2$	**Moyenne**	**Biais**	**Variance**	**MSE**	**Taux de convergence**
MCM	0.1993	−0.0007	0.0004	0.0004	99%
Poisson	0.2026	0.0026	0.0004	0.0004	80%
Curegeo	0.2140	0.0140	0.0005	0.0006	57%
Binomial Négatif	0.2071	0.0071	0.0004	0.0004	61%

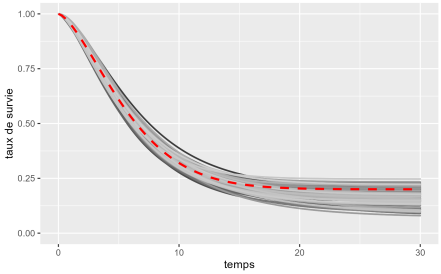
TABLE B.3 – Résultats des estimations de la proportion de cure pour $1 - \psi = 0.2$ avec une censure exponentielle pour les différents modèles

Annexe C

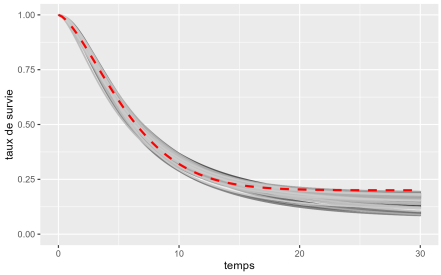
$1 - \psi = 0.2$



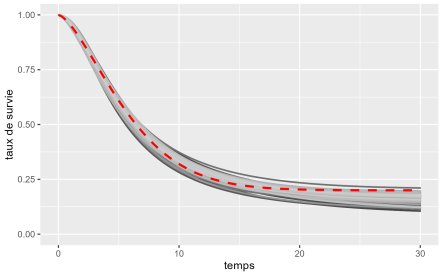
(a) ZMP



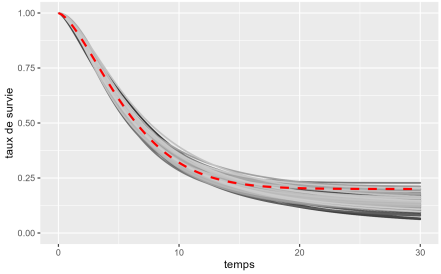
(b) ZMP*



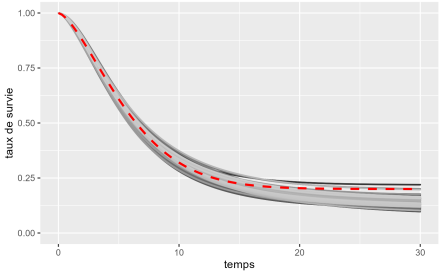
(c) ZMG



(d) ZMG*



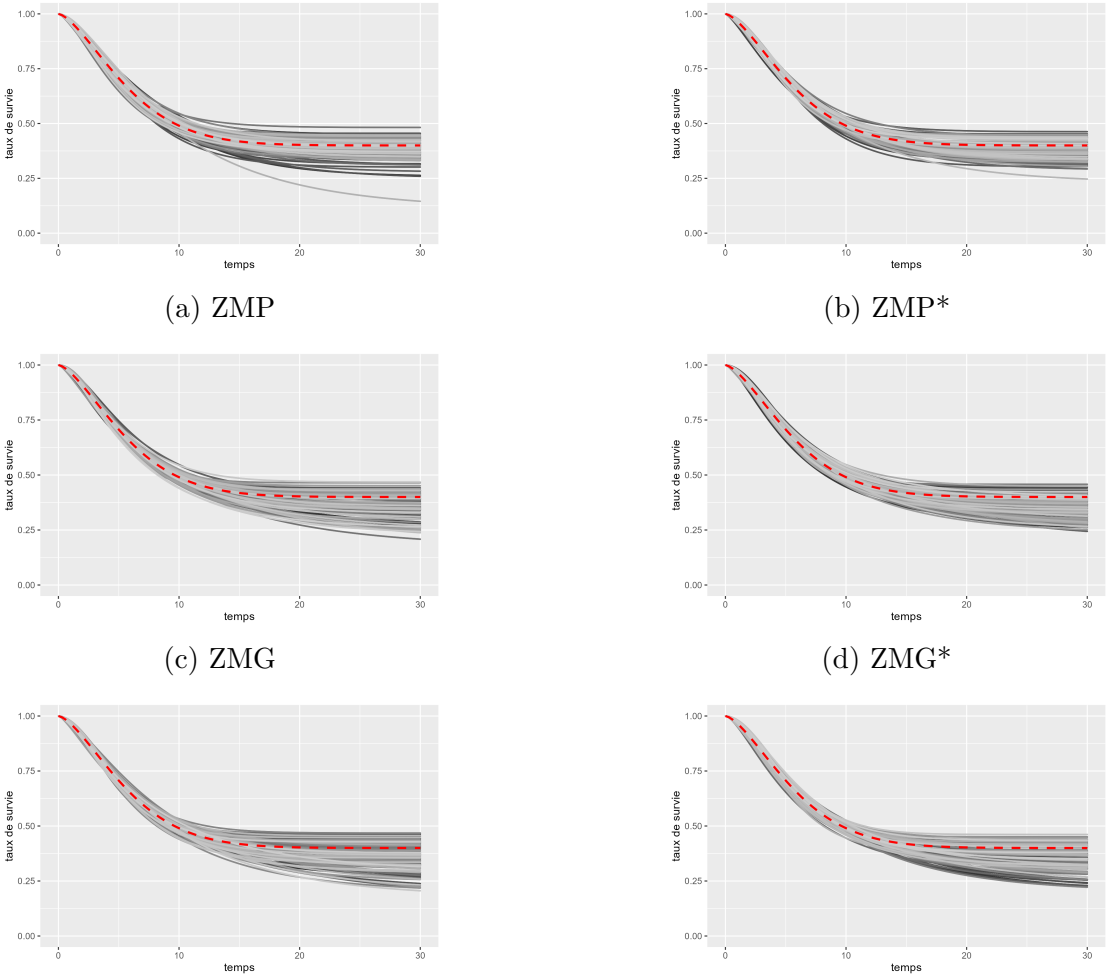
(e) ZMBN



(f) ZMBN*

FIGURE C.1 – Courbes de survie des modèles ZMP, ZMG, ZMBN, ZMP*, ZMG* et ZMBN* avec la proportion de cure fixée à $1 - \psi = 0.2$

$1 - \psi = 0.4$



(a) ZMP

(b) ZMP*

(c) ZMG

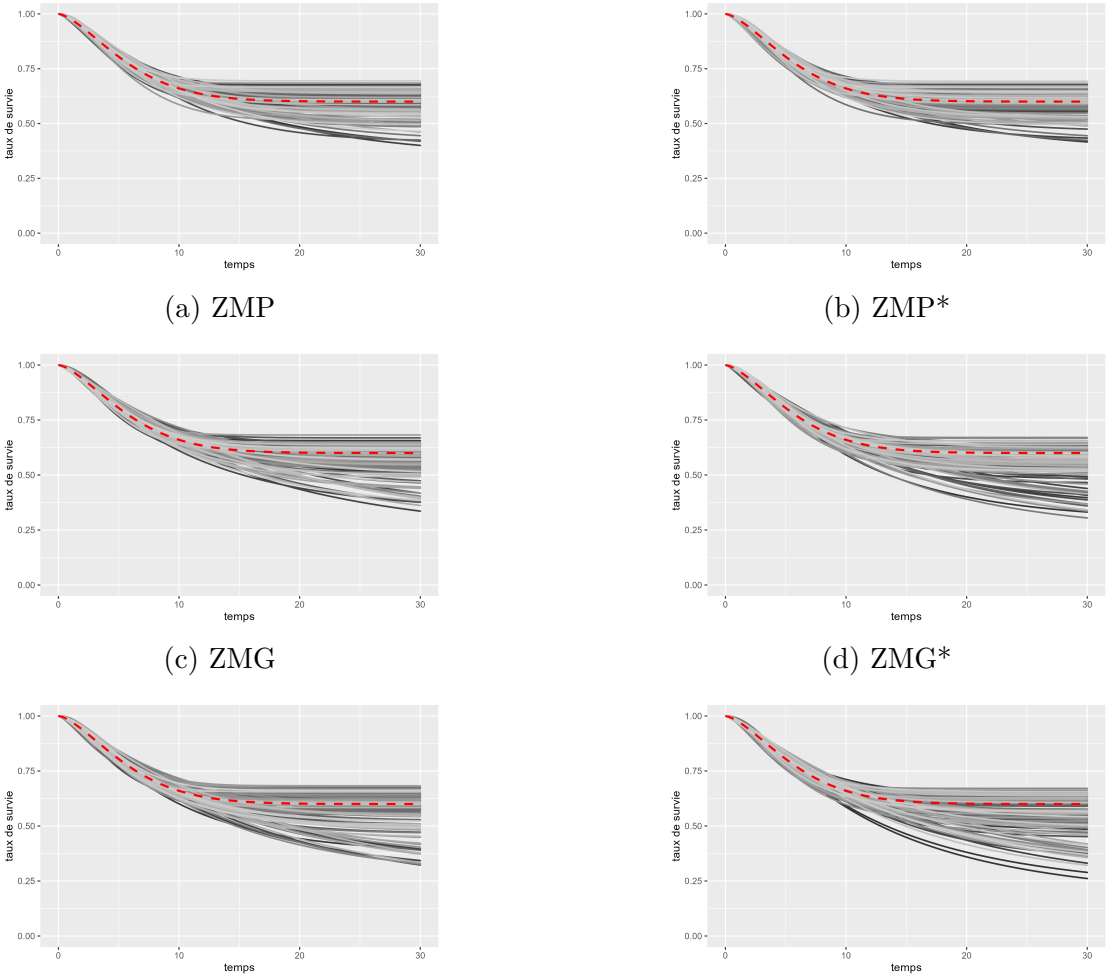
(d) ZMG*

(e) ZMBN

(f) ZMBN*

FIGURE C.2 – Courbes de survie des modèles ZMP, ZMG, ZMBN, ZMP*, ZMG* et ZMBN* avec la proportion de cure fixée à $1 - \psi = 0.4$

$1 - \psi = 0.6$



(a) ZMP

(b) ZMP*

(c) ZMG

(d) ZMG*

(e) ZMBN

(f) ZMBN*

FIGURE C.3 – Courbes de survie des modèles ZMP, ZMG, ZMBN, ZMP*, ZMG* et ZMBN* avec la proportion de cure fixée à $1 - \psi = 0.6$

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Faculté des sciences
Place des Sciences, 2 bte L6.06.01, 1348 Louvain-la-Neuve, Belgique | www.uclouvain.be/sc