Analyse de survie

Chapitre 1 : Introduction générale

Introduction

L’analyse de survie est la science d’analyser les données de durées temps-à-événement, depuis une origine de temps bien définie jusqu’à un point d’intérêt bien déterminé (événement). Par exemple, dans l’étude d’un type particulier de cancer, le temps de premier diagnostic peut être définie comme le temps d’origine, et le décès par ce même cancer peut définir l’évènement. Cependant, certaines études peuvent suivre des individues depuis la naissance (temps d’origine), jusqu’à la manifestation d’une maladie (l’événement). Les données temps à événement sont généralement collectées de manière prospective dans le temps, comme c’est le cas pour les données collectées pour une expérience clinique ou les données d'une étude de cohorte potentielle. Parfois, les données peuvent également être collectées rétrospectivement en accédant à des dossiers médicaux ou en interrogeant des patients atteints de cette maladie.

La mort est l'événement d'intérêt dans la plupart des études médicales. Mais dans de nombreux cas, le cancer par exemple, le temps entre une réponse au traitement et la réapparition de la maladie, ou le temps sans maladie, est une autre mesure essentielle. De plus, l'événement et la durée de l'observation sont importants à exprimer. Par exemple, intervalle de temps entre la réponse confirmée et la première rechute du cancer. Les données de type temps à événement peuvent inclure le temps de survie, la réponse à un traitement donné, les attributs du patient associés à la réponse, la survie et la croissance de la maladie.

Un problème particulier lié à l'analyse du temps à événement apparaît lorsque tous les individus n’expérimentent pas l'événement, de sorte que les temps de survie ne seront finalement pas connus pour une partie du groupe d'étude. Par exemple, les individus pourraient avoir différents événements, lorsque l'événement d'intérêt est le décès dû au cancer mais que le patient est décédé suite à un accident ou qu'il a abandonné l’étude. Une autre situation est que l'étude doit se terminer à un certain moment et que certains individus n'aient pas encore eu leur événement et que l'heure de leur événement n'ait donc pas été remarquée. C'est ce qu'on appelle la censure.

Ces observations incomplètes doivent être traitées de manière appropriée. C'est pourquoi des techniques "spéciales" sont nécessaires dans l'analyse du temps jusqu'à l'événement.

De plus, les données de temps à événement sont généralement biaisées et rarement distribuées normalement, par conséquent, des techniques simples établies sur une distribution normale ne seront souvent pas précises.

Concepts de base

Censure

Les données censurées sont très communes et nécessitent des traitements spéciaux. La censure résulte de différentes causes et se manifeste sous beaucoup de formes. Parmi elles la censure à gauche et la censure à droite.

Censure à droite

Dans l'analyse de survie, T dénote le temps depuis le départ de l'observation jusqu'à l'occurrence de l'évènement. Or dans plusieurs cas, les observations s'arrêtent avant l'apparition de celui-ci. Ceci est appelé la censure à droite. En conséquence, si T correspond à l'Age en années de l'individu au moment du décès, l'événement est censuré à droite à l'Age 60 si tout ce que nous pouvons savoir est que T>60. Ce concept ne concerne pas que le temps à événement. Le revenu, par exemple, est censuré à droite si tout ce que nous en savons est qu'il dépasse 15000 Dh par mois.

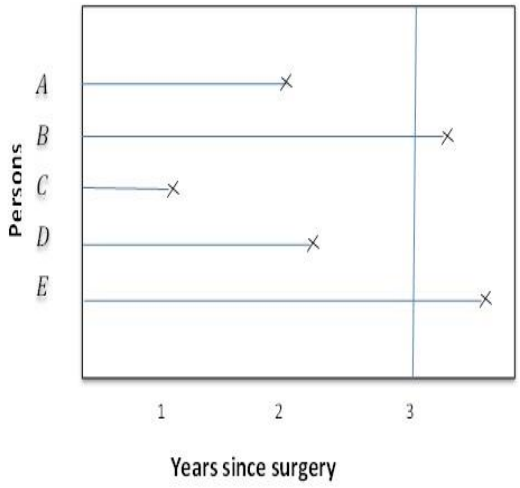

Figure 1: Image montrant la censure à droite

Censure à gauche

Il y à censure à gauche lorsque tout ce que nous pouvons savoir est que T est inférieure à une valeur donnée. Ceci aussi n'est pas réservé que pour le temps à événement. Pour l'analyse de survie, la censure à gauche apparait souvent lorsqu'un individu expérimente l'événement avant le début des observations.

Un exemple de censure à gauche ; une étude de ménarche (le début de la menstruation), si nous prenons un échantillant de filles de 12 ans et que certaines ont déjà commencé à avoir leurs règles sans pouvoir en déterminer le temps de la première apparition, l'Age de la ménarche chez elles est censuré à 12 ans.

Censure bornée

La censure bornée est plus commune. Elle apparait lorsque nous avons à la fois une censure à gauche et à droite. Tout ce que nous savons est que la variable T est comprise entre deux valeurs. Dans l'analyse de survie, la censure bornée (interval censoring) apparait lorsque les observations sont collectées selon des intervalles de temps donnés, et que le temps d'occurrence de l'événement ne peut pas être déterminé avec plus de précision.

Par exemple, pour un échantillant d'individus suivis pour infection de AIDS, des analyses sont réalisées chaque trimestre depuis l'admission. Si un individu s'avère séro-positif lors du troisième test, son temps à événement est censuré par intervalle entre la valeur 3 et 6 (en mois).

Troncature

La troncature est un autre facteur qui affecte les données de survie quand il y a lieu d'observation incomplètes. Un intervalle de temps durant lequel le sujet n'a pas été suivit mais n'a pas eu l'événement non plus est appelé une troncature. Il y a trois types différents de troncature, celle à gauche en est la plus commune.

Troncature à gauche

La période d'ignorance dans la troncature à gauche s'étale à partir du, ou avant le, début de l'étude (à t=0) jusqu'à quelque temps après le temps t=0. La Figure 2.2 explique la troncature à gauche.

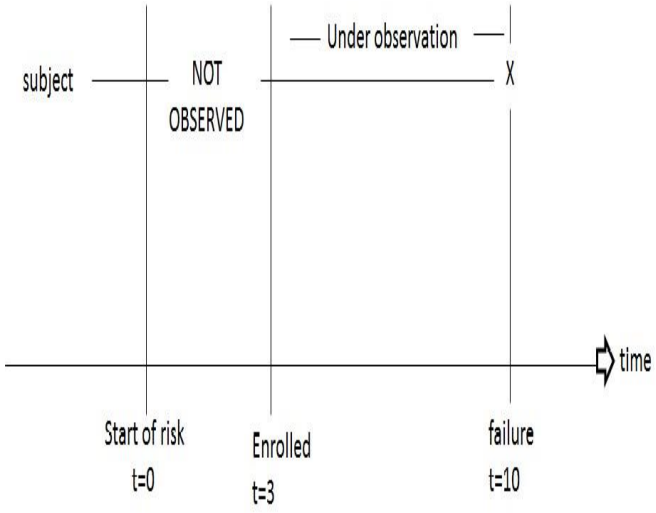

Figure 2: Représentation d'une troncature à gauche

Le sujet n'est pas observé durant une période après le départ des observations, ensuite il s'enrôle après s'il n'a pas encore vécu l'événement. C'est ce qui rend la troncature à gauche plus commune vue que certains individus rejoignent l'étude après le temps origine.

Par exemple, seules les personnes qui survivent au stade initial de l'infarctus du myocarde et arrivent à l'hôpital seront incluses dans l'étude. Si un individu a été admis à l'hôpital il est ajouté à l'étude avec comme temps d'origine le temps de l'infarctus. Pour différents patients, cela peut se produire à des moments différents, mais ces patients ne seront jamais admis à l'étude s'ils meurent avant d'arriver à l'hôpital. La saisie différée est parfois utilisée pour les données tronquées à gauche.

Troncature d'intervalle :

La troncature d'intervalle n'est qu'une adoption de la troncature à gauche où un individu entre dans l'étude au temps zéro mais disparaît pendant un certain temps puis revient à l'étude en générant un écart entre les observations. Le problème est que cette personne aurait pu mourir lorsqu'elle a disparu.

Troncature à droite :

La troncature à droite se produit quand l'information concernant un sujet n'est obtenue que lorsqu'il vit l'événement. Les sujets qui survivent après la fin de l'étude ne sont pas diagnostiqués et ne sont donc pas inclus dans l'échantillon de l'étude, ce qui donne un échantillon biaisé en faveur des sujets avec des temps de survie plus courts.

Fonctions de temps de survie :

Avant d'analyser les données de survie, certaines fonctions doivent être décrites telles que la fonction de survie, la fonction de densité, la fonction de risque et la fonction de risque cumulé parmi lesquelles

les fonctions de survie et de risque présentent un intérêt particulier. La fonction de densité et la fonction de distribution cumulatives sont utilisées dans les modèles statistiques traditionnels, mais en raison des observations incomplètes dans les données de survie (données censurées et tronquées), ces fonctions standard ne sont pas appropriées. Par conséquent, les fonctions de survie et de risque sont considérées comme plus appropriées.

Fonction de survie.

La fonction de survie est définie comme étant la probabilité de survivre au-delà d'un temps spécifié t. La fonction de survie est notée S(t) où $0 < t < \infty$. Sa formule est donnée comme suivant :

$$S(t) = P(T \geq t) = 1 - F(t), t > 0 \quad (2.1)$$

Où T est la variable aléatoire à étudier (temps à évènement) t est un nombre fixé et F(t) est la fonction de distribution cumulée de T. S(t)= 1 à t=0 et S(t) = 0 à $t=\infty$. Le graphe de la fonction de survie S(t) est appelé courbe de survie. Elle commence à S(t)=1 et comme t tend vers $+\infty$, S(t) décroit vers 0. La courbe de survie peut être estimée par la méthode de Kaplan-Meier.

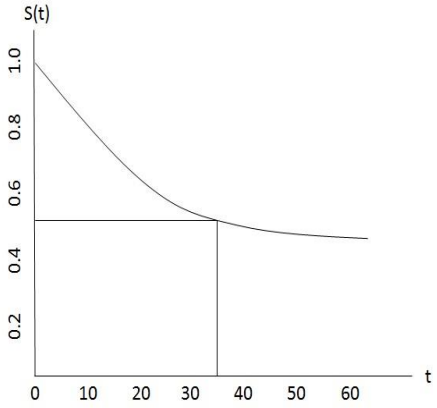

Figure 3: Un exemple de courbe de survie

Fonction de densité.

La fonction de densité de probabilité $f(t)$ est définie comme le taux d'occurrence de l'événement pour chaque unité de temps. On peut calculer la fonction de densité en dérivant la fonction de survie, comme suivant :

$$\frac{d}{dt}S(t) = \frac{d}{dt}(1 - F(t))$$

De la définition de la fonction de distribution, on a :

$$\frac{d}{dt}S(t) = -f(t)$$

$$f(t) = -\frac{d}{dt}S(t) \quad (2.2)$$

L'équation (2.2) montre la relation entre la fonction de densité de probabilité et la fonction de survie.

La fonction de densité de probabilité, également connue sous le nom de taux d'échec inconditionnel est intuitivement définie comme :

$$P(t \leq T < t + \Delta t) \approx \Delta t.f(t)$$

L'équation (2.3) est la définition mathématique traditionnelle de la fonction de densité de probabilité en tant que limite.

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t}, t > 0 \quad (2.3)$$

La définition décrite par la formule de l'équation (2.3) est bien illustrée par la figure 4, qui montre que la probabilité qu'une observation soit dans l'intervalle (t, t + Δt) est assez approximée par l'aire du rectangle dont les côtés ont une longueur Δt et f(t) [9].

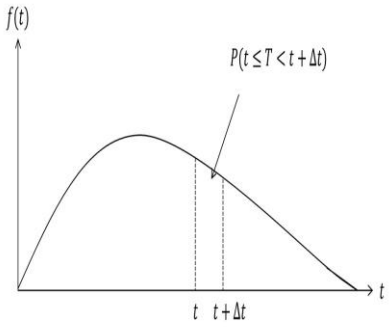

Figure 4: Courbe de fonction de densité de probabilité.

Fonction de risque.

Pour comprendre l'analyse de survie, la fonction de risque est un concept important qu'on peut définir comme une sorte de fonction de densité f(t). La différence est que la fonction de risque est conditionnelle tandis que la fonction de densité est une probabilité inconditionnelle.

La fonction de risque, également connue sous le nom de taux instantané d'échec, est définie comme la probabilité que l'événement est observé dans un intervalle [t, t + Δt[, étant donné qu'il ne s'est pas produit avant t.

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t}, T \geq t$$

$$h(t)\Delta t \approx P(t \leq T < t + \Delta t | T \geq t) \quad (2.4)$$

L'équation (2.4) explique que la fonction de hasard pour un temps t est la probabilité que la personne observée meure juste après le temps t ([t, t + Δt[) sachant que l'individu a déjà survécu jusqu'au temps t. L'interprétation graphique de la fonction de risque peut être vue en figure 5

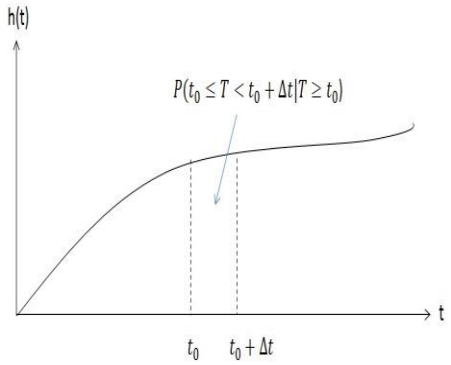

Figure 5: Courbe d'une fonction de risque

Fonction de risque Cumulative.

Le risque cumulé intègre le taux de risque (instantané) par rapport au temps. C'est comme si on additionnait des probabilités, mais comme Δt est très petit, ces probabilités ont aussi des petites valeurs. En calculant l'intégrale de la fonction de risque, nous obtenons la fonction de risque cumulé qui estime plus facilement les modèles non paramétriques comparée aux fonctions de risque et de densité. C'est pourquoi elle est considérée comme une fonction importante [9].

$$H(t) = \int_0^t h(x)dx \, , t \geq 0 \quad (2.5)$$