ANALYSE DE SURVIE

KPAM

2023-04-28

Sommaire

1. DEFINITION

L'analyse de survie peut être définie comme les méthodologies utilisées pour explorer le temps nécessaire pour qu'une occasion ou un événement se produise. Un modèle de régression normal peut échouer dans l'analyse de la prédiction précise, car le « temps écoulé avant l'événement » n'est généralement pas distribué normalement et rencontre des problèmes de gestion de la censure qui peuvent modifier le résultat prévu.
L'idée de base que l'on retient est qu'elle représente principalement les événements négatifs de sa vie ou de son scénario. Par exemple, prédire la mort d'une personne, une rechute dans l'état de santé d'une personne, un taux de désabonnement d'un employé dans une organisation ou une panne de machine. Cependant, cette méthodologie peut également être utilisée pour prédire les événements positifs dans la vie des sujets, tels que l'obtention d'un emploi après l'obtention du diplôme, le mariage, l'achat d'une maison ou d'un nouveau produit comme une voiture.

Objectifs à atteindre

-Visualiser les courbes de survie : Graphiques de Kaplan-Meier
-comparer les courbes de survie de deux groupes ou plus : Test du log-rank
-décrire l'effet des variables sur la survie : Régression des risques proportionnels de Cox.

Concepts de base

Définissons quelques termes fondamentaux de l'analyse de survie tels que :

- Temps de survie et évènement
- Censure
- Fonction de survie et fonction de risque

Le temps de survie est le temps entre « la réponse au traitement » et l'apparition de l'évènement d'intérêt.

Une caractéristique clé des données de survie est la censure.
L'analyse de survie se concentre sur la durée prévue jusqu'à la survenue d'un événement d'intérêt (rechute ou décès). Cependant, l'événement peut ne pas être observé pour certains individus au cours de la période d'étude, produisant les soi-disant observations censurées.

Un sujet peut être censuré en raison de :

- Perte de suivi
- Abandon des études
- Aucun événement à la fin de la période d'études déterminée

Différents types de censure:

-censure à droite : Quand l'évènement d'intérêt n'est pas toujours observé
-censure à gauche : on ne connaît pas toujours la date exacte d'entrée dans l'étude
-censure par intervalle : on ne connaît qu'un intervalle de temps par individu et on sait que l'évènement d'intérêt s'est produit dans cet intervalle.

2.FONCTION DE SURVIE ET FONCTION DE RISQUE

Deux probabilités liées sont utilisées pour décrire les données de survie : la probabilité de survie et la probabilité de risque.

- La probabilité de survie, également connue sous le nom de fonction de survie S (t), est la probabilité qu'un individu survive depuis l'origine du temps (par exemple le diagnostic de cancer) jusqu'à un temps futur spécifié t.
- La probabilité qu'un sujet survive au-delà d'un temps donné.

$S(t) = P(T > t) = 1 - F(t)$

$S(t)$: fonction de survie : fonction de distribution cumulative

$F(t) = P(T <= t)$

- Le danger, désigné par $h(t)$, est la probabilité qu'un individu observé à un instant t subisse un événement à cet instant.

3.1. Méthode de KAPLAN-MEIER

La méthode de Kaplan-Meier (KM) est une méthode non paramétrique utilisée pour estimer la probabilité de survie à partir des durées de survie observées.

Hypothèses :

Une analyse KM est valide dans les six conditions suivantes :

- Résultat binaire. Il n'y a que deux états de résultat (par exemple, mort ou vivant).
- Temps de survie précis. Le temps de survie est enregistré sous la forme d'un nombre et non d'un intervalle.
- Censure minimale à gauche. Les points de départ inconnus sont minimes. S'applique lorsque le point de départ de l'expérience n'est pas bien défini. Par exemple, pour une maladie, la date du diagnostic est préférable à l'apparition des symptômes.

- Censure non informative. Les causes de censure sont indépendantes de l'événement. Les sujets n'abandonnent pas l'étude à cause de quelque chose lié à leur groupe. Par exemple, un sujet n'abandonne pas une étude thérapeutique parce que la thérapie aggrave son état.

- Pas d'effets de cohorte. Il n'y a pas de tendances séculaires. Des heures de démarrage échelonnées peuvent englober l'introduction de nouvelles thérapies qui affectent la survie.
- Modèles de censure similaires. La quantité et le modèle de censure devraient être similaires.

La probabilité de survie au moment $t_i, S(t_i)$, est calculé comme suit :
$S(t_i) = S(t_i - 1)(1 - d_i/n_i)$
Où:
- $S(t_i - 1)$:la probabilité d'être en vie à $t_i - 1$
- n_i :le nombre de patients vivants juste avant t_i
- d_i : le nombre d'événements à t_i
- $t_0 = 0, S(0) = 1$

3.2. Modèle de COX

le modèle de régression de Cox étend les méthodes d'analyse de survie pour évaluer simultanément l'effet de plusieurs facteurs de risque sur le temps de survie.
Le modèle de Cox est exprimé par la fonction de danger notée h(t). En bref, la fonction de danger peut être interprétée comme le risque de mourir à l'instant t. Il peut être estimé comme suit:
$h(t) = h_0(t) * exp(\beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p)$
où:

- t représente le temps de survie
- h(t) est la fonction de danger déterminée par un ensemble de p covariables $(X_1, X_2, ..., X_p)$
- les coefficients $(\beta_1, \beta_2, ..., \beta_p)$ mesurent l'impact (c.-à-d. l'ampleur de l'effet) des covariables.

- Le terme h_0 est appelé le danger de base. Elle correspond à la valeur du danger si tous les X_i sont égaux à zéro. Le « t » dans $h(t)$ nous rappelle que le danger peut varier au fil du temps.
- Les quantités $exp(\beta_i)$ sont appelés hazard ratios (HR) **NB**:
- $HR = 1$: Aucun effet
- $HR < 1$: Réduction du danger
- $HR > 1$: Augmentation du danger

Une hypothèse clé du modèle de Cox est que les courbes de risque pour les groupes d'observations (ou de patients) doivent être proportionnelles et ne peuvent pas se croise.

4. MODELES DE DEFAILLANCE ACCELERE

Un modèle de temps de défaillance accéléré est un modèle paramétrique dont les covariables et les temps de défaillance suivent une fonction de survie de la forme :
$S(x/Z) = S_0(x * exp[\beta * Z])$ où S_o est une fonction pour le taux de survie initial et le terme $exp[\beta * Z]$ est le facteur accélérateur. Ce modèle peut être réécrit sous forme log-linéaire du temps de défaillance (logX) qui est linéairement lié à la moyenne μ, au facteur d'accélération $\beta * Z$ et au terme d'erreur $\theta * w$; $logX = \mu - \beta * Z + \theta * w$.

Sommaire

II- PRATIQUE AVEC R

Nous allons utiliser deux packages R :

- `survival` pour le calcul d'analyses de survie
- `survminer` pour résumer et visualiser les résultats de l'analyse de survie

```
library("survival")
library("survminer")
```

Nous utiliserons les données sur le cancer du poumon disponibles dans le package survival.

1. Préparation des données

```r
data(cancer, package="survival")
head(lung)
#View(lung)
```

2. calcul du temps de survie

Dans le cas où nous avons des données avec de dates de debut et de fin ,la prémière etape consiste à s'assurer qu'elles sont du type date. La durée de survie est égale à la différence entre la date sortie et la date d'entrée

3. Création d'un objet de survie

Une fonction clé pour l'analyse des données de survie dans R est la fonction `Surv()`. Ceci est utilisé pour spécifier le type de données de survie que nous avons, à savoir, censuré à droite, censuré à gauche, censuré par intervalle.

$Surv(time, event)$, $Surv(time, time2, event, type)$

La `Surv()`fonction du package `survival` crée un objet de survie à utiliser comme réponse dans une formule modèle.Pour définir notre objet de survie, il nous faudra deux variables. Une première, temporelle, indiquant la durée à laquelle survient l'évènement étudié pour ceux ayant vécu l'évènement et la durée d'observation pour ceux n'ayant pas vécu l'évènement (censure à droite). Par ailleurs, une seconde variable indiquant si les individus ont vécu l'évènement.Il y aura une entrée pour chaque sujet qui est le temps de survie, qui est suivi d'un $(+)$si le sujet a été censuré.

4. ESTIMATION DE KAPLAN-MEIER

La fonction `survfit()` dans le package de `survival` peut être utilisée pour calculer l'estimation de survie de Kaplan-Meier. Ses principaux arguments incluent :

- un objet de survie créé à l'aide de la fonction `Surv()`
- et l'ensemble de données contenant les variables.

4.1. Fonction de survie

Son premier argument est `formula`. Le côté gauche de cette formule spécifie les informations sur les temps de survie à l'aide de la fonction Surv(), et le côté droit est utilisé pour spécifier les variables de regroupement. Argument `data` spécifie le bloc de données qui contient les variables d'intérêt(dans notre cas `lung`).

```
fit <- survfit(Surv(time, status) ~ sex,
               data = lung)
```

4.1.1 MEDIAN ET MOYENNE ESTIMES

Les temps de survie médians pour chaque groupe représentent le moment auquel la probabilité de survie, S(t), est de 0,5.

```
print(survfit(Surv(time, status) ~ sex, data = lung),
      print.rmean = TRUE)
```

4.1.2 QUANTILE

On peut utiliser la méthode `quantile()` pour calculer les temps de suivi correspondants auxquels la probabilité de survie prend une valeur spécifique.

4.2 COURBE DE SURVIE

Nous utiliserons la fonction ‘ggsurvplot()‘dans le package `Survminer` R pour produire les courbes de survie des deux groupes de sujets.

4.3 Fonctions de risques cumulatif

La fonction de risque cumulatif et la fonction de survie sont liées par la relation suivante :
$S(t) = exp(-H(t))$. Il correspond au nombre d'événements qui seraient attendus pour chaque individu au temps t si l'événement était un processus répétable.

-Lorsque les deux courbes sont proportionnelles l'une à l'autre (c'est-à-dire qu'elles s'éloignent régulièrement l'une de l'autre) on dit qu'il y a une différence de survie significative entre les deux groupes.

5.1 Test de 'log-rank'

Le test du log-rank est la méthode la plus largement utilisée pour comparer deux ou plusieurs courbes de survie (c'est-à-dire pour tester l'hypothèse si les fonctions de survie de différents groupes de sujets diffèrent de manière statistiquement significative). L'hypothèse nulle est qu'il n'y a pas de différence de survie entre les deux groupes.
Le test du log-rank est un test non paramétrique, qui ne fait aucune hypothèse sur les distributions de survie.

La fonction $survdiff()$ dans le package de `survival` peut être utilisée pour calculer le test log-rank comparant deux ou plusieurs courbes de survie. Pour tester avec le test du log-rank s'il existe des différences dans les taux de survie dans l'ensemble de données Lung entre les hommes et les femmes, nous utilisons le code :

```
surv_diff <- survdiff(Surv(time, status) ~ sex,
                      data = lung)
```

Pour evaluer la même hypothèse on peut utiliser le test de `Peto & Peto Gehan-Wilcoxon`, nous utilisons `survdiff()`à nouveau la fonction, mais maintenant nous définissons l'argument `rho` à 1 :

5.2 Modèle de cox

La fonction $coxph()$dans le package de `survival` peut être utilisée pour calculer le modèle de régression à risques proportionnels de Cox dans R. $coxph(formula, data, method)$ Nous ajusterons la régression de Cox en utilisant les covariables suivantes : âge, sexe, ph.ecog et wt.loss.

-*Régression de Cox univariée*

-*REGRESSION DE COX MULTIVARIEE*

6. MODELES DE DEFAILLANCE ACCELERE

La fonction qui correspond aux modèles AFT(Accelerated Failure Times) du package de survie est `survreg()`. Son premier argument est une formule et a une syntaxe similaire à la fonction `survfit()`. L' argument `dist` spécifie la distribution des temps de survie (Remarque : l' argument `dist` spécifie la distribution des temps de survie et non les temps de survie du journal). La distribution par défaut (c'est-à-dire si vous ne spécifiez pas l' argument `dist` vous-même) est la distribution de **Weibull**. Comme pour les autres fonctions d'ajustement de modèle dans R, la `summary()` fonction renvoie une sortie détaillée du modèle ajusté.

Voici le code pour la distribution de **Weibull** :

```
fit_weibull <- survreg(Surv(time, status) ~sex + age+ph.ecog,
                       data =  lung)
```

Pour ajuster le même modèle mais avec la distribution exponentielle, il faut préciser `dist = "exponential"` On distingue également les distributions log-normale, log-logistic..

QUESTION FONDAMENTALE :

si on souhaite modéliser avec une structure plus complexe devrions nous choisir un modèle paramétrique ajusté avec la fonction survreg ou un modèle de cox ajusté avec coxph. Si nous souhaitons utiliser le modèle pour la prédiction nous devons utiliser la fonction survreg car coxph n'extrapole pas au-delà de la dernière observation.
«De combien le risque de décès diminue-t-il si un nouveau traitement médical est administré à un patient ?»:utilisation de `Coxph`
«Quelle proportion de patients mourront dans 2 ans d'après les résultats des données d'une expérience qui n'a duré que 4 mois» utilisation de `Survreg`

Abstract

L'analyse de survie est un ensemble d'approches statistiques pour l'analyse des données où la variable de résultat d'intérêt est le temps jusqu'à ce qu'un événement se produise. Les données de survie sont généralement décrites et modélisées en termes de deux fonctions liées :

- la fonction de survie représentant la probabilité qu'un individu survive depuis le temps d'origine jusqu'à un certain temps au-delà du temps t. Il est généralement estimé par la méthode de Kaplan-Meier. Le test du logrank peut être utilisé pour tester les différences entre les courbes de survie des groupes, tels que les bras de traitement.

- La fonction de risque donne le potentiel instantané d'avoir un événement à un moment donné, compte tenu de la survie jusqu'à ce moment. Il est principalement utilisé comme outil de diagnostic ou pour spécifier un modèle mathématique pour l'analyse de la survie.

Abstract(suite)

-En suite, nous avons décrit le modèle de régression de Cox pour évaluer simultanément la relation entre plusieurs facteurs de risque et la durée de survie du patient. Nous avons montré comment calculer le modèle de Cox en utilisant le package de survie . De plus, nous avons décrit comment visualiser les résultats de l'analyse à l'aide du package survminer .

Sommaire

III. Recommandations et références bibliographiques

"*Rendre à César ce qui appartient à César*"
Cet document est inspiré des souces suivantes:

1-un article sur [**Survival analysis in R companion**] (https://www.drizopoulos.com/courses/emc/basic_surivival_analysis_in_r)

2- [**Survival analysis basic**]
(http://www.sthda.com/english/wiki/survival-analysis-basics)

3- Le livre de Dirk F. Moore ,**Applied survival Analysis Using R**
4- Le livre de David M. Diez ,**Survival Analysis in R**
5- Le livre de Michael J.Crawley ,**The R Book**