AGENCE NATIONALE DE LA STATISTIQUE ET DE LA DEMOGRAPHIE

ECOLE NATIONALE DE LA STATISTIQUE ET DE L'ANALYSE ECONOMIQUE PIERRE NDIAYE

EXPOSE SUR: ANALYSE DE SURVIE AVEC R

Par

KPAKOU M'Mounéné ISEP2

Sous la supervision de:

Mr. HADY DIALLO

Ingénieur des Travaux Statistiques

Contents

I- EXPLICATION DE L'ANALYSE DE SURVIE

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. ESTIMATION DE LA FONCTION DE SURVIE

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 $\theta * w$;

$logX = \mu - \beta * Z + \theta * w.$

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

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

	inst	time	status	age	sex	ph.ecog	ph.karno	pat.karno	meal.cal	wt.loss
1	3	306	2	74	1	1	90	100	1175	NA
2	3	455	2	68	1	0	90	90	1225	15
3	3	1010	1	56	1	0	90	90	NA	15
4	5	210	2	57	1	1	90	60	1150	11
5	1	883	2	60	1	0	100	90	NA	0
6	12	1022	1	74	1	1	50	80	513	0

```
#View(lung)
```

QUELQUES DETAILS SUR LES VARIABLES DE LA BASE

• inst : code de l'établissement

• time : Temps de survie en jours

• status : statut de censure 1=censuré, 2=mort

• age : âge en années

• sex : Masculin=1 Féminin=2

• ph.ecog : score de performance ECOG (0=bon 5=mort)

• ph.karno : score de performance de Karnofsky (mauvais=0-bon=100) noté par un médecin

• pat.karno : score de performance de Karnofsky évalué par le patient

• repas.cal : Calories consommées aux repas

• wt.loss : Perte de poids au cours des six derniers mois

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

```
library("tibble")
date_ex <-
  tibble(
    sx_date = c("2007-06-22", "2004-02-13", "2010-10-27"),
    last_fup_date = c("2017-04-15", "2018-07-04", "2016-10-31")
  )

date_ex
```

Calcul de la durée de survie

```
date_ex <-
  date_ex %>%
  mutate(
    os_yrs = as.duration(sx_date %--% last_fup_date) / dyears(1)
  )

date_ex
```

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.

```
library("survival")
library("survminer")
surv_objet<-Surv(lung$time, lung$status)[1:10]
surv_objet
```

```
 [1]  306   455  1010+  210   883  1022+  310   361   218   166
```

Nous voyons que le sujet 1 a eu un événement au temps 306 jours, le sujet 2 a eu un événement au temps 455 jours, le sujet 3 a été censuré au temps 1010 jours.

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)


print(fit)
```

```
Call: survfit(formula = Surv(time, status) ~ sex, data = lung)
```

```
        n events median 0.95LCL 0.95UCL
sex=1 138    112    270     212     310
sex=2  90     53    426     348     550
```

4.1.1 MEDIAN ET MOYENNE ESTIMES

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

```
Call: survfit(formula = Surv(time, status) ~ sex, data = lung)

        n events rmean* se(rmean) median 0.95LCL 0.95UCL
sex=1 138    112    326      22.9    270     212     310
sex=2  90     53    461      34.7    426     348     550
    * restricted mean with upper limit =  1022
```

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

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. Par exemple, nous voulons trouver à combien de jours la probabilité de survie est égale à 0,7 et à combien de jours elle est égale à 0,6 ;le code est :

```
quantile(fit, probs = 1 - c(0.7, 0.6))
```

NB Dans l'argument `probs` de `quantile()`nous devons spécifier un moins nos probabilités de survie cibles ; en effet, la fonction fonctionne selon la convention de la fonction de distribution cumulative (CDF) et la CDF est égale à un moins la probabilité de survie. De plus, notez également que nous obtenons les bornes inférieure et supérieure des intervalles de confiance à 95% pour les quantiles que nous avons demandés.

differentes composantes accessibles

```
d <- data.frame(time = fit$time,
                n.risk = fit$n.risk,
                n.event = fit$n.event,
                n.censor = fit$n.censor,
                surv = fit$surv,
                upper = fit$upper,
                lower = fit$lower
)
head(d)
```

	time	n.risk	n.event	n.censor	surv	upper	lower
1	11	138	3	0	0.9782609	1.0000000	0.9542301
2	12	135	1	0	0.9710145	0.9994124	0.9434235
3	13	134	2	0	0.9565217	0.9911586	0.9230952
4	15	132	1	0	0.9492754	0.9866017	0.9133612
5	26	131	1	0	0.9420290	0.9818365	0.9038355
6	30	130	1	0	0.9347826	0.9768989	0.8944820

```
summary(fit)$time
summary(fit)$surv  #estimation de kaplan
str(summary(fit))  #plus de détails sur le contenu de la liste
```

En utilisant la méthode `summary()` et son argument `times`, nous pouvons obtenir les probabilités de survie à des moments de suivi spécifiques.

• n : nombre total de sujets dans chaque courbe.

• time : les points temporels sur la courbe.

• n.risk : le nombre de sujets à risque au temps t

• n.event : le nombre d'événements qui se sont produits à l'instant t.

• n.censor : le nombre de sujets censurés, qui sortent de l'ensemble de risques, sans événement, à l'instant t.

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.

```
ggsurvplot(fit,
           pval = TRUE, conf.int = TRUE,
           risk.table = TRUE, # risk table
           risk.table.col = "strata", # Change risk table color by groups
           linetype = "strata", # Change line type by groups
           surv.median.line = "hv", # Specify median survival
           ggtheme = theme_bw(), # Change ggplot2 theme
           palette = c("#E7B800", "#2E9FDF"))
```

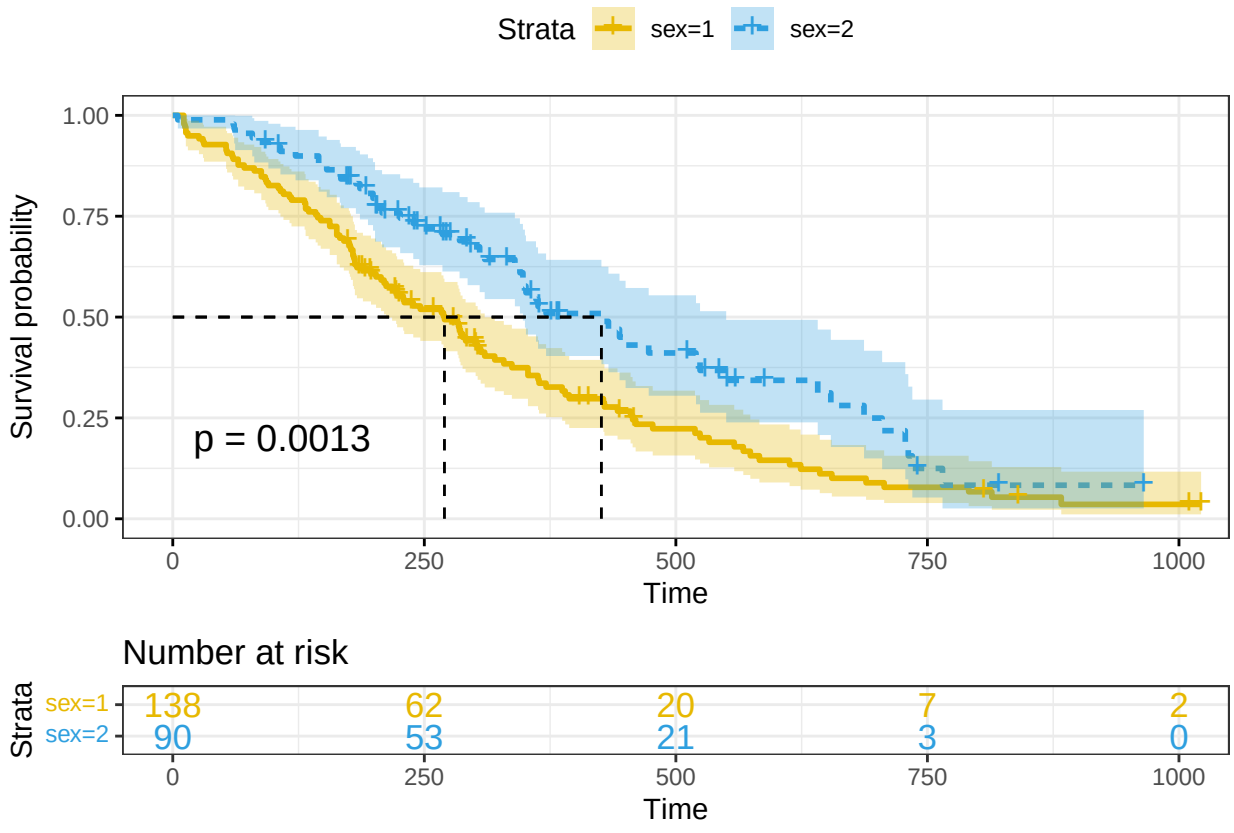

Figure 1: Courbes de survie

INTERPRETATION DU GRAPHIQUE

L'axe horizontal (axe des x) représente le temps en jours et l'axe vertical (axe des y) montre

la probabilité de survie ou la proportion de personnes survivantes. Les lignes représentent les courbes de survie des deux groupes. Une chute verticale dans les courbes indique un événement. La coche verticale sur les courbes signifie qu'un patient a été censuré à ce moment.

• Au temps zéro, la probabilité de survie est de 1,0 (ou 100 % des participants sont en vie).

• Au temps 250, la probabilité de survie est d'environ 0,55 (ou 55%) pour le sexe=1 et de 0,75 (ou 75%) pour le sexe=2.

• La survie médiane est d'environ 270 jours pour le sexe=1 et de 426 jours pour le sexe=2, suggérant une bonne survie pour le sexe=2 par rapport au sexe=1

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

Le temps de survie médian pour le sexe = 1 (groupe masculin) est de 270 jours, contre 426 jours pour le sexe = 2 (féminin). Il semble y avoir un avantage de survie pour les femmes atteintes d'un cancer du poumon par rapport aux hommes. Cependant, pour évaluer si cette différence est statistiquement significative, il faut un test statistique formel, un sujet qui est abordé dans les sections suivantes.

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.

```
ggsurvplot(fit,
           conf.int = TRUE,
           risk.table.col = "strata", # Change risk table color by groups
           ggtheme = theme_bw(), # Change ggplot2 theme
           palette = c("#E7B800", "#2E9FDF"),
           fun = "cumhaz")
```

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.

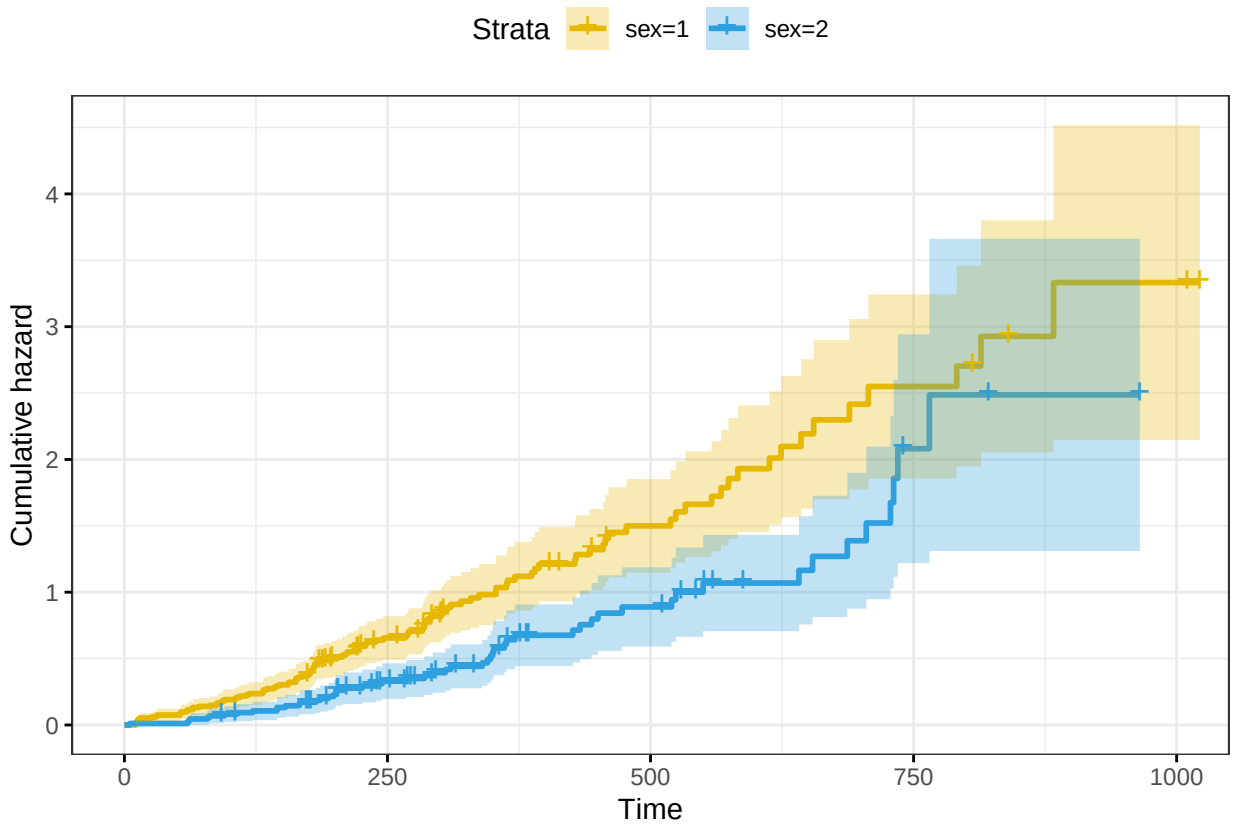

Figure 2: Courbes de survie cumulatif

5. COMPARAISON DES COURBES DE SURVIES

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)
surv_diff
```

```
Call:
survdiff(formula = Surv(time, status) ~ sex, data = lung)

        N Observed Expected (O-E)^2/E (O-E)^2/V
sex=1 138      112     91.6      4.55      10.3
sex=2  90       53     73.4      5.68      10.3

 Chisq= 10.3  on 1 degrees of freedom, p= 0.001
```

Le test du log-rank pour la différence de survie donne une valeur p de p = 0,0013, indiquant que les groupes de sexe diffèrent significativement en termes de survie.

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 :

```
peto_peto <- survdiff(Surv(time, status == 2) ~ sex, data = lung, rho = 1)
peto_peto
```

```
Call:
survdiff(formula = Surv(time, status == 2) ~ sex, data = lung,
    rho = 1)

        N Observed Expected (O-E)^2/E (O-E)^2/V
sex=1 138     70.4     55.6      3.95      12.7
sex=2  90     28.7     43.5      5.04      12.7

 Chisq= 12.7  on 1 degrees of freedom, p= 4e-04
```

La conclusion reste la meme avec une pvalue faible.

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

```
res.cox <- coxph(Surv(time, status) ~ sex, data = lung)
res.cox
```

```
Call:
coxph(formula = Surv(time, status) ~ sex, data = lung)

       coef exp(coef) se(coef)      z       p
sex -0.5310    0.5880   0.1672 -3.176 0.00149

Likelihood ratio test=10.63  on 1 df, p=0.001111
```

```
n= 228, number of events= 165
```

```
summary(res.cox)
```

```
Call:
coxph(formula = Surv(time, status) ~ sex, data = lung)

  n= 228, number of events= 165

       coef exp(coef) se(coef)      z Pr(>|z|)
sex -0.5310    0.5880   0.1672 -3.176  0.00149 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    exp(coef) exp(-coef) lower .95 upper .95
sex     0.588      1.701    0.4237     0.816

Concordance= 0.579  (se = 0.021 )
Likelihood ratio test= 10.63  on 1 df,   p=0.001
Wald test            = 10.09  on 1 df,   p=0.001
Score (logrank) test = 10.33  on 1 df,   p=0.001
```

INTERPRETATION DES RESULTATS

1. Signification statistique. La colonne marquée « z » donne la valeur statistique de Wald. Il correspond au rapport de chaque coefficient de régression à son erreur type $(z = coef/se(coef))$. La statistique de wald évalue si le bêta (β) d'une variable donnée est statistiquement significativement différent de 0. D'après la sortie ci-dessus, nous pouvons conclure que les variables sexe ont des coefficients statistiquement très significatifs.

2. Coefficients de régression. La deuxième caractéristique à noter dans les résultats du modèle de Cox est le signe des coefficients de régression $(coef)$. Un signe positif signifie que le danger (risque de décès) est plus élevé, et donc le pronostic pire, pour les sujets ayant des valeurs plus élevées de cette variable. La variable sexe est codée sous forme de

vecteur numérique. 1: mâle, 2: femelle. Le résumé R du modèle de Cox donne le hazard ratio (HR) pour le deuxième groupe par rapport au premier groupe, c'est-à-dire les femmes par rapport aux hommes. Le coefficient bêta pour le sexe = -0,53 indique que les femmes ont un risque de décès plus faible (taux de survie plus faible) que les hommes, dans ces données.

3. Rapports de risque. Les coefficients exponentiels $(exp(coef) = exp(-0,53) = 0,59)$, également appelés hazard ratios, donnent l'ampleur de l'effet des covariables. Par exemple, le fait d'être une femme (sexe =2) réduit le risque d'un facteur de 0,59, ou 41 %. Être une femme est associé à un bon pronostic.

4. Intervalles de confiance des hazard ratios. La sortie sommaire donne également des intervalles de confiance supérieurs et inférieurs à 95 % pour le hazard ratio (exp(coef)), la borne inférieure à 95 % = 0,4237, la borne supérieure à 95 % = 0,816.

5. Signification statistique globale du modèle. Enfin, la sortie donne des valeurs p pour trois tests alternatifs pour la signification globale du modèle: le test du rapport de vraisemblance, le test de Wald et les statistiques de logrank de score. Ces trois méthodes sont asymptotiquement équivalentes. Pour un N suffisamment grand, ils donneront des résultats similaires. Pour les petits N, ils peuvent différer quelque peu. Le test du rapport de vraisemblance a un meilleur comportement pour les échantillons de petite taille, il est donc généralement préféré.

6. ANALYSE DE REGRESSION DE COX MULTIVARIEE

```
res.cox <- coxph(Surv(time, status) ~ age + sex + ph.ecog, data =  lung)
summary(res.cox)
```

```
Call:
coxph(formula = Surv(time, status) ~ age + sex + ph.ecog, data = lung)

  n= 227, number of events= 164
   (1 observation effacée parce que manquante)
```

```
             coef exp(coef)  se(coef)      z Pr(>|z|)
age      0.011067  1.011128  0.009267  1.194 0.232416
sex     -0.552612  0.575445  0.167739 -3.294 0.000986 ***
ph.ecog  0.463728  1.589991  0.113577  4.083 4.45e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

        exp(coef) exp(-coef) lower .95 upper .95
age        1.0111     0.9890    0.9929    1.0297
sex        0.5754     1.7378    0.4142    0.7994
ph.ecog    1.5900     0.6289    1.2727    1.9864

Concordance= 0.637  (se = 0.025 )
Likelihood ratio test= 30.5  on 3 df,   p=1e-06
Wald test            = 29.93  on 3 df,   p=1e-06
Score (logrank) test = 30.5  on 3 df,   p=1e-06
```

EXPLICATION DU MODELE

La valeur de p pour les trois tests globaux (probabilité, Wald et score) est significative, ce qui indique que le modèle est significatif. Ces tests évaluent l'hypothèse nulle omnibus selon laquelle tous les bêtas (β) sont nuls. Dans l'exemple ci-dessus, les statistiques de test sont en étroite concordance et l'hypothèse nulle omnibus est fermement rejetée.

Dans l'analyse de Cox multivariée, les covariables sexe et ph.ecog restent significatives (p< 0,05).

Cependant, l'âge de la covariable n'est pas significatif (p=0,23, ce qui est supérieur à 0,05).

La valeur p pour le sexe est de 0,000986, avec un hazard ratio HR = exp(coef) = 0,58, indiquant une forte relation entre le sexe des patients et une diminution du risque de décès.

Les hazard ratios des covariables peuvent être interprétés comme des effets multiplicatifs sur le danger. Par exemple, le fait de maintenir les autres covariables constantes, étant une femme (sexe = 2), réduit le risque d'un facteur de 0,58, ou 42 %. Nous concluons qu'être une femme est associé à un bon pronostic. De même, la valeur p pour ph.ecog est de 4,45e-05, avec un hazard ratio HR = 1,59, indiquant une forte relation entre la valeur ph.ecog et un

risque accru de décès. En maintenant les autres covariables constantes, une valeur plus élevée de ph.ecog est associée à une faible survie. En revanche, la valeur de p pour l'âge est maintenant p = 0,23. Le hazard ratio HR = exp(coef) = 1,01, avec un intervalle de confiance à 95 % de 0,99 à 1,03. Étant donné que l'intervalle de confiance pour la FC comprend 1, ces résultats indiquent que l'âge contribue moins à la différence de HR après ajustement pour les valeurs ph.ecog et le sexe du patient, et seulement la tendance vers la signification. Par exemple, en maintenant les autres covariables constantes, un âge supplémentaire induit un risque quotidien de décès d'un facteur $exp(\beta)$= 1,01, ou 1%, ce qui n'est pas une contribution significative.

7. 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)


summary(fit_weibull)
```

```
Call:
survreg(formula = Surv(time, status) ~ sex + age + ph.ecog, data = lung)
               Value Std. Error     z       p
(Intercept)  6.27344    0.45358 13.83 < 2e-16
sex          0.40109    0.12373  3.24  0.0012
age         -0.00748    0.00676 -1.11  0.2690
ph.ecog     -0.33964    0.08348 -4.07 4.7e-05
```

```
Log(scale)  -0.31319    0.06135 -5.11 3.3e-07


Scale= 0.731


Weibull distribution
Loglik(model)= -1132.4   Loglik(intercept only)= -1147.4
    Chisq= 29.98 on 3 degrees of freedom, p= 1.4e-06
Number of Newton-Raphson Iterations: 5
n=227 (1 observation effacée parce que manquante)
```

Pour ajuster le même modèle mais avec la distribution exponentielle, le code est :

```r
fit_exp <- survreg(Surv(time, status) ~sex + age+ph.ecog, data =  lung,
                   dist = "exponential")


summary(fit_exp)
```

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`

RESUME

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.

• 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 .

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**