SYNTHESE DU PACKAGE SURVIVAL DE R

Axel LOUKOU _ Master 2 RAD _ USPN

2024-12-26

```
options(repos = c(CRAN = "https://cloud.r-project.org"))
```

APPLICATION : Analyse de survie avec le jeu de données lung

Le jeu de données **lung** inclus dans le package **survival** contient des informations sur des patients atteints d'un cancer du poumon. Il est couramment utilisé pour illustrer des techniques d'analyse de survie, notamment dans le domaine médical.

```
#installation des packages
install.packages(c( "ggplot2", "ggpubr", "survival", "survminer"))
```

```
##
## The downloaded binary packages are in
##  /var/folders/2l/m3xc9lx53_lbdm1wk1cw72_w0000gn/T//RtmpLzSmNH/downloaded_packages
```

```
library(ggplot2)
library(ggpubr)
library(survival)
library(survminer)
```

```
##
## Attaching package: 'survminer'

## The following object is masked from 'package:survival':
##
##     myeloma
```

1. Description du jeu de données "lung"

```
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
```

- **inst** : Code de l'institution
- **time** : Temps de survie en jours
- **Status** : Censure Statut (1=censuré, 2=mort)
- **age** : Age en années
- **sex** : Sexe (homme=1, femme=2)
- **ph.ecog** : score de performance ECOG (0 = bon, 5 = mort)
- **ph.karno** : Score de performance de Karnofsky (mauvais=0-bon=100) noté par le médecin
- **pat.karno** : Score de performance de Karnofsky évalué par le patient
- **meal.cal** : Calories consommées aux repas
- **wt.loss** : Perte de poids au cours des six derniers mois

2. Gestion des données censurées

La colonne **status** doit être transformée pour indiquer si le patient est censuré (0) ou décédé (1)

```
lung$status <- ifelse(lung$status == 2, 1, 0)  # Transformer '2' en '1' pour indiquer le décès
# Ne lancer qu'une fois, si dans les données la colonne status est déjà 0  et 1
# ne plus lancer une seconde fois cette partie du code
# Parce qu'auparavant, on avait : Status : Censure Statut (1=censuré, 2=mort)
```

```
#verification du changement de la colonne "status"
head(lung)
```

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

3. Calculer les courbes de survie : survfit()

La fonction **survfit()** peut être utilisée pour calculer l'estimation de survie de Kaplan-Meier. Ses principaux arguments sont les suivants :

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

Pour calculer la probabilité de survie par sexe :

```
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
```

Les résultats montrent une différence notable de survie entre les sexes dans le jeu de données **lung**.

Les hommes (**138 participants**) ont une médiane de survie de **270 jours**, avec **112 décès** observés, tandis que les femmes (**90 participantes**) présentent une médiane de survie plus élevée, à **426 jours**, avec **53 décès**. Cela suggère que les femmes ont une meilleure survie globale que les hommes. Ces différences peuvent être confirmées statistiquement par un test du log-rank pour évaluer leur significativité. Voici la commande pour le test :

```
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 révèle une différence statistiquement significative ($p = 0.001$) dans les courbes de survie entre les hommes et les femmes. Les hommes ont un **nombre de décès observés 112 supérieur à celui attendu 91.6**, tandis que les femmes ont un **nombre de décès observés 53 inférieur à celui attendu 73.4**. Cela confirme que les femmes ont une meilleure survie globale que les hommes.

4. Visualisation des courbes de survie

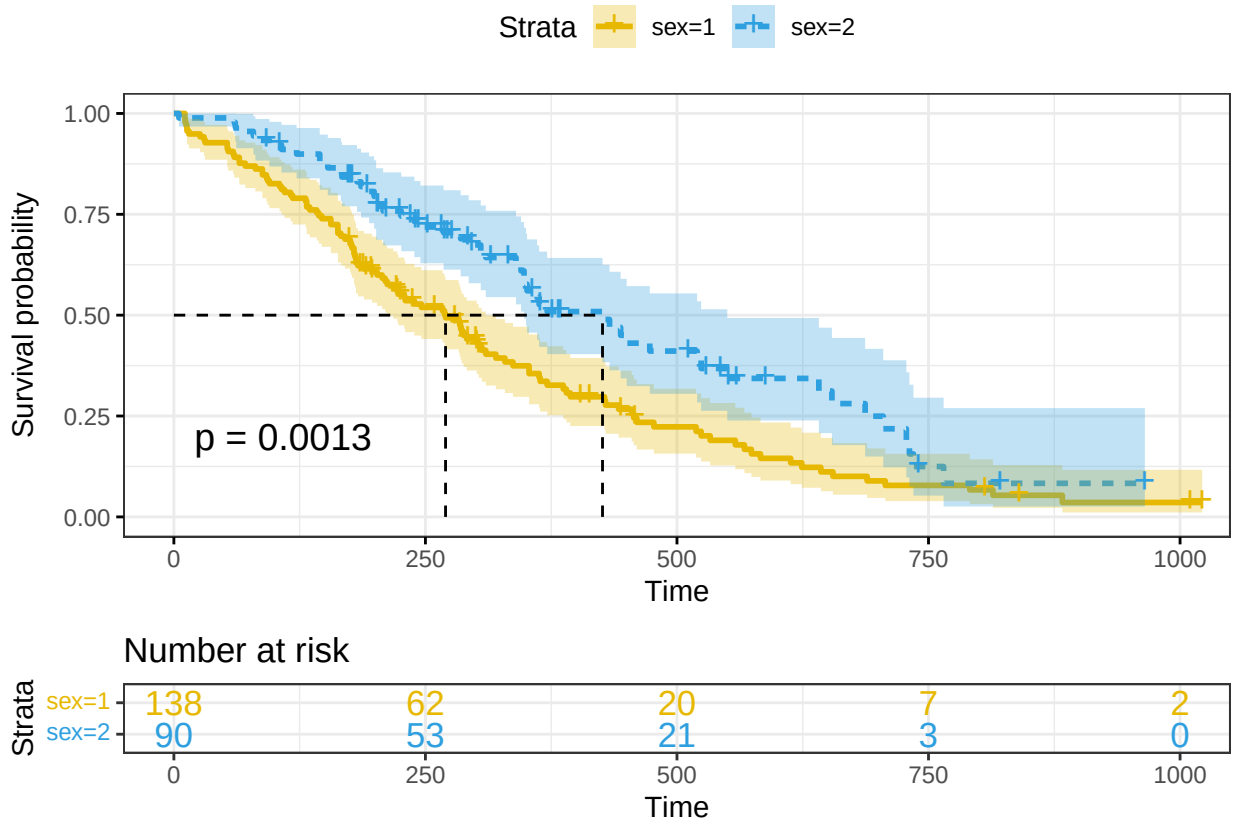
a. Graphique de Kaplan-Meier

Nous allons utiliser la fonction **ggsurvplot()** dans le package R de **Survminer** pour produire les courbes de survie pour les hommes et les femmes.

```
ggsurvplot(fit,
           pval = TRUE,              # Affiche la p-valeur du test log-rank.
           conf.int = TRUE,          # Affiche les intervalles de confiance autour des courbes de survie
           risk.table = TRUE,        # Ajoute une table des sujets à risque sous le graphique.
           risk.table.col = "strata",# Colore la table de risque par groupe (strates).
           linetype = "strata",      # Modifie le style des lignes (type de trait) en fonction des group
           surv.median.line = "hv",  # Ajoute des lignes horizontale et verticale pour indiquer la média
           ggtheme = theme_bw(),     # Utilise un thème clair et minimaliste (type `black & white`) pour
           palette = c("#E7B800", "#2E9FDF") # Définit une palette de couleurs personnalisée
                                             # (jaune pour le premier groupe, bleu pour le second).
)
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
```

```
## i The deprecated feature was likely used in the ggpubr package.
##   Please report the issue at <https://github.com/kassambara/ggpubr/issues>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Le graphique de Kaplan-Meier peut être interprété comme suit :

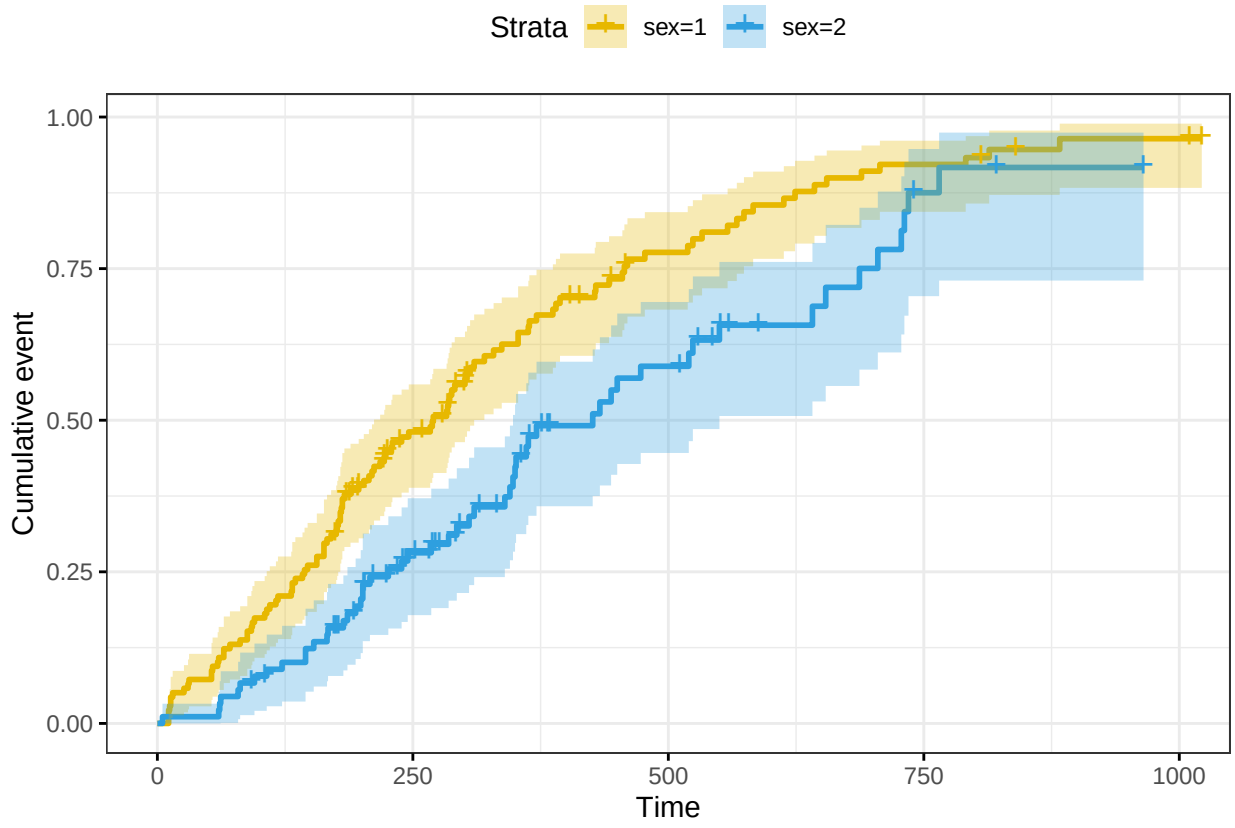
L'axe horizontal (axe des x) représente le temps en jours, et l'axe vertical (axe des y) montre la probabilité de survie ou la proportion de personnes qui survivent. Les lignes représentent les courbes de survie des deux groupes. Une chute verticale dans les courbes indique un événement. La coche verticale sur les courbes signifie qu'un patient a été censuré à ce moment-là.

- Au temps 0, la probabilité de survie est de 1,0 (soit 100% des participants sont en vie).
- Au temps 250, la probabilité de survie est d'environ 0,50 (ou 50 %) pour les hommes et de 0,75 (ou 75 %) pour les femmes.
- La survie médiane est d'environ 270 jours pour les hommes et de 426 jours pour les femmes, ce qui suggère une bonne survie pour les femmes par rapport au hommes.

b. Tracer des événements cumulatifs

```
ggsurvplot(fit,
           conf.int = TRUE,          # Affiche les intervalles de confiance autour des courbes.
```

```
          risk.table.col = "strata",# Colore la table des sujets à risque en fonction des groupes.
          ggtheme = theme_bw(),      # Utilise un thème clair et minimaliste pour le graphique.
          palette = c("#E7B800", "#2E9FDF"), # Définit une palette de couleurs personnalisée
                                             # (jaune pour le premier groupe, bleu pour le second).
          fun = "event"              # Transforme les courbes pour afficher la probabilité cumulative de
)
```



L'aléa cummulatif est couramment utilisé pour estimer la probabilité d'aléa. Il est défini comme suit :

$\mathbf{H(t) = -\log(S(t))}$ avec S(t) = survival function

Le risque cumulatif H(t) peut être interprétée comme la force cumulative de la mortalité. En d'autres termes, il correspond au nombre d'événements qui seraient attendus pour chaque individu par le temps t si l'événement était un processus reproductible.

- Le sexe a un effet significatif sur le risque de décès, avec un coefficient négatif. Le hazard ratio exp(coef) est 0.575, ce qui indique que les femmes (catégorie de référence) ont un risque de décès réduit de 42.5% par rapport aux hommes (1 - 0.575=0.425).
- L'indice d'état de performance physique ph.ecog a un effet très significatif sur le risque de décès, avec un coefficient de 0.464. Le hazard ratio est 1.59, indiquant qu'une augmentation d'une unité de cet indice est associée à une augmentation de 59% du risque de décès.

5. Ajustement du modèle de Cox

```r
# Ajustement du modèle de Cox avec l'âge et le sexe comme covariables
fit_cox <- coxph(Surv(time, status) ~ age + sex, data = lung)

# Résumé des résultats
summary(fit_cox)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ age + sex, data = lung)
##
##   n= 228, number of events= 165
##
##          coef exp(coef)  se(coef)      z Pr(>|z|)
## age  0.017045  1.017191  0.009223  1.848  0.06459 .
## sex -0.513219  0.598566  0.167458 -3.065  0.00218 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##     exp(coef) exp(-coef) lower .95 upper .95
## age    1.0172     0.9831    0.9990    1.0357
## sex    0.5986     1.6707    0.4311    0.8311
##
## Concordance= 0.603  (se = 0.025 )
## Likelihood ratio test= 14.12  on 2 df,   p=9e-04
## Wald test            = 13.47  on 2 df,   p=0.001
## Score (logrank) test = 13.72  on 2 df,   p=0.001
```

Interpretation :

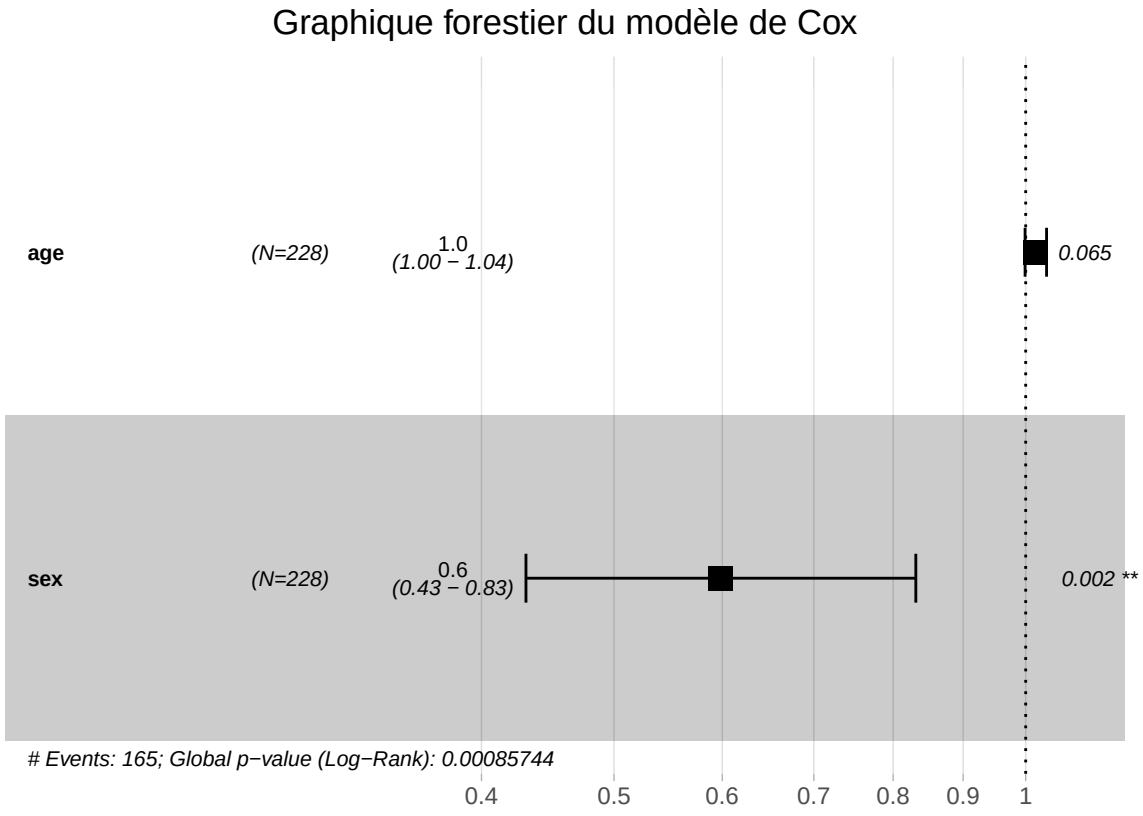
- **Âge** : L'effet de l'âge sur le risque est représenté par un coefficient de 0.017, suggérant une légère augmentation du risque avec chaque année supplémentaire. Le **hazard ratio (HR)**, donné par exp(coef)=1.017, indique une augmentation de 1.7% du risque pour chaque année supplémentaire. Cependant, la p-valeur est légèrement supérieure au seuil de significativité de 0.05, ce qui suggère que l'effet de l'âge n'est pas statistiquement significatif dans ce modèle. L'intervalle de confiance pour le HR [0.999,1.036] inclut 1, ce qui corrobore cette conclusion.
- **Sexe** : Le sexe a un effet significatif sur le risque, avec un coefficient de -0.513, indiquant une réduction du risque pour les femmes par rapport aux hommes (référence). Le **hazard ratio (HR)**, exp(coef)=0.599, montre que le risque pour les femmes est réduit de 40.1% par rapport aux hommes. Cet effet est statistiquement significatif, comme le montre la p-valeur. De plus, l'intervalle de confiance [0.431,0.831] ne contient pas 1, confirmant la significativité et la direction de l'effet.

```r
# Vérification de l'hypothèse des risques proportionnels
test_rp <- cox.zph(fit_cox)
print(test_rp)
```

```
##        chisq df    p
## age    0.209  1 0.65
## sex    2.608  1 0.11
## GLOBAL 2.771  2 0.25
```

Les résultats des tests des risques proportionnels montrent que le modèle de Cox ajusté respecte l'une de ses hypothèses clés, à savoir que les effets des covariables sur le risque sont constants au fil du temps. Cela valide l'utilisation de ce modèle pour les données analysées.

```r
# Visualisation du modèle sous forme de graphique forestier
ggforest(fit_cox, data = lung, main = "Graphique forestier du modèle de Cox")
```



CONCLUSION

Le package **survival** de R constitue un outil incontournable pour l'analyse de survie, offrant des fonctionnalités complètes pour traiter une large gamme de scénarios. Ce document a permis de couvrir les concepts fondamentaux de l'analyse de survie, notamment les modèles paramétriques, semi-paramétriques et non paramétriques, et leur implémentation sous R.

Les fonctions principales, telles que **Surv()**, **survfit()**, **coxph()** et **survreg()**, permettent de modéliser et d'analyser efficacement des données de survie, même en présence de censure. Des outils de visualisation comme **plot** et **ggsurvplot** enrichissent l'analyse en fournissant des graphiques clairs et interprétables, idéaux pour la présentation des résultats.

Grâce à des exemples pratiques basés sur le jeu de données lung, nous avons démontré l'application des modèles dans un cadre médical. Les résultats obtenus, notamment les différences significatives de survie entre hommes et femmes, illustrent la puissance et la flexibilité des outils fournis par le package **survival**.

En conclusion, le package **survival** offre une base solide pour explorer et modéliser les données de survie dans divers contextes, allant de la recherche médicale aux analyses actuarielles. Cependant, une compréhension approfondie des hypothèses et des limites des modèles reste essentielle pour interpréter correctement les résultats et garantir leur validité dans des applications réelles.