COMPTE RENDU
Biostatistique
Analyse des durées de survie

El Hadrami N'DOYE et Ismaïl RAMDÉ

14 Janvier 2022

1 Introduction

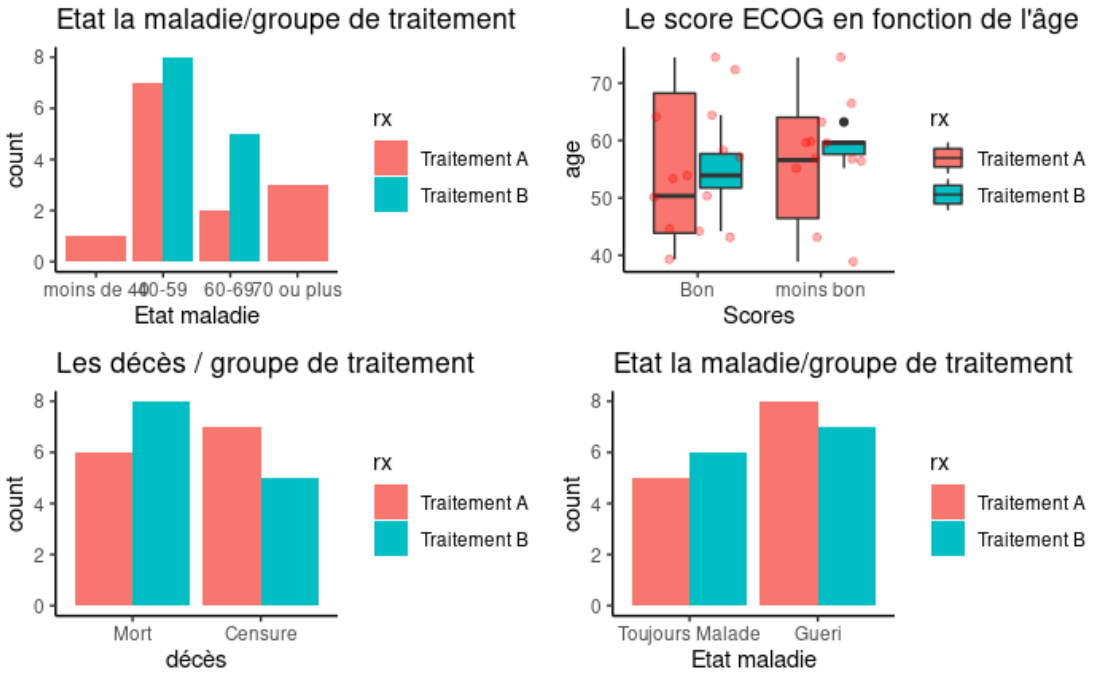
Dans le cadre du module de formation Biostatistique (Survie), nous nous intéressons à une étude portant sur le cancer de l'ovaire qui est une croissance de cellules qui se forme dans les ovaires. Elle se manifeste par une multiplication rapide (anormale) des cellules et peut envahir et détruire les tissus corporels sains. Pour ce faire nous disposons du jeu de donnée "ovarian" composé de 26 patients et de 6 variables. L'enjeu de cette étude est de d'analyser la durée de survie des patients tout en apportant des réponses quant à l'influence de certains facteurs comme l'âge ainsi que l'effet de 2 traitements (A et B). Dans la suite de ce rapport nous utiliserons diverses méthodes d'analyses afin de comprendre au mieux ce cancer.

2 Étude descriptive des données

2.1 Pré-traitement

Avant d'analyser les données, nous avons procédé à un certain nombre de pré-traitements notamment la transformation de variables en facteurs (groupe de traitement, la valeur du score ECOG et la présence résiduelle de la maladie) avec des modalités plus explicites et la transformation de la variable âge en classes.

2.2 Visualisation des données



Les patients de cette étude ont un âge moyen de 56.17. La classe d'âge comportant le grand nombre de patients est celle de "40 - 59". En ce qui concerne l'état résiduel de la maladie, les patients du traitement A sont les plus nombreux à ne plus présenter de résidus de la maladie. On observe plus de morts et moins de censures chez les patients ayant suivi le traitement B par rapport aux patients ayant suivi le traitement A. Ces visualisations ne nous permettent pas de tirer

toute de suite des conclusions. En effet nous devons tenir compte du temps en utilisant la courbe de Kaplan-Meier qui estime la probabilité de survie conditionnelle au temps t.

3 Résultat

3.1 Estimation de Kaplan-Meier et tests de comparaison

3.1.1 Estimation globale

Au cours de l'étude 12 cas (soit 46.15%) de censures ont été recensées.

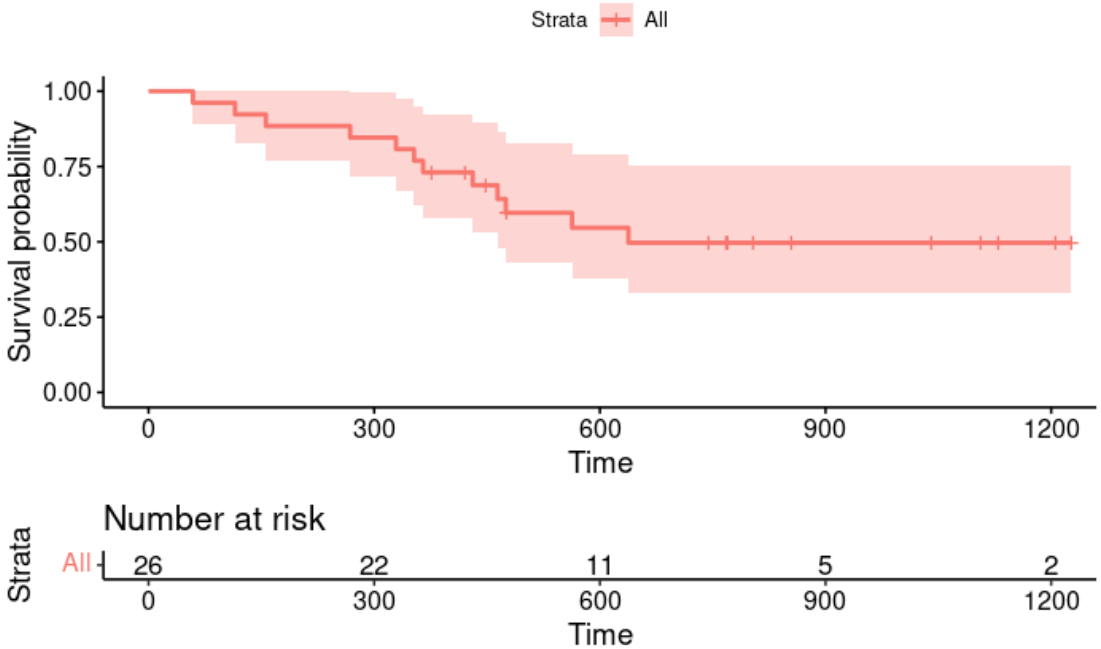
```
Call: survfit(formula = base ~ 1, data = ovarian, type = "kaplan-meier")

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   59     26       1    0.962  0.0377        0.890        1.000
  115     25       1    0.923  0.0523        0.826        1.000
  156     24       1    0.885  0.0627        0.770        1.000
  268     23       1    0.846  0.0708        0.718        0.997
  329     22       1    0.808  0.0773        0.670        0.974
  353     21       1    0.769  0.0826        0.623        0.949
  365     20       1    0.731  0.0870        0.579        0.923
  431     17       1    0.688  0.0919        0.529        0.894
  464     15       1    0.642  0.0965        0.478        0.862
  475     14       1    0.596  0.0999        0.429        0.828
  563     12       1    0.546  0.1032        0.377        0.791
  638     11       1    0.497  0.1051        0.328        0.752
```

Table 1: Résumé de l'estimateur global de Kaplan-Meier

La probabilité de vivre moins de 200 jours vaut 1 - 0.846 = 0.154 soit 15.4% et celle de vivre plus de 600 jours vaut 0.546 soit 54.6%.

Figure 1: La courbe de survie globale



On remarque que le nombre de patients à risque ainsi que la probabilité de survivre décroît avec le temps. Par contre la courbe de survie devient constante autour de la probabilité de 0.50 et après 600 jours.

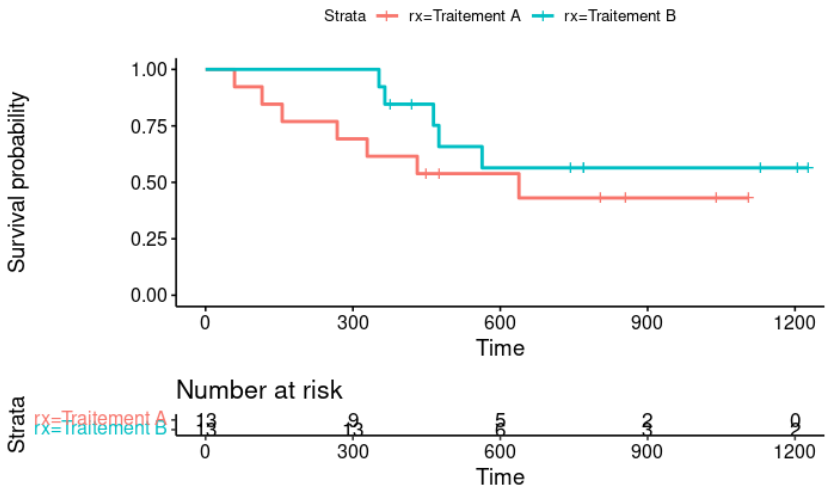
25	**50**	**75**
365	638	NA

Table 2: Estimation des quartiles

Pour 50% des observations, l'événement étudié est survenue à l'instant 638, pour 75% des observations l'événement est survenue à l'instant 365. On observe pas de temps de survenue de l'évènement pour 25% de des observations.

3.1.2 Estimation selon le traitement (rx) et la présence résiduelle de la maladie (resid.ds)

Figure 2: La courbe de survie des deux groupes de traitement



La courbe des patients du traitement A se situe toujours en dessous de celle des patients du traitement B. Cela signifie que les patients du groupe traitement B ont une probabilité de survie plus élevée que ceux des patients du groupe traitement A. Même si le traitement B semble plus efficace que le traitement A, il est difficile d'affirmer avec certitude à ce niveau son efficacité.

Le test de du **log-rank** nous permettra de comparer la survie des patients en fonction du traitement.

Hypothèses :

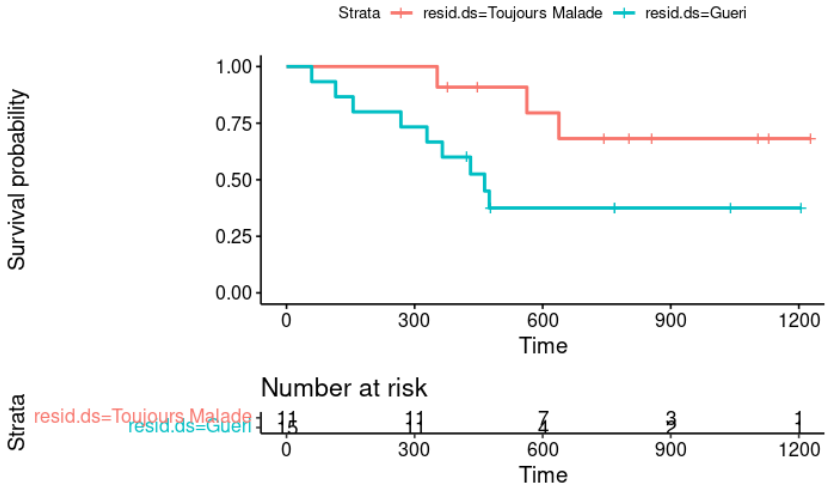
H_0 : pas de différence de survie entre les deux groupes étudiés

H_1 : différence de survie entre les deux groupes étudiés

Le test du log-rank nous indique une p-valeur égale à 0.3. Elle est supérieure au seuil $\alpha = 0.05$. On rejette H_0, il y a donc une différence significative de survie entre les deux groupes étudiés.

Intéressons nous à l'effet de la présence résiduelle de la maladie au cours du temps :

Figure 3: La courbe de survie de la présence résiduelle de la maladie



Visuellement les deux courbes ont chacune des trajectoires différentes. La probabilité de survie des patients toujours malade semble plus élevé que celle des patients guéris.

Le test du **Log-rank** sous les mêmes hypothèses mentionnées dans le cas précédent nous confirme une différence entre les deux groupes car la p-valeur égale à 0.06 est supérieur à $\alpha = 0.05$.

Les deux tests peuvent être illusoires pour tirer des conclusions car nous avons un faible effectif de chaque groupe pour les variables traitement (rx) et présence résiduelle de la maladie (resid.ds).

3.2 Modèle de Cox

Pour aller plus loin dans l'analyse de l'effet ou l'influence d'une variable (les groupes) sur la survie en fonction du temps, on peut utiliser le modèle de cox. Elle consiste à faire une régression logistique.

On peut écrire le modèle à risques proportionnels de Cox ainsi :

$h(t) = h_0(t)e^{\beta_k x_k}$

Où $h_0(t)$ est le risque de décès pour un individu qui a survécu jusqu'au temps t et a une valeur de 0 pour toutes les co-variables x_k.

Un rapport de hasards désigne le risque de survenue d'un résultat dans une analyse réalisée à l'aide du modèle de régression de cox. Il s'agit d'un risque relatif (instantané) tenant compte de la durée de présence dans l'étude.

Un risque relatif est le quotient de deux risques (absolus), le risque dans le groupe exposé ou intervention et le risque dans le groupe contrôle. Dans une étude d'intervention, le risque relatif est une estimation de la probabilité que le résultat (par exemple survenue d'un décès) dans le groupe intervention soit autant de fois supérieur ($RR > 1$) ou inférieur ($RR < 1$) à celui observé dans le groupe contrôle.

Variable de survie traitement (rx) :

On peut écrire le modèle comme suit: $ln(h(t)/h_0(t)) = \beta traitement(rx)$

On obtient une p-valeur à l'issue du modèle de cox égale à $0.3 > 0.05$. Néanmoins, avec un $\beta_1 = -0.5964 < 0$ l'effet du traitement B diminue le risque h de décès d'un facteur de $e^{-0.6} = 0.55$. On a donc un risque de décès du groupe de Traitement A qui est de $1/0.55 = 1.81$. Il est 1.81 fois supérieur à celui du groupe de Traitement B.

Variable âge :

```
Concordance= 0.752  (se = 0.07 )
Likelihood ratio test= 14.51  on 3 df,   p=0.002
Wald test            = 1099  on 3 df,   p=<2e-16
Score (logrank) test = 30.44  on 3 df,   p=1e-06
```

Table 3: Résumé des p-valeurs

Les p-valeurs des différents test (Likelihood ratio test, Wald test et logrank) sont tous inférieur au seuil $\alpha = 0.05$. L'âge a donc un effet statistiquement significatif sur le risque de décès h des patients. Ce risque augmente d'un facteur de $1.322e^{+01}$ par année pour les patients de la classe $[40 - 59]$, de $1.402e^{+01}$ par année pour les patients de la classe $[60 - 69]$ et de $1.732e^{+01}$ par année pour les patients de la classe 70 ans et plus.

Variable valeur du score ECOG :

```
Concordance= 0.521  (se = 0.078 )
Likelihood ratio test= 0.47  on 1 df,   p=0.5
Wald test            = 0.46  on 1 df,   p=0.5
Score (logrank) test = 0.47  on 1 df,   p=0.5
```

Table 4: Résumé des p-valeurs

On obtient une $p - valeur > .05$. Le score ECOG n'a pas d'effet statistiquement significatif sur le risque de décès des patients.

Étude de toutes les variables simultanément :

```
Call:
coxph(formula = Surv(futime, fustat) ~ age_classe + resid.ds +
    rx + ecog.ps, data = ovarian, method = "breslow")

  n= 26, number of events= 12

                          coef  exp(coef)   se(coef)      z Pr(>|z|)
age_classe40-59       1.492e+01  3.032e+06  6.512e-01 22.919   <2e-16 ***
age_classe60-69       1.525e+01  4.189e+06  6.528e-01 23.358   <2e-16 ***
age_classe70 ou plus  1.771e+01  4.923e+07  1.163e+00 15.227   <2e-16 ***
resid.dsGueri         1.371e+00  3.941e+00  7.166e-01  1.914   0.0557 .
rxTraitement B       -8.815e-01  4.142e-01  6.540e-01 -1.348   0.1777
ecog.psmoins bon      5.501e-01  1.733e+00  6.071e-01  0.906   0.3649
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                      exp(coef) exp(-coef) lower .95 upper .95
age_classe40-59      3.032e+06  3.299e-07 8.460e+05 1.086e+07
age_classe60-69      4.189e+06  2.387e-07 1.165e+06 1.506e+07
age_classe70 ou plus 4.923e+07  2.031e-08 5.036e+06 4.812e+08
resid.dsGueri        3.941e+00  2.538e-01 9.673e-01 1.605e+01
rxTraitement B       4.142e-01  2.415e+00 1.149e-01 1.492e+00
ecog.psmoins bon     1.733e+00  5.769e-01 5.274e-01 5.697e+00

Concordance= 0.819  (se = 0.066 )
Likelihood ratio test= 17.6  on 6 df,   p=0.007
Wald test            = 1309  on 6 df,   p=<2e-16
Score (logrank) test = 32.88  on 6 df,   p=1e-05
```

Table 5: Résumé du modèle de Cox

$$ln(h(t)/h_0(t)) = \beta_1 age + \beta_2 resid.ds + \beta_3 rx + \beta_4 ecog.ps(rx)$$

À l'issue du modèle complet, les trois tests affichent une p-valeur inférieure à 0.05. Tous les co-variables du modèle semblent significatifs. Par contre en s'intéressant aux résultats de façon séparée, seul l'âge a un effet sur le risque de décès h (significatif).

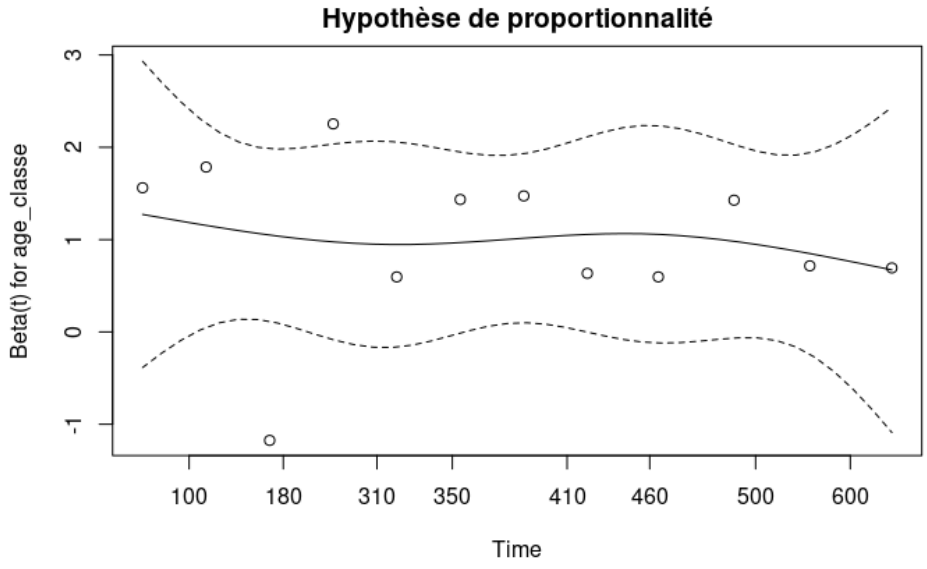
Nous pouvons donc effectuer une sélection de variables à partir du résultat de modèle de Cox complet. En effet nous utilisons la méthode de sélection **descendante de variable pas à pas** qui consiste à enlever celle dont la p-valeur est la plus élevée c'est-à-dire dont la statistique de Wald est la plus faible. On refait tourner le modèle et on recommence jusqu'à obtention de toutes les variables significatives.

Les variables explicatives qui ont des p-valeurs les plus élevées sont la valeur du score ECOG (ecog.ps), la présence résiduelle de maladies (resid.ds) et le traitement (rx). Nous les enlevons du modèle. **L'âge est la seule variable sélectionnée**.

Les hypothèses de modélisation étant :

- Le rapport des risques instantanés ("hazard rate" en anglais) de deux patients est indépendant du temps. C'est l'hypothèse des risques proportionnels.
- $log(h(t|Z_{i1}, \ldots, Z_{ip})) = log(h_0(t)) + \theta_0 Zi$. Le logarithme du risque instantané est une fonction linéaire des Zij. C'est l'hypothèse de log-linéarité.

Vérifions les pour le modèle final impliquant que l'âge des patients.



	chisq	df	p
age_classe	1.24	3	0.74
GLOBAL	1.24	3	0.74

En se basant sur l'étude des résidus de Schoenfield, on voit que la probabilité du test est supérieure à 0.05. La condition d'indépendance est donc vérifiée pour la variable âge.

Graphiquement on constate que les estimations bêta changent avec le temps. L'hypothèse des risques de proportion est donc vérifiée.

4 Conclusion

En somme, l'étude du cancer de l'ovaire que nous avons effectué nous a montré qu'au cours du temps la survie des patients diminue pour ensuite stationner. Le traitement B ainsi que l'âge du patient se sont avérés avoir un effet significatif sur leur survie. En effet le suivi du traitement B permettrait de diminuer le risque de décès tant disque que l'âge aurait un impact négatif sur le risque de décès des patients.