Projet BIOSTATISTIQUE

AGBLODOE Komi/ M2 SSD

28 décembre 2019

```
library(survival)

head(ovarian)
```

```
##   futime fustat     age resid.ds rx ecog.ps
## 1     59      1 72.3315        2  1       1
## 2    115      1 74.4932        2  1       1
## 3    156      1 66.4658        2  1       2
## 4    421      0 53.3644        2  2       1
## 5    431      1 50.3397        2  1       1
## 6    448      0 56.4301        1  1       2
```

```
dim(ovarian)
```

```
## [1] 26  6
```

```
attach(ovarian)
```

I. Estimateur de Kaplan-Meier et tests de comparaison

1. Quel est le pourcentage d'observations censurées dans le jeu de données?

```
length(fustat[fustat==0])/length(fustat)
```

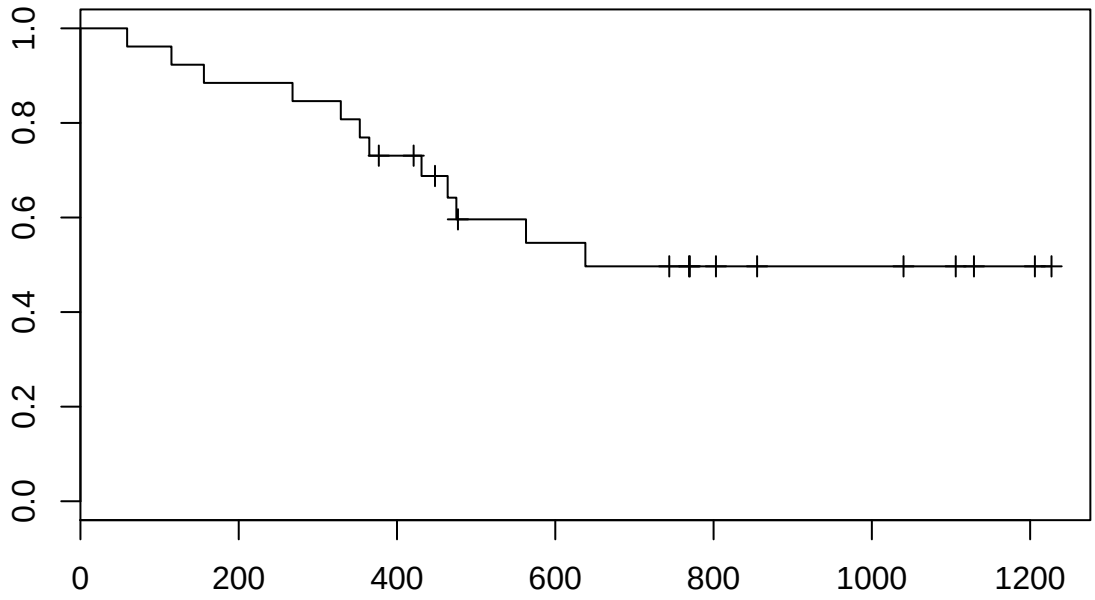
```
## [1] 0.5384615
```

53,84% des données sont censurées. Soit environ 54%.

2. Quelle quantité est estimée par l'estimateur de Kaplan-Meier ?

La fonction de survie ou la probabilité de survie est la quantité estimée par l'estimateur de Kaplan-Meier.

```
base<-Surv(futime,fustat)
fit<-survfit(base~1,data=ovarian)
plot(fit,mark.time=TRUE,conf.int=FALSE,col=c(1,2))
```



Commenter le graphique représentant l'estimateur de Kaplan-Meier pour tous les groupes confondus:

Le temps de survie médian est estimé à 638 jours pour tous les groupes.

On donnera notamment les estimations des quartiles de la loi de T (c'est à dire les quantiles d'ordre 25%, 50% et 75%)

```
quantile(fit,c(.25,.5,.75))$quantile
```

```
##  25  50  75
## 365 638  NA
```

3. A combien estimez-vous la probabilité qu'une patiente atteinte du cancer des ovaires vive moins de 200 jours ? Qu'elle vive plus de 600 jours ?

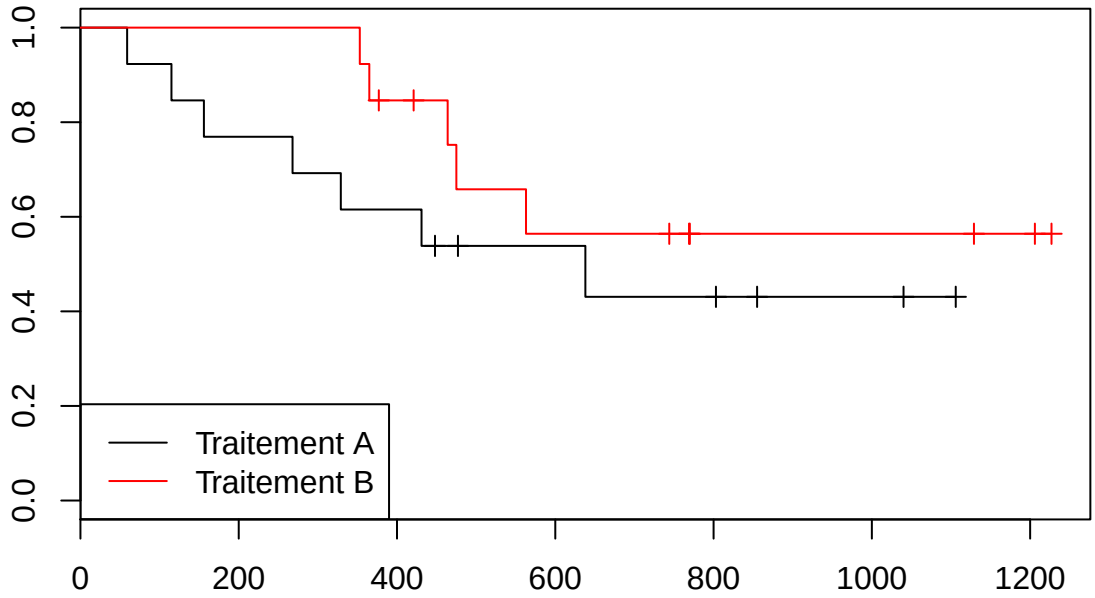
```
fn<-stepfun(fit$time,c(0,fit$surv))

proba_200_jours<-1-fn(200)
proba_600_jours<-fn(600)
```

la probabilité qu'une patiente atteinte du cancer des ovaires vive moins de 200 jours est de 0.115 et de 0.546 qu'elle vive plus de 600 jours.

4. On s'intéresse dans un premier temps à l'influence du traitement sur la survie des patients. D'un point de vue descriptif vous semble-t-il que le traitement permette d'augmenter la durée de vie des patients ?

```
base<-Surv(futime,fustat)
fit<-survfit(base~rx,data=ovarian)
plot(fit,mark.time=TRUE,conf.int=FALSE,col=c(1,2))
legend("bottomleft", c("Traitement A","Traitement B"),col=c(1,2),lty=1)
```



D'un point de vue descriptif, il semble que le traitement permette d'augmenter la durée de vie des patients. Le traitement B semble augmenter la durée de vie des patients comparé au traitement A.

5. Quel test statistique proposez-vous pour comparer la survie des patients en fonction du traitement ? On donnera les hypothèses nulle et alternative.

Pour comparer la survie des patients, on propose pour test statistique celui du log-rank

Notons S_A et S_B les fonctions de survie des groupes A et B. On souhaite tester:

$(H_0) : S_A = S_B$ contre $(H_1) : S_A \neq S_B$

```r
base<-Surv(futime,fustat)

survdiff(base~rx,data=ovarian)
```

```
## Call:
## survdiff(formula = base ~ rx, data = ovarian)
##
##       N Observed Expected (O-E)^2/E (O-E)^2/V
## rx=1 13        7     5.23     0.596      1.06
## rx=2 13        5     6.77     0.461      1.06
##
##  Chisq= 1.1  on 1 degrees of freedom, p= 0.3
```

Commenter le résultat du test:

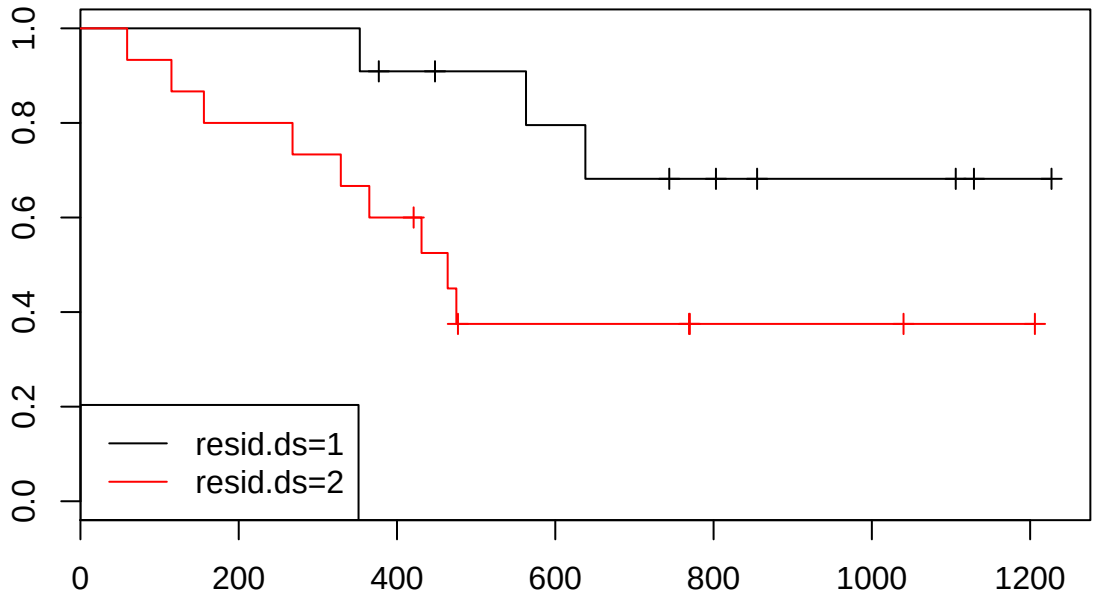
Au regard de la p-valeur obtenue, on ne rejette pas l'hypothèse nulle H_0. Les deux courbes de survie ne sont pas significativement différentes au seuil $\alpha = 5\%$. On ne peut pas conclure à un effet de traitement.

6. On s'intéresse enfin à la survie des patients en fonction de la variable resid.ds.

Commenter le graphique ainsi que le test associé permettant de comparer la survie des patients en fonction de cette variable.Quel est l'impact de la présence de la maladie sur la durée de vie d'un patient ?

```r
base<-Surv(futime,fustat)

fit2<-survfit(base~resid.ds,data=ovarian)
plot(fit2,mark.time=TRUE,col=c(1,2))
legend("bottomleft", c("resid.ds=1","resid.ds=2"),col=c(1,2),lty=1)
```



Graphiquement, il semble avoir un écart significatif entre les fonctions de survie. Graphiquement, on peut affirmer que l'absence de résidus permet d'augmenter la durée de survie.

Faisons le test du log-rank:

```
survdiff(base~resid.ds,data=ovarian) #(test du log-rank)
```

```
## Call:
## survdiff(formula = base ~ resid.ds, data = ovarian)
##
##              N Observed Expected (O-E)^2/E (O-E)^2/V
## resid.ds=1 11        3     6.26      1.70      3.62
## resid.ds=2 15        9     5.74      1.85      3.62
##
##  Chisq= 3.6  on 1 degrees of freedom, p= 0.06
```

Au regard de la p-valeur obtenue (p=0.06), on ne rejette pas l'hypothèse nulle d'égalité des fonctions de survie au seuil $\alpha = 5\%$.

Par exemple, au bout de 400 jours, donner une estimation de la probabilité d'être toujours en vie en fonction de la présence au nom de residus de maladie.

```
summary(fit2)
```

```
## Call: survfit(formula = base ~ resid.ds, data = ovarian)
##
##                 resid.ds=1
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   353     11       1    0.909  0.0867        0.754            1
##   563      8       1    0.795  0.1306        0.577            1
##   638      7       1    0.682  0.1536        0.438            1
##
##                 resid.ds=2
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    59     15       1    0.933  0.0644        0.815        1.000
##   115     14       1    0.867  0.0878        0.711        1.000
##   156     13       1    0.800  0.1033        0.621        1.000
##   268     12       1    0.733  0.1142        0.540        0.995
##   329     11       1    0.667  0.1217        0.466        0.953
##   365     10       1    0.600  0.1265        0.397        0.907
##   431      8       1    0.525  0.1310        0.322        0.856
##   464      7       1    0.450  0.1321        0.253        0.800
##   475      6       1    0.375  0.1296        0.190        0.738
```

Au bout de 400 jours, la probabilité d'être toujours en vie est estimée à 0.6 s'il y a présence de résidus et 0.909 s'il n'y a pas de présence de résidus.

II. Modèle de Cox

1. On propose maintenant d'expliquer la variable temps de survie en fonction de la variable rx par un modèle de Cox. Ecrire le modèle de Cox correspondant

```
base<-Surv(futime,fustat)
fit3<-coxph(base ~ rx,data=ovarian)
```

2.

Le rapport des risques (ou hazard ratio) noté RR correspond au fait que le risque instantané est multiplié par RR pour un individu appartenant au groupe ayant eu une thérapie antérieure.

Donner l'estimation du rapport de risque instantané comparant le groupe ayant eu le traitement B par rapport au traitement A

```
fit3
```

```
## Call:
## coxph(formula = base ~ rx, data = ovarian)
##
##       coef exp(coef) se(coef)     z    p
## rx -0.596     0.551    0.587 -1.02 0.31
##
## Likelihood ratio test=1.05  on 1 df, p=0.3
## n= 26, number of events= 12
```

Le rapport de risque instantané comparant le groupe ayant eu le traitement B par rapport au traitement A est estimé à 0.55

Commenter:

Le taux de risque instantané des patients sous le traitement B est 0.55 fois plus petit que celui des patients suivant le traitement A.

3. Donner les hypothèses de test de l'influence de la variable rx sur le temps de survie.

Proposer trois tests de l'influence de la variable rx sur le temps de survie. Commenter le

programme correspondant et les résultats obtenus. Que concluez-vous ?

Notons β la mesure d'influence de la variable rx. Les hypothèses du test sont:

$(H_0) : \beta = 0$ contre $(H_1) : \beta \neq 0$

On propose les tests de wald, du rapport de vraisemblance et du score pour tester l'influence de la variable rx sur le temps de survie.

```
summary(fit3)
```

```
## Call:
## coxph(formula = base ~ rx, data = ovarian)
##
##   n= 26, number of events= 12
##
##       coef exp(coef) se(coef)      z Pr(>|z|)
## rx -0.5964    0.5508   0.5870 -1.016     0.31
##
##    exp(coef) exp(-coef) lower .95 upper .95
## rx    0.5508      1.816    0.1743      1.74
##
## Concordance= 0.608  (se = 0.078 )
## Rsquare= 0.04   (max possible= 0.932 )
## Likelihood ratio test= 1.05  on 1 df,   p=0.3
## Wald test            = 1.03  on 1 df,   p=0.3
## Score (logrank) test = 1.06  on 1 df,   p=0.3
```

Au regard des p-valeurs obtenues pour chacun de ces trois tests (p=0.3), on ne rejette pas l'hypothèse nulle au seuil $\alpha = 5\%$. On conclut que $\beta = 0$.

4. Commenter brièvement les deux modèles de Cox univariés pour les variables age et ecog.ps.

Ces deux variables ont-elles un effet sur la survie des patients ?

```
base<-Surv(futime,fustat)
fit_age<-coxph(base ~ age,data=ovarian)
fit_ecog.ps<-coxph(base ~ ecog.ps,data=ovarian)
```

Faisons les trois tests sur les modèles précédents

```
summary(fit_age)
```

```
## Call:
## coxph(formula = base ~ age, data = ovarian)
##
##   n= 26, number of events= 12
##
##        coef exp(coef) se(coef)     z Pr(>|z|)
## age 0.16162   1.17541  0.04974 3.249  0.00116 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##     exp(coef) exp(-coef) lower .95 upper .95
## age     1.175     0.8508     1.066     1.296
##
## Concordance= 0.784  (se = 0.091 )
## Rsquare= 0.423   (max possible= 0.932 )
## Likelihood ratio test= 14.29  on 1 df,   p=2e-04
## Wald test            = 10.56  on 1 df,   p=0.001
## Score (logrank) test = 12.26  on 1 df,   p=5e-04
```

Au regard des p-valeurs des trois tests (toutes inférieures à 0.05), on conclut que la variable age a un effet sur la survie des patients.

```
summary(fit_ecog.ps)
```

```
## Call:
## coxph(formula = base ~ ecog.ps, data = ovarian)
##
##   n= 26, number of events= 12
##
##           coef exp(coef) se(coef)     z Pr(>|z|)
## ecog.ps 0.3984    1.4894   0.5864 0.679    0.497
##
##         exp(coef) exp(-coef) lower .95 upper .95
## ecog.ps     1.489     0.6714    0.4719       4.7
##
## Concordance= 0.521  (se = 0.079 )
## Rsquare= 0.018   (max possible= 0.932 )
## 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
```

Au regard des p-valeurs des trois tests (toutes supérieures à 0.05), on conclut que la variable ecog.ps n'a pas d'effet sur la survie des patients.

5. Écrire le modèle de Cox complet (avec toutes les variables explicatives). Discuter les effets des différentes variables.

Modèle de Cox complet

```r
fit4<-coxph(base ~ age+resid.ds+rx+ecog.ps,data=ovarian)
summary(fit4)
```

```
## Call:
## coxph(formula = base ~ age + resid.ds + rx + ecog.ps, data = ovarian)
##
##   n= 26, number of events= 12
##
##              coef exp(coef) se(coef)      z Pr(>|z|)
## age       0.12481   1.13294  0.04689  2.662  0.00777 **
## resid.ds  0.82619   2.28459  0.78961  1.046  0.29541
## rx       -0.91450   0.40072  0.65332 -1.400  0.16158
## ecog.ps   0.33621   1.39964  0.64392  0.522  0.60158
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##          exp(coef) exp(-coef) lower .95 upper .95
## age         1.1329     0.8827    1.0335     1.242
## resid.ds    2.2846     0.4377    0.4861    10.738
## rx          0.4007     2.4955    0.1114     1.442
## ecog.ps     1.3996     0.7145    0.3962     4.945
##
## Concordance= 0.807  (se = 0.091 )
## Rsquare= 0.481   (max possible= 0.932 )
## Likelihood ratio test= 17.04  on 4 df,   p=0.002
## Wald test            = 14.25  on 4 df,   p=0.007
## Score (logrank) test = 20.81  on 4 df,   p=3e-04
```

Au regard des différentes p-valeurs pour les différentes variables, seule la variable age a un effet significatif.

6. Proposer une méthode de sélection de variables pour le modèle de Cox. Quelles sont les

variables sélectionnées par ce critère ?

```r
model_final<-step(fit4, dir = "backward")
```

```
## Start:  AIC=60.93
## base ~ age + resid.ds + rx + ecog.ps
##
##            Df    AIC
## - ecog.ps   1 59.202
## - resid.ds  1 60.055
## - rx        1 60.868
## <none>        60.927
## - age       1 69.939
##
## Step:  AIC=59.2
## base ~ age + resid.ds + rx
##
##            Df    AIC
## - resid.ds  1 58.084
```

```
## - rx        1 58.942
## <none>        59.202
## - age       1 68.537
##
## Step:  AIC=58.08
## base ~ age + rx
##
##        Df    AIC
## - rx    1 57.676
## <none>    58.084
## - age   1 70.918
##
## Step:  AIC=57.68
## base ~ age
##
##        Df    AIC
## <none>    57.676
## - age   1 69.970
```

La variable age est la seule qui est sélectionnée par cette méthode de sélection de variables.

7. Quelles sont les hypothèses de modélisation que doit vérifier un modèle de Cox ? Détaillez

votre réponse.

Un modèle de Cox doit vérifier:

-l'hypothèse des risques proportionnels: l'effet de chacune des covariables du modèle est indépendant du temps (risque relatif constant au cours du temps).

-l'hypothèse de log-linéarité pour une covariable continue: le risque relatif d'une augmentation d'une unité de la covariable est constant.

8. Implémenter sous R cette validation pour le modèle final. Commenter.

Test de l'hypothèse des risques proportionnels:

```
model_final_test<-cox.zph(model_final)
model_final_test
```

```
##        rho chisq     p
## age -0.209 0.647 0.421
```

On ne rejette pas l'hypothèse des risques proportionnels compte tenu de la p-valeur du test (supérieure à 0.05)