PROJET DONNÉE CENSURÉE

NOM : SALL

PRÉMOM : ADAMA

CODE PERMANENT : P33-65-53

Master 2 : SDA(Statistique)

Année : 2023-2024

Objectif :

L'objectif de ce travail est de modéliser la durée de survie des patients atteints d'un cancer de l'estomac, en identifiant les facteurs pronostiques associés à la survie et en analysant aussi l'influence des variables explicatives , telles que l'âge, le sexe, le type de traitement, ainsi que d'autres paramètres cliniques pertinents. Cette étude vise aussi à montrer l'impact de ces facteurs sur la mortalité du cancer l'estomac, tout en tenant compte des données censurées (*DECES*), afin de proposer des modèles prédictifs.

Dans un premier temps nous allons l'analyse descriptive pour la comprehension de la base de donnée.

Statistique descriptive

variables	Min	Q1	Median	Moyenne	Q3	Max
Duree de survie	2	82,5	3600	2097,4	3600	3600
Age	26	40	53,50	53,52	60	82
DuréeSymptom (mois)	1	2	4	7	22	24

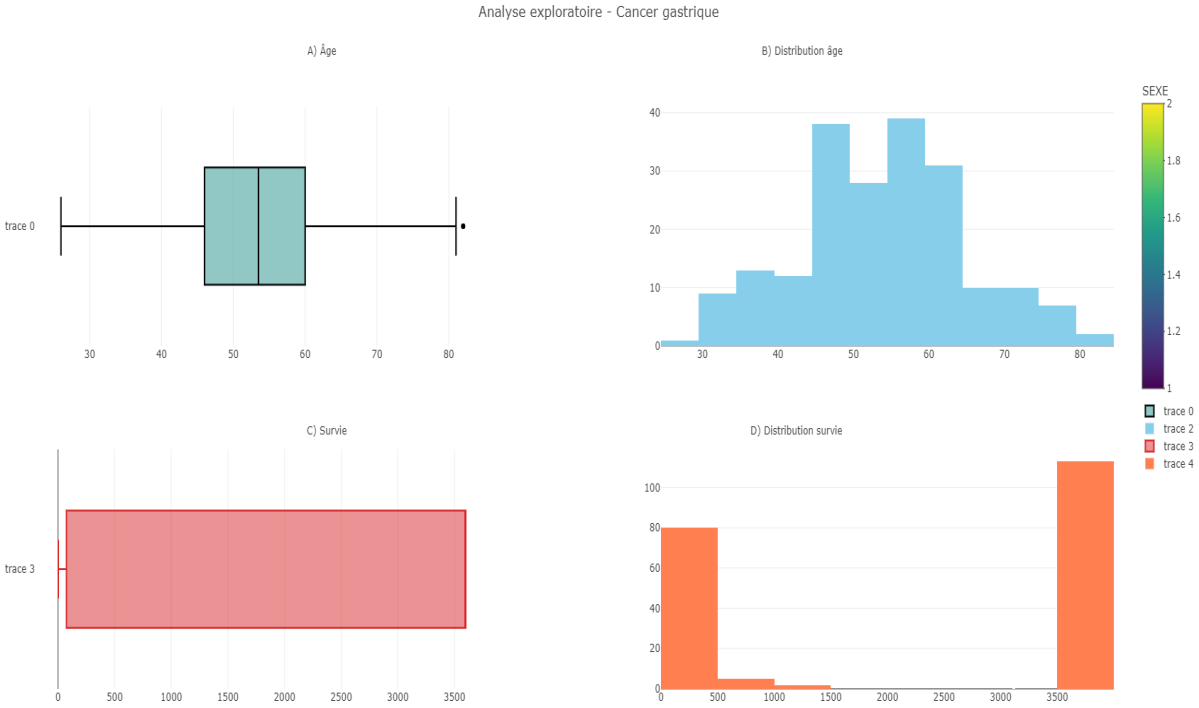
Pour la variable sexe :

Homme (1) = 99
Femme (2) = 101

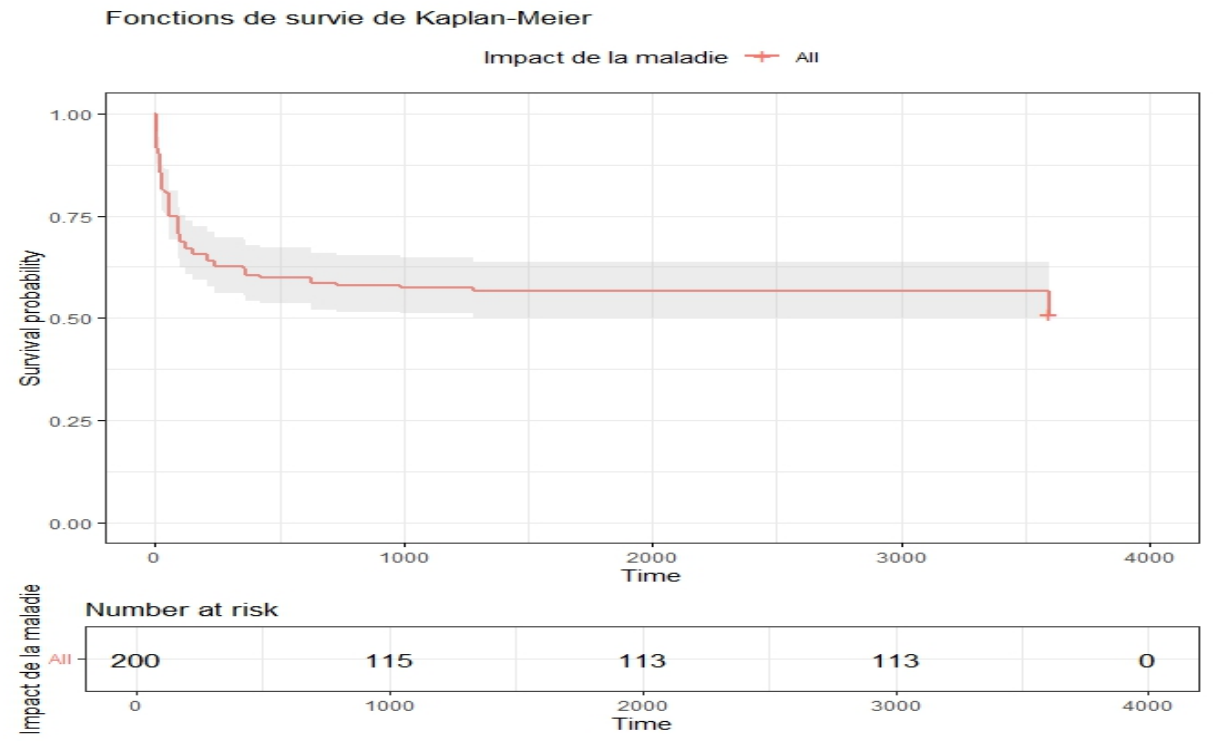
Pour la variable traitement:

Traitement 0 = 9
Traitement 1 = 129
Traitement 2 = 1
Traitement 3 = 61

Visualisation



1. **Estimation de la fonction de survie de Kaplan Meier avec les intervalles de confiance.**



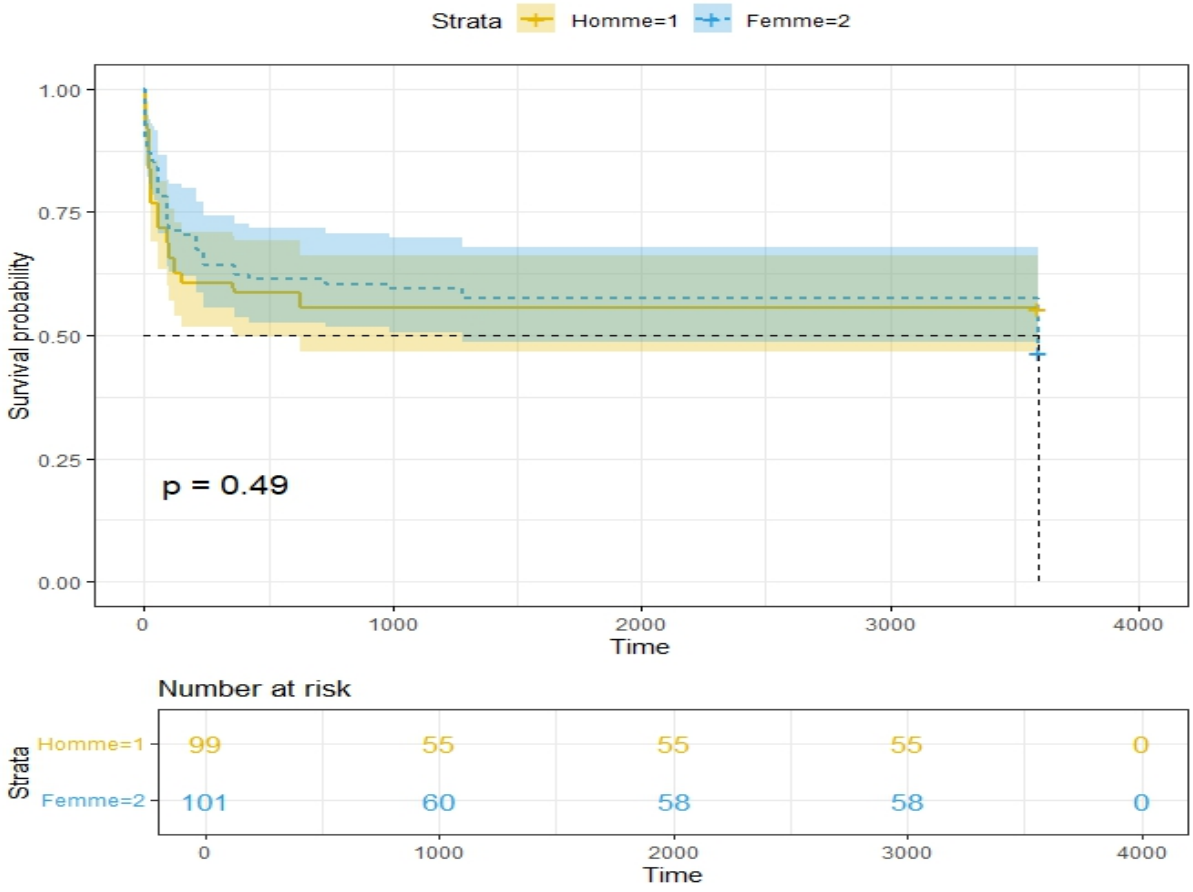
n	events	median	0.95LCL	0.95UCL
200	98	NA	3600	NA

L'analyse de survie réalisée avec la méthode de Kaplan-Meier montre l'évolution de la probabilité de survie au fil du temps pour un nombre de 200 patients suivis dont **98 décès** ont été observés au cours du suivi sur une période de 3600 jours.
Graphiquement la courbe montre une diminution rapide de la survie au cours des premiers mois . Nous allons faire un bref résumé donné par la fonction summary appliquée au modele.

- Au **debut** (2 jours), la probabilité de survie est d'environ **98,5 %**.
- À **un mois** (30 jours), elle diminue à **81,5%** marquée par plusieurs pics de décès, notamment autour du 7e et du 30e jour.
- À **2 mois** (60 jours): La survie est à 75%, avec un nombre important d'événements (11 décès).
- Après **1 an** (365 jours), la probabilité de survie est environ **60,5 %**.
- À la fin (3600 jours), la probabilité de survie atteint **51 %**, ce qui veut dire que la moitié des patients ont survécu jusqu'à la fin du suivi.

2. Estimer la fonction de survie pour les hommes et pour les femmes.

La figure ci-dessous présente la comparaison des courbes de survie de Kaplan-Meier pour deux groupes définis par le sexe (Homme=1 , Femme=2).



	n	events	median	0.95LCL	0.95UCL
SEXE=1	99	44	NA	365	NA
SEXE=2	101	54	3600	1280	NA

3. Comparaison des deux fonctions de survie (hommes et femmes).

Au debut de l'étude, il y'a 99 hommes et 101 femmes et le nombre de patient diminue avec le temps. Nous allons faire l'analyse graphique et théorique.

De façon graphique :

Visuellement, les courbes de survie pour les hommes et les femmes sont proches l'une de l'autre avec des intervalles de confiance confondue (potentiellement absence de significativité), mais la courbe des femmes est légèrement supérieure à celle des hommes ce qui veut dire nous avons des taux de survie similaires entre les deux groupes. On peut dire aussi que les femmes semblent avoir plus de chance de survie que les hommes.
Le tableau sous le graphique montre le nombre de patients à risque à différents moments. On remarque que le nombre de patients à risque diminue avec le temps dans les deux groupes. À la fin du suivi, il n'y a plus de patients à risque.

De façon théorique :

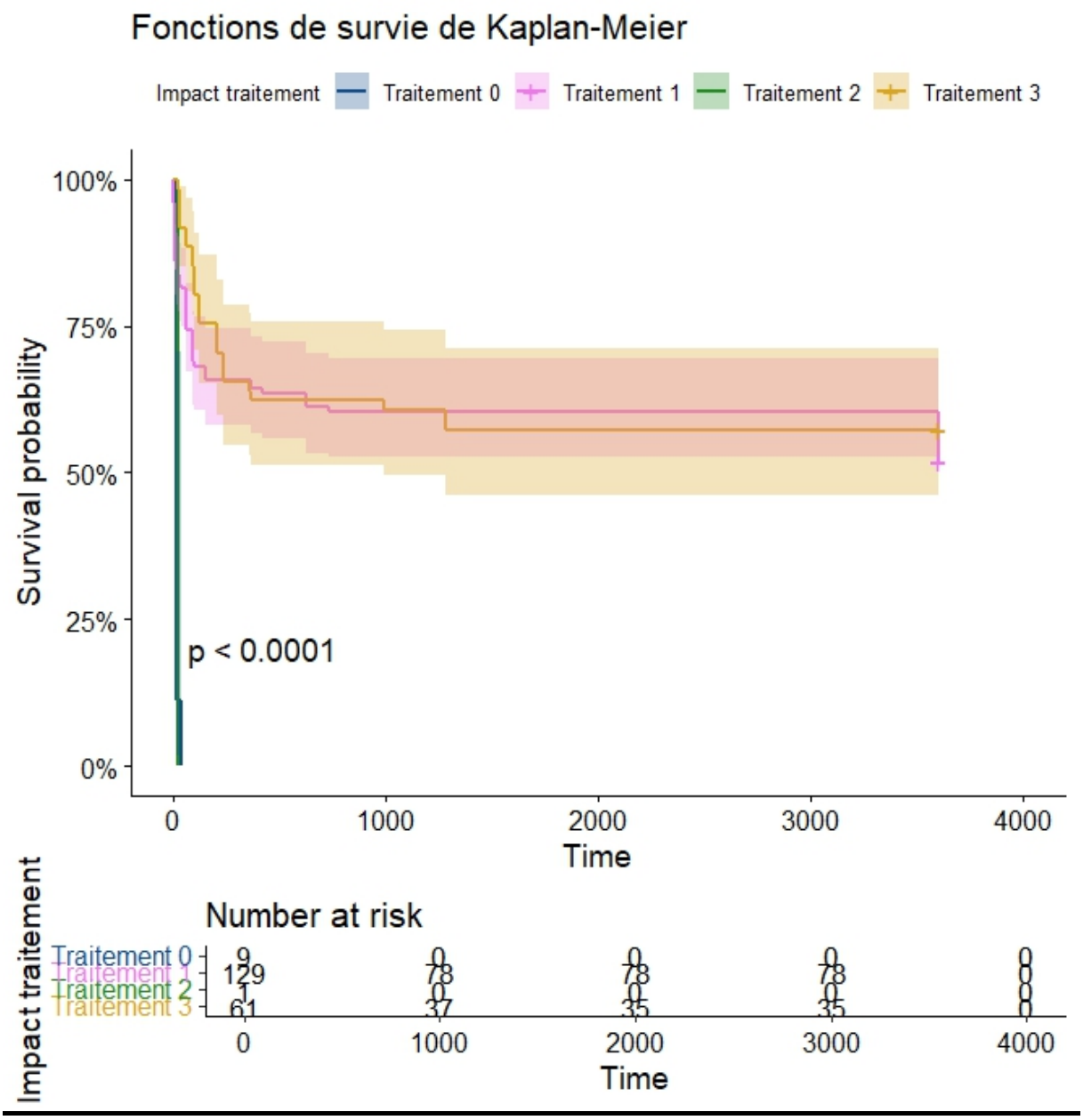
Nous allons faire le test de log-rank.

Résultat du test :

Effectifs et Événements Observés
SEXE=1 (Hommes) : 99 patients, 44 décès observés
SEXE=2 (Femmes) : 101 patients, 54 décès observés
chisq = 0.5
p-value=0.5

Conclusion et interprétation : Le resultat du test log-rank nous donne une p-value largement supérieure à 0,05 alors nous ne pouvons pas rejetter l'hypothèse nulle selon laquelle il n'y a pas de différence statistiquement significative dans les courbes de survie entre les hommes et les femmes.

4. **Estimer la fonction de survie pour les différents types de traitement.**



	n	events	median	0.95LCL	0.95UCL
Traitement=0	9	9	19	19	NA
Traitement=1	129	62	NA	3600	NA
Traitement=2	1	1	26	NA	NA
Traitement=3	61	26	NA	990	NA

5. Comparer les trois fonctions de survie (pour les différents types de traitement)

Le graphique de Kaplan-Meier compare les courbes de survie de quatre groupes de traitement différents (Traitement 0, Traitement 1, Traitement 2 et Traitement 3).

De façon graphique

Visuellement les fonctions de survie de Kaplan-Meier montrent des différences entre les groupes de traitement. Le groupe Traitement 0 présente une survie extrêmement faible, avec une chute rapide à 0, ce qui suggère un impact négatif significatif. Les groupes Traitement 1 et Traitement 3 montre une survie plus élevée et plus progressive, indiquant des taux de survie supérieurs. Le Traitement 2 a une très petite taille d'échantillon, donc on ne peut pas l'interpréter d'une maniere fiable. On voit aussi que le nombre de patients à risque diminue avec le temps, mais cette diminution est beaucoup plus rapide pour le Traitement 0.

De façon théorique

Nous allons faire le test de log-rank

Résultat du test :

Effectifs et Événements Observés
Traitement=0 : 9 patients, 9 décès observés
Traitement=1 : 129 patients, 62 décès observés
Traitement=2 : 1 patients, 1 décès observés
Traitement=3 : 61 patients, 26 décès observés
chisq = 52.1
p-value=0.0001

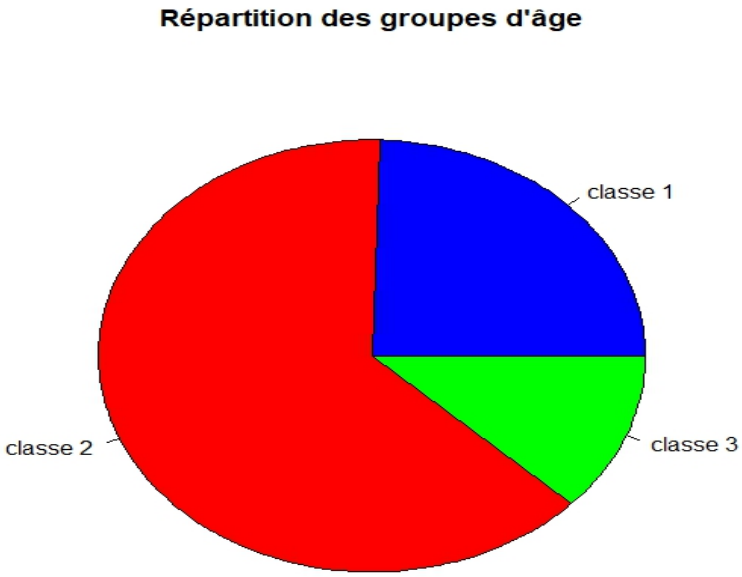
Conclusion et interprétation :

D'aprés l'analyse graphique et les résultats du test on peut en conclure qu'il existe une différence statistiquement significative entre les courbes de survie entre les groupes de traitement ($p < 0.0001$). Le Traitement 0 est inefficace, quant à les Traitements 1 et 3 montrent une survie significativement meilleure, bien qu'il puisse y avoir une certaine incertitude.On note une différence entre ces deux traitements car il existe un chevauchement sur les intervalles de confiance. Le Traitement 2, en raison de sa très petite taille d'échantillon, on ne peut pas de tirer des conclusions. Donc les Traitements 1 et 3 sont préférables au Traitement 0 et 2.

6. **Décomposer la variable « AGE »interpréter le graphe.**

Classe	Tranche d'âge	Effectif (n)	Proportion (%)
Classe 1	25-45 ans	49	24,5%
Classe 2	45-65 ans	127	63,5%
Classe 3	65-82 ans	24	12%

Diagramme en camembert

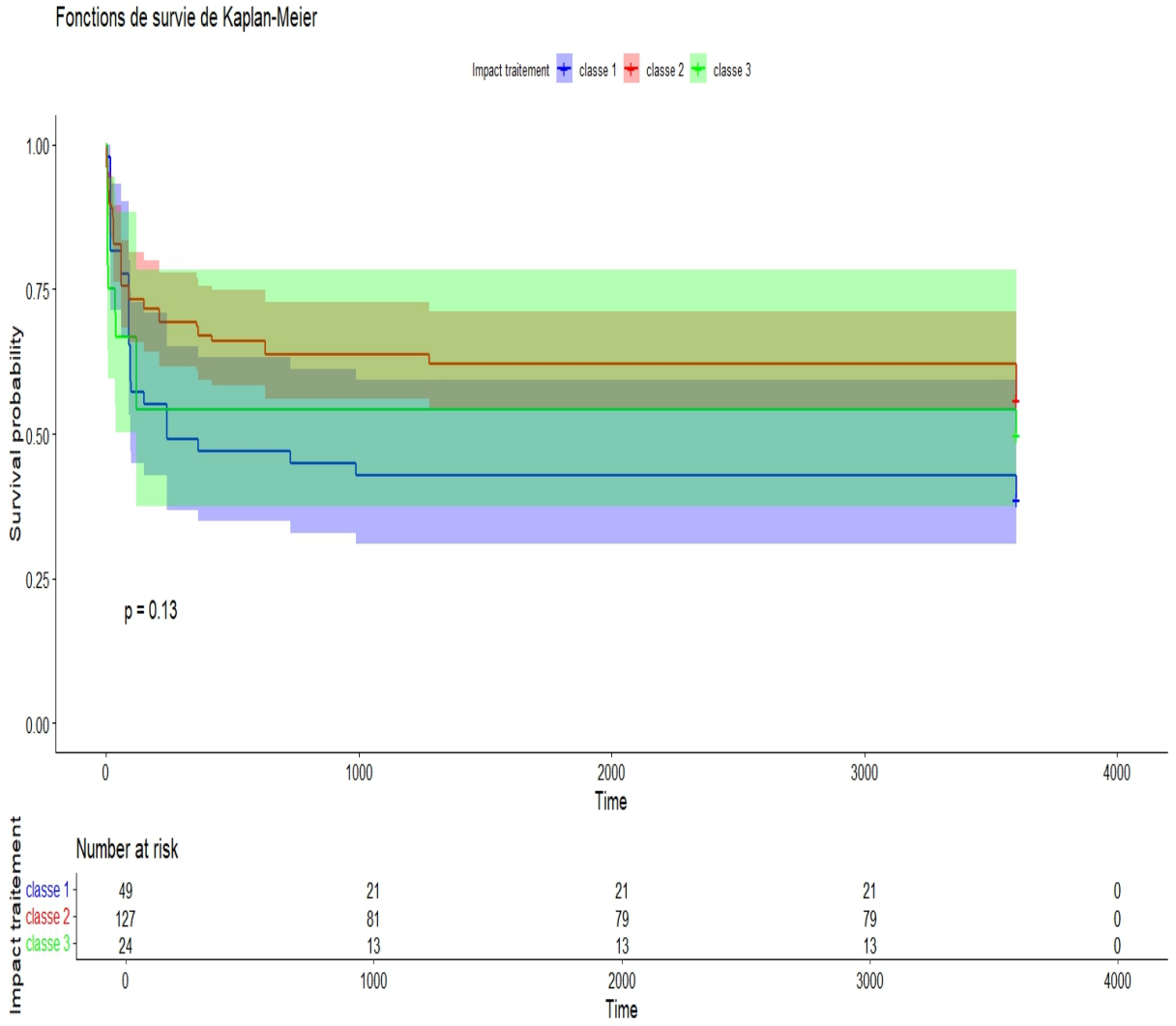


-

La classe 2 représente la part la plus importante de l'échantillon environ 63,5 %, suivie par la classe 1 environ 24%, et enfin la classe 3 qui est la moins représentée environ 12%.

7. **Estimer la fonction de survie pour les différentes classes d'âge obtenues.**

Le graphique de Kaplan-Meier suivant compare les courbes de survie pour les trois classes d'âge : classe 1 (bleu), classe 2 (rouge) et classe 3 (vert).



	n	events	median	0.95LCL	0.95UCL
classe_age=classe 1	49	30	240	98	NA
classe_age=classe 2	127	56	NA	3600	NA
classe_age=classe 3	24	12	3600	120	NA

8. Comparer les trois fonctions de survie (pour les différentes classes d'âge)

De façon graphique :

Visuellement les fonctions de survie de Kaplan-Meier pour les classes d'âge révèle des trajectoires de survie distinctes.
La classe 1 montre la probabilité de survie la plus faible par rapport aux autres classes.
La classe 2 montre la survie la plus élevée tout au long de la période.
La classe 3 présente une survie intermédiaire entre les deux autres classes.
On note aussi une réduction importante des effectifs à 1000 jours dans toute les classe d'age.
À 3000 jours: 21 patients restants en classe 1, 79 en classe 2, et 13 en classe 3.

De façon théorique.

Nous allons faire le test de log-rank

Résultat du test :

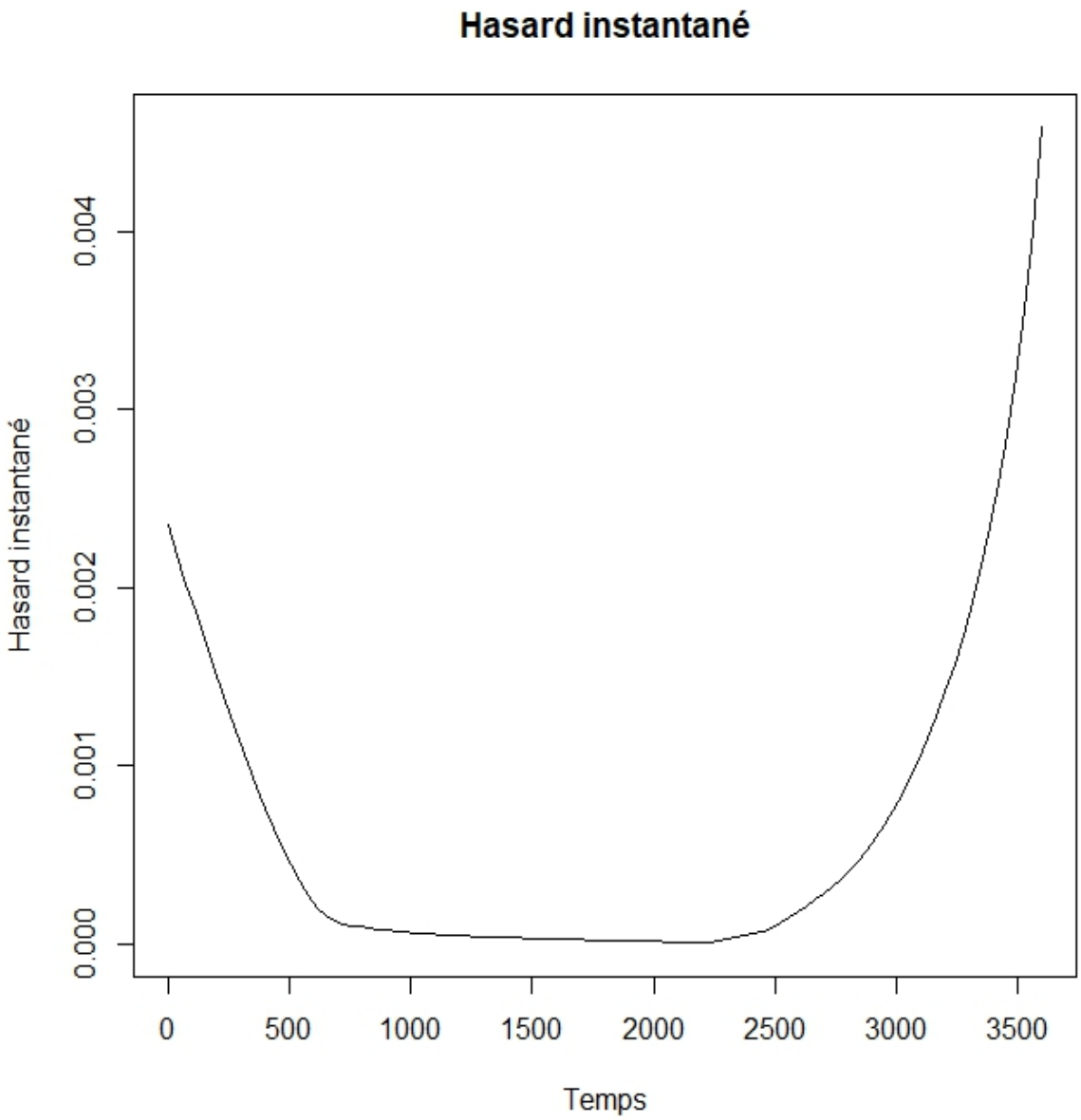
Effectifs et Événements Observés
Classe 1 : 49 patients, 30 décès observés
Classe 2 : 127 patients, 56 décès observés
Classe 3 : 24 patients, 12 décès observés
chisq = 4.1
p-value=0.13

Conclusion et interprétation.

La valeur de p = 0.13 qui est supérieure à 0,05 indique qu'il n'y a pas de différence statistiquement significative entre les courbes de survie des trois classes d'âge bien qu'il semble y avoir des différences visuellement, mais on peut en dire que la classe 2 semble avoir une meilleure survie globale comparativement aux autres classes.

9. Estimer la fonction de hasard instantané et tracer son graphe

Le graphe suivant montre l'évolution du risque instantané de deces en fonction du temps.



Le risque instantané est élevé au début puis diminue rapidementet se stabilise à un niveau faible entre les temps 1 000 et 2 500, puis augmente fortement après le temps 2 500.

10. Ajustement de modèle de régression de Cox.

Modèle 1: effet du traitement et de l'âge

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
Traitement1	-2.189155	0.112011	0.402485	-5.439	5.36e-08	***
Traitement2	-0.201369	0.817611	1.060613	-0.190	0.849	
Traitement3	-2.372688	0.093230	0.429259	-5.527	3.25e-08	***
AGE	0.006795	1.006818	0.009452	0.719	0.472	

Le modèle identifie qu le traitement1 et le traitement 3 sont des facteur hautement protecteurs et réduisent le risque de décès de 89% et de 91% par rapport au traitement 2 (20%), Les traitements sont significatifs avec un $p < 0.001$ et RR <1 contrairement à l'age.

Modèle 2 : Interaction Traitement × age

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
Traitement1	-2.998840	0.049845	1.632384	-1.837	0.0662	.
Traitement2	-0.083714	0.919694	1.105596	-0.076	0.9396	
Traitement3	-2.554181	0.077756	1.855769	-1.376	0.1687	
AGE	-0.004060	0.995948	0.027915	-0.145	0.8844	
Traitement1:AGE	0.015192	1.015308	0.030245	0.502	0.6155	
Traitement2:AGE	NA	NA	0.000000	NA	NA	
Traitement3:AGE	0.003199	1.003204	0.035208	0.091	0.9276	

Les traitements ont des hasard ratio inferieur à 1 on peut dire que les traitement sont des facteurs protecteurs mais ils ne sont pas significatifs car aucune p-value ne dépasse le seuil de significativité (5%).

Modèle 3: effet du traitement

	coef	exp(coef)	se(coef)		Pr(>|z|)
Traitement	-0.2502	0.7786	0.1165	-2.148	0.0317 *

Le traitement montre un effet protecteur et réduit significativement le risque de décès de 22 % (HR = 0,7786, p = 0,0317).

Modèle 4: effet des antécedents familliaux

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
AntFamil	1.5001	4.4822	0.2552	5.878	4.15e-09 ***

Les antécédents familiaux un impact significatif sur la durée de survie et multiplient le risque de décès . Les individus ayant des antécédents familiaux de la maladie ont un risque de décès environ 4.48 fois plus élevé que ceux qui n'en ont pas.

Modèle 5: effet du traitement, de l'âge et leur interaction

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
Traitement	-0.121577	0.885523	0.625996	-0.194	0.846
AGE	0.008107	1.008140	0.018500	0.438	0.661
Traitement:AGE	-0.002337	0.997666	0.011968	-0.195	0.845

Ni le traitement ,ni l'âge, ni leur interaction a un effet significatif sur la survie bien que l'age est un facteur de risque (HR > 1) ce qui peut augmenter le risque de deces.

Modèle 6: effet de l'âge ,du traitement et de l'antècedent famillial

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
AGE	0.007901	1.007932	0.008895	0.888	0.374
Traitement	-0.217410	0.804600	0.111423	-1.951	0.051 .
AntFamil	1.499108	4.477693	0.256510	5.844	5.09e-09 ***

Seul le traitement montre un effet protecteur (HR<1) et que son impact est presque significatif (p = 0,051). Les antécédents familiaux multiplient par 4.478 le risque de décès.

L'age aussi est un facteur risque maiss son impact n'est pas significatif (p-value = 0,374).

Comparaison des modeles

Nous allons faire un test d'anova entre les modèles pour la sélection du modele en fonction de sa complexité et en fonction de son critère d'information d'Akaike (AIC).

Résultat du test de l'anova.

Comparaison	combinaison	combinaison	p-value	Conclusion
model3 vs model1	Traitement	Traitement + AGE	0.000173 ***	L'age améliore significativement le modèle
model1 vs model2	Traitement + AGE	Traitement * AGE	0.8072	Traitement×AGE n'améliore pas le modèle
model3 vs model5	Traitement	Traitement + AGE + Traitement:AGE	0.8539	l'age et l'interaction n'ameliore pas le modèle
model5 vs model6	Traitement + AGE + Traitement:AGE	AGE + Traitement + AntFamil	<2.2e-16 ***	AntFamil améliore très significativement le modèle

Modèle	AIC
Modèle 1	966.8
Modèle 2	970.3715
Modèle 3	980.7602
Modèle 4	959.4222
Modèle 5	984.4443
Modèle 6	958.3907

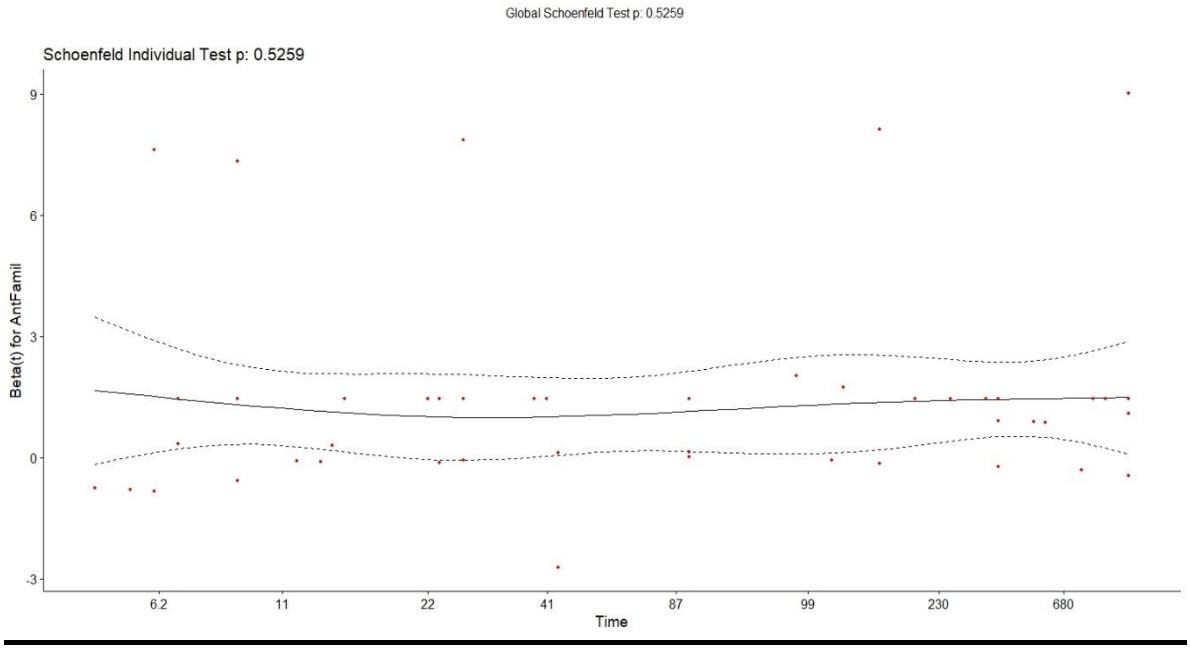
Le modèle 6 combinant l'âge ,le traitement et les antécédents familiaux est retenue comme un bon modèle car il présente le plus faible AIC mais celui-ci ne verifie pas les hypothèses proportionnalité des risques relatifs. Ce pendant nous avons stratifier la variable age et traitement. L'lajout de ces dernieres a améliorer significativement le modèle avec un AIC plus optimale (617.1486) ainsi que la verification des hypothèses.

Les résultats somt présentés ci-dessous.

11. Etude de l'adéquation du modèle final obtenu.

Le tableau suivant montre les resultats du test de cox.zph appliqués sur le modèle final.

Variable	chisq	p-value
AntFamil	0,402	0,53
Global	0,402	0,53



12. Interpréter les résultats de cette modélisation.

En regardant le graphe et le resultat du le test de cox.zph, il est clair que la p-value est supérieur à 0,05 globalement et que la ligne centrale est stable d'où l'hypothèse de proportionnalité des risques est respectée pour la covariable AntFamil. Alors notre modèle est très bien ajusté aux données.

Annexe

```
#############################################################################
### Packages
library(readxl)
library(survival)
library(CoxR2)
library(timereg)
library(survminer)
library(stringi)
library(muhaz)
library(knitr)
library(Hmisc)
library(ggplot2)
library(plotly)
### importation des données
donnee <- read_excel("C:/Users/Hp/Desktop/Master 2/Donnee censurée/Base_Projet.xlsx")
head(donnee)
View(donnee)
attach(donnee)
### Reconversion des donnees
Traitement=as.factor(Traitement)
SEXE = as.factor(SEXE)
### vérifier les types
str(donnee)
### statistiques descriptives
summary(donnee)
kable(summary(donnee))
describe(donnee)
### variable sexe
describe(SEXE)
describe(donnee)
#############################################################################################
#
###Visualisation
### Pour la variable durrée de survie
# Boxplot et Histogramme avec Plotly
p1 <- plot_ly(data = donnee,
              x = ~AGE,
              color = ~SEXE,
              type = "box") %>%
  layout(title = "Boxplot de l'âge en fonction du sexe",
         xaxis = list(title = "Âge"),
         yaxis = list(title = "Sexe"))
p2 <- plot_ly(data = donnee,
              x = ~AGE,
              type = "histogram",
              marker = list(color = 'skyblue')) %>%
  layout(title = "Histogramme de l'âge",
         xaxis = list(title = "Âge"),
         yaxis = list(title = "Fréquence"))
fig1 <- subplot(p1, p2, nrows = 2, margin = 0.05, heights = c(0.5, 0.5)) %>%
  layout(title = "Distribution de l'âge combinée")
fig1

### Pour la variable durrée de survie
p3 <- plot_ly(data = donnee,
```

```
            x = ~DureeSurvieJr,
            type = "box") %>%
  layout(title = "Boxplot de la durée de survie",
         xaxis = list(title = "Âge"),
         yaxis = list(title = "Sexe"))
p4 <- plot_ly(data = donnee,
            x = ~DureeSurvieJr,
            type = "histogram",
            marker = list(color = 'skyblue')) %>%
  layout(title = "Histogramme de la durée de survie",
         xaxis = list(title = "Âge"),
         yaxis = list(title = "Fréquence"))
fig2 <- subplot(p3, p4, nrows = 2, margin = 0.05, heights = c(0.5, 0.5)) %>%
  layout(title = "Distribution de la durée de survie")
fig2

##############################################################################
#
### 1)la fonction de survie de Kaplan Meier avec les intervalles de confiance.
fit1 <- survfit(Surv(DureeSurvieJr, DECES) ~ 1, data = donnee)
fit1
summary(fit1)
x11()
ggsurvplot(fit1,
          data = donnee,
          conf.int = TRUE,
          pval = TRUE,
          risk.table = TRUE,
          legend.title = "Impact de la maladie",
          title = "Fonctions de survie de Kaplan-Meier",
          risk.table.height = 0.2,
          ggtheme = theme_bw())

##############################################################################
#
### 2) Estimer la fonction de survie pour les hommes et pour les femmes.
# risque cumule pour deux groupes
km_sexe <- survfit(Surv(DureeSurvieJr, DECES) ~ SEXE, data = donnee)
km_sexe
summary(km_sexe)
## Comparaison de fonction de survie par sexe.
ggsurvplot(km_sexe,
          pval = TRUE, conf.int = TRUE,
          risk.table = TRUE,
          risk.table.col = "strata",
          linetype = "strata",
          surv.median.line = "hv",
          ggtheme = theme_bw(),
          legend.labs=c("Homme=1","Femme=2"),
          palette = c("#E7B800", "#2E9FDF"))
#### test de comparaison de courbe de survie.
test.survie = survdiff(Surv(DureeSurvieJr, DECES) ~ SEXE, data = donnee)
test.survie
##############################################################################
#
### 4) Estimer la fonction de survie pour les différents types de traitement.
km_traitement <- survfit(Surv(DureeSurvieJr, DECES) ~ Traitement, data = donnee)
```

```
km_traitement
## Comparaison de fonction de survie par sexe.
x11()
ggsurvplot(km_traitement,
          conf.int = TRUE,
          pval = TRUE,
          risk.table = TRUE,
          surv.scale = "percent",
          legend.labs = c("Traitement 0", "Traitement 1", "Traitement 2", "Traitement 3"),
          legend.title = "Impact traitement",
          palette = c("dodgerblue4", "orchid2", "forestgreen", "goldenrod"), # 4 couleurs
          title = "Fonctions de survie de Kaplan-Meier",
          risk.table.height = .2)
### 5)  Comparer les trois fonctions de survie (pour les différents types de traitement) de
façon
#graphique et de façon théorique. Interpréter les résultats.
test.survie_traitement =  survdiff(Surv(DureeSurvieJr, DECES) ~ Traitement, data = donnee)
test.survie_traitement
################################################################################
#
###  6) Décomposer la variable « AGE » en trois classes
classe_age = cut(AGE, breaks=c(25,45,65,82))
levels(classe_age) = c("classe 1","classe 2","classe 3")
eff_age = table(classe_age)
eff_age
### proportion
prop = prop.table(table(classe_age))
prop
###  diagramme en camembert
x11()
pie(prop,
   main = "Répartition des groupes d'âge",
   col = c("blue", "red", "green"), # Ajustez selon le nombre de groupes
   border = "black")
################################################################################
#
### 7) Estimer la fonction de survie pour les différentes classes d'âge obtenues.
# estimation des fonctions de survie
km_age = survfit(Surv(DureeSurvieJr, DECES)~classe_age,data=donnee)
km_age
summary(km_age)
################################################################################
#
### 8) Comparaison des trois fonctions de survie (graphique et théorique) .
ggsurvplot(km_age,
          conf.int = TRUE,
          pval = TRUE,
          risk.table = TRUE,
          legend.labs = c("classe 1", "classe 2", "classe 3"),
          legend.title = "Impact traitement",
          palette = c("blue", "red", "green"),
          title = "Fonctions de survie de Kaplan-Meier",
          risk.table.height = .2)
### test
test.survie_age =  survdiff(Surv(DureeSurvieJr, DECES) ~ classe_age, data = donnee)
test.survie_age
```

```
##############################################################################
####
### 9) Estimer la fonction de hasard instantané et tracer son graphe.
hasard1 = muhaz(DureeSurvieJr,DECES,min.time = 2,max.time = 3600)
hasard1
# Tracer les graphiques
plot(hasard,
    main = "Hasard instantané",
    xlab = "Temps",
    ylab = "Hasard instantané")

##############################################################################
#
### 10) Ajuster sur les données un modèle de régression de Cox.
### modele 1: effet du traitement et de l'âge
model1 = coxph(Surv(DureeSurvieJr, DECES)~Traitement+AGE, data=ovarian,
method="breslow")
summary(model1)
coxr2(model1)
### Modèle 2 : Interaction Traitement × age
model2 = coxph(Surv(DureeSurvieJr, DECES)~Traitement*AGE, data=ovarian,
method="breslow")
summary(model2)
coxr2(model2)
### Modele 3: effet du traitement
model3 = coxph(Surv(DureeSurvieJr, DECES)~ Traitement, data=donnee, method="breslow")
summary(model3)
coxr2(model3)
### Modele 4: effet des antécedents familliaux
model4 = coxph(Surv(DureeSurvieJr, DECES)~AntFamil, data=donnee, method="breslow")
summary(model4)
coxr2(model4)
### Modele 5: effet du traitement, de l'âge et leur interaction
model5 = coxph(Surv(DureeSurvieJr, DECES)~Traitement+AGE+Traitement:AGE,data=donnee,
method="breslow")
summary(model5)
coxr2(model5)
### Modele 6: effet del'âge du traitement et de l'antecedent familliale
model6 = coxph(Surv(DureeSurvieJr, DECES)~AGE+ Traitement +AntFamil, data = donnee,
method = "breslow")
summary(model6)
coxr2(model6)
### Test de l'Anova
anova(model3, model1)
anova(model1, model2)
anova(model3, model5)
anova(model5, model6)
### Comparaison des modeles
AIC(model1)
AIC(model2)
AIC(model3)
AIC(model4)
AIC(model5)
AIC(model6)
##############################################################################
#
### 11) Etudier l’adéquation du modèle final obtenu.
```

```
model_final = model6
### Test des résidus de Schoenfeld
test_zph <- cox.zph(model_final)
test_zph
par(mfrow = c(2,3))
### graphe des residus
ggcoxzph(test_zph)
#################################################################################
#
### Modèle avec ajustement dépendant du temps
model_corrige <- coxph(
  Surv(DureeSurvieJr, DECES) ~ strata(Traitement, classe_age) + AntFamil,
  data = donnee)
### Comparaison AIC
AIC(model6, model_corrige)
### test des residus de Schoenfeld pour le modele corrigé
test_zph <- cox.zph(model_corrige)
test_zph
par(mfrow = c(2,3))
### Graphe des residus
ggcoxzph(test_zph)
```