[Title Here, up to 12 Words, on One to Two Lines]

[Author Name(s), First M. Last, Omit Titles and Degrees]

[Institutional Affiliation(s)]

Author Note

[Include any grant/funding information and a complete correspondence address.]

Abstract

[The abstract should be one paragraph of between 150 and 250 words. It is not indented. Section titles, such as the word Abstract above, are not considered headings so they don’t use bold heading format. Instead, use the Section Title style. This style automatically starts your section on a new page, so you don’t have to add page breaks. Note that all of the styles for this template are available on the Home tab of the ribbon, in the Styles gallery.]

Keywords: [Click here to add keywords.]

[Title Here, up to 12 Words, on One to Two Lines]

[The body of your paper uses a half-inch first line indent and is double-spaced. APA style provides for up to five heading levels, shown in the paragraphs that follow. Note that the word Introduction should not be used as an initial heading, as it’s assumed that your paper begins with an introduction.]

# [Heading 1]

[The first two heading levels get their own paragraph, as shown here. Headings 3, 4, and 5 are run-in headings used at the beginning of the paragraph.]

## [Heading 2]1

[To add a table of contents (TOC), apply the appropriate heading style to just the heading text at the start of a paragraph and it will show up in your TOC. To do this, select the text for your heading. Then, on the Home tab, in the Styles gallery, click the style you need.]

[Heading 3]. [Include a period at the end of a run-in heading. Note that you can include consecutive paragraphs with their own headings, where appropriate.]

[Heading 4]. [When using headings, don’t skip levels. If you need a heading 3, 4, or 5 with no text following it before the next heading, just add a period at the end of the heading and then start a new paragraph for the subheading and its text.] (Last Name, Year)

[Heading 5]. [Like all sections of your paper, references start on their own page. The references page that follows is created using the Citations & Bibliography feature, available on the References tab. This feature includes a style option that formats your references for APA 6th Edition. You can also use this feature to add in-text citations that are linked to your source, such as those shown at the end of this paragraph and the preceding paragraph. To customize a citation, right-click it and then click Edit Citation.] (Last Name, Year)

References

Last Name, F. M. (Year). Article Title. *Journal Title*, Pages From - To.

Last Name, F. M. (Year). *Book Title.* City Name: Publisher Name.

Footnotes

1[Add footnotes, if any, on their own page following references. For APA formatting requirements, it’s easy to just type your own footnote references and notes. To format a footnote reference, select the number and then, on the Home tab, in the Styles gallery, click Footnote Reference. The body of a footnote, such as this example, uses the Normal text style. (Note: If you delete this sample footnote, don’t forget to delete its in-text reference as well. That’s at the end of the sample Heading 2 paragraph on the first page of body content in this template.)]

Tables

Table 1

[Table Title]

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Column Head | Column Head | Column Head | Column Head | Column Head |
| Row Head | 123 | 123 | 123 | 123 |
| Row Head | 456 | 456 | 456 | 456 |
| Row Head | 789 | 789 | 789 | 789 |
| Row Head | 123 | 123 | 123 | 123 |
| Row Head | 456 | 456 | 456 | 456 |
| Row Head | 789 | 789 | 789 | 789 |

Note: [Place all tables for your paper in a tables section, following references (and, if applicable, footnotes). Start a new page for each table, include a table number and table title for each, as shown on this page. All explanatory text appears in a table note that follows the table, such as this one. Use the Table/Figure style, available on the Home tab, in the Styles gallery, to get the spacing between table and note. Tables in APA format can use single or 1.5 line spacing. Include a heading for every row and column, even if the content seems obvious. A default table style has been setup for this template that fits APA guidelines. To insert a table, on the Insert tab, click Table.]

Figures

Figure 1. [Include all figures in their own section, following references (and footnotes and tables, if applicable). Include a numbered caption for each figure. Use the Table/Figure style for easy spacing between figure and caption.]

For more information about all elements of APA formatting, please consult the APA Style Manual, 6th Edition.

# STA215

#### Sebastien DIAZ

#### 13 mars 2016

## Introduction

Le projet qui est exposé dans ce document porte sur les délais de grossesse, dont une partie est censurée. Le contenu métier de ces données n’est pas très bien connu et pourra paraitre déroutant pour un non expert du domaine. L’étude sera faite sans poser d’apriori sur les événements arrivant pendant la grossesse.  
Le début de l’étude essaiera d’analyser les données selon des méthodes classiques puis par des méthodes adaptées a ce genre de données de délai avec censure. Nous verrons de surprenantes spécificités lors de cette analyse.  
Nous continuerons ensuite par essayer de trouver un modèle adéquate à nos données. Deux types de modèle seront étudiés : un modèle paramétrique et un modèle à risque proportionnelle. Ces deux modèles quoique travaillant sur des aspects différents permettrons de comprendre qu’ils sont très similaires dans leur composition. Pour simplifier les modèles, il sera effectué une sélection de variable.

## Description des données

Les données décrivent le délai de grossesse selon les co-variables suivantes :

* d2g : delai de grossesse en jours
* indic : 1 = grossesse , 0 = censure
* bmiF : bmi de la femme
* bmiH : bmi de l’homme
* alcF : nombre de boisson alcolisée par semaine chez la femme
* alcH : nombre de boisson alcolisée par semaine chez l’homme
* fumF : 1 = fumeuse, 0= pas fumeuse
* fumH : 1 = fumeur, 0= pas fumeur
* ageF : age en année
* ageH : age en année
* sperm : spermatozoïdes en million
* testo : niveau de testosterone

## Chargement de la librairie survival

library(survival)

## Chargement des données

Les données sont chargées et on corrige les données manquante ou les type de données comme numérique ou facteur.

d2g<-read.csv("d2g.txt", header = TRUE, sep = " ",na.strings = "<NA>")

d2g[d2g$sperm == "NA",]$sperm=NA

d2g$sperm=as.numeric(d2g$sperm)

d2g[d2g$testo == "NA",]$testo=NA

d2g$fumF=as.factor(d2g$fumF)

d2g$fumH=as.factor(d2g$fumH)

d2g$testo=as.numeric(gsub("NA", "", d2g$testo))

## Sommaire statistique des variables

nrow(d2g)

## [1] 423

summary(d2g)

## id d2g indic bmiF

## Min. : 1.0 Min. : 15.31 Min. :0.0000 Min. :-1111.11

## 1st Qu.:106.5 1st Qu.: 62.92 1st Qu.:0.0000 1st Qu.: 20.31

## Median :212.0 Median :119.97 Median :1.0000 Median : 21.89

## Mean :212.0 Mean :116.46 Mean :0.6052 Mean : 17.31

## 3rd Qu.:317.5 3rd Qu.:174.43 3rd Qu.:1.0000 3rd Qu.: 24.34

## Max. :423.0 Max. :194.65 Max. :1.0000 Max. : 37.64

##

## bmiH alcF alcH fumF fumH

## Min. :-1111.11 Min. : 0.000 Min. : 0.000 0:300 0:290

## 1st Qu.: 22.53 1st Qu.: 0.000 1st Qu.: 3.000 1:123 1:133

## Median : 24.11 Median : 2.000 Median : 7.000

## Mean : 19.08 Mean : 3.995 Mean : 9.416

## 3rd Qu.: 25.83 3rd Qu.: 6.000 3rd Qu.:13.000

## Max. : 38.57 Max. :39.000 Max. :84.000

##

## ageF ageH sperm testo

## Min. :19.82 Min. :18.33 Min. : 1.00 Min. :0.000

## 1st Qu.:24.15 1st Qu.:26.18 1st Qu.: 52.25 1st Qu.:0.800

## Median :25.95 Median :27.90 Median : 89.00 Median :1.100

## Mean :26.10 Mean :28.20 Mean : 85.31 Mean :1.206

## 3rd Qu.:27.87 3rd Qu.:30.06 3rd Qu.:120.00 3rd Qu.:1.400

## Max. :35.19 Max. :37.19 Max. :156.00 Max. :2.900

## NA's :113 NA's :388

Le délai de grossesse (d2g) s’étale de 15 à 195 jours. Il y a 40% de censure (indic). Nous avons certainement à faire a une population plutot a risque de grossese prématuré. Le délai de grosses normal étant de 9 mois,plus de 260 jours.

Le bmi des femmes et des hommes contient une incohérence sur la valeur minimum -1111.11 qui est bien entendu impossible et doit représenté une donnée manquante.

nrow(d2g[d2g$bmiF<0,])

## [1] 2

nrow(d2g[d2g$bmiH<0,])

## [1] 2

nrow(d2g[d2g$bmiH<0&d2g$bmiF<0,])

## [1] 2

Seulement deux indivdus sont concernés pour le bmi des femmes et des hommes. En les enlevement, le sommaire est:

summary(d2g[d2g$bmiF>0,]$bmiF)

## Min. 1st Qu. Median Mean 3rd Qu. Max.

## 15.24 20.32 21.94 22.67 24.34 37.64

summary(d2g[d2g$bmiH>0,]$bmiH)

## Min. 1st Qu. Median Mean 3rd Qu. Max.

## 18.20 22.53 24.15 24.44 25.83 38.57

Le bmi des hommes dans l’échantillon est plus important aussi sur la valeur minimum, la médiane, la moyenne ou la valeur maximale.

Le nombre de boisson alcoolisé (alcF et alcH) est très largement supérieur pour les hommes. Il y a dans les deux sexes des quantités minimum à 0. Les hommes peuvent consommer jusqu’a 84 boissons par semaine ce qui est très important.

Il y a 29 % de fumeur femme(fumF) et 31 % d’hommes(fumH) dans cet échantillon.

L’âge des hommes et des femmes est sensiblement comparable. L’âge minimum est de 20 ans pour les femmes et 18 pour les hommes. L’âge maximale est de 35 ans pour les femmes et 37 ans pour les hommes. La moyenne quant a elle est de 26 pour les femmes et 28 ans pour les hommes.

Les spermatozoïdes possèdent près de 113 values manquantes. Evoluant de 1 à 156 millions de spermatozoïdes, il y a en moyenne 85 millions de spermatozoïdes.

Les taux de testostérones sont particulièrement mal renseignés avec près de 388 valeurs manquantes soit près de 92% de l’échantillon.

La répartition des événements montrent que les censures sont plutot situé après 150 jours. Les événements de fin de grosesse sont eux plutot situé aux alentour de 75 jours. On peut dire que les événement et les censures sont significativement séparé, permettant de dire que les censures sont de type plutôt à droite et quels sont quasi systématiques aprés 150 jours.

## Gestion des données manquantes

On change les deux individus ayants des données de bmi manquante par NA.

d2g[d2g$bmiH<0,]$bmiH<-NA

d2g[d2g$bmiF<0,]$bmiF<-NA

On charge les libraries utiles par imputation multiple.

library(mice)

## Warning: package 'mice' was built under R version 3.2.4

## Loading required package: Rcpp

## mice 2.25 2015-11-09

library(lattice)

On lance l’algorithme d’imputation des données manquantes. Puis on impute

imp1<-mice(d2g[,c(2,4,5,6,7,10,11,12)],m=1,seed=500)

##

## iter imp variable

## 1 1 bmiF bmiH sperm

## 2 1 bmiF bmiH sperm

## 3 1 bmiF bmiH sperm

## 4 1 bmiF bmiH sperm

## 5 1 bmiF bmiH sperm

densityplot(imp1)

Comme il n’y pas que deux données d’imputation de données pour les variables bmi, on ne visualisera pas leur distribution.

Au contraire du nombre de spermatozoide, ou les variables manquantes sont tres frequentes. Lorsque l’on regarde la distribution des spermatozoides, les nouvelles données ont une distribution proche des données originales.

Puis, on impute les données sur notre jeux original.

comp<-complete(imp1, 1)

d2g$sperm<-comp$sperm

d2g$bmiH<-comp$bmiH

d2g$bmiF<-comp$bmiF

## Correlation des variables quantitatives

cor(d2g[,c(2,4,5,6,7,10,11,12)])

## d2g bmiF bmiH alcF alcH

## d2g 1.00000000 0.06519858 7.811374e-02 8.884950e-02 0.068360252

## bmiF 0.06519858 1.00000000 2.862884e-01 -9.211860e-02 -0.105893854

## bmiH 0.07811374 0.28628839 1.000000e+00 -1.158473e-05 0.001822953

## alcF 0.08884950 -0.09211860 -1.158473e-05 1.000000e+00 0.417793875

## alcH 0.06836025 -0.10589385 1.822953e-03 4.177939e-01 1.000000000

## ageF 0.03842423 -0.15551915 -3.609474e-02 1.686418e-01 -0.025267557

## ageH 0.03844639 -0.01636724 1.686170e-02 1.506552e-02 -0.030977262

## sperm -0.05598857 0.04659407 1.031938e-01 -7.803540e-02 -0.022789013

## ageF ageH sperm

## d2g 0.03842423 0.03844639 -0.05598857

## bmiF -0.15551915 -0.01636724 0.04659407

## bmiH -0.03609474 0.01686170 0.10319380

## alcF 0.16864185 0.01506552 -0.07803540

## alcH -0.02526756 -0.03097726 -0.02278901

## ageF 1.00000000 0.48114778 -0.03661488

## ageH 0.48114778 1.00000000 0.15150992

## sperm -0.03661488 0.15150992 1.00000000

Il n’y a aucune forte correlation entre nos variables. La corrélation la plus forte est sur le nombre de boisson alcoolisé par semaine entre les hommes et les femmes.

## Variables qualitatives

table(d2g$fumF,d2g$fumH)

##

## 0 1

## 0 241 59

## 1 49 74

La proportion de non fumeur est plus importante.

## Analyse des fonctions de survie et de risque cumulé

On commence par decoupé nos variables quantitatives en variables qualitative de facon équitable en prenant les quantiles (<25%,25%-50%,50%-75% et >75%).

d2g$bmiFQual <- cut(d2g$bmiF,breaks=quantile(d2g$bmiF),include.lowest=TRUE)

levels(d2g$bmiFQual)

## [1] "[15.2,20.3]" "(20.3,21.9]" "(21.9,24.3]" "(24.3,37.6]"

d2g$bmiHQual <- cut(d2g$bmiH,breaks=quantile(d2g$bmiH),include.lowest=TRUE)

levels(d2g$bmiHQual)

## [1] "[18.2,22.5]" "(22.5,24.1]" "(24.1,25.9]" "(25.9,38.6]"

d2g$alcFQual <- cut(d2g$alcF,breaks=unique(quantile(d2g$alcF)),include.lowest=TRUE)

levels(d2g$alcFQual)

## [1] "[0,2]" "(2,6]" "(6,39]"

d2g$alcHQual <- cut(d2g$alcH,breaks=quantile(d2g$alcH),include.lowest=TRUE)

levels(d2g$alcHQual)

## [1] "[0,3]" "(3,7]" "(7,13]" "(13,84]"

d2g$ageFQual <- cut(d2g$ageF,breaks=quantile(d2g$ageF),include.lowest=TRUE)

levels(d2g$ageFQual)

## [1] "[19.8,24.1]" "(24.1,25.9]" "(25.9,27.9]" "(27.9,35.2]"

d2g$ageHQual <- cut(d2g$ageH,breaks=quantile(d2g$ageH),include.lowest=TRUE)

levels(d2g$ageHQual)

## [1] "[18.3,26.2]" "(26.2,27.9]" "(27.9,30.1]" "(30.1,37.2]"

d2g$spermQual <- cut(d2g$sperm,breaks=quantile(d2g$sperm),include.lowest=TRUE)

levels(d2g$spermQual)

## [1] "[1,54]" "(54,89]" "(89,120]" "(120,156]"

fit <- survfit(Surv(d2g, indic) ~ 1, data = d2g)

plot(fit )

title("Fonction de survie")

plot(fit, fun="cloglog")

title("Risque cumulé")

La courbe de survie décroit régulierement en fonction du nombre de jour. Notre fonction de survie s’arrete autour de 200 jours. Les censures du jeu de données sont plus nombreuses à partir de 150 jours. Eb dessous de 50 jours de grosesses, les événements sont font plus rare. Le risque instantané, traduit le risque de présenter l’événement sur un intervalle de temps infinitésimal, conditionnellement au fait de ne pas l’avoir présenté auparavant. On représente ce risque par la courbe de risque cumulé. Le risque cumulé commence par croitre fortement. Ce qui indique que les risques instannées d’interruption de grosses est extremement importante. Puis, la courbe montre rapidement une stabilisation ascendante permettant de constater que le risque instantanné ce stabilise à une valeur constante plus faible.

colList<-c("red","blue","green","black")

#bmiF

fit <- survfit(Surv(d2g, indic) ~ bmiFQual, data = d2g)

plot(fit, lty = 2:3,col=colList)

title("Fonction de survie - BMI Femme")

legend(130, .97, levels(d2g$bmiFQual), lty = 2:3,col=colList)

plot(fit, lty = 2:3,col=colList,fun="cloglog")

title("Risque cumulé - BMI Femme")

legend(110, -1.17, levels(d2g$bmiFQual), lty = 2:3,col=colList)

Jusqu’a 100 jours et quelque soit l’indice de masse corporelle, il n’y a pas grande différence. Les courbes sont assez proche. Après 100 jours de grosses, les femmes, dont le BMI est plus important ont tendance à avoir une survie plus longue. Au vu de la courbe des risque cumulé, on ne peut tenter un test du log rank pour voir si les courbes de survie sont identiques.

#bmiH

fit <- survfit(Surv(d2g, indic) ~ bmiHQual, data = d2g)

plot(fit, lty = 2:3,col=colList)

title("Fonction de survie - BMI Homme")

legend(130, .97, levels(d2g$bmiHQual), lty = 2:3,col=colList)

plot(fit, lty = 2:3,col=colList,fun="cloglog")

title("Risque cumulé - BMI Homme")

legend(110, -1.17, levels(d2g$bmiHQual), lty = 2:3,col=colList)

La fonction met en évidence que le l’indice de masse corporel moyen [22-24] chez l’homme a une survie légèrement plus forte. Cependant, ce constat est plus nuancé en regardant la courbe des risques cumulés.Les courbes restantes sont paralleles et ne presume pas d’une forte différence. Les courbes de risque cumulés qui sont très chahuté,ne sont pas indiqué pour tester l’égalité des courbes de survie.

#alcF

fit <- survfit(Surv(d2g, indic) ~ alcFQual, data = d2g)

plot(fit, lty = 2:3,col=colList)

title("Fonction de survie - Alcool Femme")

legend(130, .97, levels(d2g$alcFQual), lty = 2:3,col=colList)

plot(fit, lty = 2:3,col=colList,fun="cloglog")

title("Risque cumulé - Alcool Femme")

legend(120, -1.17,levels(d2g$alcFQual), lty = 2:3,col=colList)

On s’appercoit que les femmes ayant une alcoolemie plus forte (>6 verres), on une courbe se survie pendant la grosesse plus importante. Cela se traduit sur la courbe des risques cumulés par un risque instantanné plus faible aux alentours de 50 jours. Après 120 jours, les risques instantannés croient plus rapidement que les autres consomations. Une l’acoolemie inférieure a six verres par semaine conduit à une coubre plus faible. La courbe de risque cumulé fait apparaitre une non proportionnalité des courbes du nottament aux croisements entre les variables.

#alcH

fit <- survfit(Surv(d2g, indic) ~ alcHQual, data = d2g)

plot(fit, lty = 2:3,col=colList)

title("Fonction de survie - Alcool Homme")

legend(130, .97,levels(d2g$alcHQual), lty = 2:3,col=colList)

plot(fit, lty = 2:3,col=colList,fun="cloglog")

title("Risque cumulé - Alcool Homme")

legend(120, -1.17, levels(d2g$alcHQual), lty = 2:3,col=colList)

Les courbes de survie sur l’alcoolemie des hommes sont tres proches. Le test du log rank n’est pas approprié au vue de des risques cumulés dont les courbes s’entrecroisent. On ne peut conclure à l’égalité des courbes de survie bien quels soient proche

#ageF

fit <- survfit(Surv(d2g, indic) ~ ageFQual, data = d2g)

plot(fit, lty = 2:3,col=colList)

title("Fonction de survie - Age Femme")

legend(130, .97, levels(d2g$ageFQual), lty = 2:3,col=colList)

plot(fit, lty = 2:3,col=colList,fun="cloglog")

title("Risque cumulé - Age Femme")

legend(110, -1.17, levels(d2g$ageFQual), lty = 2:3,col=colList)

La fonction de survie des femmes d’age de 24 à 26 ans est legerement plus faible comparer aux autres ages. Les courbes de risque cumulé et de fonction de survie sont très paralleles.

#ageH

fit <- survfit(Surv(d2g, indic) ~ ageHQual, data = d2g)

plot(fit, lty = 2:3,col=colList)

title("Fonction de survie - Age Homme")

legend(130, .97, levels(d2g$ageHQual), lty = 2:3,colList)

plot(fit, lty = 2:3,col=colList,fun="cloglog")

title("Risque cumulé - Age Homme")

legend(100, -1.17, levels(d2g$ageHQual), lty = 2:3,colList)

Les courbes de survie et de risque cumulé de l’age des hommes sont très proches et se croisent à de nombreux endroits. Le test du log rank n’étant pas forcement le plus indiqué dans ce cas de non respect de proportionnalité des risques, nous ne pouvons conclure à l’égalité des courbes de survie selon les ages. Le risque instanné des hommes dont l’age est supérieur à 30 ans est très supérieur durant les 20 premiers jours de grossesse. Ensuite la courbe décroit et se mélange aux autres.

#fumF

fit <- survfit(Surv(d2g, indic) ~ fumF, data = d2g)

plot(fit, lty = 2:3,col=colList)

title("Fonction de survie - Fumeur Femme")

legend(130, .97, c("Non fumeuse","Fumeuse"), lty = 2:3,col=colList)

plot(fit, lty = 2:3,col=colList,fun="cloglog")

title("Risque cumulé - Fumeur Femme")

legend(90, -1.17, c("Non fumeuse","Fumeuse"), lty = 2:3,col=colList)

summary(fit)$table

## records n.max n.start events median 0.95LCL 0.95UCL

## fumF=0 300 300 300 194 120.3313 109.5757 136.4402

## fumF=1 123 123 123 62 182.6003 132.8533 NA

Les fumeuses ont un survie plus importante que les non fumeuses. L’écart s’agrandi en fonction du délai. Les courbes de risque cumulé sont très proche et presque paralleles avec quelques entrecroisements.

#fumH

fit <- survfit(Surv(d2g, indic) ~ fumH, data = d2g)

plot(fit, lty = 2:3,col=colList)

title("Fonction de survie - Fumeur Homme")

legend(130, .97, c("Non fumeur","Fumeur"), lty = 2:3,col=colList)

plot(fit, lty = 2:3,col=colList,fun="cloglog")

title("Risque cumulé - Fumeur Homme")

legend(100, -1.17, c("Non fumeur","Fumeur"), lty = 2:3,col=colList)

survdiff(Surv(d2g, indic) ~ fumH, data = d2g)

## Call:

## survdiff(formula = Surv(d2g, indic) ~ fumH, data = d2g)

##

## N Observed Expected (O-E)^2/E (O-E)^2/V

## fumH=0 290 183 172.3 0.663 2.04

## fumH=1 133 73 83.7 1.365 2.04

##

## Chisq= 2 on 1 degrees of freedom, p= 0.154

Comme pour les femmes les hommes fumeurs ont une courbe de survie légèremment plus importante. Les courbes de risque sont parrallèle avec un risque des non fumeurs plus important durant les 20 premiers jours de grosesse.

#Sperm

fit <- survfit(Surv(d2g, indic) ~ spermQual, data = d2g)

plot(fit, lty = 2:3,col=colList)

title("Fonction de survie - Sperm")

legend(130, .97, levels(d2g$spermQual), lty = 2:3,col=colList)

plot(fit, lty = 2:3,col=colList,fun="cloglog")

title("Risque cumulé - Sperm")

legend(110, -1.17, levels(d2g$spermQual), lty = 2:3,col=colList)

Les courbes de survie correspondant aux millions de spermes chez les hommes sont tres proche et très chahutés. Les fonctions de risque cumulées sont proche mais s’entrecroisent et ne permettent de valider une égalité des fonctions de survies. Il apparait que les quantités de sperme in férieur à 54 millions , ont un risque instantanné plus faible les 50 premiers jours de grosesse.

#fumF & alcF Croisement de facteur

d2g$fumAlcF=as.factor(paste(d2g$fumF,d2g$alcFQual,' '))

fit <- survfit(Surv(d2g, indic) ~ fumAlcF, data = d2g)

plot(fit, lty = 2:10,col=colList)

title("Fonction de survie - Fumeuse/Alcool des femmes")

legend(130, .97, levels(d2g$fumAlcF), lty = 2:10,col=colList)

plot(fit, lty = 2:10,col=colList,fun="cloglog")

title("Risque cumulé - Fumeuse/Alcool des femmes")

legend(110, -1, levels(d2g$fumAlcF), lty = 2:10,col=colList)

L’étude du mélange des critères des fumeuses et du nombre de boisson alcoolisé pour les femmes permet de mettre en évidence un fort effet séparateur. On peut voir que les femmes non fumeuses et peu alcoolisées (0, 2-6) sont à l’opposé des femmes fumeuses et fortement alcoolisées (1 , 6-39 ). Les vingt cinq premiers jours sont peu différencier puis l’écart se creuse rapidement jusqu’a 75 jours pour resté assez parallèle jusqu’au dernier jours du jeu de données. La courbe du risque cumulé a de nombreux croisements et en même temps très parallèle rendant l’analyse très difficile.

# Modélisation

Après les contestations effectuées précédemment, nous entrons dans la partie modèlisation. Nous testons deux type de modèle. Un modèle paramétrique et un modèle à risque proportionnel (Cox). Pour chaque modèle nous faisons une sélection de variable en éliminant successivement les variables qui ont le mmoins d ’effet et qui ne sont pas significative. Le modèle paramétrique permettra de modéliser la fonction de survie. Le modèle de Cox quand a lui permettra la moélisation du risque instantannée. ## Modèle paramétrique Dans notre modèle paramétrique, nous inserons toutes nos variables excepté celle dont les données manquanttes sont en très grand nombre.

On commencera par étudier la forme de la fonction de survie globale en essayant de la modéliser par une loi de weibull (en bleu) qui sera comparer à l’estimation graphique de Kaplan Meier. Puis nous ajouterons les covariables afin d’ideintifier un modèle utilisable.

fit <- survfit(Surv(d2g, indic) ~ 1, data = d2g)

plot(fit,ylim=c(0.05,1),xlab="Délai de grosesse (jours)", ylab="Survie")

reg<-survreg(Surv(d2g, indic) ~ 1, data = d2g,

dist="weibull")

curve(exp(-(exp(-reg$coef[1]) \* x)^(1/reg$scale)),

col="blue", add=TRUE)

lines(predict(reg,d2g,type="quantile",col="red"))

title("Fonction de survie - modèle de weibull.")

summary(reg)

##

## Call:

## survreg(formula = Surv(d2g, indic) ~ 1, data = d2g, dist = "weibull")

## Value Std. Error z p

## (Intercept) 5.172 0.0469 110.27 0.00e+00

## Log(scale) -0.316 0.0550 -5.75 9.04e-09

##

## Scale= 0.729

##

## Weibull distribution

## Loglik(model)= -1587.9 Loglik(intercept only)= -1587.9

## Number of Newton-Raphson Iterations: 5

## n= 423

Le choix du modèle weibull conforte l’idée que la loi exponentielle n’est pas meilleure. Le scale de 0.73 qui est à 1 pour une loi exponentielle, nous précise cette finalité.

reg<-survreg(Surv(d2g,indic) ~ bmiF + bmiH+alcF+alcH+fumF+fumH+ageF+ageH+sperm , d2g, dist='weibull')

summary(reg)

##

## Call:

## survreg(formula = Surv(d2g, indic) ~ bmiF + bmiH + alcF + alcH +

## fumF + fumH + ageF + ageH + sperm, data = d2g, dist = "weibull")

## Value Std. Error z p

## (Intercept) 3.260873 0.70589 4.6195 3.85e-06

## bmiF 0.018942 0.01344 1.4091 1.59e-01

## bmiH 0.029523 0.01685 1.7521 7.98e-02

## alcF 0.009393 0.01047 0.8976 3.69e-01

## alcH 0.004615 0.00537 0.8597 3.90e-01

## fumF1 0.256904 0.11436 2.2464 2.47e-02

## fumH1 0.070384 0.10887 0.6465 5.18e-01

## ageF -0.000707 0.02010 -0.0352 9.72e-01

## ageH 0.025558 0.01884 1.3562 1.75e-01

## sperm -0.001369 0.00108 -1.2642 2.06e-01

## Log(scale) -0.333372 0.05472 -6.0928 1.11e-09

##

## Scale= 0.717

##

## Weibull distribution

## Loglik(model)= -1578.1 Loglik(intercept only)= -1587.9

## Chisq= 19.64 on 9 degrees of freedom, p= 0.02

## Number of Newton-Raphson Iterations: 5

## n= 423

La probabilité du test du Khi² est signicatif pour valider le modèle. L’intercepte est le paramètre le plus significatif. L’effet des covariables est faible. Seul la variable fumeuse possède un effet significatif. La valeur fumeuse de 0.257 qui mis à l’exponentiel donne 1.29, permet de quantifier que la fumeuse à une survie 29% plus importante qu’une non fumeuse.

### Sélection de variables

Pour améliorer le modèle, une sélection de variable est recherchée par supression paramètre après paramètre.

reg<-survreg(Surv(d2g,indic) ~ bmiH+fumF , d2g, dist='weibull')

summary(reg)

##

## Call:

## survreg(formula = Surv(d2g, indic) ~ bmiH + fumF, data = d2g,

## dist = "weibull")

## Value Std. Error z p

## (Intercept) 4.2985 0.3906 11.01 3.59e-28

## bmiH 0.0326 0.0160 2.04 4.14e-02

## fumF1 0.2766 0.1059 2.61 9.04e-03

## Log(scale) -0.3258 0.0549 -5.94 2.88e-09

##

## Scale= 0.722

##

## Weibull distribution

## Loglik(model)= -1582.3 Loglik(intercept only)= -1587.9

## Chisq= 11.27 on 2 degrees of freedom, p= 0.0036

## Number of Newton-Raphson Iterations: 5

## n= 423

Le résultat de cette étape extrait deux variables que sont l’indice de masse corporelle des hommes et la variable fumeuse. La valeur fumeuse de 0.277 qui mis à l’exponentiel donne 1.32, permet d’indiquer que la une survie d’une fumeuse est plus importante de 32% par rapport à une non fumeuse. La variable indice de masse corporelle de 0.033 qui mis à l’exponentiel donne 1.034, permet de dire qu’à chaque unité de cette mesure la survie augmente de 3%.

## Modèle à risque proportionnel

Le modèle de Cox permet de modéliser l’effet des covariables sur les risques instantannés. Cela permet de modèliser toutes les personnes à risque qui n’ont pas encore eut l’événement de l’arret de grosesse.

cox <-coxph(Surv(d2g,indic) ~ bmiF + bmiH+alcF+alcH+fumF+fumH+ageF+ageH+sperm , d2g)

summary(cox)

## Call:

## coxph(formula = Surv(d2g, indic) ~ bmiF + bmiH + alcF + alcH +

## fumF + fumH + ageF + ageH + sperm, data = d2g)

##

## n= 423, number of events= 256

##

## coef exp(coef) se(coef) z Pr(>|z|)

## bmiF -0.0268718 0.9734860 0.0188330 -1.427 0.1536

## bmiH -0.0379620 0.9627496 0.0235271 -1.614 0.1066

## alcF -0.0132395 0.9868477 0.0145909 -0.907 0.3642

## alcH -0.0066115 0.9934103 0.0074167 -0.891 0.3727

## fumF1 -0.3278366 0.7204807 0.1594101 -2.057 0.0397 \*

## fumH1 -0.0943757 0.9099408 0.1520723 -0.621 0.5349

## ageF -0.0007165 0.9992838 0.0281158 -0.025 0.9797

## ageH -0.0335035 0.9670516 0.0261183 -1.283 0.1996

## sperm 0.0017655 1.0017671 0.0015163 1.164 0.2443

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## exp(coef) exp(-coef) lower .95 upper .95

## bmiF 0.9735 1.0272 0.9382 1.0101

## bmiH 0.9627 1.0387 0.9194 1.0082

## alcF 0.9868 1.0133 0.9590 1.0155

## alcH 0.9934 1.0066 0.9791 1.0080

## fumF1 0.7205 1.3880 0.5271 0.9847

## fumH1 0.9099 1.0990 0.6754 1.2259

## ageF 0.9993 1.0007 0.9457 1.0559

## ageH 0.9671 1.0341 0.9188 1.0178

## sperm 1.0018 0.9982 0.9988 1.0047

##

## Concordance= 0.579 (se = 0.019 )

## Rsquare= 0.041 (max possible= 0.999 )

## Likelihood ratio test= 17.82 on 9 df, p=0.0373

## Wald test = 17.08 on 9 df, p=0.04745

## Score (logrank) test = 17.18 on 9 df, p=0.04603

De toutes les covariables, seule la variable fumeuse comme dans le modèle paramétrique à un effet significatif. Etre fumeuse permet de réduire le risque instantané d’interruption de grosesse de 28%. Les probabilités de validité du modèle que sont les tests de Wald, de Maximum de vraisemblance et du score sont tous positives et inférieur à 5%.

### Vérification de l’hypotèse de proportionnalité

Un modèle de cox demande un certain nombre de condition pour être utiliser comme l’hypothès de proportionnalité qui n’est pas toujours facile à identifier.

#vérification HP

coxzph <- cox.zph(cox, transform="km", global=TRUE )

plot(coxzph[1]); abline (h =0 , col =" blue ")

print(coxzph)

## rho chisq p

## bmiF -0.03255 0.25607 0.613

## bmiH 0.00481 0.00656 0.935

## alcF 0.09617 2.57730 0.108

## alcH 0.02729 0.19462 0.659

## fumF1 -0.10189 2.66415 0.103

## fumH1 0.08444 1.83097 0.176

## ageF -0.04097 0.46439 0.496

## ageH -0.02495 0.18564 0.667

## sperm -0.04119 0.39887 0.528

## GLOBAL NA 9.24487 0.415

La représentation des résidus de schoenfeld, permet de mettre en évidence l’indépendance entre les résidus et le temps ce qui valide l’hypothèse de proportionnalité.

Toutes les variables sont significatifs sur la proportionnalité des risques. Le test Global renforce ce résultat de proportionnalité des risques. Notre modèle de Cox est valide. Il reste encore à l’améliorer et à le simplifier.

### Sélection de variables

L’amélioration du modèle est fait par une sélection de variable par supression de paramètre après paramètre.

cox<-coxph(Surv(d2g,indic) ~ fumF , d2g)

summary(cox)

## Call:

## coxph(formula = Surv(d2g, indic) ~ fumF, data = d2g)

##

## n= 423, number of events= 256

##

## coef exp(coef) se(coef) z Pr(>|z|)

## fumF1 -0.3452 0.7081 0.1461 -2.362 0.0182 \*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## exp(coef) exp(-coef) lower .95 upper .95

## fumF1 0.7081 1.412 0.5317 0.9429

##

## Concordance= 0.531 (se = 0.015 )

## Rsquare= 0.014 (max possible= 0.999 )

## Likelihood ratio test= 5.9 on 1 df, p=0.01514

## Wald test = 5.58 on 1 df, p=0.01817

## Score (logrank) test = 5.64 on 1 df, p=0.0176

Cette sélection de variable montre une seule variable significative: la variable fumeuse. Au contraire du modèle paramétrique, la variable d’indice de masse corporelle des hommees qui a été la dernière variable à être retiré, n ’a pas été suffisamment significative pour être gardé. Ce résultat passe l’effet de la variable de 28 à 29 % de la réduction du risque instantané lorsque c’est une fumeuse.

# Conclusion

Après avoir analysé les données en regardant aussi bien le coté Survie que le coté Risque cumulé, nous avons put appréhender les problématique de ces données. Nous avons trouver des données manquantes que nous avons put corrigé autant que possible. Nous avons pu remarquer, que les censures et les événements n’étaient pas organiser de la même facon. Nous avons pu remarquer aussi que le risque instanné était globalement très important durant les premiers jours de grossesse pour se réduire fortement par la suite. Et que la courbe survie avait une courbe descendante assez linéaire. Nous avon remarqué aussi que les variables dont certaines différences pouvaient apparaitre étaient le BMI des hommes, l’alcool chez les femmes, l’age des femmes, les fumeuses et fumeurs.

La modélisation a permis de mieux estimer les effets des variables et a permis de nuancer fortement les analyses des courbe de survie et des risques cumulés. Ainsi par exemple, l’effet de l’alcool n’est pas paru aussi significatif qu’espéré.

La modélisation paramétrique a permis de mettre en oeuvre un loi de weibull. On a pu estimé très précisément l’effet de chaque variable sur la survie du délai de grossesse. La sélection des variables de ce modèle a permis de choisir le bmi des hommes et les femmes fumeuse comme éléments déterminant.

La modélisation à risque proportionnelle a permis de mesurer précisément l’effet de chaque variable sur le risque intantanné. Le résultat a été conforté sur le fait que le modèle respectait l’hypothèse de proportionnalité. La sélection de variable y a été plus stricte en ne retenant que les femmes fumeuses.