

SIMULATION DE COVARIABLES DÉPENDANTES DU TEMPS VIA UNE DISTRIBUTION DE WEIBULL

Projet Master M1

Enseignant responsable : BRUNEL-PICCININI Elodie

par

AARAB AYOUB
BURGAT PAUL
LAABSI ZAKARIA

MASTER 1 - MATHÉMATIQUES DE L'INFORMATION ET DE LA
DÉCISION / BIOSTATISTIQUE



FACULTÉ DES SCIENCES
UNIVERSITÉ DE MONTPELLIER

Année 2020-2021



Résumé



Ce rapport se veut une introduction à l'analyse de la survie avec des jeux de données simulées comportant des covariables dépendantes du temps. Après présentation du modèle semi-paramétrique de survie de Cox standard, nous introduisons une approche permettant l'analyse de la survie en s'affranchissant des hypothèses classiques du modèle de Cox.

Remerciements

Nous tenons à remercier Élodie Brunel-Piccinini, professeure à l'Université de Montpellier, pour son investissement et ses précieux conseils qui nous ont guidés tout au long de la rédaction de ce rapport.

Table des matières

1	Introduction	4
2	Modèle de Cox standard	5
2.1	Hypothèses du modèle de Cox	6
2.2	Hypothèse des risques proportionnels	7
2.2.1	Vérifications graphiques	7
2.2.2	Résidus de Schoenfeld (standardisés)	7
2.3	Génération d'une durée de survie avec covariables indépen- dantes du temps	12
2.3.1	Loi de Weibull	12
2.3.2	Première méthode	14
2.3.3	Deuxième méthode : modèle log-linéaire	15
2.4	Simulations de jeu de données 	18
2.4.1	Première méthode - méthode de la <i>transformée inverse</i>	18
2.4.2	Deuxième méthode - modèle <i>log-linéaire</i>	20
3	Extension du modèle de Cox : durée de survie avec des co- variables dépendantes du temps	24
3.1	Une nouvelle approche du modèle de Cox	24
3.2	Extension du modèle via une distribution de Weibull	25
3.3	Générer des données longitudinales via un modèle linéaire mixte	27
3.3.1	Modèle linéaire à effets fixes (modèle linéaire simple)	27
3.3.2	Modèle linéaire mixte	28
3.3.3	Ecriture matricielle	30
3.3.4	Génération des données longitudinales via un modèle linéaire mixte	31
3.4	Générer des durées de survie avec la fonction de Lambert et une distribution de Weibull	34
3.4.1	Fonction W de Lambert	34
3.4.2	Génération de durées de survie avec covariable dépen- dante du temps	34
3.5	Simulation d'un jeu de données sur  commenté	36
3.5.1	Courbes de survie avec dépendance du temps	41
3.5.2	Limites	43
4	Estimation des coefficients	44
4.1	Coefficients estimés	44
5	Conclusion	45

6	Annexes	46
6.1	Simulation des jeux de données 	46
6.2	Tableaux de données de survie avec indépendance et dépendance du temps	53
6.3	Coefficients estimés 	55
6.3.1	Tableaux des coefficients estimés	58

1 Introduction

Ce document traite de techniques permettant de générer des durées de survie pour les études de simulation concernant les modèles à risques proportionnels de Cox. Les études de simulation sont couramment utilisées pour évaluer les performances de modèles statistiques actuels et nouvellement développés.

En effet les études de simulation représentent un outil statistique important pour étudier les performances, les propriétés et l'adéquation des modèles statistiques, des statistiques de test et des techniques d'estimation en tenant compte de conditions pré-spécifiées.

La première section introduit le modèle de survie semi-paramétrique de Cox standard qui est utilisé ici pour l'analyse des durées de vie des individus que l'on simule dans des jeux de données. L'objectif de ce chapitre est d'illustrer la vérification des hypothèses du modèle de Cox standard sur des jeux de données simulés avec une méthode particulière de génération des durées de survie via la distribution de Weibull.

La seconde section introduit la notion de covariables dépendantes du temps et la limite du modèle de Cox standard lorsque des covariables sont dépendantes du temps. Dans ce chapitre nous verrons une approche développée par Peter C Austin [1] qui nous permettra de simuler des jeux de données avec covariables dépendantes du temps. L'objectif étant de pouvoir générer des durées de survie pour des individus que nous simulons de telle sorte que l'analyse de ces données soit possible avec un modèle de régression de Cox.

2 Modèle de Cox standard

Le modèle de Cox est un modèle de régression semi-paramétrique classé comme modèle de survie où un évènement temporel est relié à certaines covariables / variables explicatives. Soient Z_1, \dots, Z_n des covariables. Soit h la fonction de risque instantané de décès en fonction du temps, représentant la probabilité d'apparition de l'évènement dans un intervalle de temps $[t, t + \delta t]$.

On a alors :

$$h(t | \mathbf{Z}) = h_0(t) \exp \left(\sum_{i=1}^n \beta_i Z_i \right) = h_0(t) \exp (\mathbf{Z}' \beta) \quad (1)$$

avec :

$\mathbf{Z} = (Z_1, \dots, Z_n)'$ le vecteur des covariables

$\beta = (\beta_1, \dots, \beta_n)'$ le vecteur des constantes

β_i est le paramètre pour la i -ème covariable Z_i

Remarques

- Les covariables Z_1, \dots, Z_n ne dépendent pas du temps t .
- Le rapport des risques instantanés $\exp(\mathbf{Z}'\beta)$ dépend des covariables Z_1, \dots, Z_n et non du temps t .

2.1 Hypothèses du modèle de Cox

Le modèle de Cox possède deux hypothèses :

— **Hypothèse des risques proportionnels :**

— Le rapport des risques relatifs RR entre deux individus i et j est indépendant du temps t .

$$\frac{h_i(t \mid \mathbf{Z}_i)}{h_j(t \mid \mathbf{Z}_j)} = \frac{h_0(t) \exp(\mathbf{Z}'_i \beta)}{h_0(t) \exp(\mathbf{Z}'_j \beta)} = \exp((\mathbf{Z}_i - \mathbf{Z}_j)' \beta)$$

avec :

$\mathbf{Z}_i = (Z_{i1}, \dots, Z_{in})'$ le vecteur des n covariables de l'individu i

— **Hypothèse de log-linéarité :**

$\log(h(t \mid \mathbf{Z})) = \log(h_0(t)) + \mathbf{Z}'\beta$ est une fonction linéaire des Z_i

2.2 Hypothèse des risques proportionnels

2.2.1 Vérifications graphiques

Une façon de tester si l'hypothèse des risques proportionnels (HRP) est valable pour une covariable donnée est d'observer graphiquement les fonctions de logarithme du risque cumulé de base pour les modalités de la covariable. Si l'on a proportionnalité entre les fonctions entre chaque modalité, alors l'HRP est valable. A partir de l'estimation des paramètres, la fonction de survie individuelle (probabilité de survie jusqu'à l'instant t) telle que $\forall t \geq 0, S(t) = \mathbb{P}(X > t)$ peut s'écrire de la façon suivante :

$$\begin{aligned} S(t|\mathbf{Z}_i) &= \exp \left(- \exp(\mathbf{Z}'_i \beta) \int_0^t h_0(u) du \right) \\ &= \exp \left(- \int_0^t h_0(u) du \right)^{\exp(\mathbf{Z}'_i \beta)} \\ &= S_0(t)^{\exp(\mathbf{Z}'_i \beta)} \end{aligned}$$

Une estimation de cette fonction peut être faite par la méthode de Breslow ou de Kalbfleisch-Prentice. On a donc, par quelconque des deux méthodes citées, l'estimation de la fonction de survie pour un individu i :

$$\hat{S}(t|\mathbf{Z}_i) = \hat{S}_0(t)^{\exp(\mathbf{Z}'_i \beta)}$$

Le logarithme de cette estimation donne l'estimation de la fonction de risque cumulée pour l'individu i :

$$\log(-\log(\hat{S}(t|\mathbf{Z}_i))) = \log(H_0(t)) + \mathbf{Z}'_i \beta$$

Les courbes des fonctions pour chaque individu différent pour une modalité d'une covariable ont la même tendance et ne se croisent pas dans le cas de l'HRP validée pour la dite covariable. Si elles se croisent ou bien n'ont pas la même tendance, on peut à priori rejeter l'HRP.

2.2.2 Résidus de Schoenfeld (standardisés)

L'hypothèse de risques proportionnels (HRP) peut être vérifiée à l'aide de tests statistiques et de diagnostics graphiques basés sur les résidus de Schoenfeld (standardisés).

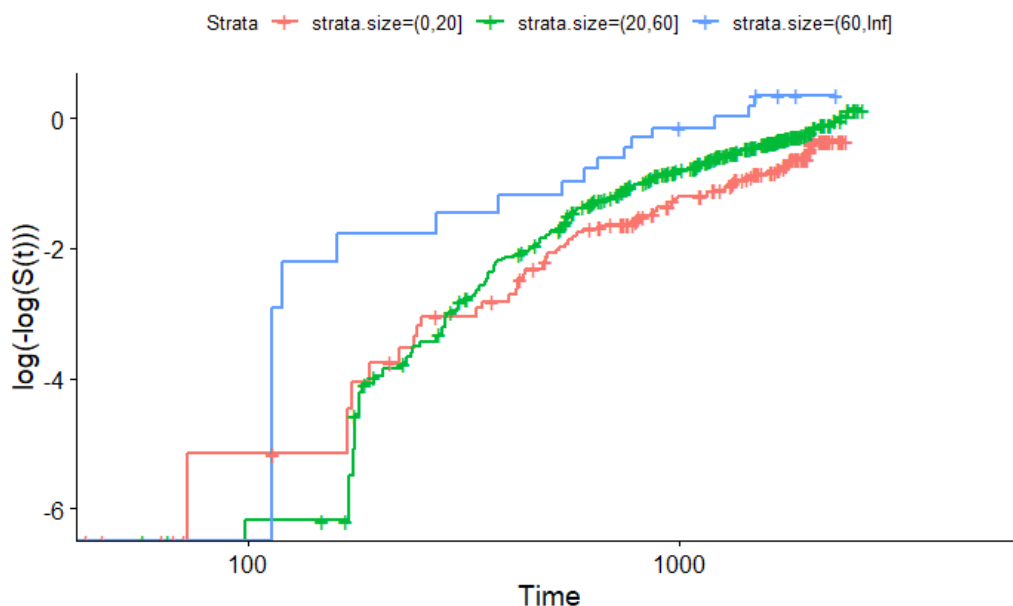


FIGURE 1 – Courbes respectant l’hypothèse de log-linéarité
Données issus du dataset `gbsg` du package [R survival](#)

En principe, les résidus de Schoenfeld sont indépendants du temps. Un graphique montrant un modèle non aléatoire en fonction du temps est une preuve de violation de l’hypothèse des risques proportionnels.

La fonction `cox.zph()` (dans le package `survival`) fournit une solution pratique pour tester l’hypothèse des risques proportionnels pour chaque covariable incluse dans un ajustement du modèle de régression de Cox.

Pour chaque covariable, la fonction `cox.zph()` met en corrélation l’ensemble correspondant de résidus de Schoenfeld standardisés (mis à l’échelle) avec le temps, afin de tester l’indépendance entre les résidus et le temps. En outre, elle effectue un test global pour le modèle dans son ensemble.

L’hypothèse de risque proportionnel est soutenue par une relation non significative entre les résidus et le temps, et réfutée par une relation significative :

$$H_0 : \mathbf{Z}_i(t) = \mathbf{Z}_i$$

$$H_1 : \mathbf{Z}_i(t) \neq \mathbf{Z}_i$$

avec

$$\mathbf{Z}_i = \begin{pmatrix} Z_{i1} \\ \vdots \\ Z_{ij} \\ \vdots \\ Z_{in} \end{pmatrix}$$

Le vecteur des résidus de Schoenfeld s'obtient comme la différence entre la valeur prise par l'individu i pour la j -ème covariable Z_{ij} la moyenne pondérée de chaque covariable \mathbf{Z}_i calculé pour chaque temps d'événement t_i , $\overline{Z}_j(t_i, \hat{\beta})$.

$$\mathbf{r}_j = \begin{pmatrix} r_{1j} \\ \vdots \\ r_{ij} \\ \vdots \\ r_{nj} \end{pmatrix} \quad \text{avec } r_{ij} = Z_{ij} - \overline{Z}_j(t_i, \hat{\beta})$$

avec :

- Z_{ij} la valeur de la covariable j pour l'individu i décédé au temps t_i
- $\overline{Z}_j(t_i, \hat{\beta})$ la moyenne pondérée des valeurs de covariable j pour l'ensemble des individus exposés au risque en t_i telle que :

$$\overline{Z}_j(t_i, \hat{\beta}) = \frac{\sum_{j \in R_{t_i}} Z_j \exp(\mathbf{Z}'_j \hat{\beta})}{\sum_{j \in R_{t_i}} \exp(\mathbf{Z}'_j \hat{\beta})}$$

Remarque : la somme de tous les résidus de Schoenfeld est nul :

$$\sum_{j=1}^n r_{ij} = 0$$

Résidus de Schoenfeld standardisés

On définit un résidu de Schoenfeld standardisé tel que :

- $\mathbf{r}_j^* = \frac{\mathbf{r}_j}{V(\mathbf{r}_j)}$ est le vecteur des résidus de Schoenfeld standardisés
- $r_{ij}^* = (V(\beta, t))^{-1} r_{ij}$ est le résidu de Schoenfeld standardisé avec $V(\beta, t)$ la matrice de variance-covariance de β à un temps t .

Pour valider l'hypothèse des risques proportionnels, nous pouvons effectuer une régression linéaire des résidus pour chaque covariable Z_{ij} .

La droite de régression est :

$$r_{ij} = at_i + \epsilon_i \quad \forall i$$

Si $a \neq 0$ (droite de régression non constante) alors l'hypothèse des risques proportionnels sera rejetée.

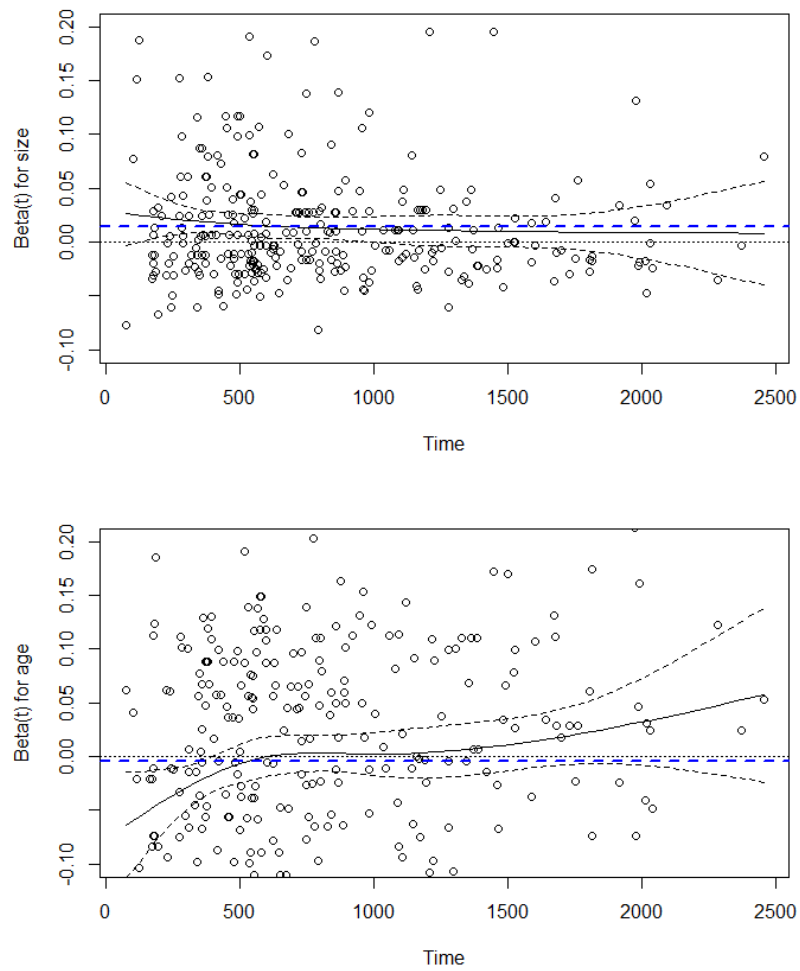


FIGURE 2 – En haut : HRP validée, en bas : HRP rejetée.
Données issus du dataset `gbsg` du package [R survival](#)

2.3 Génération d'une durée de survie avec covariables indépendantes du temps

2.3.1 Loi de Weibull

La loi de Weibull regroupe un ensemble de lois qui sont des approximations utiles de distribution rencontrées dans la nature. Parmi elles, on retrouve la loi exponentielle ou encore la loi de Rayleigh.

La densité de probabilité de la loi de Weibull, pour $x > 0$ est :

$$f(x; k, \lambda) = \frac{k}{\lambda} \left(\frac{x}{\lambda}\right)^{k-1} \exp\left(-\left(\frac{x}{\lambda}\right)^k\right)$$

où :

- $k > 0$ est le paramètre de forme ("shape" dans la documentation anglaise)
- $\lambda > 0$ est le paramètre d'échelle de la distribution

Sa fonction de répartition est :

$$F(x; k, \lambda) = 1 - \exp\left(-\left(\frac{x}{\lambda}\right)^k\right)$$

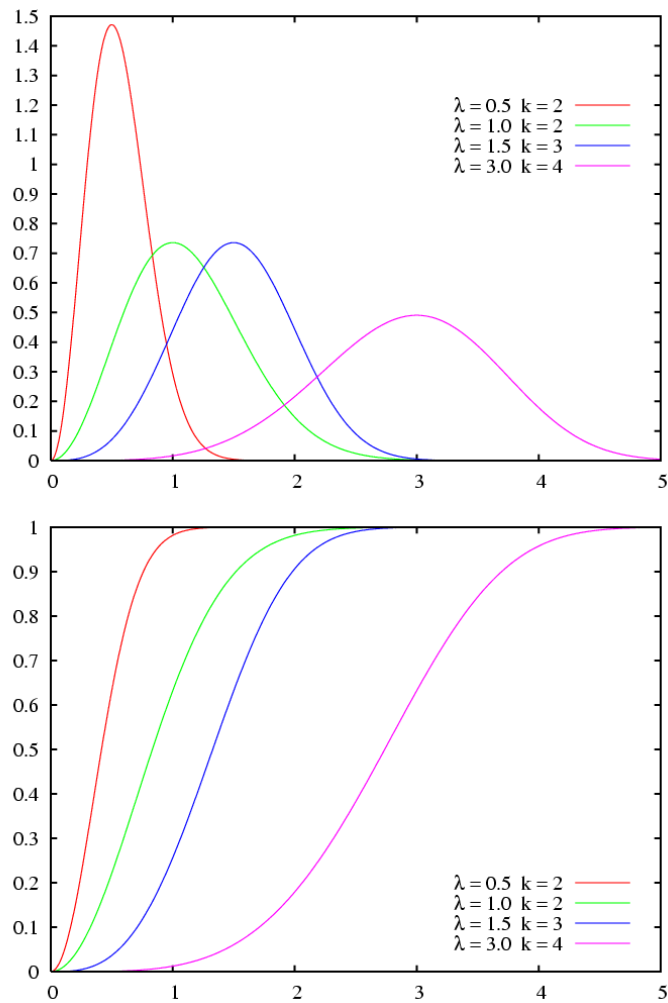


FIGURE 3 – En haut : Densité de probabilité, en bas : Fonction de répartition.
Source : [Wikipédia](#)

Dans ce qui suit, nous utilisons la distribution de Weibull pour modéliser les temps de survie.

Proposition 2.1 Soit S la fonction de survie. Si $T \sim S(\cdot | \mathbf{Z})$ alors $V = S(T | \mathbf{Z}) \sim U(0, 1)$

2.3.2 Première méthode

Soit $V \sim U(0, 1)$ une variable aléatoire.

Soit la fonction de survie

$$S(t | \mathbf{Z}) = \exp \left(-H_0(t) \exp(\mathbf{Z}'\beta) \right)$$

avec

$$H_0(t) = \int_0^t h_0(u) du$$

Soit T le temps de survie qui a pour fonction de survie $S(\cdot | \mathbf{Z})$. Pour générer un temps de survie $T \sim S(\cdot | \mathbf{Z})$, il suffit de prendre une valeur v de $V \sim U(0, 1)$ et de faire la transformation inverse $t = S^{-1}(v | \mathbf{Z})$. On obtient ainsi

$$T = S^{-1}(V | \mathbf{Z}) = H_0^{-1} \left(-\frac{\log(V)}{\exp(\mathbf{Z}'\beta)} \right)$$

On considère la fonction de risque de base de Weibull

$$h_0(t) = \lambda \alpha t^{\alpha-1} \quad \text{avec } \alpha \geq 0 \text{ et } \lambda \geq 0$$

On obtient ainsi

$$H_0(t) = \lambda t^\alpha \text{ et } H_0^{-1}(t) = \left(\frac{t}{\lambda} \right)^{\frac{1}{\alpha}}$$

On génère $T \sim S(\cdot | \mathbf{Z})$ via la transformation

$$t = \left(-\frac{\log(v)}{\lambda \exp(\mathbf{Z}'\beta)} \right)^{\frac{1}{\alpha}}$$

Ainsi, on a

$$\begin{aligned}
T &= H_0^{-1} \left(-\frac{\log(V)}{\exp(\mathbf{Z}'\beta)} \right) \\
&= \left(\frac{-\frac{\log(V)}{\exp(\mathbf{Z}'\beta)}}{\lambda} \right)^{\frac{1}{\alpha}} \\
&= \left(-\frac{\log(V)}{\lambda \exp(\mathbf{Z}'\beta)} \right)^{\frac{1}{\alpha}}
\end{aligned}$$

Finalement, on a la fonction de risque

$$\begin{aligned}
h(t \mid \mathbf{Z}) &= h_0(t) \exp(\mathbf{Z}'\beta) \\
&= \lambda \alpha t^{\alpha-1} \exp(\mathbf{Z}'\beta) \\
&= \underbrace{\lambda \exp(\mathbf{Z}'\beta)}_{= \lambda(\mathbf{Z})} \alpha t^{\alpha-1}
\end{aligned}$$

avec $\lambda(\mathbf{Z})$ le paramètre d'échelle variant (*varying scale parameter*)

2.3.3 Deuxième méthode : modèle log-linéaire

Soit X une variable aléatoire positive représentant une durée de vie telle que :

$$\log(X) = \mu + \gamma Z + \sigma W$$

où $Z \sim F_Z = U[a, b]$ avec $a, b \in \mathbf{R}_+$ tels que $a < b$ et $W \sim F_W$ suit la loi des extrêmes.

En multipliant des deux côtés par l'exponentielle, on obtient :

$$X = \exp(\mu + \gamma Z + \sigma W)$$

La fonction de répartition d'une loi des extrêmes est de la forme :

$$F_W(w) = 1 - \exp(-\exp(w))$$

Par la méthode de la transformée inverse, on obtient que :

$$W \sim \log(-\log(1-U))$$

où $U \sim U[0, 1]$, ce résultat va être utile pour simuler des données de survie sur \mathbb{R} .

Nous allons maintenant montrer que $X \sim Weibull(\lambda, \alpha)$ où l'on va déterminer les paramètres λ et α .

$$X = \exp(\mu + \gamma Z + \sigma W) \sim Weibull(\lambda, \alpha)$$

Déterminer λ et α :

$$\begin{aligned} \mathbb{P}(X \leq t) &= \mathbb{P}(\exp(\mu + \gamma Z + \sigma W)) \\ &= \mathbb{P}(\mu + \gamma Z + \sigma W \leq \ln t) \\ &= \mathbb{P}(\sigma W \leq \ln(t) - \mu - \gamma Z) \\ &= \mathbb{P}\left(W \leq \frac{\ln(t) - \mu - \gamma Z}{\sigma}\right) \\ \mathbb{P}(X \leq t) &= 1 - \exp\left(-\exp\left(\frac{\ln(t) - \mu - \gamma Z}{\sigma}\right)\right) \end{aligned}$$

Or, la fonction de distribution de la loi de Weibull de paramètre (λ et α) est donnée par :

$$F_{Weibull}(x) = 1 - \exp\left(-\left(\frac{x}{\lambda}\right)^\alpha\right)$$

Si on fait correspondre les deux équations, on a :

$$\begin{aligned}
1 - \exp\left(-\exp\left(\frac{\ln(t) - \mu - \gamma Z}{\sigma}\right)\right) &= 1 - \exp\left(-\left(\frac{t}{\lambda}\right)^\alpha\right) \\
\exp\left(-\exp\left(\frac{\ln(t) - \mu - \gamma Z}{\sigma}\right)\right) &= \exp\left(-\left(\frac{t}{\lambda}\right)^\alpha\right) \\
\exp\left(\frac{\ln(t) - \mu - \gamma Z}{\sigma}\right) &= \alpha \ln\left(\frac{t}{\lambda}\right) \\
\exp\left(\frac{\ln(t) - \mu - \gamma Z}{\sigma}\right) &= \alpha(\ln(t) - \ln(\lambda))
\end{aligned}$$


D'où :


$$\begin{cases} \alpha = \frac{1}{\sigma} \\ \ln(\lambda) = \mu + \gamma Z \iff \lambda = \exp(\mu + \gamma Z) \end{cases}$$

Ainsi, nous avons

$$X = \exp(\mu + \gamma Z + \sigma W) \sim \text{Weibull}\left(\exp(\mu + \gamma Z), \frac{1}{\sigma}\right)$$

2.4 Simulations de jeu de données

Afin d'illustrer les propos mentionnés en section 2.3, il peut être intéressant de simuler des jeux de données. Ce faisant, nous utilisons le langage .

En effet à l'aide des résultats mathématiques conclus en section 2.3, on peut générer des temps de survie sur , et créer à l'aide de cet outil un jeu de données complet auquel on va appliquer la méthode de régression de Cox. Ceci nous permettra entre autres d'illustrer les méthodes de vérification des hypothèses de la section 2.2.

On produit ainsi les deux simulations suivantes à l'aide des deux méthodes de génération de temps de survie.

2.4.1 Première méthode - méthode de la *transformée inverse*

```
set.seed(40)

# Risque de base : Weibull

# N = taille de l'échantillon
# lambda = parametre d'échelle dans h0() (scale parameter)
# alpha = parametre de forme dans h0() (shape parameter)
# beta = parametre d'effet fixe
# rateC = parametre du taux de la distribution
# exponentielle de C

simulWeibull <- function(N, lambda, alpha, beta_sexe,
  beta_age, rateC)
{
  # covariable

  sexe=rbern(N,0.540)
  for (i in 1:N){if (sexe[i] == 0) sexe[i]=1 else sexe[i]
    =2 }

  age=round(rnorm(N,35,5))

  # Temps d'évenements latents de Weibull
  v <- runif(n=N) #n = nombre d'observations
  T <- (- log(v)/(lambda * exp(sexe*beta_sexe+age*beta_age)))^(1 / alpha)
```

```

# Temps de censure suivant distribution exponentielle

C <- rexp(n=N, rate=rateC)

#C<-runif(n=N, min=0, max=N)

# Temps de suivi et indicateurs d'évenements
time <- pmin(T, C) # Z = min(T,C)
status <- as.numeric(T <= C) #      = 1(T < C)

# Observations (Zi, i ) (dataset)
data.frame(id=1:N,
            time=time,
            status=status,
            sexe =sexe,
            age =age)
}

#status 0 = censure
#status 1 = non censure

datasimul <- simulWeibull(N=500, lambda=0.01, alpha=3,
                           beta_sexe=-0.5, beta_age=0.05, rateC=0.001)

```

Le tableau de données généré est disponible au tableau [4](#)

2.4.2 Deuxième méthode - modèle *log-linéaire*

```
#log(X)=mu+gammaZ+sigmaW
#Ou Z ~ unif[15,80]
#W=log(-log(1-U)) ou U ~ unif[0,1]
#mu=10
#gamma=1
#sigma=0.5
n=1000
mu=0.2
gamma=0.05
sigma=0.5
Z <- runif(n, 15, 80)
U <- runif(n, 0, 1)
W <- log(-log(1-U))
X <- exp(mu+ gamma*Z+ sigma*W)

T=rweibull(n, 1/sigma, exp(mu+gamma*Z))
#View(T)
#View(X)

mu=0.05
C=rexp(n,mu)

time=pmin(X,C)
#indicateur de l'événement d'intérêt
#delta vaut 1 si X<C et 0 sinon
#delta=as.numeric(TT==X)
#ou bien
status=as.numeric(X<C)
#View(time)
pourcentage_cens=1-sum(status)/n
D<-data.frame(id=1:n, time = time, status=status, Z=Z)
D
View(D)

library(survival)
library(survminer)

strata_Z<-cut(D$Z,breaks=c(15,50,80))

f<- survfit(Surv(time, status) ~ strata_Z, type = "kaplan"
```

```

    -meier", conf.type = "plain", D)
ggsurvplot(f, data = D)

```

Le tableau de données généré est le tableau 3

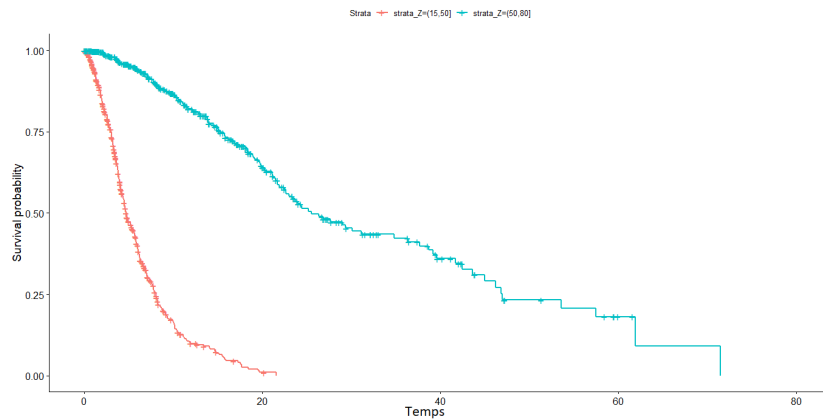


FIGURE 4 – Courbe de survie par tranche d'âge

On a généré des temps de survie tels que la tranche d'âge la plus jeune est celle qui survit le moins longtemps.

```

ggsurvplot(f, fun = "cloglog", data = D, xlim=c(-50,50))

```

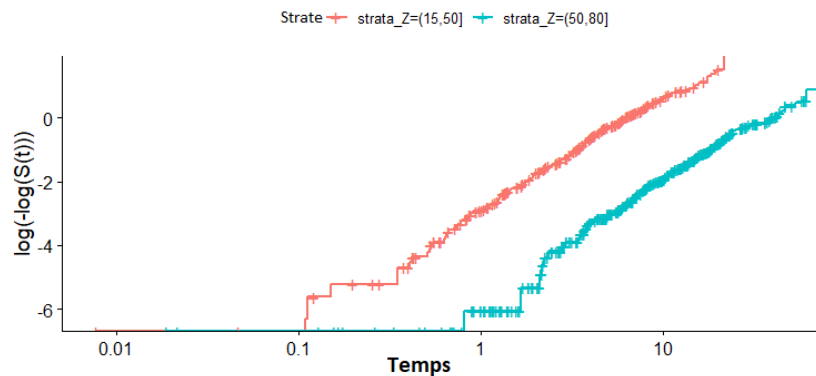


FIGURE 5 – Hypothèse de log-linéarité validée

La tendance des deux courbes pour les deux tranches d'âge est la même, sans intersections. On peut supposer que l'hypothèse des risques proportionnels est vérifiée.

```

fit<-coxph(Surv(time,status)~Z, data = D)
# summary(fit)

```

```
cox.zph(fit)
plot(cox.zph(fit))
abline(h=0, col='red')
```

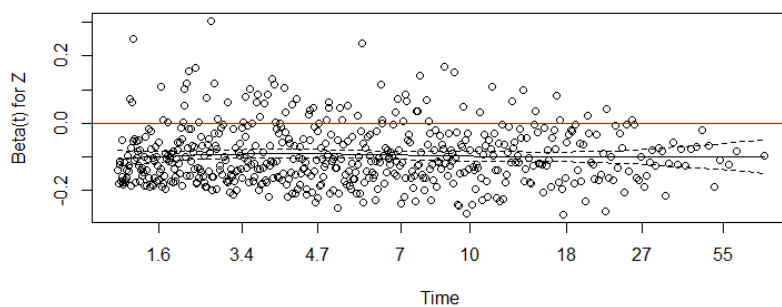


FIGURE 6 – Courbe test des résidus de Schoenfeld

	chisq	df	p
Z	1.07	1	0.3
GLOBAL	1.07	1	0.3

TABLE 1 – Test d’hypothèse de la dépendance au temps d’une covariable

On remarque dans ce tableau que la p-value est trop élevée (>0.05) pour rejeter l’hypothèse nulle qui est que les covariables sont indépendantes du temps. On peut donc conclure que la covariable Z est non corrélée avec le temps.

Dans cette première partie, nous avons défini ce qu'est un modèle de Cox et ses hypothèses de validités. Puis nous avons généré des temps de survie à l'aide d'une distribution de Weibull avec des covariables indépendantes du temps simulées via deux méthodes : la méthode de la transformée inverse ainsi que la méthode du modèle log-linéaire. En récupérant les données obtenues par ces méthodes, on s'aperçoit que les hypothèses de proportionnalité du modèle de Cox sont vérifiées.

Dans une deuxième partie, après avoir montré la procédure pour générer des temps de survie avec des covariables indépendantes du temps, nous allons générer des temps de survie avec des covariables dépendantes du temps via la méthode de la transformée inverse. Pour cela, nous allons introduire les notions de données longitudinales qui font office de covariables dépendantes du temps via le modèle linéaire mixte. Puis nous verrons comment générer des temps de survie avec des covariables dépendantes du temps via la fonction de Lambert.

3 Extension du modèle de Cox : durée de survie avec des covariables dépendantes du temps

3.1 Une nouvelle approche du modèle de Cox

Peter C Austin [1] a décrit une approche pour simuler les temps d'événements lorsque les covariables varient avec le temps. Dans ce scénario, la valeur de certaines covariables pour un individu peut changer tout au long du temps de suivi. Afin d'analyser ces covariables, un modèle de risques proportionnels peut être étendu afin d'incorporer des variables dépendantes du temps et de s'affranchir de l'hypothèse des risques proportionnels.

Cependant, lorsque nous nous intéressons au temps d'évènement et que nous souhaitons prendre en compte l'effet de la variable longitudinale en tant que covariable dépendante du temps, les approches traditionnelles d'analyse ne sont pas applicables.

Ainsi cette extension du modèle de Cox ne nécessite plus que les risques inter-individus soient proportionnels.

Ici, nous considérons un échantillon de sujets de taille n :

$$\left\{ (T_i, \mathbf{Y}_i(t), \mathbf{Z}_i, \Delta_i) \mid 0 \leq t \leq T_i \quad \forall i = 1, 2, \dots, n \right\}$$

Avec :

- T_i : le temps d'évènement pour l'individu i
- Δ_i : l'indicatrice d'évènement de l'individu i
- \mathbf{Z}_i : le vecteur des covariables indépendantes du temps de l'individu i
- $\mathbf{Y}_i(t)$: le vecteur des covariables dépendantes du temps de l'individu i avec m_i mesures au cours du temps tel que :

$$\mathbf{Y}_i(t) = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{im_i} \end{pmatrix}$$

Ainsi le modèle de Cox des risques proportionnels a une fonction de risque au temps t qui peut être écrit comme suit :

$$\begin{aligned} h(t \mid \mathbf{Z}_i, \mathbf{Y}_i(t)) &= h_0(t) \exp(\mathbf{Z}_i' \beta + \gamma \mathbf{1} \cdot \mathbf{Y}_i(t)) \\ &= h_0(t) \exp\left(\mathbf{Z}_i' \beta + \gamma \sum_{j=1}^{m_i} Y_{ij}(t)\right) \end{aligned}$$

où $h_0(t)$ représente la fonction de risque de base, β le vecteur des coefficients pour des covariables invariantes dans le temps. Le vecteur des covariables dépendantes du temps $\mathbf{Y}_i(t)' = (Y_{i1}(t), \dots, Y_{ij}(t), \dots, Y_{im_i}(t))$ est l'ensemble de covariables pour le nombre de mesures longitudinales m_i pour l'individu i . Enfin le scalaire $\gamma \in \mathbf{R}$ est le paramètre qui relie les covariables dépendantes du temps au risque.

La fonction des risques proportionnels $h(t)$ prend en compte deux résultats : la réponse longitudinale et le temps de survie.

Afin d'estimer ce type de modèles, nous devons d'abord ajuster un modèle pour la réponse longitudinale via un modèle linéaire mixte et ensuite un modèle pour le temps de survie via une distribution de Weibull.

3.2 Extension du modèle via une distribution de Weibull

Dans l'un des scénarios envisagés par Austin [1], nous supposons une seule covariable variant dans le temps, désignée par $z(t)$, proportionnelle à t : $z(t) = kt$, avec $k > 0$. Les covariables invariantes dans le temps sont désignées par Z . La fonction de risque peut s'écrire comme suit :

$$h(t \mid z(t)) = h_0(t) \exp(Z' \beta + \gamma z(t))$$

Ainsi, la fonction de risques cumulés peut être écrite comme suit :

$$H(t, x, z(t)) = \int_0^t \exp(Z'\beta + \gamma z(r)) h_0(r) dr$$

Si des temps de survie suivent une loi de Weibull, alors $h_0(t) = \lambda \alpha t^{\alpha-1}$.
Ainsi

$$\begin{aligned} H(t, x, z(t)) &= \int_0^t \exp(Z'\beta + \gamma z(r)) h_0(r) dr \\ &= \int_0^t \exp(Z'\beta + \gamma z(r)) \lambda \alpha r^{\alpha-1} dr \\ &= t^\alpha (-\gamma kt)^{-\alpha} (\Gamma(\alpha) - \Gamma(\alpha, \gamma kt)) \end{aligned} \tag{1}$$

Dans (1), Γ représente la fonction Gamma telle que $\forall x > 0$:

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$$

et $\Gamma(a, x)$ la fonction Gamma incomplète supérieure telle que pour $a \in \mathbb{C}$ avec $\Re(a) > 0$:

$$\Gamma(a, x) = \int_x^\infty t^{a-1} e^{-t} dt$$

Austin a conclu qu'une expression pour la fonction de risques cumulés n'existe pas dans ce scénario particulier. Cependant, l'intégrale (1) peut être évaluée numériquement. [1] [3]

3.3 Générer des données longitudinales via un modèle linéaire mixte

Dans une régression linéaire classique, tous les individus i ont le même comportement au cours du temps.

Nous souhaitons prendre en compte la dépendance par rapport au temps. Pour cela, nous introduisons les effets aléatoires, qui permettent de refléter la corrélation entre les unités statistiques. Les unités statistiques peuvent être groupées ensemble et avoir une corrélation intra-groupe, par exemple une mutation observée au cours du temps pour différents groupes de virus. Ce sont des données dites **longitudinales** : les individus/sujets sont mesurés dans le temps en fonction d'un événement de départ. Elles s'opposent aux données transversales qui s'intéressent à un événement à un instant t .

3.3.1 Modèle linéaire à effets fixes (modèle linéaire simple)

Le modèle linéaire simple suppose une dépendance linéaire entre les individus et la covariable :

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \varepsilon_{ij} \quad \text{avec les erreurs } \varepsilon_{ij}$$

Avec :

- Y_{ij} est la réponse de l'individu i au temps d'évènement j avec $i \in \{1, \dots, n\}$ et $j \in \{1, \dots, m_i\}$
- X_{ij} est la variable variant dans le temps de l'individu i .
- β_0 et β_1 les coefficients identiques pour tous les individus (effets fixes)

Ici tous les individus ont le même comportement avec la covariable X_{ij} . Une des limites du modèle à effet fixes est qu'il suppose l'existence d'une indépendance entre les réponses longitudinales Y_{ij} . Les Y_{ij} d'un même individu peuvent présenter une dépendance à une covariable.[\[2\]](#)

3.3.2 Modèle linéaire mixte

"Le modèle linéaire à effets mixtes est une extension du modèle linéaire qui prend en compte la variabilité liée aux individus. Ce modèle est composé d'une partie fixe et d'une partie aléatoire. La partie fixe est identique pour chaque individu et représente l'effet groupe. La partie aléatoire est propre à chacun des individus et traduit la variabilité liée à chaque sujet." [2]

Introduisons de façon formelle la représentation des données longitudinales.

individu	mesure	réponse	covariables
1	1	$Y_{1,1}$	$X_{1,1}^{(1)} \dots X_{1,1}^{(p)}$
1	2	$Y_{1,2}$	$X_{1,2}^{(1)} \dots X_{1,2}^{(p)}$
\vdots	\vdots	\vdots	\vdots
1	m_1	Y_{1,m_1}	$X_{1,m_1}^{(1)} \dots X_{1,m_1}^{(p)}$
\vdots	\vdots	\vdots	\vdots
n	1	$Y_{n,1}$	$X_{n,1}^{(1)} \dots X_{n,1}^{(p)}$
\vdots	\vdots	\vdots	\vdots
n	m_n	Y_{n,m_n}	$X_{n,m_n}^{(1)} \dots X_{n,m_n}^{(p)}$

TABLE 2 – Tableau des données longitudinales

Tout d'abord, nous avons le modèle comme suit :

$$Y_{ij} = \beta_{i,0} + \beta_{i,1}X_{ij} + \varepsilon_{ij} \text{ avec } \beta_i = \begin{pmatrix} \beta_{i,0} \\ \beta_{i,1} \end{pmatrix}$$

$$\text{avec } E(Y_{ij}|\beta_i) = \beta_{i,0} + \beta_{i,1}X_{ij}$$

où

- Y_{ij} est la réponse de l'individu i au temps d'observation j avec $i \in \{1 \dots n\}$ et $j \in \{1 \dots m_i\}$

- X_{ij} est la variable variant dans le temps de l'individu i .
- $\beta_{i,0}$ et $\beta_{i,1}$ les coefficients pour l'individu i
- $\beta_{i,0} + \beta_{i,1}X_{ij}$ représente l'observation pour l'individu i au temps j .

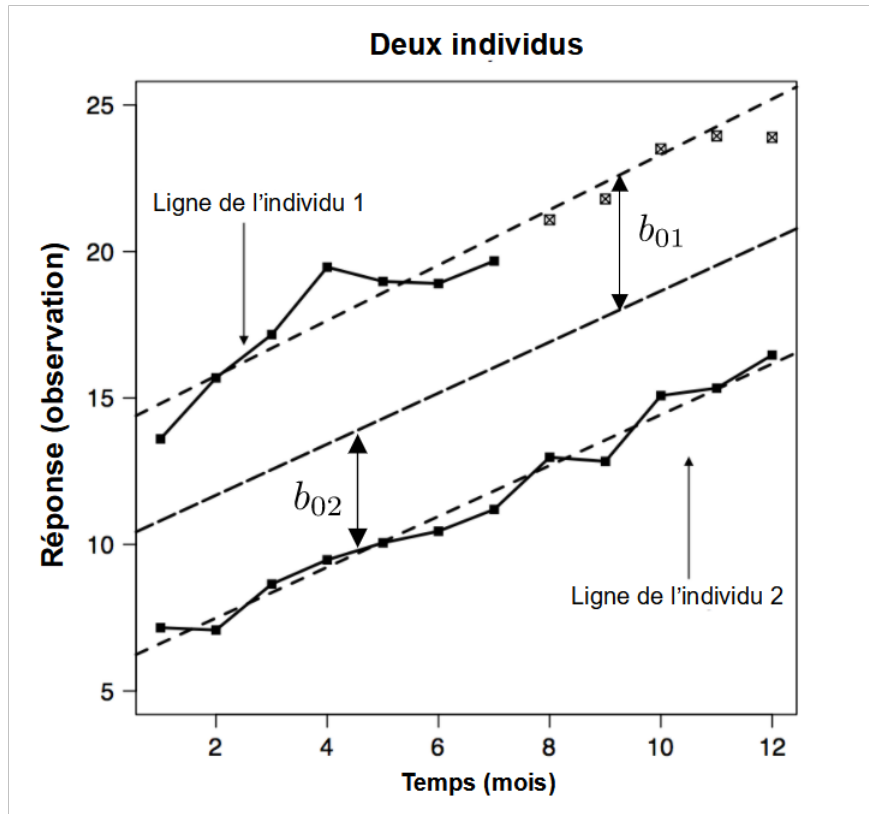


FIGURE 7 – Données longitudinales hypothétiques pour deux individus. D'après Belle & al. [6]

Si nous posons $b_{i,0} = (\beta_{i,0} - \beta_0)$ et $b_{i,1} = (\beta_{i,1} - \beta_1)$, alors le modèle peut s'écrire comme suit :

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + b_{i,0} + b_{i,1} X_{ij} + \varepsilon_{ij}$$

avec $b_{i,0}$ et $b_{i,1}$ les effets aléatoires représentant respectivement les déviations par rapport à l'intercept et à la pente moyenne du groupe d'individus (voir la figure 7). Ce sont les effets aléatoires du modèle.

Avec p observations, nous avons :

$$Y_{ij} = \beta_0 + \beta_1 X_{ij,1} + \dots + \beta_p X_{ij,p} + b_{i,0} + b_{i,1} X_{ij,1} + \dots + b_{i,p} X_{ij,p} + \varepsilon_{ij}$$

Écrit autrement :

$$Y_{ij} = X'_{ij} \beta + Z'_{ij} b_i + \varepsilon_{ij}$$

avec

$$X'_{ij} = (X_{ij,1}, X_{ij,2}, \dots, X_{ij,p})$$

$$Z'_{ij} = (X_{ij,1}, X_{ij,2}, \dots, X_{ij,q}) \quad \text{où } q < p$$

Le modèle est appelé modèle linéaire à effets mixtes car il s'agit d'une combinaison linéaire d'effets fixes et aléatoires.

Les effets aléatoires sont supposés être normalement distribués dans un modèle linéaire à effets mixtes. [5]

3.3.3 Ecriture matricielle

$$Y_{ij} = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{im_i} \end{pmatrix}, \quad X_{ij} = \begin{pmatrix} 1 & X_{i1}^{(1)} & \dots & X_{i1}^{(p)} \\ 1 & X_{i2}^{(1)} & \dots & X_{i2}^{(p)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{in}^{(1)} & \dots & X_{in}^{(p)} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{pmatrix}$$

$$b_i = \begin{pmatrix} b_{i0} \\ b_{i1} \\ \vdots \\ b_{im_i} \end{pmatrix}, \quad e_i = \begin{pmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{im_i} \end{pmatrix}$$

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{im_i} \end{pmatrix} = \begin{pmatrix} 1 & X_{i1}^{(1)} & X_{i1}^{(2)} \\ 1 & X_{i2}^{(1)} & X_{i2}^{(2)} \\ \vdots & \vdots & \vdots \\ 1 & X_{in}^{(1)} & X_{in}^{(2)} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} 1 & X_{i1}^{(1)} \\ 1 & X_{i2}^{(1)} \\ \vdots & \vdots \\ 1 & X_{in}^{(1)} \end{pmatrix} \begin{pmatrix} b_{i0} \\ b_{i1} \end{pmatrix} + \begin{pmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{in} \end{pmatrix}$$

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{im_i} \end{pmatrix} = \begin{pmatrix} 1 & X_{i1}^{(1)} & X_{i1}^{(2)} \\ 1 & X_{i2}^{(1)} & X_{i2}^{(2)} \\ \vdots & \vdots & \vdots \\ 1 & X_{in}^{(1)} & X_{in}^{(2)} \end{pmatrix} \begin{pmatrix} \beta_{i0} \\ \beta_{i1} \\ \beta_{i2} \end{pmatrix} + \begin{pmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{in} \end{pmatrix}$$

Dans ce qui suivra, $X_{ij}^{(1)} = t_{ij}$ désignera le temps d'observation pour chaque sujet i (la covariable dépendante du temps). Nous ferons en sorte qu'une des covariables générée soit dépendante du temps via la génération de données longitudinales à partir des temps t_{ij}

3.3.4 Génération des données longitudinales via un modèle linéaire mixte

Nous allons faire une procédure en deux temps (TSA ou Two-Step Approach dans la littérature scientifique [4]) :

- Dans un premier temps, nous générons des données longitudinales Y_i afin d'avoir les paramètres estimés voulus.
- Nous utilisons ces paramètres estimés pour générer des durées de survie avec dépendance au temps.

Voici la procédure pour générer des données longitudinales :

- Nous simulons deux covariables indépendantes Z_1 et Z_2 :

$$Z_1 \sim N(\mu, \sigma^2) \text{ et } Z_2 \sim \text{Bernouilli}(p) \text{ avec } p \in \{0, 1\}$$


Nous pouvons bien sûr générer des covariables via d'autres lois de probabilités

- Nous générons par la suite les trajectoires longitudinales $\phi(t_{ij})$ tels que

$$\phi(t_{ij}) = \beta_{i,0} + \beta_{i,1} \times t_{ij} + \beta_{i,2} Z_1$$

pour chaque sujet $i \in \{1, 2, \dots, n\}$ et chaque temps d'observation $j \in \{1, 2, \dots, m_i\}$ en utilisant un modèle linéaire pour $t_{ij} \in \{0, 1, 2, 3, 4, 5, \dots\}$

Pour cela :

- Les paramètres estimés des effets aléatoires mixtes β_i de la moyenne et de la matrice de variance-covariance peuvent être obtenus en adaptant un modèle linéaire à effets mixtes (LME) à des données longitudinales connues. Dans notre code  (section 6.1), nous simulons ces paramètres de moyenne et de matrice de variance-covariance.

$$\beta_i = \begin{pmatrix} \beta_{i,0} \\ \beta_{i,1} \end{pmatrix} \sim N(\beta, \Sigma)$$

avec

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \quad \Sigma = \begin{pmatrix} \text{var}(b_{i0}) & \text{cov}(b_{i0}, b_{i1}) \\ \text{cov}(b_{i0}, b_{i1}) & \text{var}(b_{i1}) \end{pmatrix}$$

- Nous générons les effets aléatoires mixtes (β_{i1}, β_{i2}) à partir d'une distribution normale bivariée avec une moyenne et une matrice de variance-covariance obtenues à l'étape précédente. Les effets aléatoires (b_{i1}, b_{i2}) représentent les déviations ou parties aléatoires de l'intercept et la pente de chaque individu i .

$$b_{i1} = \beta_{i,1} - \beta_1$$

$$b_i = \begin{pmatrix} \beta_{i,1} - \beta_1 \\ \beta_{i,2} - \beta_2 \end{pmatrix} \sim N(0_p, \Sigma)$$

- Enfin, nous générons les mesures observées des covariables dépendantes du temps $Y_{ij}(t_{ij}) = \phi(t_{ij}) + \epsilon_{ij} = \beta_{i,1} + \beta_{i,2} \times t_{ij} + \epsilon_{ij}$ à partir d'une distribution normale multivariée de moyenne $\phi(t_{ij})$ et de variance V_i

$$Y_{ij} \mid \beta_i \sim N(\phi(t_{ij}), V_i)$$

tel que :

$$V_i = Z_i \Sigma Z_i' + R_i$$

avec :

- Σ la matrice de variance-covariance des effets aléatoires
- $Z_i = (\mathbf{1} \mid t_{ij})$ la matrice des temps d'observation longitudinale t_{ij} (covariables dépendantes du temps)
- $R_i = \sigma^2 I_{m_i}$ la matrice des régions de risque de l'individu i où $\epsilon_i \sim N(0, R_i)$ i.e $\epsilon_{ij} \sim N(0, \sigma^2)$.
- Nous ajustons enfin le modèle à effets aléatoires aux mesures longitudinales observées pour obtenir le couple $(\beta_{i,0}, \beta_{i,1})$ pour chaque individu i . Ces effets aléatoires seront utilisés pour générer les données de survie. [4]

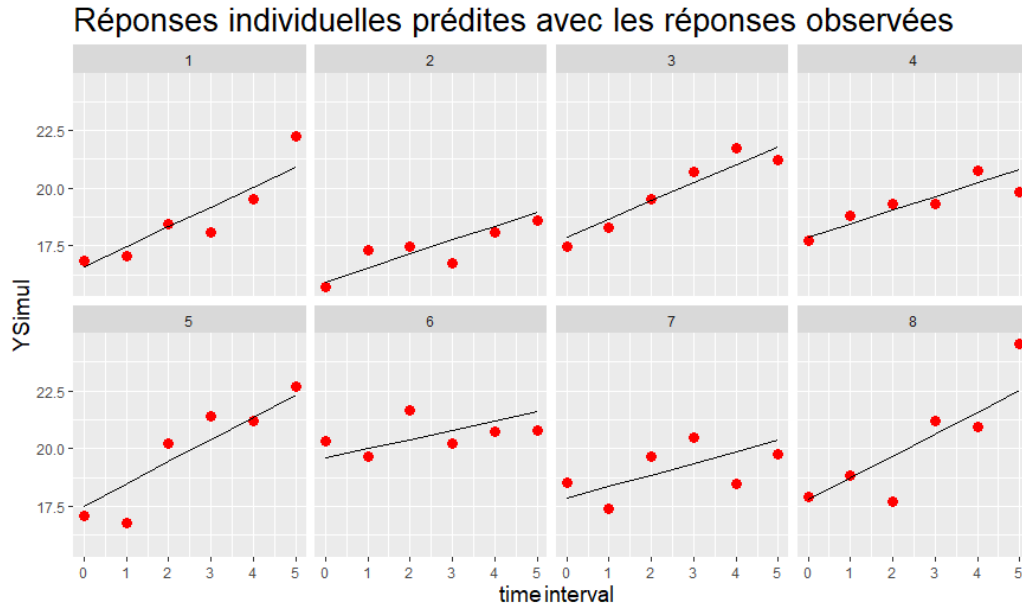


FIGURE 8 – Mesures longitudinales générées via une simulation pour les 8 premiers individus (section 6.1)

3.4 Générer des durées de survie avec la fonction de Lambert et une distribution de Weibull

Dans cette section, nous présentons une approche pour simuler les temps de survie avec une covariable dépendante du temps à l'aide de la fonction Lambert W (LWF).

3.4.1 Fonction W de Lambert

La fonction W de Lambert est la fonction réciproque de la fonction $f(w) = we^w$ tel que

$$z = f(w) = we^w \iff w = W(z) \quad \text{avec } z, w \in \mathbf{C}$$

3.4.2 Génération de durées de survie avec covariable dépendante du temps

Pour une distribution de Weibull des temps de survie avec covariable dépendante du temps, nous notons Y_{ij} les mesures longitudinales observées, qui suivent un modèle linéaire mixte avec j le nombre d'observations de l'individu i . Nous considérons le vecteur des effets aléatoires mixtes $\beta_i = (\beta_{i,0}, \beta_{i,1})$ pour le modèle linéaire mixte suivant : $Y_{ij} = \phi(t) + \epsilon_{ij} = \beta_{i,0} + \beta_{i,1} \times t + \epsilon_{ij}$ où les $\epsilon_{ij} \sim N(0, \sigma^2)$ sont les termes d'erreur de mesure.

Pour une loi de Weibull, la fonction de hasard cumulé est $H_0(t) = \lambda t^\alpha$, si $h_0(t) > 0 \forall t$, alors H_0 peut être inversée et à partir de l'équation (1) ci-dessus nous avons pour $Q \sim U[0, 1]$:

$$\frac{-\log Q}{\lambda \exp(Z'\beta + \gamma \cdot (\beta_{i,1} \times t))} = t^\alpha \times \exp(\gamma \cdot (\beta_{i,1} \times t)) \quad (2)$$

Une forme équivalente de (2) peut s'écrire comme :

$$\left(\frac{-\log(Q)}{\lambda \exp(Z'\beta + \gamma \cdot (\beta_{i,0}))} \right)^{\frac{1}{\alpha}} = t \times \exp(\gamma \cdot (\beta_{i,1} \times t))^{\frac{1}{\alpha}}$$

En appliquant la fonction de Lambert, nous obtenons :

$$\gamma \cdot \left(\beta_{i,1} \times t \times \frac{1}{\alpha} \right) = W \left(\gamma \cdot \left(\beta_{i,1} \times \frac{1}{\alpha} \right) \times \left(\frac{-\log(Q)}{\lambda \exp(Z'\beta + \gamma \cdot (\beta_{i,0}))} \right)^{\frac{1}{\alpha}} \right)$$

En résolvant l'équation ci-dessus, le temps de survie peut s'écrire :

$$t = \frac{1}{\gamma \cdot (\beta_{i,1} \times \frac{1}{\alpha})} \times W \left(\gamma \cdot \left(\beta_{i,1} \times \frac{1}{\alpha} \right) \times \left(\frac{-\log(Q)}{\lambda \exp(Z'\beta + \gamma(\beta_{i,0}))} \right)^{\frac{1}{\alpha}} \right) \quad (3)$$

Résumons la procédure :

- Nous spécifions les valeurs pour les paramètres (coefficients) estimés des covariables Z_i et le paramètre de liaison γ qui mesure la force de l'association entre les données longitudinales $Y_{ij}(t_{ij})$ et le temps d'évènement T_i
- Nous générons le temps d'évènement T_i via la transformation inverse de Weibull avec dépendance du temps (3)
- Nous générons une variable du temps de censure $C_i \sim U[a, b]$ avec $a, b \in \mathbf{R}_+$ tels que $a < b$. A partir des temps de survie et de la censure, nous obtenons l'indicateur de censure Δ_i tel que :

$$\Delta_i = \begin{cases} 1 & \text{si } T_i \leq C_i \\ 0 & \text{si } T_i > C_i \end{cases}$$

Ainsi, nous obtenons le jeu de données voulu. [3]

3.5 Simulation d'un jeu de données sur commenté

Ces données de survie générées se trouvent au tableau 5.

```
### Creation des variables globales

time <- c(0, 1, 2, 3, 4, 5)

N=500

M=6

### Creation des effets aleatoires mixtes

# On simule un B_i1 et B_i2 avec une variance pour chacun
  et une covariance (il s'agit des effets aleatoires
    mixtes Beta_i)

beta=c(16.761, 0.6601)

#Matrice de variance covariance simulee

S=structure(.Data=c(5.4149, -0.3214, -0.3214, 0.0513), .
  Dim=c(2, 2))
```

Attribution des valeurs des coefficients fixes et de la matrice de variance-covariance des effets aléatoires :

$$\beta = \begin{pmatrix} 16.701 \\ 0.6601 \end{pmatrix} \quad \text{et} \quad \Sigma = \begin{pmatrix} 5.4149 & -0.3214 \\ -0.3214 & 0.0513 \end{pmatrix}$$

```
RMVN=rmvnorm(N, mean=beta, sigma=S, method = "chol")

RMVNB=rmvnorm(N, mean= matrix(0,1,2), sigma=S, method = "
  chol")

Z=matrix(0,M,length(dim(S)))

Z[,1] <- 1
```

```

Z[,2] <- c(time)

muy <- matrix(0,N,M)

Y <- matrix(0,N,M)

R <- diag(rnorm(M,1.000,0.00001), M)

V=Z%*%S%*%t(Z) + R

```

Estimation des effets aléatoires mixtes β_i et effets aléatoires b_i :

$$\beta_i = \begin{pmatrix} \beta_{i,0} \\ \beta_{i,1} \end{pmatrix} \sim N(\beta, \Sigma)$$

$$b_i = \begin{pmatrix} \beta_{i,1} - \beta_1 \\ \beta_{i,2} - \beta_2 \end{pmatrix} \sim N(0_p, \Sigma)$$

Création des $Y_{ij}(t_{ij})$ et de la variance V_i avec :

$$V_i = Z_i \Sigma Z_i' + R_i$$

```

### Covariables

sexe=rbern(N,0.540)

for (i in 1:N){if (sexe[i] == 0) sexe[i]=1 else sexe[i]=2
}

age=round(rnorm(N,35,5))

```

Simulation des deux covariables indépendantes Z_1 (âge) et Z_2 (sexe) de telle sorte que

$$Z_1 \sim N(35, 5) \text{ et } Z_2 \sim \text{Bernouilli}(0.54)$$

Ces covariables ont été générées afin de correspondre le plus possible à des covariables issues de l'étude de Framingham ([Framingham Heart Study](#))

```

### Generation des trajectoires longitudinales

for (j in 1:M){

  for (i in 1:N){

    muy[i, j] <- RMVN[i,1] + RMVN[i,2]*(time[j])

  }

}

muy <- data.frame(muy)

names(muy) <- c("X1", "X2", "X3", "X4", "X5", "X6")

```

Génération des trajectoires longitudinales $\phi(t_{ij})$ tels que

$$\phi(t_{ij}) = \beta_{i,0} + \beta_{i,1} \times t_{ij} + \beta_{i,2}Z_1$$

pour chaque sujet et chaque temps d'observation.

```

### Simulation des valeurs de Y

Y <- rmvnorm(n=N, mean=colMeans(muy), sigma=V, method = "
  chol")

Y <- data.frame(Y)

names(Y) <- c("Y1", "Y2", "Y3", "Y4", "Y5", "Y6")

erreur <- rnorm(N, 0, 0.001)

for (i in 1:N){

  Y[i, ] <- Y[i, ]+0.05*age[i] + erreur[i]

}

```

Simulation des Y_{ij} telle que

$$Y_{ij} \mid \beta_i \sim N(\phi(t_{ij}), V_i)$$

Ajout des erreurs résiduelles :

$$Y_{ij} = \phi(t) + \epsilon_{ij} = \beta_{i,0} + \beta_{i,1} \times t + \epsilon_{ij} \quad \text{où les } \epsilon_{ij} \sim N(0, \sigma^2)$$

```
### Regression du modele lineaire mixte

timeinterval <- rep(time, N)

YSimul <- c(t(as.matrix(Y)))

ID <- rep(1:N, each=M)

Cage <- rep(age, each=M)

Long <- data.frame(ID, YSimul, timeinterval, Cage)

LME <- lme(YSimul~timeinterval+Cage, random = ~
  timeinterval | ID,data=Long,
  control=lmeControl(msMaxIter = 100, msVerbose=
    TRUE, opt = "optim"))

U <- coef(LME)

names(U) <- c("X1_1", "X2_1")

LME_Coeff <- data.frame(coefficients(LME))

colMeans(LME_Coeff)
```

Ajustement du modèle à effets aléatoires aux mesures longitudinales observées pour obtenir le couple $(\beta_{i,0}, \beta_{i,1})$ pour chaque individu i .

```
#### Specification des estimations de param tres

Gamma <- 0.500

agebeta <- 0.050

sexebeta <- -0.500

### Temps de survie en utilisant la distribution de
  Weibull
```



```

wshape = 3

# Lambda (Scale) = (1/Lambda)^v

wikiscale = 150

wscale = 1/(wikiscale)^(wshape)

Uni <- runif(N, min = 0, max = 1)

Numweib <- -log(Uni)

Denweib <- wscale*exp(agebeta*age + sexebeta*sexe + Gamma
*(U[,1]))

ratioweib <- Gamma*(U[,2])*(1/wshape)*((Numweib/Denweib)
^(1/wshape))

Lweib <- LambertW(ratioweib)

Survweib <- Lweib*1/(Gamma*(U[,2])*(1/wshape))

```

Génération de la durée de survie T via la transformée inverse et la fonction de Lambert W :

$$T = \frac{1}{\gamma \cdot (\beta_{i,1} \times \frac{1}{\alpha})} \times W \left(\gamma \cdot \left(\beta_{i,1} \times \frac{1}{\alpha} \right) \times \left(\frac{-\log(Q)}{\lambda \exp(Z'\beta + \gamma(\beta_{i,0}))} \right)^{\frac{1}{\alpha}} \right)$$

```

### Censure des temps via une distribution uniforme

Censoring <- runif(n=N, min = 0, max = 5)

### Variable de temps de survenue de l'evenement

timeevent=pmin(Survweib, Censoring)

status=as.numeric(Survweib <= Censoring)

```

Génération du temps de censure C_i et de l'indicateur de censure Δ_i

3.5.1 Courbes de survie avec dépendance du temps

```
strata_age<-cut(LongSurv$age,breaks=c(15,35,50))  
  
f<- survfit(Surv(timeevent, status) ~ strata_age, type =  
  "kaplan-meier", conf.type = "plain", LongSurv)  
ggsurvplot(f, data = LongSurv)  
ggsurvplot(f, fun = "cloglog", data = LongSurv, xlim=c(0.  
  1,10))
```

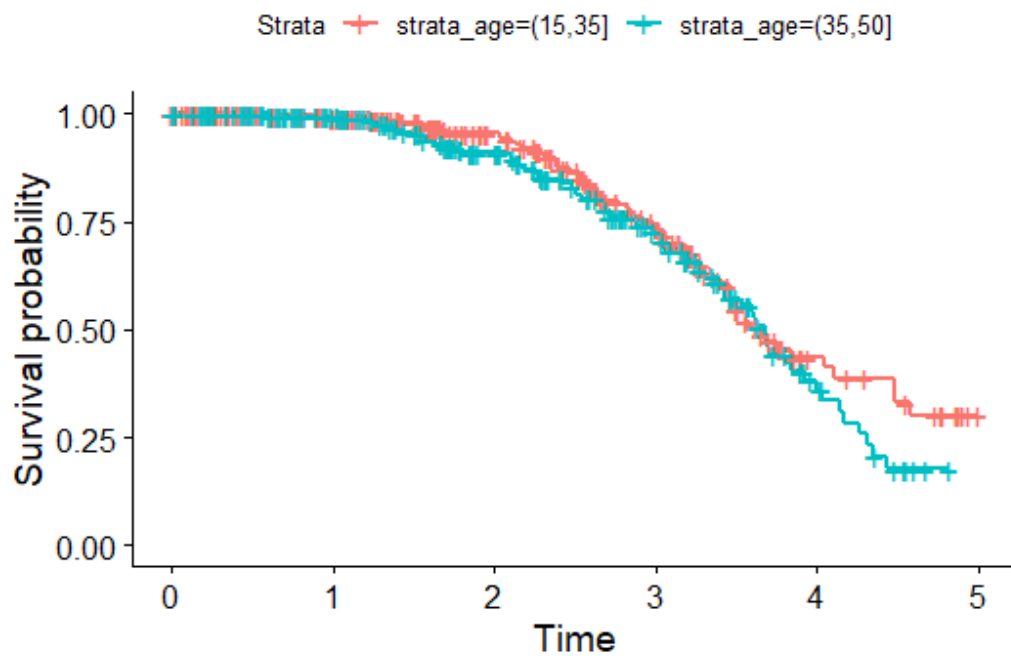


FIGURE 9 – Courbe de survie avec dépendance du temps

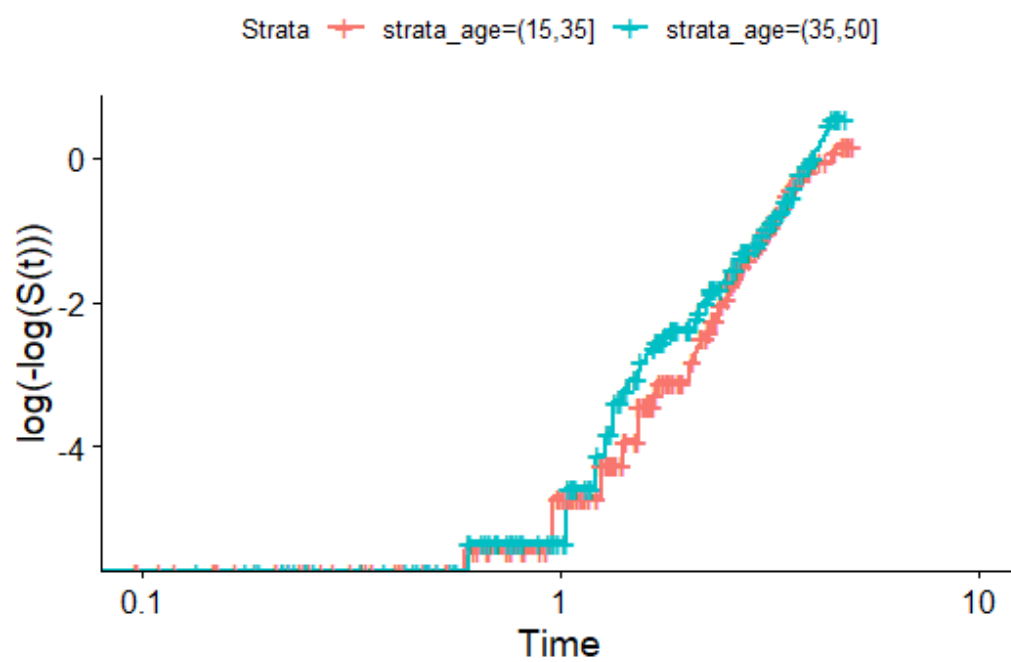


FIGURE 10 – Hypothèse de log-linéarité rejetée

3.5.2 Limites

Finalement, il n'est pas possible d'illustrer la réfutation de l'hypothèse des risques proportionnels du modèle de Cox classique à l'aide de la fonction `cox.zph` tel qu'effectué en section 2.2.2.

En effet "les termes d'effets aléatoires tels que les effets aléatoires dans un modèle linéaire mixte ne sont pas vérifiés pour les risques proportionnels, ils sont plutôt traités comme un décalage fixe dans le modèle."

Source : [Documentation](#)  de la fonction `cox.zph`

.

4 Estimation des coefficients

Les coefficients du modèle linéaire mixte peuvent être estimés en utilisant les packages `R nlme` et `lme4`. Le code `R` utilisé afin d'obtenir ces valeurs se trouve en annexe en section 6.3

4.1 Coefficients estimés

Paramètres statistiques observés

$$\beta = \begin{pmatrix} 16.701 \\ 0.6601 \end{pmatrix}$$
$$\Sigma = \begin{pmatrix} 5.4149 & -0.3214 \\ -0.3214 & 0.0513 \end{pmatrix}$$

Paramètres statistiques estimés

Les effets fixes estimés sont :

$$\hat{\beta}_1 = 15.608489 \quad , \quad \hat{\beta}_2 = 0.675938 \quad , \quad \hat{\beta}_3 = 0.082484$$

Les erreurs standards et la corrélation de ces estimations sont les suivantes :

$$se(\hat{\beta}_1) = 0.5918143 \quad , \quad se(\hat{\beta}_2) = 0.0149335, \quad se(\hat{\beta}_3) = 0.0166120$$
$$corr(\hat{\beta}_1, \hat{\beta}_2) = -0.096 \quad corr(\hat{\beta}_1, \hat{\beta}_3) = -0.985 \quad corr(\hat{\beta}_2, \hat{\beta}_3) = 0$$

Les écarts types et la corrélation estimés des effets aléatoires sont les suivants :

$$\widehat{sd}(b_{i1}) = 2.150225 \quad , \quad \widehat{sd}(b_{i2}) = 0.2416097 \quad , \quad \widehat{corr}(b_{i1}, b_{i2}) = -0.56$$

La matrice de variance-covariance estimée des effets aléatoires est donc :

$$\widehat{\Sigma} = \begin{pmatrix} 4.62345698 & -0.29112163 \\ -0.29112163 & 0.0583752 \end{pmatrix}$$

Enfin la variance estimée des erreurs résiduelles est

$$\hat{\sigma}^2 = 0.9642480$$

5 Conclusion

Dans une première partie nous avons présenté le modèle de régression de Cox standard et les hypothèses qui y sont rattachées. Nous avons pu par la suite simuler des jeux de données avec covariables indépendantes du temps via deux méthodes distinctes faisant toutes deux intervenir la distribution de Weibull pour la génération des durées de survie des observations simulées. Ces méthodes nous ont permises d'illustrer la vérification des hypothèses du modèle de régression de Cox à partir des jeux de données simulées.

Dans une seconde partie, nous nous sommes penchés sur la simulation de jeu de données avec covariables dépendantes du temps. Pour ce faire, nous avons généré des données longitudinales à l'aide d'un modèle linéaire mixte que nous avons exploitées par la suite pour la génération de durées de survie via une distribution de Weibull en faisant intervenir la méthode de la transformée inverse et la fonction de Lambert.

6 Annexes

6.1 Simulation des jeux de données

```
# Taille echantillon = 500; Taux de survenue de l'
  evenement = 21%; gamma = 0.500;

library(MCMCpack)

library(survminer)

library(stats)

library(Rlab)

library(MASS)

library(Matrix)

library(mvtnorm)

library(survival)

library(ggplot2)

library(coda)

library(lattice)

library(boa)

library(nlme)

library(car)

set.seed(33)

### Creation des variables globales

time <- c(0, 1, 2, 3, 4, 5)

N=500
```

```

M=6

### Creation des effets aleatoires mixtes

# On simule un B_i1 et B_i2 avec une variance pour chacun
  et une covariance (il s'agit des effets aleatoires
    mixtes Beta_i)

beta=c(16.761, 0.6601)

#Matrice de variance covariance simulee

S=structure(.Data=c(5.4149, -0.3214, -0.3214, 0.0513), .
  Dim=c(2, 2))

RMVN=rmvnorm(N, mean=beta, sigma=S, method = "chol")
RMVNB=rmvnorm(N, mean= matrix(0,1,2), sigma=S, method = "
  chol")

Z=matrix(0,M,length(dim(S)))

Z[,1] <- 1

Z[,2] <- c(time)

muy <- matrix(0,N,M)

Y <- matrix(0,N,M)

R <- diag(rnorm(M,1.000,0.00001), M)

V=Z%*%S%*%t(Z) + R

### Covariables

sexe=rbern(N,0.540)

for (i in 1:N){if (sexe[i] == 0) sexe[i]=1 else sexe[i]=2
}

age=round(rnorm(N,35,5))

### Generation des trajectoires longitudinales

```



```

for (j in 1:M){
  for (i in 1:N){
    muy[i, j] <- RMVN[i,1] + RMVN[i,2]*(time[j])
  }
}

muy <- data.frame(muy)

names(muy) <- c("X1", "X2", "X3", "X4", "X5", "X6")

### Simulation des valeurs de Y

Y <- rmvnorm(n=N, mean=colMeans(muy), sigma=V, method = "
chol")

Y <- data.frame(Y)

names(Y) <- c("Y1", "Y2", "Y3", "Y4", "Y5", "Y6")

erreur <- rnorm(N, 0, 0.001)

for (i in 1:N){
  Y[i, ] <- Y[i, ]+0.05*age[i] + erreur[i]
}

boxplot(Y, main = "Boxplot des covariables dependantes du
temps mesurees")

### Regression du modele lineaire mixte

timeinterval <- rep(time, N)

YSimul <- c(t(as.matrix(Y)))

ID <- rep(1:N, each=M)

```

```

Cage <- rep(age, each=M)

Long <- data.frame(ID, YSimul, timeinterval, Cage)

LME <- lme(YSimul~timeinterval+Cage, random = ~
  timeinterval | ID,data=Long,
  control=lmeControl(msMaxIter = 100, msVerbose=
    TRUE, opt = "optim"))

U <- coef(LME)

names(U) <- c("X1_1", "X2_1")

LME_Coeff <- data.frame(coefficients(LME))

colMeans(LME_Coeff)

### Fonction W de Lambert

LambertW=function(z, b=0, maxiter=10, eps=.Machine$double
  .eps, min.imag = 1e-9) {

  if (any(round(Re(b)) != b))

    stop("W doit tre un entier")

  if (!is.complex(z) && any(z < 0)) z = as.complex(z)

  w = (1 - 2*abs(b))*sqrt(2*exp(1)*z + 2) - 1

  ## expansion asymptotique 0 et Inf

  v = log(z + as.numeric(z==0 & b==0)) + 2*pi*b*1i;

  v = v - log(v + as.numeric(v==0))

  ## choisir une strategie pour la premi re estimation

  c = abs(z + exp(-1));

  c = (c > 1.45 - 1.1*abs(b));

  c = c | (b*Im(z) > 0) | (!Im(z) & (b == 1))

```

```

w = (1 - c)*w + c*v

## Halley iteration

##

for (n in 1:maxiter) {

  p = exp(w)

  t = w*p - z

  f = (w != -1)

  t = f*t/(p*(w + f) - 0.5*(w + 2.0)*t/(w + f))

  w = w - t

  if (abs(Re(t)) < (2.48*eps)*(1.0 + abs(Re(w)))
      && abs(Im(t)) < (2.48*eps)*(1.0 + abs(Im(w))))
    break

}

if (n==maxiter) warning(paste("iteration limit (",
  maxiter,") reached, result of W may be inaccurate",
  sep=""))

if (all(Im(w) < min.imag)) w = as.numeric(w)

return(w)

}

#### Specification des estimations de param tres

Gamma <- 0.500

agebeta <- 0.050

```

```

sexebeta <- -0.500

### Temps de survie en utilisant la distribution de
  Weibull

wshape = 3

# Lambda (Scale) = (1/Lambda)^v

wikiscale = 150

wscale = 1/(wikiscale)^(wshape)

Uni <- runif(N, min = 0, max = 1)

Numweib <- -log(Uni)

Denweib <- wscale*exp(agebeta*age + sexebeta*sexe + Gamma
  *(U[,1]))

ratioweib <- Gamma*(U[,2])*(1/wshape)*((Numweib/Denweib)
  ^(1/wshape))

Lweib <- LambertW(ratioweib)

Survweib <- Lweib*1/(Gamma*(U[,2])*(1/wshape))

hist(Survweib)

### Censure des temps via une distribution uniforme

Censoring <- runif(n=N, min = 0, max = 5)

### Variable de temps de survenue de l'evenement

timeevent=pmin(Survweib, Censoring)

status=as.numeric(Survweib <= Censoring)

table(status)

### Donnees generees

```

```
ID <- rep(1:N)

LongSurv <- data.frame(ID, timeevent, status, age, sexe)

strata_age<-cut(LongSurv$age,breaks=c(15,35,50))

f<- survfit(Surv(timeevent, status) ~ strata_age, type =
  "kaplan-meier", conf.type = "plain", LongSurv)
ggsurvplot(f, data = LongSurv)
ggsurvplot(f, fun = "cloglog", data = LongSurv, xlim=c(0.
  1,10))
```

6.2 Tableaux de données de survie avec indépendance et dépendance du temps

id	time	status	Z
1	4.30350447583891	1	34.9416159535758
2	33.3937575671256	1	53.6635086522438
3	1.91465907184548	1	16.4665146975312
4	13.8402382284403	0	49.3133641185705
5	7.02675018464967	1	35.6287161633372
6	1.61917428349337	0	40.4571199009661
7	7.10778385151036	1	27.1263798605651
8	0.0631177052855492	0	66.4912935683969
9	5.76664806820572	1	49.1585529618897

TABLE 3 – Jeu de données simulées avec la méthode du modèle log-linéaire

id	time	status	sexe	age
1	3.78131989677095	1	2	33
2	3.20401032611256	1	1	39
3	0.606578774750233	0	2	31
4	2.29220402100908	0	1	34
5	4.57843154879496	1	2	30
6	1.8531678280215	1	2	33
7	4.15312609239579	1	2	35
8	2.16125916757321	1	2	40
9	3.0561903249519	1	2	36

TABLE 4 – Jeu de données simulées avec la méthode de la transformée inverse

ID	timeevent	status	age	sexe
1	0.450336593203247	0	38	1
2	1.39453388284892	0	29	1
3	2.32733723474666	0	38	1
4	2.53727840608917	0	35	1
5	1.28621067618951	0	42	2
6	0.662009874358773	0	39	1
7	3.34438525256701	0	31	1
8	4.08358773682266	0	33	1
9	0.689597429009154	0	27	2
10	1.12496465910226	0	45	1
11	2.73812527768314	0	33	2
12	2.49274279245924	1	34	2

TABLE 5 – Tableau du jeu de données de survie avec dépendance du temps

6.3 Coefficients estimés

```
# Recuperer les coefficients estimates et la matrice de
  variance-covariance

> summary(LME)
Linear mixed-effects model fit by REML
  Data: Long
      AIC      BIC    logLik
10251.37 10293.41 -5118.685

Random effects:
  Formula: ~timeinterval | ID
  Structure: General positive-definite, Log-Cholesky
    parametrization
              StdDev      Corr
(Intercept)  2.150225 (Intr)
timeinterval 0.2416097 -0.56
Residual      0.9642480

Fixed effects: YSimul ~ timeinterval + Cage
              Value Std.Error   DF  t-value p-
              value
(Intercept)  15.608489 0.5918143 2499 26.37396
              0
timeinterval  0.675938 0.0149335 2499 45.26313      0
Cage          0.082484 0.0166120  498  4.96533
              0
Correlation:
              (Intr) tmvntn
timeinterval -0.096
Cage         -0.985  0.000

Standardized Within-Group Residuals:
      Min      Q1      Med      Q3
      Max
-3.1233163905 -0.5916319752  0.0001073076  0.5704319220
 2.8652059149

Number of Observations: 3000
```



```

Number of Groups: 500

> VarCorr(LME)
ID = pdLogChol(timeinterval)
      Variance StdDev   Corr
(Intercept)  4.62345698 2.1502225 (Intr)
timeinterval 0.05837523 0.2416097 -0.56
Residual      0.92977420 0.9642480

#####

# Idem, avec les covariances des effets aléatoires en
# plus (non disponibles sur le package nlme)

library(lme4)

> LMER <- lmer(YSimul~timeinterval+Cage + (timeinterval |
  ID) ,data=Long)
> summary(LMER)
Linear mixed model fit by REML ['lmerMod']
Formula:
YSimul ~ timeinterval + Cage + (timeinterval | ID)
Data: Long

REML criterion at convergence: 10237.4

Scaled residuals:
      Min       1Q   Median       3Q      Max
-3.12332 -0.59163  0.00011  0.57042  2.86519

Random effects:
  Groups      Name                Variance Std.Dev. Corr
ID           (Intercept)          4.62332  2.1502
            timeinterval 0.05838  0.2416   -0.56
Residual                        0.92977  0.9642
Number of obs: 3000, groups: ID, 500

Fixed effects:
              Estimate Std. Error t value
(Intercept)    15.60849    0.59181  26.374
timeinterval    0.67594    0.01493  45.263
Cage             0.08248    0.01661   4.965

```

```

Correlation of Fixed Effects:
              (Intr)  tmvntn
timvntntrvl  -0.096
Cage         -0.985   0.000

> VcoV <- VarCorr(LMER)$ID[,]
> VcoV

              (Intercept)  timeinterval
(Intercept)      4.6233195      -0.29112163
timeinterval    -0.2911216       0.05837693

#####

#Tracer les donnees longitudinales pour les 9 premiers
  individus

Long$pred.LMER <- fitted(LMER)

ggplot(data= subset(Long, ID<9))+ geom_point(aes(x=
  timeinterval,y= YSimul), color="red", size=3) +
geom_line(aes(x= timeinterval,y=pred.LMER)) + facet_wrap(
  ~ID, ncol=4)

```

6.3.1 Tableaux des coefficients estimés

individu	$\widehat{\beta}_{i,0}$	$\widehat{\beta}_{i,1}$
1	16.4447044861127	0.671550226696584
2	19.1125179723687	0.492116763112939
3	11.7425179685612	1.04748217505712
4	15.0035512056871	0.904175763944465
5	18.516898196738	0.358501072524506
6	17.1504223171826	0.63172929762078
7	21.12607969126	0.444925838420388
8	18.3949072512263	0.560411845741483
9	16.4273787164964	0.737464457112656
10	17.0456706410845	0.536455851579321

TABLE 6 – Tableau des valeurs des β_i estimés

individu	\widehat{b}_{i0}	\widehat{b}_{i1}
1	-0.178078096457943	0.151023288452372
2	1.76603114800033	-0.302545930667445
3	0.495372799767418	-0.12044957126205
4	2.56641351688972	0.089657331537682
5	-1.20634860253935	0.115193636972383
6	-2.44638214500153	-0.0864945653945225
7	-0.812035892293177	0.358561011974835
8	-0.53817423468855	0.287660751547043
9	-3.84003826518824	0.23173685541057
10	0.351688557070841	0.079497580265165

TABLE 7 – Tableau des valeurs des b_i estimés

Références

- [1] Peter C Austin. Generating survival times to simulate cox proportional hazards models with time-varying covariates. *Statistics in medicine*, 31(29) :3946–3958, 2012.
- [2] JC Thalabard D Thiam S Whengang J Gaudart, R Giorgi. Modèles linéaires à effets mixtes. 2010.
- [3] Julius Ngwa, Howard Cabral, Debbie Cheng, David Gagnon, Michael Lavalley, and L. Cupples. Generating survival times with time-varying covariates using the lambert w function. *Communications in Statistics - Simulation and Computation*, 2019 :1–19, 08 2019.
- [4] Julius Ngwa, Howard Cabral, Debbie Cheng, David Gagnon, Michael Lavalley, and L. Cupples. Revisiting methods for modeling longitudinal and survival data : Framingham heart study. *BMC Medical Research Methodology*, 21, 02 2021.
- [5] Dimitris Rizopoulos. Joint models for longitudinal and time-to-event data. *CRC Press*, 06 2012.
- [6] G. van Belle, L.D. Fisher, P.J. Heagerty, and T. Lumley. *Biostatistics : A Methodology For the Health Sciences*. Wiley Series in Probability and Statistics. Wiley, 2004.