Université de Montpellier

Compte rendu de TP

TP 2 : Modèles paramétriques pour les durées de vie avec ou sans covariables

---

Akkouh Maryam

14 Février 2024

Table des matières

1 Modèle exponentiel

Soit X une variable aléatoire représentant une durée de vie, de loi exponentielle de paramètre $\lambda = 1$.

1.1 Simulation d'un échantillon avec contrôle du taux d'observations censurées

On commence par simuler avec rexp() un échantillon de taille $n = 50$ de couples (X_i, C_i), $i = 1, ..., n$ avec X_i indépendant de C_i, de lois exponentielles de paramètres respectifs $\lambda = 1$ et $\mu = 0.5$.

On définit ensuite un vecteur $T = min(X, C)$ et un vecteur $\delta = \mathbb{1}(X \leq C)$.

Nous pouvons calculer la probabilité de censure théorique et la comparer par la suite à celle observée dans l'échantillon que nous venons de générer.

Calcul de $\mathbb{P}(X > C)$:

$$\begin{aligned}
\mathbb{P}(X > C) &= \int_{\mathbb{R}_+}\int_{\mathbb{R}_+} \mathbb{1}_{(x>c)}\lambda e^{-\lambda x}\mu e^{-\mu c}\,dx\,dc \\
&= \int_0^\infty \int_c^\infty \lambda e^{-\lambda x}\mu e^{-\mu c}\,dx\,dc \\
&= \int_0^\infty \mu e^{-\mu c}\left(\int_c^\infty \lambda e^{-\lambda x}\,dx\right)\,dc \\
&= \int_0^\infty \mu e^{-\mu c}\left[-e^{-\lambda x}\right]\Big|_c^\infty\,dc \\
&= \int_0^\infty \mu e^{-\mu c}e^{-\lambda c}\,dc \\
&= \int_0^\infty \mu e^{-c(\lambda+\mu)}\,dc \\
&= \left(-\frac{\mu}{\lambda+\mu}e^{-c(\lambda+\mu)}\right)\Big|_0^\infty \\
&= \frac{\mu}{\lambda+\mu} \\
&= 1 - \frac{\lambda}{\lambda+\mu}
\end{aligned}$$

Avec $\lambda = 1$ et $\mu = 0.5$ on a $\mathbb{P}(X > C) = \frac{0.5}{1+0.5} = \frac{1}{3}$

On calcule ensuite la proportion $1 - \frac{1}{n}\sum_{i=1}^n \delta_i$ de données censurées dans notre échantillon. Cela donne $prop\_obs = 0.42$ et $|prop\_obs - prop\_theorique| = 0.0867$. Nous remarquons que le taux de censure peut être contrôlé en paramétrant μ et λ.

Nous pouvons répéter ce calcul pour plusieurs réalisations de l'échantillon et observer la fluctuation de la proportion de censure.

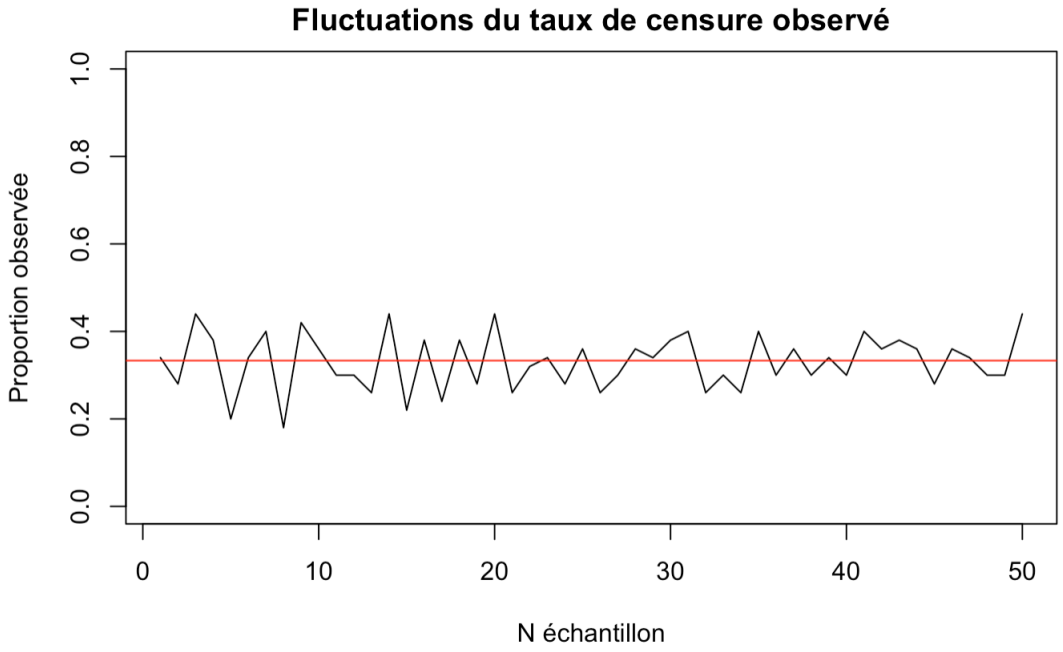


FIGURE 1 – Fluctuations du taux de censure

En rouge on représente la droite $y = \frac{1}{3}$. On remarque que la proportion observée est toujours très proche de cette dernière, avec des fluctuations assez limitées autour de la droite $y = \frac{1}{3}$. L'écart moyen entre la proportion observée et la théorique est de 0.0029. Le taux de censure est donc relativement stable quand la taille de l'échantillon varie.

Nous avons donc vu que le taux de censure peut être contrôlé en ajustant les paramètres de la distribution exponentielle, et que la proportion observée de censure reste généralement proche de la probabilité théorique prédite, ce qui indique une certaine robustesse du processus de simulation.

1.2 Estimation du paramètre λ

On commence par calculer une estimation ponctuelle du paramètre lambda pour l'échantillon simulé (cas des données censurées) en utilisant l'estimateur du maximum de vraisemblance :

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} t_i}$$

On obtient ainsi $\hat{\lambda} = 0.869$. On rappelle que la valeur théorique est $\lambda = 1$. Il n'y a donc pas un grand écart entre la valeur théorique et la valeur estimée.

On reprend la simulation de l'échantillon censuré de la partie précédente et on fait varier la taille n de l'échantillon. Pour n allant de 10 à 5000 on calcule les estimateurs $\hat{\lambda}_n$. On observe donc le résultat suivant :

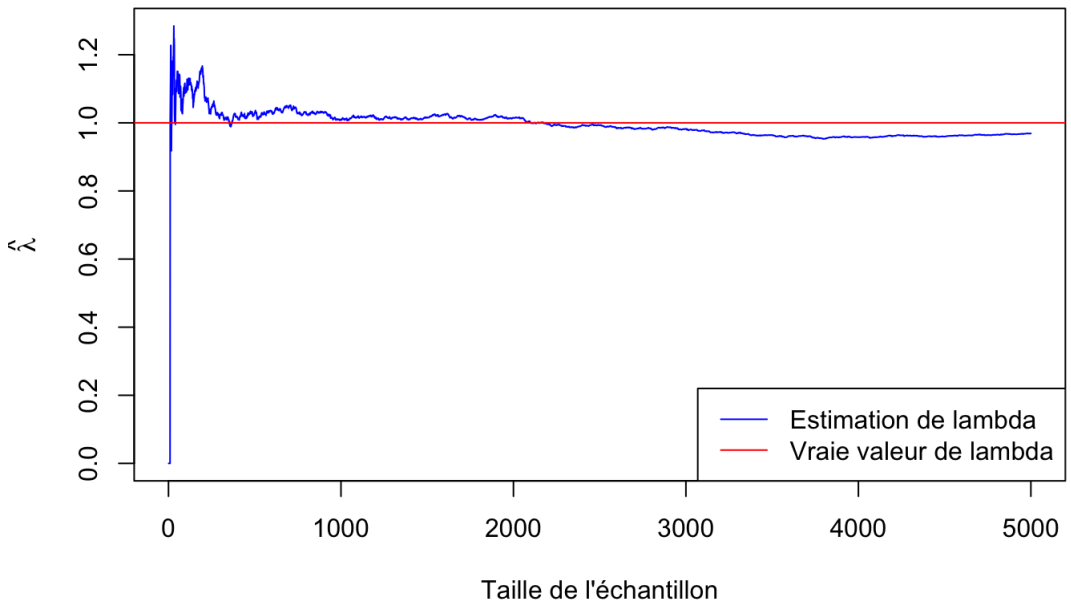


FIGURE 2 – $\hat{\lambda}_n$ en fonction de n

Ici nous avons simulé un seul échantillon de taille $n = 5000$ et fait varier la taille de ce dernier en prenant des sous-échantillons et en recalculant l'estimation de $\hat{\lambda}_n$ pour chaque n.

Pour n allant de 10 à 5000 on calcule les bornes de l'intervalle de confiance, pour chaque valeur de n et on les ajoute au graphique. En effet, on a :

$$\mathbb{P}\left(\lambda \in \left[\hat{\lambda}_n + \frac{1}{\sqrt{n}} \times \frac{\hat{\lambda}}{\frac{1}{n}\sum \delta_i}\right]\right) \approx 0.95$$

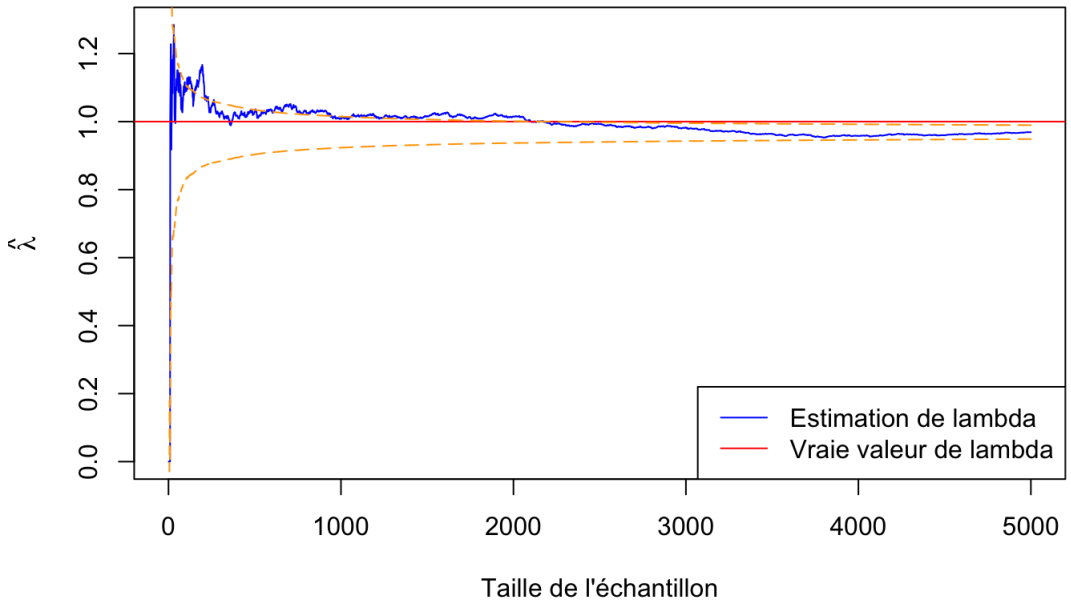


FIGURE 3 – $\hat{\lambda}_n$ en fonction de n avec IC

On voit que les valeurs de $\hat{\lambda}_n$ convergent très rapidement au fur et à mesure que n augmente vers la vraie valeur de λ. De plus, les valeurs estimées sont quasiment tout le long dans l'intervalle de confiance calculé. La largeur de l'intervalle de confiance diminue à mesure que n augmente, l'estimation devient donc plus précise. Même avec des échantillons de taille modérée, $n = 1000$ par exemple, l'estimation de λ est très proche de la vraie valeur.

On veut ensuite voir l'effet du taux de censure sur la précision de l'estimation de λ. On recommence alors cette fois-ci avec un échantillon simulé dont le taux de censure est de 50%. Comme vu précédemment, λ et μ permettent de maîtriser ce taux. On sait que $\lambda = 1$ et $\mathbb{P}(X > C) = \frac{\mu}{\mu+\lambda}$. Ainsi, avec $\mu = 1$ on obtient $\mathbb{P}(X > C) = \frac{1}{2}$.

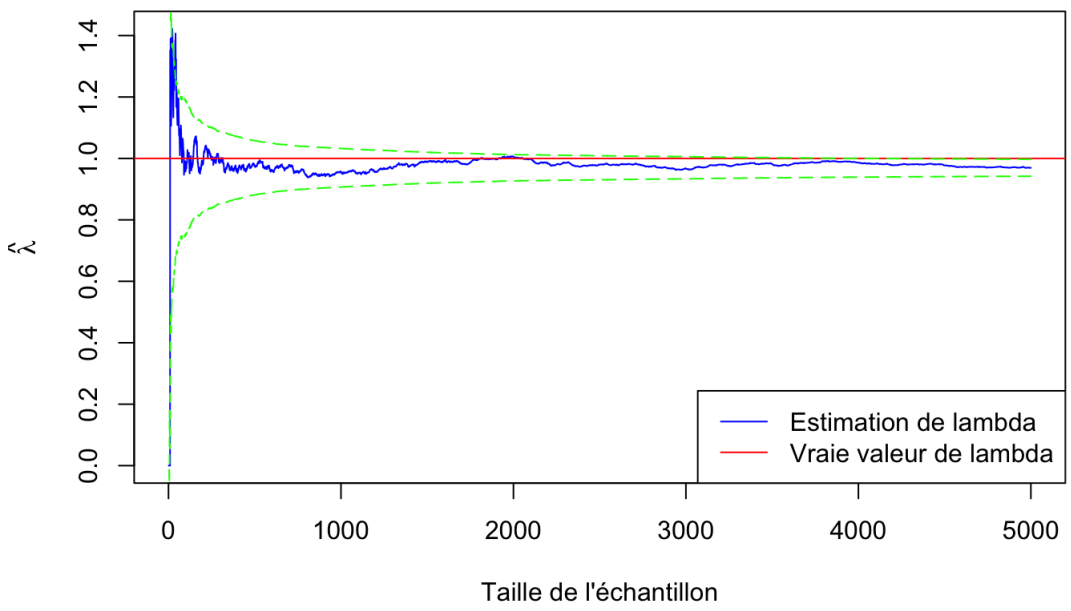


FIGURE 4 – $\hat{\lambda}_n$ en fonction de n pour $\mu = 1$

On voit que cette fois-ci les valeurs estimées de λ avec une taille plus petite de l'échantillon sont moins précises (un écart très élevé au début par exemple, la valeur estimée est $1,4$). Cependant, la convergence est légèrement plus lente au cours du temps,

En conclusion, dans cette partie on remarque qu'il n'est pas forcément utile d'avoir une taille d'échantillon très élevée ou un taux de censure très bas pour avoir de bons résultats grâce aux modèles, notamment concernant l'estimation des coefficients du modèle.

1.3 Estimation paramétrique de la fonction de survie

On construit un estimateur paramétrique $\hat{S}_n(t) = \exp(-\hat{\lambda}_n t)$ de la fonction de survie $S(t)$ dans le modèle exponentiel. Pour un échantillon de taille n, on calcule les valeurs de cet estimateur avec t un vecteur de pas 0.01 allant de 0 à 3. Pour un échantillon de taille n, on calcule les valeurs de l'estimateur. On trace alors la courbe représentative de $\hat{S}_n(t)$ et on superpose la courbe théorique $S(t)$.

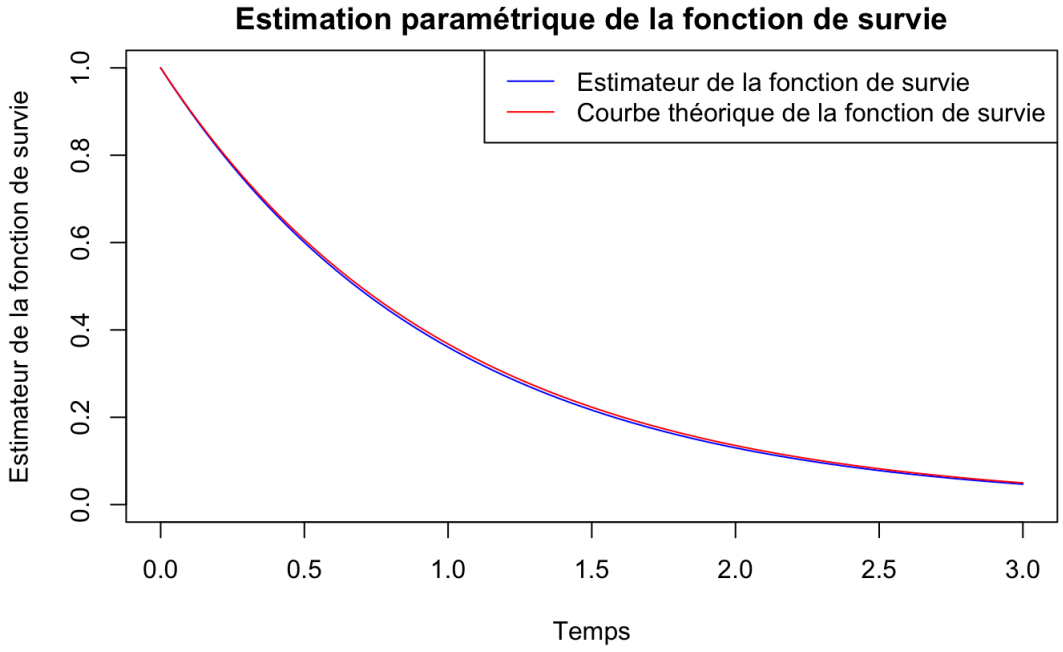


FIGURE 5 – $\hat{S}_n(t)$ et $S(t)$

Les résultats sont très satisfaisants, les deux courbes sont pratiquement confondues. Les calculs ont été faits pour $n = 5000$, ce qui explique en partie ces très bons résultats. Nous pouvons tester les mêmes estimations pour d'autres tailles afin de voir les résultats :

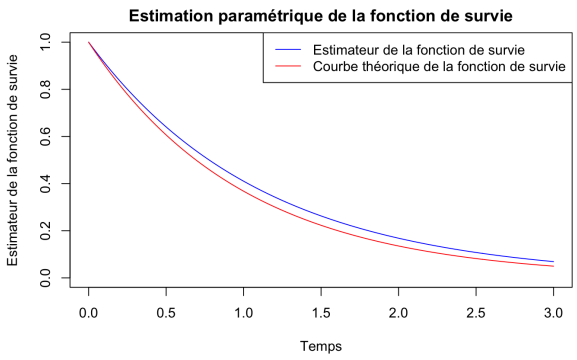


FIGURE 6 – Comparaison pour $n = 50$

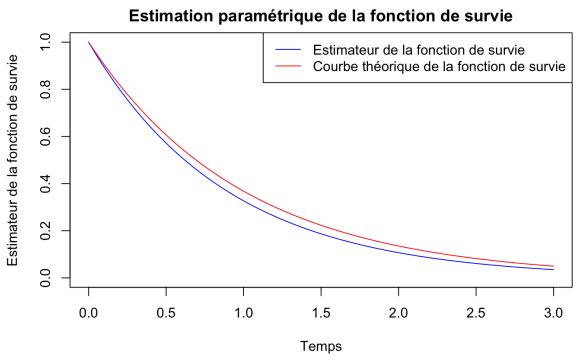


FIGURE 7 – Comparaison pour $n = 500$

On veut ensuite montrer que :

$$\sup_{t\geq 0}\left|\hat{S}_n(t) - S(t)\right| \leq \frac{|\hat{\lambda}_n - \lambda|}{\max(\lambda; \hat{\lambda}_n)}$$

pour en déduire ainsi une bande de confiance pour $S(t)$ uniforme en t.

Pour montrer que

$$\sup_{t\geq 0}\left|\hat{S}_n(t) - S(t)\right| \leq \frac{|\hat{\lambda}_n - \lambda|}{\max(\lambda, \hat{\lambda}_n)}$$

nous pouvons étudier la fonction $f(t) = \exp(-\lambda t) - \exp(-\hat{\lambda}_n t)$ pour $\lambda, \hat{\lambda}_n > 0$.

Tout d'abord, remarquons que $|\hat{S}_n(t) - S(t)| = |f(t)|$, où $f(t) = \exp(-\lambda t) - \exp(-\hat{\lambda}_n t)$.

Évaluons $|f(t)|$:

$$\begin{aligned}
|f(t)| &= |\exp(-\lambda t) - \exp(-\hat{\lambda}_n t)| \\
&= |\exp(-\lambda t)| - |\exp(-\hat{\lambda}_n t)| \\
&= \exp(-\lambda t) - \exp(-\hat{\lambda}_n t) \\
&= \exp(-\min(\lambda, \hat{\lambda}_n) t)
\end{aligned}$$

Donc, pour tout $t \geq 0$, nous avons :

$$|f(t)| \leq \exp(-\min(\lambda, \hat{\lambda}_n) t)$$

La borne maximale de $|f(t)|$ est donc obtenue pour $t = 0$, ce qui donne : $|f(0)| = 1$

Donc, nous avons montré que $|f(t)|$ est bornée par 1 pour tout $t \geq 0$.

En utilisant cette borne, nous obtenons :

$$\sup_{t\geq 0} |f(t)| \leq 1$$

En utilisant la définition de $f(t)$, nous avons : $\sup_{t\geq 0} |\hat{S}_n(t) - S(t)| \leq 1$

Maintenant, en divisant des deux côtés par $\max(\lambda, \hat{\lambda}_n)$, nous obtenons :

$$\frac{\sup_{t\geq 0} |\hat{S}_n(t) - S(t)|}{\max(\lambda, \hat{\lambda}_n)} \leq 1$$

En multipliant les deux côtés de l'inégalité par $|\hat{\lambda}_n - \lambda|$, nous obtenons :

$$\frac{|\hat{\lambda}_n - \lambda|}{\max(\lambda, \hat{\lambda}_n)} \geq \sup_{t \geq 0} |\hat{S}_n(t) - S(t)|$$

Ainsi, nous avons montré que

$$\sup_{t \geq 0} |\hat{S}_n(t) - S(t)| \leq \frac{|\hat{\lambda}_n - \lambda|}{\max(\lambda, \hat{\lambda}_n)}$$

Puisque $|\hat{S}_n(t) - S(t)|$ est une mesure de l'écart entre l'estimateur $\hat{S}_n(t)$ et la vraie fonction de survie $S(t)$ au temps t, nous pouvons utiliser cet écart pour construire une bande de confiance pour $S(t)$.

$$\text{Définissons } \epsilon = \frac{|\hat{\lambda}_n - \lambda|}{\max(\lambda, \hat{\lambda}_n)}.$$

Alors, pour tout $t \geq 0$, nous avons :

$$\sup_{t \geq 0} |\hat{S}_n(t) - S(t)| \leq \epsilon$$

Cela signifie que pour tout t, l'écart entre $\hat{S}_n(t)$ et $S(t)$ est borné par ϵ.

Par conséquent, nous pouvons construire une bande de confiance uniforme en t pour $S(t)$ en ajoutant et soustrayant ϵ à partir de $\hat{S}_n(t)$. Cela donne :

$$\hat{S}_n(t) - \epsilon \leq S(t) \leq \hat{S}_n(t) + \epsilon$$

Ainsi, pour tout t, nous avons une bande de confiance $[\hat{S}_n(t) - \epsilon, \hat{S}_n(t) + \epsilon]$ qui contient la vraie fonction de survie $S(t)$ avec une probabilité élevée. Cette bande de confiance est uniforme en t.

On compare ensuite graphiquement l'estimateur de Kaplan-Meier de la fonction de survie et l'estimateur $\hat{S}_n(t)$ pour $n = 20, 50, 100, 200$.

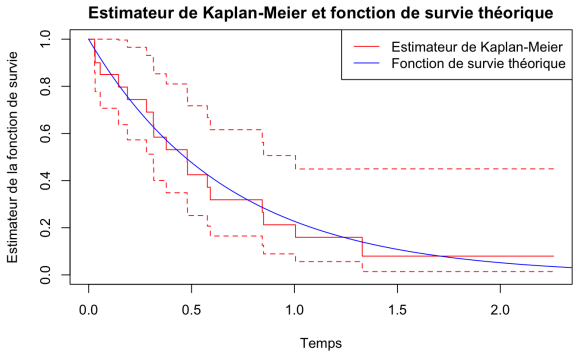


FIGURE 8 – Comparaison pour $n = 20$

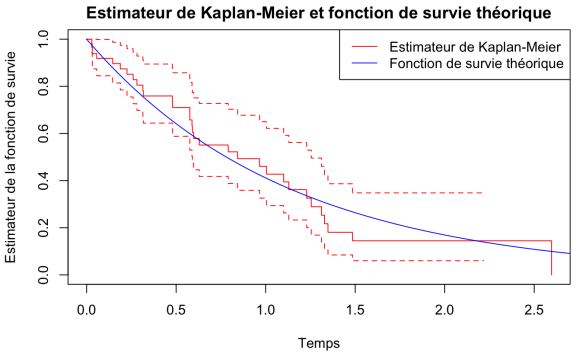


FIGURE 9 – Comparaison pour $n = 50$

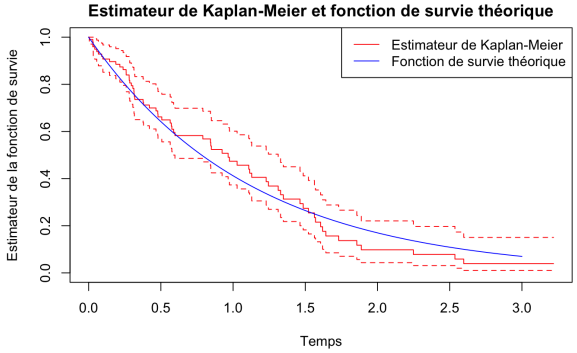


FIGURE 10 – Comparaison pour $n = 100$

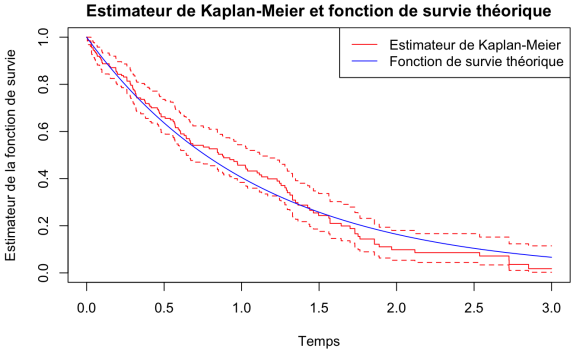


FIGURE 11 – Comparaison pour $n = 200$

2 Modèle Weibull

On s'intéresse maintenant à une variable aléatoire positive qui suit la loi de Weibull $\mathcal{W}(a, b)$ de densité :

$$f(x) = \frac{a}{b} \cdot (\frac{x}{b})^{a-1} \cdot \exp(-\frac{x}{b})^a, \quad x > 0$$

où $a, b > 0$ sont les paramètres de forme et d'échelle.

2.1 Estimation des paramètres avec la fonction "survreg"

On simule à l'aide de la fonction rweibull() un échantillon de X_i, $i = 1, ..., n$ de loi de Weibull $\mathcal{W}(a, b)$ de paramètres $a = 2$, $b = 5$ et $n = 50$.

On pose $Y = log(X) = \mu + \sigma \cdot W$ où W admet pour fonction de répartition $F_w(x) = 1 - e^{-e^x}$, $x \in \mathbb{R}$. On veut alors exprimer les paramètres a et b de la loi de X en fonction de μ et σ.

On calcule alors $\mathbb{P}(X > x)$:

$$\begin{aligned}
\mathbb{P}(X > x) &= \mathbb{P}(\log X > \log x) \\
&= \mathbb{P}(\mu + \sigma\ W > log\ x) \\
&= \mathbb{P}(W > \frac{log\ x - \mu}{\sigma} \\
&= \exp\left(-e^{\frac{-log\ \ x-\mu}{\sigma}}\right) \\
&= S(x) = \exp\left(-(\frac{x}{b})^a\right)
\end{aligned}$$

On identifie les coefficients et on obtient : $a = \frac{1}{\sigma}$ et $b = \exp(\mu)$.

On utilise la fonction survreg de R pour estimer μ et σ. Pour cela, on regarde ce que contiennent :

- `model$coefficients`
- `model$scale`
- `model$var`

On obtient ainsi :

```
(Intercept)
   1.709235
[1] 0.4295924
             (Intercept)   Log(scale)
(Intercept)  0.002047815 -0.001113391
Log(scale)  -0.001113391  0.006127180
```

Avec ces résultats on peut identifier μ et σ :

$$\hat{\sigma} = \log(scale) = 0.4295924$$

$$\hat{\mu} = \text{intercept} = 1.709235$$

Grâce au lien établi précédemment entre a, b, σ et μ on peut estimer a et b, les paramètres de la loi $\mathcal{W}(a, b)$:

$$\hat{a} = \frac{1}{\hat{\sigma}} = \frac{1}{0.4295924} \approx 2.3280$$

$$\hat{b} = \exp(\hat{\mu}) = \exp(1.709235) \approx 5.52$$

Les valeurs estimées sont très proches des vraies valeurs. On veut ensuite donner des intervalles de confiance asymptotiques de μ et σ pour un niveau de confiance de 95%.

Matrice de variance-covariance : à partir de la fonction `deltamethod` sur `R` on estime la matrice de variance-covariance du vecteur $g(U_n)$ à partir de celle de U_n.

Si $g : (x_1, x_2) \to (g_1(x_1, x_2), g_2(x_1, x_2))$ alors on a le code :

```
deltamethod(list(~g1(x1,x2),~g2(x1,x2),Un,Var(Un),ses=FALSE)
```

Dans notre cas :

$$g : (x_1, x_2) \to (e^{-x_2}, e^{-x_1})$$

$$(\hat{\mu}, \hat{\log\sigma}) \to (\hat{a} = e^{-\hat{\log\sigma}}, \hat{b} = e^{\hat{\mu}})$$

On applique alors cette formule sur `R` :

```
coeff0 <- model$icoef  #(avec mu_chapeau et log(sigma_chapeau))
hat_a <- exp(-coeff0[2])
hat_b <- exp(coeff0[1])
deltamethod(list(~exp(-x2),~exp(x1)), coeff0, model$var,ses=TRUE)
```

On obtient alors :

```
[1] 0.1822106 0.2500095
```

Dans notre cas $U_n = \begin{pmatrix} \hat{\mu} \\ \hat{\log(\sigma)} \end{pmatrix}$ et on a alors $\text{Var}(\hat{a}) \approx 0.18$ et $\text{Var}(\hat{b}) \approx 0.25$.

Les variances des estimateurs sont relativement faibles ce qui laisse penser à une certaine robustesse du modèle et des estimations.

2.2 Estimation des paramètres avec la fonction `flexsurvreg`

On reprend la même simulation en fixant la graine et on utilise une autre fonction du package `flexsurv` :

```
library(flexsurv)
model2 <- flexsurvreg(Surv(X)~1, dist="weibull")
model2
plot(model2,type="survival", est=TRUE, ci=TRUE)
```

On obtient alors :

```
Call:
flexsurvreg(formula = Surv(X) ~ 1, dist = "weibull")

Estimates:
       est    L95%   U95%   se
shape  2.328  1.997  2.714  0.182
scale  5.525  5.056  6.037  0.250

N = 100,  Events: 100,  Censored: 0
Total time at risk: 489.1741
Log-likelihood = -219.3447, df = 2
AIC = 442.6894
```

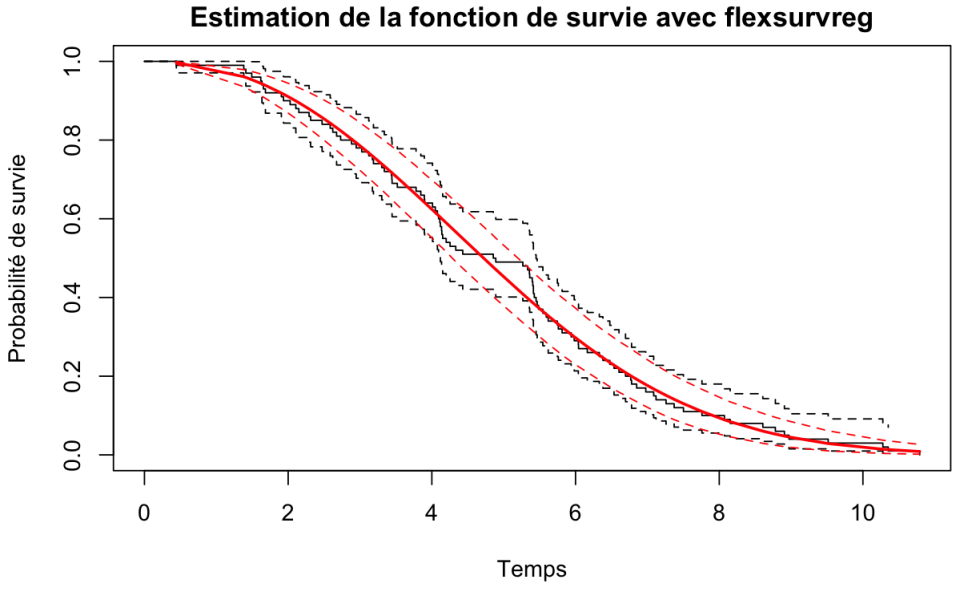


FIGURE 12 – Courbe de survie avec flexsurvreg

`flexsurvreg` nous donne directement les estimations de $\hat{a}$ et $\hat{b}$: $\hat{a} \approx 2.328$ et $\hat{b} \approx 5.525$ avec les variances correspondantes $\text{Var}(\hat{a}) \approx 0.182$ et $\text{Var}(\hat{b}) \approx 0.250$. Elle donne également les intervalles de confiance obtenus avec la δ-méthode.

En rouge, la figure 12 représente la fonction de survie estimée, en pointillées rouges il y a les intervalles de confiance correspondants. Elle superpose également par défaut l'estimateur de Kaplan-Meier de la fonction de survie. Elle discrimine alors graphiquement entre plusieurs modèles paramétriques. L'utilité pratique de superposer l'estimateur de Kaplan-Meier sur la fonction de survie estimée dans le modèle de Weibull est de pouvoir comparer visuellement les deux estimations de la fonction de survie et évaluer ainsi la performance du modèle paramétrique par rapport à une approche non paramétrique. On voit que l'estimation de la courbe de survie avec `flexsurvreg` est assez proche de celle de l'estimateur de Kaplan-Meier.

Échantillon censuré

Dans la partie précédente les estimations sont faites pour un échantillon non censuré. Nous reprenons alors la simulation pour un échantillon censuré en conservant la même loi $\mathcal{W}(a, b)$ pour l'échantillon $(X_i)_{i=1,\dots,n}$ et on simule l'échantillon $(C_i)_{i=1,\dots,n}$ selon une loi exponentielle de paramètre μ calibré de façon à obtenir environ 25% de censure pour $n = 50$ puis ensuite pour $n = 100$.

On remarque que pour $\mu = 0.07$ on obtient un taux de censure de 26%. On choisit donc ce paramètre pour la loi de (C_i).

On simule alors un échantillon censuré et on répète les étapes précédentes. Pour le modèle on obtient :

```
Call:
survreg(formula = Surv(time, status) ~ 1, data = data_censure,
    dist = "weibull")
             Value Std. Error     z       p
(Intercept)  1.5627     0.0886 17.65 < 2e-16
Log(scale)  -0.6242     0.1293 -4.83 1.4e-06

Scale= 0.536

Weibull distribution
```

```
Loglik(model)= -86.6   Loglik(intercept only)= -86.6
Number of Newton-Raphson Iterations: 6
n= 50
```

et les coefficients donnent :

```
(Intercept)
    1.56267
[1] 0.5357163
             (Intercept)   Log(scale)
(Intercept)  0.007841660 -0.001192964
Log(scale)  -0.001192964  0.016719547
```

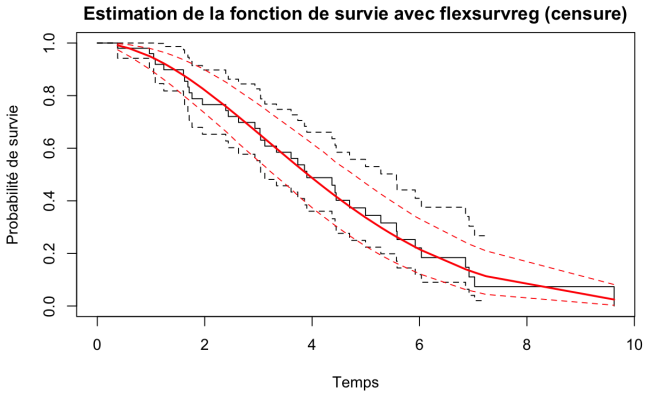


FIGURE 13 – Estimation pour $n = 50$ censuré

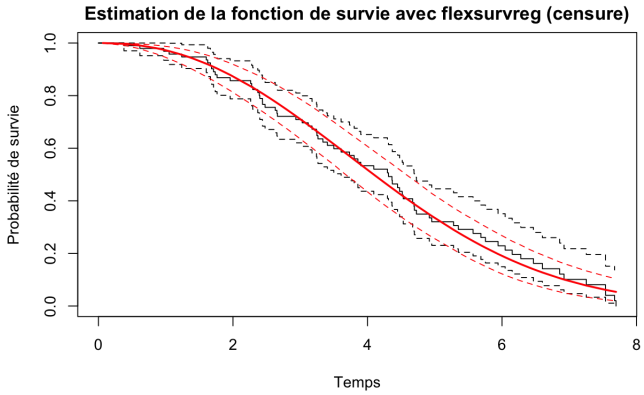


FIGURE 14 – Estimation pour $n = 100$ censuré

Les estimations de a et b donnent avec la δ-méthode :

— $n = 50$: $\hat{a} \approx 1.87$ et $\hat{b} \approx 4.77$

— $n = 100$: $\hat{a} \approx 2.28$ et $\hat{b} \approx 4.81$

On voit qu'avec un taux de censure beaucoup plus élevé les estimations des paramètres sont bien moins précises. La précision s'améliore quand la taille de l'échantillon augmente.

3 Autres modèles paramétriques

On veut ensuite estimer dans cette partie les paramètres de la loi de X suivant cette fois-ci une loi log-normale, et cela à l'aide de la fonction `flexsurvreg`.

Nous faisons l'estimation des paramètres de la loi pour des échantillons non censurés puis censurés.

X suit une loi log-normale de densité : $f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln(x)-\mu)^2}{2\sigma^2}\right)$ où :

— μ est la moyenne de $\ln(X)$,

— σ est l'écart type de $\ln(X)$.

On veut donc estimer σ et μ.

On obtient les résultats suivants sur `R` pour un échantillon de taille 100 non censuré :

```
Call:
flexsurvreg(formula = Surv(X) ~ 1, dist = "lnorm")

Estimates:
         est        L95%       U95%       se
meanlog -0.03919  -0.08816   0.00978   0.02498
sdlog    0.24984   0.21751   0.28698   0.01767

N = 100,  Events: 100,  Censored: 0
Total time at risk: 99.3577
Log-likelihood = 0.7175803, df = 2
AIC = 2.564839
```

Et pour un échantillon de taille 100 avec un taux de censure de $\frac{1}{3}$:

```
Call:
flexsurvreg(formula = Surv(time, status) ~ 1, data = data_censure,
    dist = "lnorm")

Estimates:
         est       L95%      U95%      se
meanlog -0.0283  -0.0793    0.0227    0.0260
sdlog    0.2513   0.2174    0.2905    0.0186
```

```
N = 100,  Events: 91,  Censored: 9
Total time at risk: 96.33783
Log-likelihood = -2.574316, df = 2
AIC = 9.148632
```

On obtient les graphes suivants des courbes de survie :

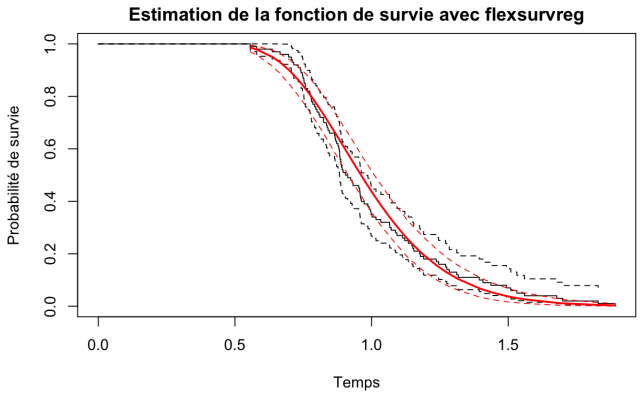


FIGURE 15 – Estimation pour $n = 100$ non censuré

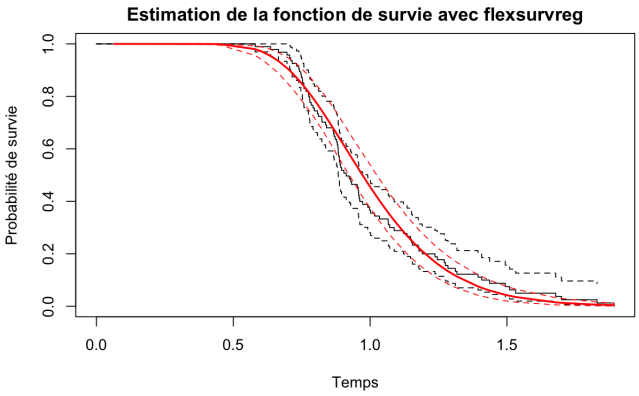


FIGURE 16 – Estimation pour $n = 100$ censuré

On remarque que, comme dans la partie précédente, les estimations des coefficients de la loi de X sont assez proches des vrais coefficients avec la fonction `flexsurvreg`. $\hat{\mu} \approx -0.039$ et $\hat{\sigma} \approx 0.249$ pour l'échantillon non censuré et pour l'échantillon censuré $\hat{\mu} \approx -0.0283$ et $\hat{\sigma} \approx 0.2513$. Elle fournit directement $\ln(\mu)$ et $\ln(\sigma)$ sans avoir à appliquer la $\delta-$méthode.

Les deux courbes d'estimation sont assez proches de celle de l'estimateur de Kaplan-Meier encore une fois comme dans la partie précédente.

4 Ajout d'une covariable

On s'intéresse maintenant à un modèle de régression avec une covariable Z.

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

On génère des variables $(W_i)_{i=1,...,n}$ de fonction de répartition $F_W(x) = 1 - \exp(-\exp(x))$, avec $x \in \mathbb{R}$ pour n=100. Puis on génère les $(X_i)_{i=1,...,n}$ selon le modèle de régression log-linéaire avec $\mu = 2, \gamma = 3$ et $\sigma = 0.5$ pour une covariable binaire Z telle que $Z_i = 0$ pour $i = 1, ..., \frac{n}{2}$. Et $Z_i = 1$ pour $i = \frac{n}{2} + 1, ..., n$ avec n pair.

On génère un échantillon censuré avec l'échantillon $(C_i)_{i=1,...,n}$ selon une loi exponentielle de paramètre $\lambda = 0.01$. On tilise ensuite la fonction `survreg` pour donner une estimation de μ, γ et σ.

On obtient les résultats suivants avec la fonction `survreg` :

```
Call:
survreg(formula = Surv(tt, delta) ~ Z, data = mydata, dist = "weibull")
              Value Std. Error     z      p
(Intercept)  2.0264     0.0543 37.35 <2e-16
Z            2.9923     0.1112 26.91 <2e-16
Log(scale)  -0.6562     0.0666 -9.85 <2e-16

Scale= 0.519

Weibull distribution
Loglik(model)= -440.3   Loglik(intercept only)= -617
        Chisq= 353.39 on 1 degrees of freedom, p= 7.7e-79
Number of Newton-Raphson Iterations: 8
n= 200
```

On voit alors $\hat{\mu} \approx 2.03$, $\hat{\gamma} \approx 2.99$.

On remarque que les valeurs estimées de μ et γ sont très proches des vraies valeurs. Cependant la valeur de σ est proche de la vraie valeur mais en valeur absolue, en effet, le modèle a calculé $\log(\hat{\sigma})$. On a alors : $\hat{\sigma} = \exp(\log(\text{scale})) \approx \exp(-0.6562) \approx 0.52$ ce qui est proche de la vraie valeur $\sigma = 0.5$. Cela veut dire que le modèle a très bien réussi à capturer la relation entre les covariables et la variable de survie.

On modifie ensuite la valeur de $n = 100, 200, 500$ et le taux de censure (avec $\mu = 0.01, et \mu = 0.003$) et faire une étude de la sensibilité (en examinant par exemple les écarts-types des estimateurs).

TABLE 1 – Écarts types des estimateurs pour différentes valeurs de n et μ

Index	n	μ	**Std. Error (Intercept)**	**Std. Error (Z)**	**Std. Error (exp(Log(scale)))**
1	100	0.01	0.0521	0.0749	0.0567
2	100	0.003	0.0509	0.0736	0.0562
3	200	0.01	0.0366	0.0526	0.0409
4	200	0.003	0.0365	0.0519	0.0402
5	500	0.01	0.0227	0.0325	0.0255
6	500	0.003	0.0228	0.0327	0.0255

En analysant les résultats des écarts-types des estimateurs pour différentes valeurs de n et μ, nous constatons que bien qu'il y ait des variations d'un scénario à l'autre, ces variations restent généralement limitées. Plus précisément :

— Les écarts-types des estimateurs pour l'intercept , le coefficient Z, et log(scale) semblent présenter des fluctuations relativement faibles d'un cas à l'autre.

— Malgré les variations importantes de n et μ, les écarts-types restent généralement dans des plages comparables. On peut alors penser qu'il y a une certaine stabilité dans la précision des estimations du modèle, indépendamment des variations de ces paramètres.

— La cohérence des écarts-types suggère que le modèle est assez robuste face à ces variations, ce qui signifie que les estimations des coefficients ne sont pas fortement influencées par les changements de taille de l'échantillon (n) ou du taux de censure (μ).

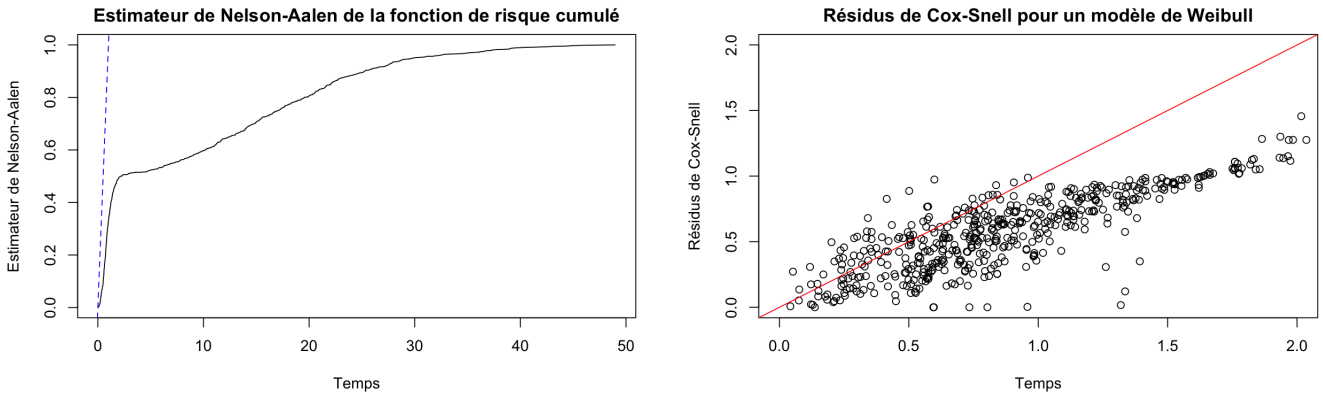
En conclusion, le modèle de régression de survie semble bien généraliser et maintenir une certaine précision même lorsque les conditions de l'échantillon sont changées. Les résultats du modèle sont donc fiables, ce qui peut-être très utile à savoir dans le cas de données réelles où les conditions de l'étude varient régulièrement et certaines fois de manière considérable.

4.1 Examen des résidus de Cox-Snell

On détermine les résidus de Cox-Snell $(R_i)_{i=1,\dots n}$ associés à l'hypothèse du modèle de Weibull où $R_i = \hat{H}_{\hat{\mu},\hat{\gamma},\hat{\sigma}}(T_i)$

A l'aide de la fonction `survfit` on représente l'estimateur de Nelson-Aalen de la fonction de risque cumulé de l'échantillon censuré (R_i, δ_i) (en noir) et on superpose la droite d'équation $y = x$ correspondant à la fonction de risque cumulé d'une loi exponentielle de paramètre 1.

On observe alors le résultat suivant :



Dans le graphe ci-dessous, la courbe noire représente l'estimateur de Nelson-Aalen partir du modèle de Weibull. La droite en pointillés bleus est celle de $y = x$. On voit que l'écart est très grand ce qui laisse supposer que l'hypothèse d'une distribution de Weibull n'est pas adaptée ici. On devrait pouvoir trouver une meilleure distribution adaptée à cet échantillon.

5 Application au jeu de données réelles *alloauto*

5.1 Présentation des données

Le dataset "alloauto" comprend des informations sur des patients ayant subi une greffe de moelle osseuse pour traiter la leucémie contenant trois variables :

— **time** : temps écoulé en mois jusqu'au décès du patient ou jusqu'à la rechute de la leucémie.

— **type** : indique le type de greffe effectuée. Elle est codée en tant que 1 pour une greffe allogénique et 2 pour une greffe autologue.

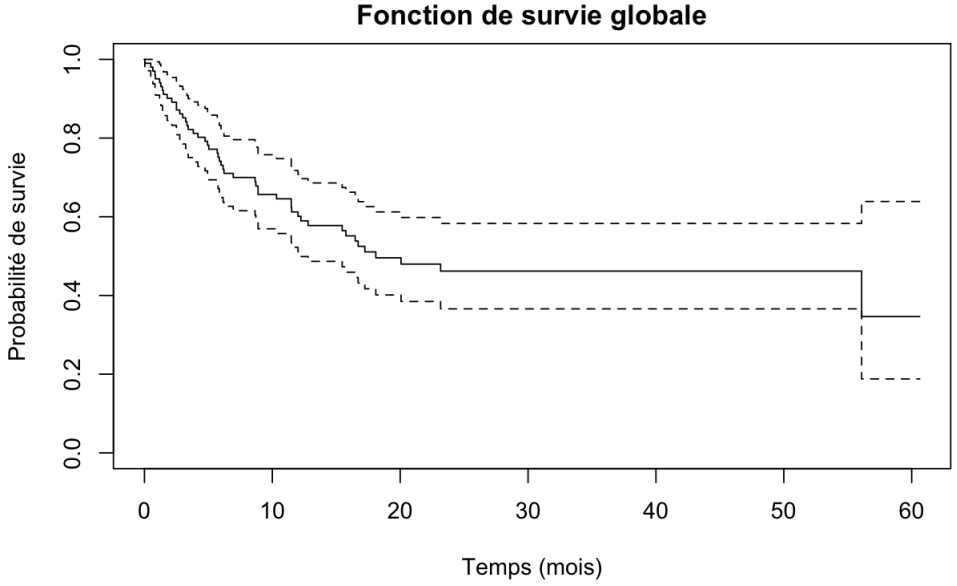
— **delta** : indicateur binaire de survie sans leucémie. Il est codé en tant que 0 si le patient est toujours en vie sans rechute et 1 si le patient est décédé ou a connu une rechute.

Ces données sont souvent utilisées dans les analyses de survie pour évaluer l'efficacité des greffes de moelle osseuse dans le traitement de la leucémie, en examinant les temps de survie des patients et les facteurs associés à la rechute ou à la mortalité.

Le taux de censure dans le dataset est de près de 50%. C'est un taux relativement élevé donc il faudra en tenir compte car la précision des analyses peut-être altérée.

5.2 Analyses de survie

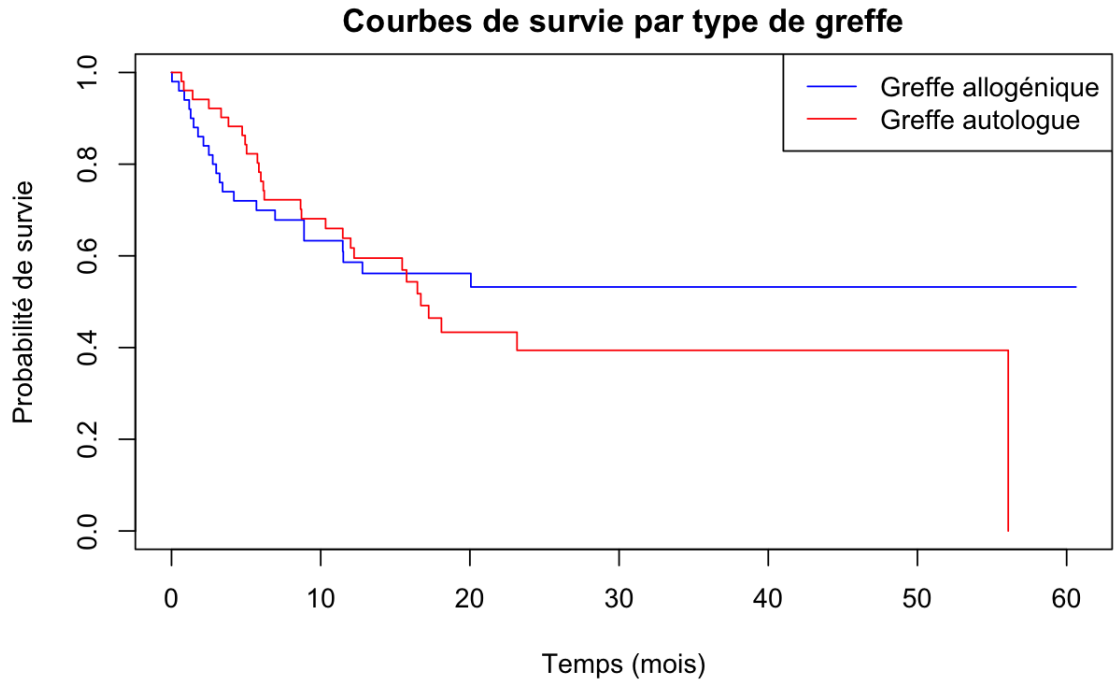
Nous commençons par effectuer quelques analyses de la survie, notamment avec la fonction de survie globale calculée avec l'estimateur de Kaplan-Meier :



Nous voyons qu'au bout de 18 mois la moitié des décès (ou rechutes) ont eu lieu, à la fin de l'étude, le taux de survie est d'environ 35%. Les intervalles de confiance associés à chaque point d'estimation illustrent la variabilité des données et la précision de l'estimation de la survie à différents moments. Par exemple, à 0.03 mois, le taux de survie estimé est de 99%, avec un intervalle de confiance de 95% allant de 97.1% à 100%, indiquant une estimation très précise à ce stade précoce de l'étude.

Cependant, à mesure que le temps progresse, l'intervalle de confiance s'élargit, ce qui reflète une incertitude croissante dans l'estimation due à une diminution du nombre d'individus à risque et à un nombre élevé d'événements (décès ou rechutes). Par exemple, à 56.086 mois, le taux de survie estimé est de 34.6%, avec un intervalle de confiance de 18.8% à 63.9%, illustrant une plus grande incertitude dans l'estimation à ce stade plus tardif de l'étude.

Ces résultats soulignent l'importance de suivre la dynamique de survie des patients sur une très longue période après une greffe de moelle osseuse. On peut maintenant s'intéresser à la survie des patients en tenant compte du type de greffe subie. On simule alors les courbes de survie des patients par type de greffe.



On remarque quelques différences entre la survie des deux groupes, notamment après 16 mois.

Pour les patients ayant subi une greffe allogénique (type=1), le taux de survie est légèrement plus élevé tout au long de l'étude. Au début, les deux groupes ont des taux de survie similaires,

mais au fil du temps, le groupe de greffe allogénique maintient un taux de survie légèrement plus élevé.

À la fin de l'étude, le taux de survie estimé pour les patients ayant subi une greffe allogénique est d'environ 39.4%, tandis que pour ceux ayant subi une greffe autologue, il est d'environ 0% (avec une seule observation restante dans ce groupe).

Ces résultats suggèrent que le type de greffe pourrait avoir un impact sur la survie des patients atteints de leucémie, avec une tendance à une meilleure survie pour ceux ayant subi une greffe allogénique.

Nous pouvons vérifier cette hypothèse en utilisant le test du log-rank :

```
Call:
survdiff(formula = Surv(time, delta) ~ type, data = alloauto)

        N Observed Expected (O-E)^2/E (O-E)^2/V
type=1 50       22     24.2     0.195     0.382
type=2 51       28     25.8     0.182     0.382

 Chisq= 0.4  on 1 degrees of freedom, p= 0.5
```

La p-value = 0.5 ce qui est très grand par rapport au seuil de 0.05, on ne peut donc pas rejeter l'hypothèse nulle e selon laquelle il n'y a pas de différence significative dans les taux de survie entre les deux groupes. Autrement dit, il n'y a pas suffisamment de preuves pour conclure que le type de greffe influence la survie des patients dans notre échantillon.

5.3 Ajustement de modèle

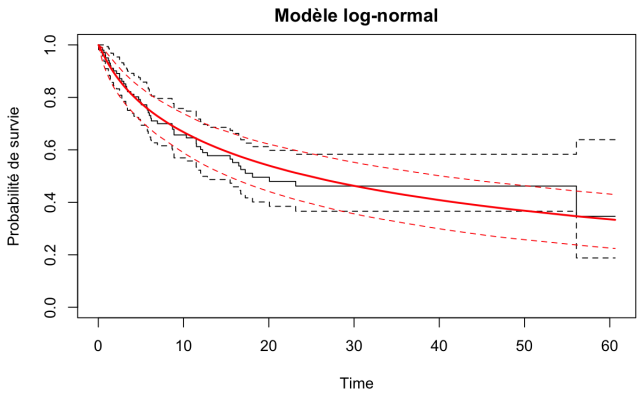
5.3.1 Sans covariable

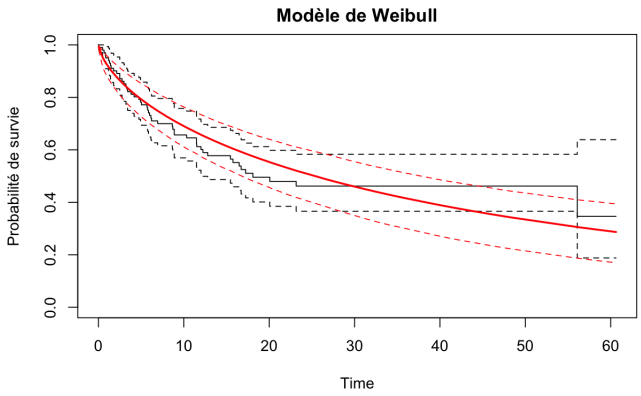
On voit que la différence de survie entre les deux types de greffe est limitée, donc dans un premier temps nous allons essayer de trouver un modèle adapté à la survie dans ces données sans tenir compte de la covariable `type`. Nous considérons dans un premier temps trois modèles : Weibull, log-logistique et log-normal dont on calcule les `AIC`.

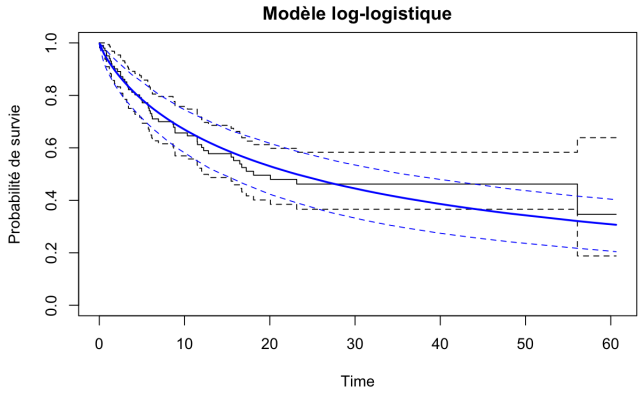
Modèle	AIC
Weibull	450.0842
Log-logistique	446.4582
Log-normal	445.9870

TABLE 2 – Table des valeurs AIC pour les différents modèles.

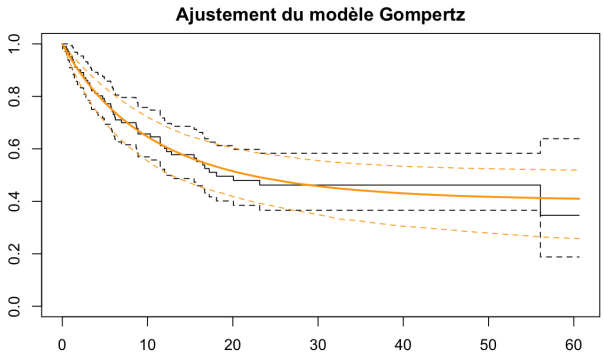
Les `AIC` sont relativement proches avec une valeur légèrement plus faible pour le modèle log-normal. Nous allons devoir faire d'autres analyses pour décider si l'un de ces modèles est adapté à notre jeu de données, et dans ce cas, lequel est le meilleur. On trace alors les courbes de survie estimées suivant les modèles et on les superpose avec la courbe de survie des données.

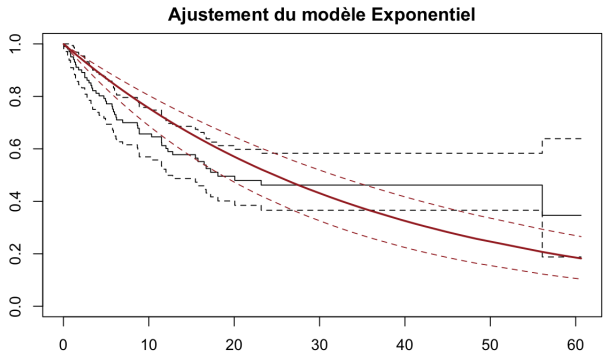




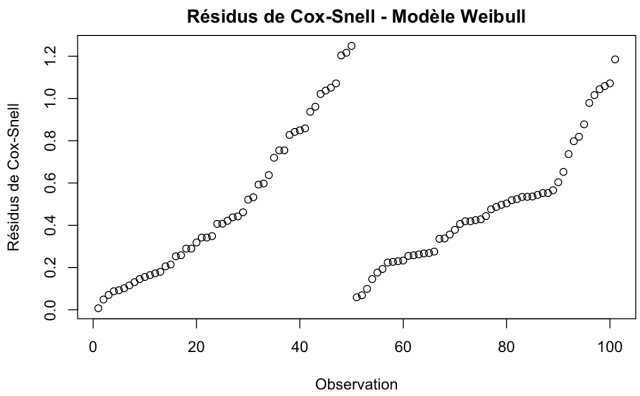


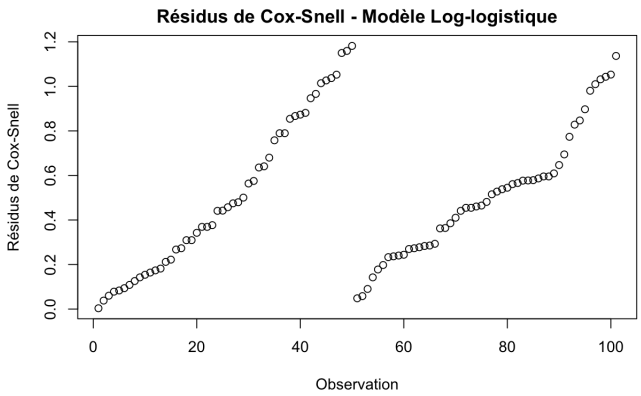
On voit que ces modèles ne s'ajustent pas très bien à la courbe de survie de nos données. On peut considérer d'autres modèles comme Gompertz ou exponentiel :

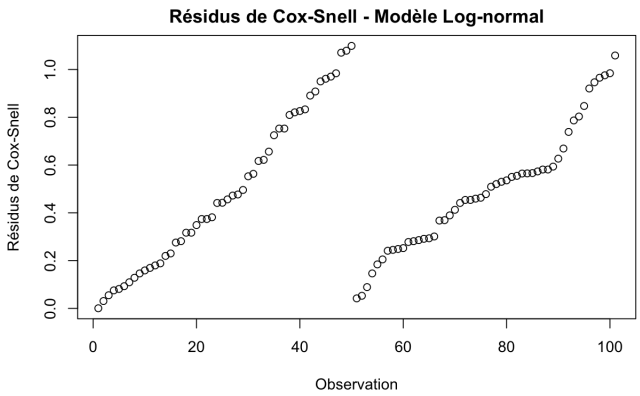


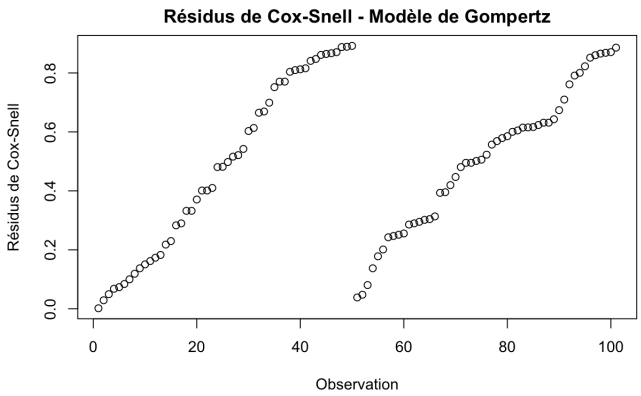


On calcule l'`AIC` du modèle de Gompertz et on obtient 440.34 environ. L'écart n'est pas très important par rapport aux modèles précédents mais reste plus faible avec l'hypothèse de ce modèle. Il semble bien s'ajuster à la courbe de survie et son `AIC` n'est pas très élevé, il est donc à privilégier. Nous pouvons également examiner les résidus de Cox-Snell :





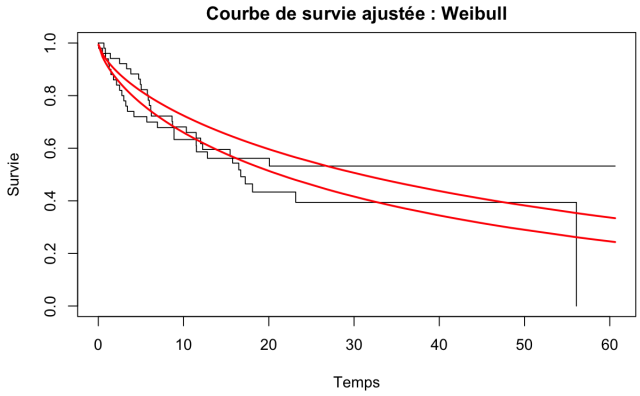


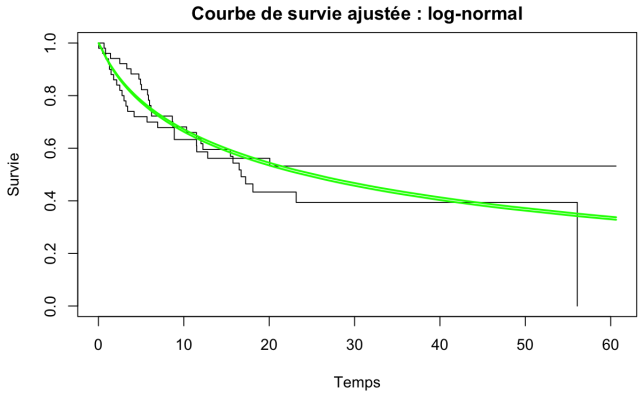


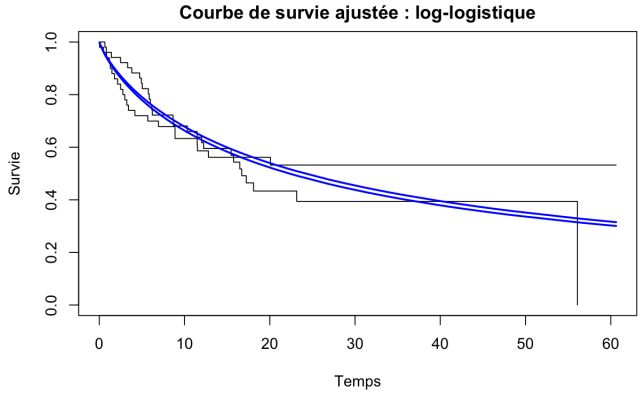
Le modèle de Gompertz est celui avec les résidus les moins élevés. Encore une fois, il semble être le meilleur modèle.

5.3.2 Avec la covariable *type*

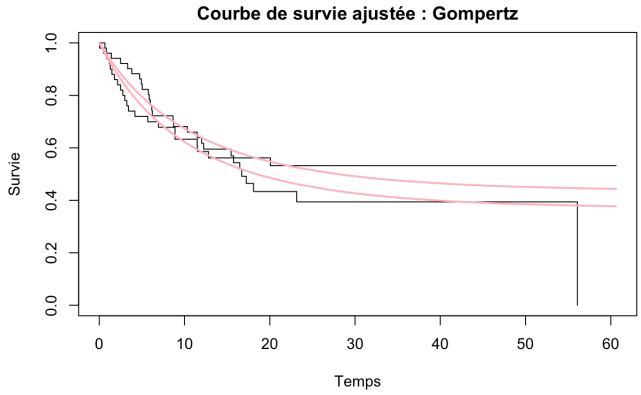
Cette fois-ci on introduit la covariable `type` dans les modèles précédents, nous avons alors les courbes de survie suviantes selon le type de greffe (0 ou 1) :







On remarque que les courbes de survie estimées selon le type de greffe dans les modèles sont très proches l'une de l'autre, à l'exception de celui du modèle de Weibull. Ils n'arrivent pas à identifier correctement les différences de survie entre les deux types de greffes. Ces dernières étant très proches dans les données au début, les courbes sont confondues dans les estimations, et un léger écart est observé en fin d'étude.



Encore une fois, le meilleur modèle graphiquement semble être celui suivant Gompertz. Il estime correctement la courbe de survie globale des patients, ainsi que les survies par type de greffe. Il a cependant du mal vers la fin de l'étude où il estime correctement la survie pour un type de greffe mais sous-estime le taux de survie final pour l'autre type.

Le modèle de Gompertz est normalement utilisé pour modéliser la croissance des tumeurs. Dans le cas des patients après une greffe de moelle osseuse, la survie peut être affectée par la progression de maladies comme le cancer, où la croissance tumorale peut jouer un rôle important. Il paraît donc logique finalement que ce modèle soit le plus adapté dans ce contexte.

En conclusion, le modèle le plus adapté à ces données de survie semble être Gompertz. La survie semble être légèrement influencée par le type de greffe subie, mais pas de manière cruciale, notamment au début de l'étude (avant 18 mois).