Etude de survie: mortalité infantilo -juvénile

Contexte

Pour aborder la problématique de la compréhension de la notion de survie, nos recherches se sont orientées vers l'analyse de la mortalité infantilo-juvénile. Cette étude nous offrira l'opportunité d'appliquer des modèles de survie, tels que le modèle de Cox, afin d'explorer ce phénomène de mortalité. L'objectif principal sera de modéliser la durée de survie et de réaliser des inférences statistiques, notamment en termes d'estimation, de tests et de prédictions.

- **Données utilisées**

Nos données proviennent de l'extension **questionr** de R, qui recense des informations sur la fécondité ainsi que les caractéristiques socio-économiques associées, dans le cadre de l'étude de la mortalité infantile et juvénile.

https://juba.github.io/questionr/reference/fecondite.html

NB : Ces données sont fictives et ont été créées à des fins pédagogiques. Ce package a été publié pour la première fois en 2014, et les données relatives à la fécondité ont été introduites à cette même époque.

En l'absence de la durée exacte de suivi des ménages, nous avons considéré la date de naissance de l'enfant comme le début du suivi, et avons fixé la durée du suivi à la date l'unique entretien réalisée.

Voici ci-dessous la dimension de notre base de données avant pré-traitement :

```
[1] 1584   25
```

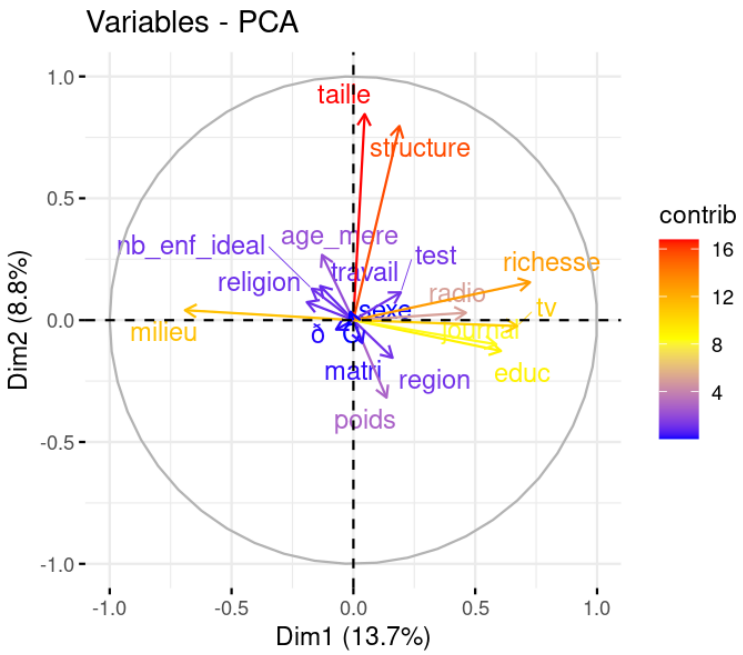
Un seul entretien est réalisé, ce qui signifie que les individus (ici, les bébés) ne sont pas suivis dans le temps, car aucun autre entretien n'est effectué. Il s'agit de censure à droite et non de censure tronquée, car nous ne savons pas ce qui se passe après la date de l'enquête pour les enfants toujours vivants au moment de l'entretien. De plus, nous n'avons pas exclu les individus décédés avant la date de l'entretien unique. Nous nous trouvons donc dans le cas d'une censure à droite de type 1 et 3, respectivement pour les individus encore vivants à la fin de l'étude et ceux entrés tardivement dans l'étude (censure régulière et censure d'entrée tardive).

Soit δ le statut de censure et T le temps de survie en mois

Statistique descriptive

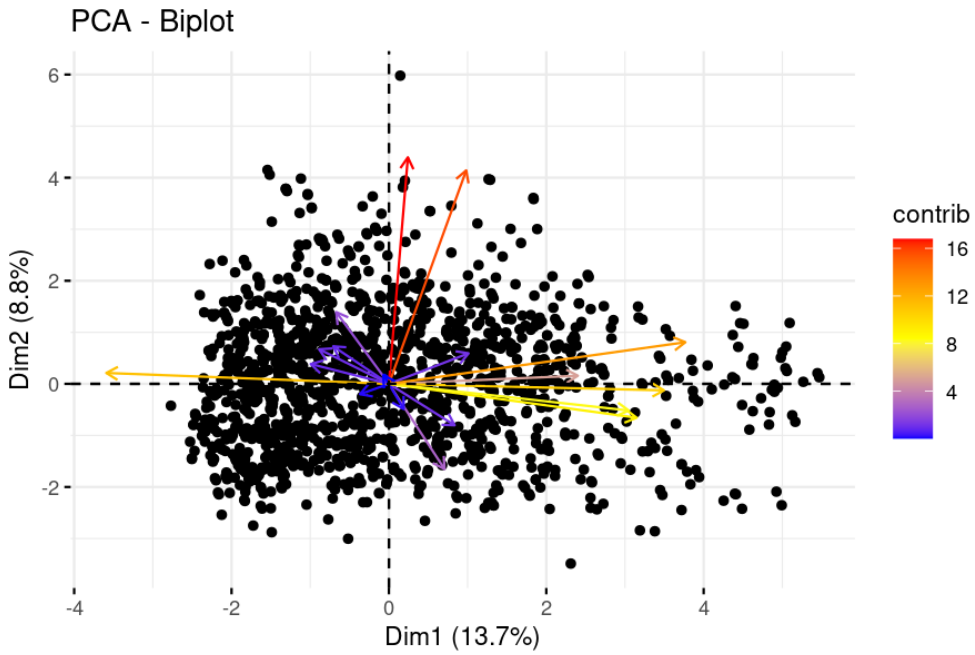
ACP

- Visualisation des variables sur les deux premiers axes



On constate que les variables “taille” et “structures” sont colinéaires, ou apportent les mêmes informations lorsqu’on prend en compte l’axe 1. De même, les variables “temps de survie”, “richesse”, “TV”, “journal” et “éducation” sont colinéaires lorsqu’on considère l’axe 2.

- Visualisation des individus sur les deux premiers axes



Nous parvenons difficilement à extraire des informations sur la distribution des individus (ici, les nouveau-nés) sous forme de clusters dans cette visualisation. Toutefois, d’après la distribution des individus, on constate une faible variance parmi eux.

Modèles non paramétriques :

- Estimation de la probabilité de survie dans le temps: **Kaplan-Meier**

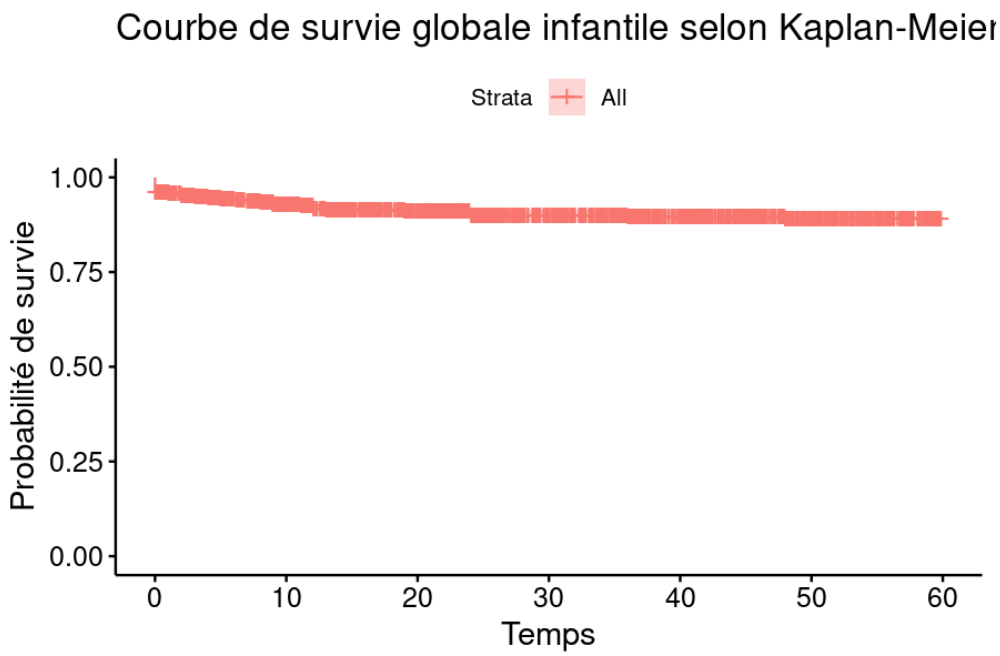
Nous allons débuter l'étude de la distribution des temps de survie sans poser d'hypothèse préalable, en estimant directement la fonction de survie à partir des données, sans contrainte de distribution.

Cette analyse porte sur la probabilité que les nouveau-nés survivent au-delà d'un temps t donné, permettant ainsi d'étudier l'évolution de la survie au sein d'une population spécifique au cours du temps.

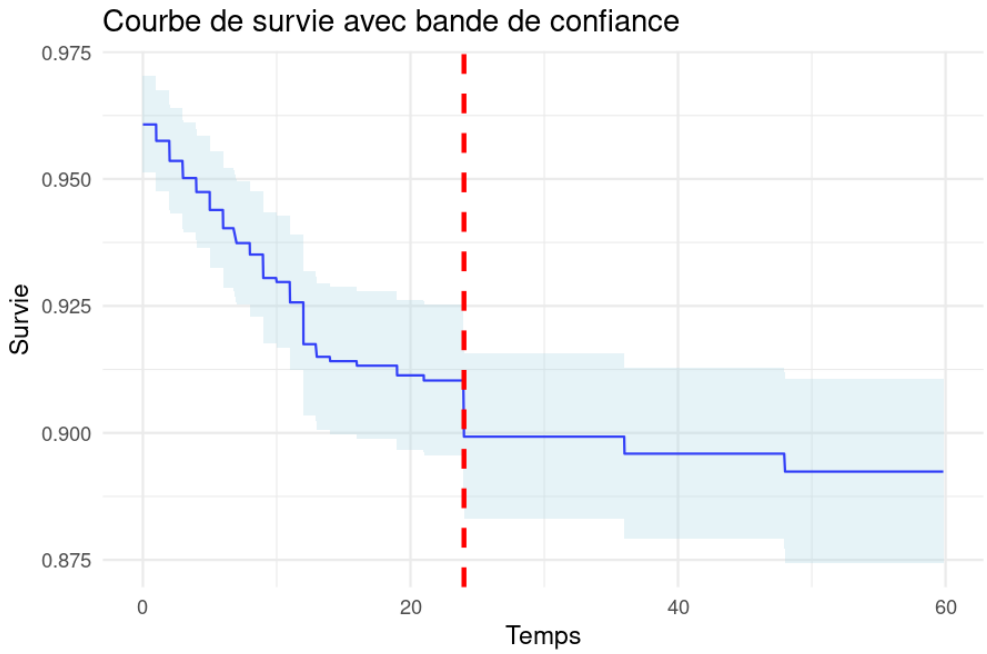
Soit d_i le nombre de décès au temps T_i et R_i le nombre d'individus à risque de mourir au temps T_i, c'est-à-dire le nombre d'individus vivants non censurés au temps T_i. La formule de l'estimation de la fonction de survie, qui représente la probabilité qu'un individu survive au-delà d'un temps t, s'exprime comme suit :

$$\hat{\bar{\mathbf{F}}}_{\mathbf{n}}(\mathbf{t}) = \prod_{\mathbf{i},\mathbf{T_i} \leq \mathbf{t}} \left(\mathbf{1} - \frac{\mathbf{d_i}}{\mathbf{R_i}} \right)$$

- Représentation de la courbe de survie infantile



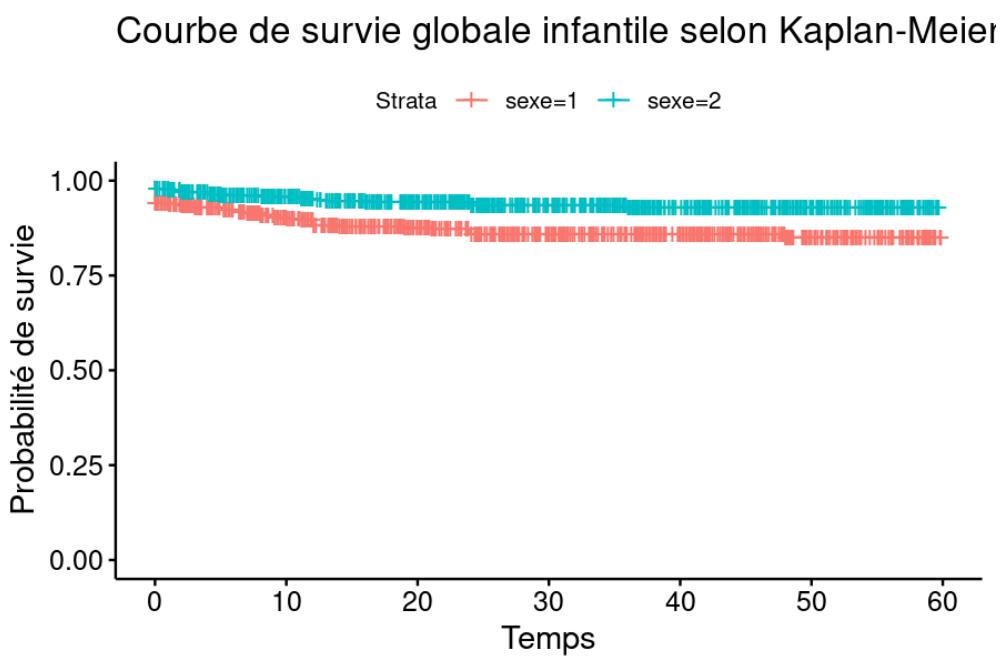
D'après la courbe de survie globale d'après **Kaplan-Meier** on constate une diminution progressive au sur 60 mois. Il est important de noter que cette probabilité reste supérieure à 0,89 pour l'ensemble de notre cohorte, indiquant ainsi un faible risque de mortalité dans cette population. Pour une analyse plus approfondie, nous avons effectué un zoom sur l'intervalle de probabilité de survie $I = [0.89, 1]$, qui englobe l'ensemble des valeurs observées, permettant une visualisation détaillée des tendances.



À partir d'une visualisation détaillée de la courbe de survie, nous observons une importante chute de la probabilité de survie au moment de la naissance. Cette baisse pourrait s'expliquer par les complications potentielles survenant lors de l'accouchement, telles que la mortalité néonatale. Par la suite, au cours des 24 premiers mois (2 ans) de vie, la probabilité de survie diminue rapidement, ce qui correspond à une

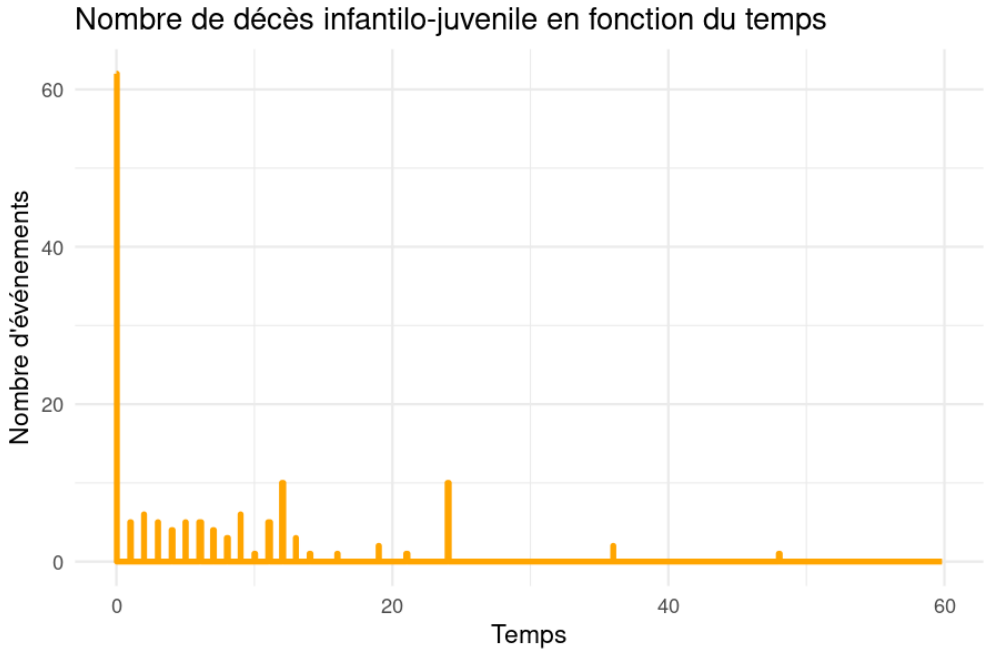
augmentation progressive du risque de mortalité. Cette tendance est probablement liée à des facteurs tels que les maladies infantiles, l'immaturité du système immunitaire des enfants et les difficultés d'accès aux soins de santé. Au-delà de ces 24 mois, la probabilité de survie diminue de manière beaucoup plus lente, reflétant une stabilisation relative du risque de mortalité.

- **Comparaison de deux groupes relatifs au sexe des enfants avec la courbe de Kaplan-Meier**



À la naissance, les nourrissons présentent des probabilités de survie similaires. Cependant, par la suite, on observe que les nourrissons de sexe masculin (sexe = 1) ont une probabilité de survie inférieure à celle des nourrissons de sexe féminin (sexe = 2)..

- Détection des tendances brutes : Courbe de distribution des décès au fil du temps.



D'après la représentation précédente, on constate trois périodes critiques, assimilables à des périodes où les décès sont les plus fréquents(décès >=7). Cela peut être dû, par exemple, à une période de risque accru de complications postnatales.

```
# La valeur 7 a été choisie de manière arbitraire dans le but d'identifier et
#de capturer les périodes où surviennent des pics de mortalité
#ou des événements de décès significatifs.

big_event_time= fit_KME$time[fit_KME$n.event>= 7]

big_event_time
```

```
[1]  0 12 24
```

Les durées identifiées précédemment corroborent notre analyse de la courbe de survie, révélant des pics de mortalité à trois périodes critiques : au moment de la naissance (t = 0), à 12 mois (1 an) après la naissance, et à 24 mois (2 ans) après la naissance.

- Estimation de la fonction de risque cumulé: **Nelson-Aalen**

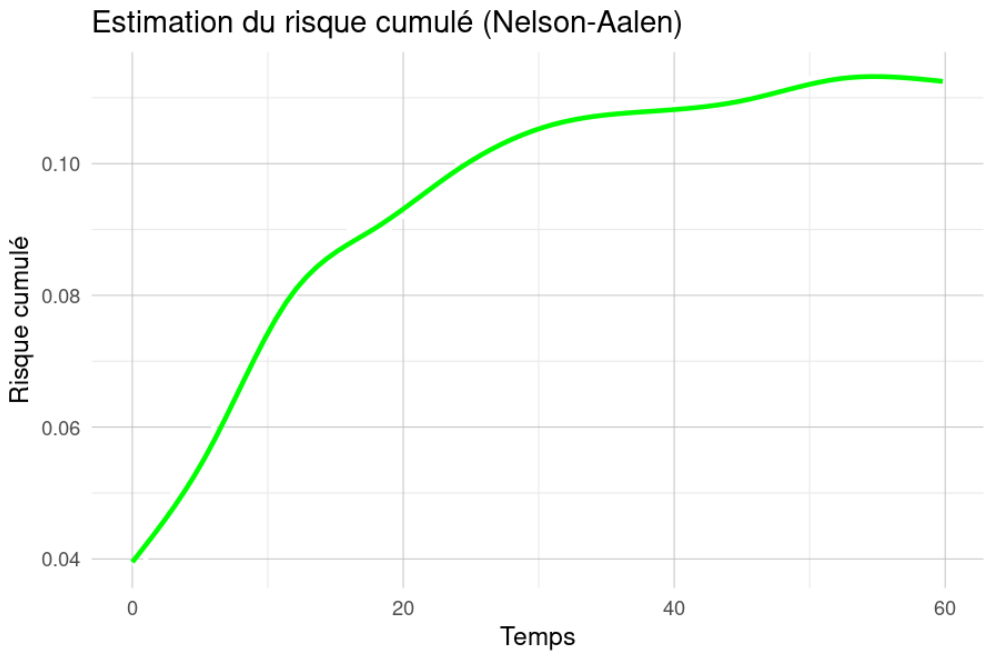
Pour **pousser plus loin notre curiosité** et explorer les informations supplémentaires que peut apporter la **fonction de risque cumulé de Nelson-Aalen** dans l'étude de survie, nous avons décidé, **à des fins pédagogiques**, de la visualiser également. Cette démarche nous permet de mieux comprendre comment le risque s'accumule au fil du temps et d'offrir une perspective complémentaire à l'analyse traditionnelle de survie.

Soit $\mathbf{\hat{RC}}$ l'estimation du risque cumulé au temps t et Ri le nombre d'individus à risque de mourir au temps Ti.

Ci dessous la formule de l'estimation du rique :

$$\mathbf{\hat{RC}_n(t) = \sum_{i,T_i \leq t} \left(\frac{d_i}{R_i} \right) = -ln(\hat{\bar{F}}_n(t))}$$

À défaut de relever les mêmes **subtilités liées à la fréquence des décès** déjà mises en évidence par la courbe de **Kaplan-Meier**, nous avons choisi de visualiser la **courbe lissée par spline** de l'estimation cumulative de **Nelson-Aalen**. Cette approche permet de mieux appréhender la **tendance globale du risque cumulé** tout en atténuant les fluctuations brutes, offrant ainsi une perspective complémentaire à l'analyse de survie.



Après un lissage par spline, on observe une tendance haussière de la la courbe de Nelson-Aalen, particulièrement une pente plus raide avant les 24 mois après la naissance ce qui indique une accumulation plus rapide du risque (c'est-à-dire une fréquence plus élevée de décès) durant cette période. En revanche, après 24 mois, la pente devient plus douce, reflétant une baisse de la fréquence des décès. Cette courbe est un complément utile à la courbe de Kaplan-Meier, car elle offre une perspective différente sur les données de survie.

On peut même aller jusqu'à dire que, dans l'ensemble, les deux courbes fournissent globalement la même information, mais avec des nuances spécifiques. La courbe de Kaplan-Meier se concentre sur la probabilité de survie au fil du temps, tandis que la courbe de Nelson-Aalen permet une visualisation cumulative des événements de décès, mettant en évidence l'accumulation du risque au cours du temps.

Test de comparaison des groupes: log-rank

Nous avons choisi de mener un test du log-rank, basé sur la comparaison des courbes de survie entre deux ou plusieurs groupes, notés J, en fonction de leur fonction de survie. $\hat{F}_1, ..., \hat{F}_J$.

- Hypothèse H_0 : les courbes de survie sont égales c.a.d $\hat{F}_1 = ... = \hat{F}_J$
- Hypothèse H_1 : les courbes de survie ne sont pas égales $j, j' \leq J, j \neq j', \hat{F}_j \neq \hat{F}_{j'}$

Et d'après le cours, une statistique de test

$$T_{log-rank} = (Z_1, ..., Z_{J-1})' \sum^{-1} (Z_1, ..., Z_{J-1})$$

Nous avons appliqué le test du log-rank à chacune des variables qualitatives explicatives afin d'évaluer les différences de survie entre les groupes d'individus.

```
    variable  p_value Decision_sur_H0
1       sexe 3.47e-06          rejet
2    travail 1.38e-02          rejet
3     milieu 3.17e-02          rejet
4     region 3.24e-02          rejet
5      matri 8.56e-02      non rejet
6      radio 2.52e-01      non rejet
```

```
7    journal 2.54e-01       non rejet
8       educ 6.05e-01       non rejet
9         tv 6.38e-01       non rejet
10 structure 6.75e-01       non rejet
11      test 7.35e-01       non rejet
12  richesse 8.67e-01       non rejet
13  religion 9.53e-01       non rejet
```

D'après le test de significativité, la p-value étant inférieure au seuil de 5% uniquement pour quatre variables, à savoir le sexe, le travail, le milieu et la région, nous rejetons l'hypothèse $H0$ pour ces variables. Cela indique qu'il existe une différence significative entre les groupes issus de chacune de ces quatre variables. En d'autres termes, ces variables ont un impact statistiquement significatif sur la survie.

Nous avons donc envisagé d'explorer des modèles paramétriques sous hypothèse de loi de distribution afin de mieux comprendre l'impact de toutes les variables explicatives sur la survie infantile, tout en prenant en compte les effets potentiels et en ajustant l'analyse pour améliorer la précision des résultats.

Modèles paramétriques

Avant de passer au modèle de Cox, nous avons envisagé par curiosité et pédagogie d'explorer les lois exponentielle et de weibull applicables à nos données de survie en mettant en place pour chacune de ces lois deux différents types de modèle nottament le modèle de base ou nul et le modele complet afin de comparer ces deux modèles. En observant la courbe des distributions brutes des flux relatifs aux décès infantilo-juvéniles, nous remarquons des pics à des moments clés de la croissance des nourrissons. Ces pics pourraient correspondre à des moments où la distribution weibull des durées de vie serait plus appropriée, suggérant ainsi l'applicabilité de ce modèle.

- **Distribution exponentielle**:
$$X \sim \mathcal{E}(\beta)$$
- Une fonction de survie équivalente à
$$\bar{F} = e^{-\beta t} \wedge 1$$

Il est intéressant de noter que l'estimation paramétrique de la loi exponentielle se fait particulièrement par l'EMV (maximum de vraissemblance).

- **Distribution de weibull**:
$$X \sim \mathcal{W}(\alpha, \beta)$$
- Une fonction de survie s'exprimant par:
$$\bar{F}(t) = e^{-\beta t^{\alpha}} \wedge 1$$

Soit α, β les paramètres de forme et d'échelle

- Comparaison des deux hypothèses de distribution sur la durée de vie infantilo-juvenile:

```
                    df      AIC
fit_expo_base        1 1162.820
fit_weibull_base     2 1152.166
fit_weibull_complet 19 1159.867
fit_expo_complet    18 1169.353
```

L'AIC de la distribution de Weibull est inférieure à celle de tout les autres modèles. Cela indique que la distribution de Weibull explique mieux nos données relatives à la durée de vie infantilo-juvénile.

Modele de cox

Ce modèle prend en compte non seulement la période d'entrée des individus dans l'étude, mais il permet également d'estimer l'effet des covariables sur le risque de survie. Ces covariables méritent une investigation approfondie dans la suite de notre étude, notamment à travers l'application d'un modèle de régression de Cox, qui nous permettra d'explorer leurs effets sur la mortalité infantile.

- **Modèle complet**

```
Call:
coxph(formula = Surv(C, ð) ~ ., data = df_survie_MIJ)

                  coef  exp(coef)   se(coef)      z        p
sexe         -8.111e-01  4.444e-01  1.794e-01 -4.520 6.17e-06
poids        -2.675e-02  9.736e-01  1.133e-01 -0.236  0.81338
age_mere      8.908e-03  1.009e+00  1.243e-02  0.716  0.47375
milieu        7.121e-01  2.038e+00  2.705e-01  2.633  0.00846
region        3.777e-02  1.038e+00  7.751e-02  0.487  0.62612
educ         -1.977e-02  9.804e-01  1.646e-01 -0.120  0.90437
travail       4.542e-01  1.575e+00  1.435e-01  3.165  0.00155
matri        -2.925e-02  9.712e-01  9.960e-02 -0.294  0.76904
religion     -1.179e-02  9.883e-01  7.596e-02 -0.155  0.87662
journal      -2.117e-01  8.092e-01  4.259e-01 -0.497  0.61917
radio        -2.948e-01  7.446e-01  1.973e-01 -1.494  0.13508
tv            1.136e-01  1.120e+00  1.971e-01  0.576  0.56432
nb_enf_ideal -5.224e-05  9.999e-01  3.014e-03 -0.017  0.98617
test          3.357e-02  1.034e+00  5.753e-02  0.584  0.55954
taille       -5.927e-02  9.424e-01  2.522e-02 -2.350  0.01878
structure     1.058e-01  1.112e+00  9.161e-02  1.155  0.24827
richesse      1.428e-01  1.154e+00  9.119e-02  1.566  0.11732

Likelihood ratio test=44.51  on 17 df, p=0.0002877
n= 1579, number of events= 142
```

D'après la mise en place du modèle complet de Cox, comprenant toutes les covariables, les variables les plus significatives sont: le **sexe** de l'enfant, le **milieu** de vie, le **travail** de la mère et la **taille** de la mère.

- **Modèle obtenu par sélection de variable backward**

Dans la suite, nous avons jugé opportun de procéder à une sélection des variables pour notre modèle en utilisant la méthode de sélection backward, qui consiste à partir d'un modèle incluant toutes les variables, puis à éliminer progressivement celles qui sont les moins significatives, en évaluant leur impact sur la qualité du modèle à chaque étape.

Ci suit le modèle que nous obtenons après sélection :

```
Call:
coxph(formula = Surv(C, ð) ~ sexe + milieu + travail + radio +
    taille + richesse, data = df_survie_MIJ)

             coef exp(coef) se(coef)      z        p
sexe     -0.80770   0.44588  0.17813 -4.534 5.78e-06
milieu    0.74111   2.09826  0.26115  2.838  0.00454
travail   0.44512   1.56068  0.13733  3.241  0.00119
radio    -0.25984   0.77117  0.18033 -1.441  0.14961
taille   -0.04069   0.96012  0.01992 -2.043  0.04107
richesse  0.17384   1.18986  0.08798  1.976  0.04818
```

```
Likelihood ratio test=41.33  on 6 df, p=2.497e-07
n= 1579, number of events= 142
```

Comparaison des modèles

Nous avons ensuite comparé tous les modèles en évaluant leurs performances à l'aide du critère AIC. Ci dessous l'évaluation:

```
                     df      AIC
fit_expo_base         1 1162.820
fit_weibull_base      2 1152.166
fit_weibull_complet  19 1159.867
fit_expo_complet     18 1169.353
modele_cox_complet   17 2026.445
modele_cox_backward   6 2007.631
```

D'après cette évaluation, le meilleur modèle serait le modèle nul paramétrique basé sur l'hypothèse d'une distribution weibull pour la durée de vie infantilo-juvénile d'après le critère **AIC**.

Mise en place du modèle d'apprentissage

Nous avons opté pour l'algorithme des forêts aléatoires de survie comme modèle d'apprentissage automatique pour l'analyse de survie.La forêt aléatoire de survie est une méthode qui repose sur la combinaison de plusieurs arbres de décision de survie. Chaque arbre est construit de manière aléatoire en utilisant un sous-ensemble des données disponibles. Les prédictions de ces arbres sont ensuite agrégées pour renforcer la robustesse et la précision du modèle global. Celui-ci intègre un estimateur de mortalité défini par :

$$\hat{S}(Z_i) = \sum_{j=1}^{n} \hat{\wedge}(T_j|Z_i)$$

telle que:

- le score de risque: $\hat{S}(Z_i)$
- une covariable: Z_i
- avec une probabilité de concordance :

$$P_c = \mathbb{P}(\hat{S}(Z_2) < \hat{S}(Z_1)|X_2 > X_1)$$

```
i Creating pre-processing data to finalize unknown parameter: mtry
```

- **Présentation du C-index de Harrell**

Pour estimer la probabilité de concordance, nous avons utilisé une méthode reposant sur le calcul de l'**indice de concordance de Harrell**. Cette métrique permet d'évaluer la capacité de discrimination d'un modèle ou d'un marqueur. En d'autres termes, elle mesure la performance du modèle à distinguer les sujets à risque élevé de ceux à faible risque en termes de survie.

Si nous considérons un sous ensemble de variables observées défini par un sous ensemble ν de $1, ..., n$ pour restreindre le calcul du C-index à un groupe spécifique, le C de Harrell s'exprime alors par:

$$CI_H^{\nu}(\hat{S}) = \frac{\sum_{(i,j \in \nu, i \neq j)} 1_{t_i < t_j} \times 1_{\hat{S}(z_i) > \hat{S}(z_j)} \times \delta_i}{\sum_{(i,j \in \nu, i \neq j)} 1_{t_i < t_j} 1 \times \delta_i}$$

avec

- t_j et t_j: Les temps de survie observés pour les individus i et j

- $\hat{S}(z_j)$ et $\hat{S}(z_j)$: Les probabilités de survie prédites par le modèle pour les individus i et j, respectivement.
- δ_i : Un indicateur d'événement (par exemple, $\delta_i = 1$ si l'individu i a eu un événement, et $\delta_i = 0$ s'il est censuré).
- $1_{t_i<t_j}$: Une fonction indicatrice qui vaut 1 si $t_i < t_j$ (l'individu i a un temps de survie plus court que j).
- $1_{\hat{S}(z_j)>\hat{S}(z_j)}$: Une fonction indicatrice qui vaut 1 si la probabilité de survie prédite pour i est supérieure à celle de j.
- **Affichage des 5 meilleurs modèles en fonction du C- index de Harrell**

```
# A tibble: 5 x 9
   mtry min_n .metric          .estimator .eval_time  mean     n std_err .config
  <int> <int> <chr>            <chr>           <dbl> <dbl> <int>   <dbl> <chr>
1     4    39 concordance_sur~ standard           NA 0.541    10 0.0109  Prepro~
2     6    30 concordance_sur~ standard           NA 0.538    10 0.0102  Prepro~
3    12    34 concordance_sur~ standard           NA 0.537    10 0.0102  Prepro~
4     7    28 concordance_sur~ standard           NA 0.533    10 0.00999 Prepro~
5     5    20 concordance_sur~ standard           NA 0.528    10 0.0104  Prepro~
```

- **Autre métrique: Brier survivial score**

Nous avons également utilisé une autre métrique, qui évalue des caractéristiques différentes de notre modèle par rapport au C-index de Harrell. Il s'agit du **Brier survival** (ou **score de Brier** adapté aux données de survie). Cette métrique mesure l'exactitude des probabilités prédites de survie à un temps donné et est sensible à la calibration des probabilités(correspondance des probabilités prédites par le modèle et celles réelement observées). De plus, elle pénalise à la fois les prédictions trop confiantes et celles qui sont erronées.

Le Brier survivial score s'exprime par:

$$BS(t) = \frac{1}{N} \sum_{i=1}^{N} \left[\hat{S}(t|X_i > t) - I(T_i > t) \right]^2 \times w_i(t)$$

avec:

N : le nombre total d'individus

$\hat{S}(t|X_i > t)$: la probabilité prédite de survie au temps t pour l'individu i, conditionnellement à ses covariables X_i.

$I(T_i > t)$: Cette fonction indicatrice représente le résultat observé : 1 si l'individu survit au-delà de t, et 0 sinon.

$w_i(t)$: Estimée par la méthode de Kaplan-Meier, ce poids ajuste pour la censure. Car il permet de donner moins de poids aux individus censurés avant le temps t, car leur statut de survie au-delà de t est inconnu.

- **Affichage des 5 meilleurs modèles en fonction du Brier survival score.**

```
# A tibble: 5 x 9
   mtry min_n .metric        .estimator .eval_time  mean     n std_err .config
  <int> <int> <chr>          <chr>           <dbl> <dbl> <int>   <dbl> <chr>
1    16    16 brier_survival standard            0     0    10       0 Preproce~
2     6    30 brier_survival standard            0     0    10       0 Preproce~
3    10    11 brier_survival standard            0     0    10       0 Preproce~
4     2     7 brier_survival standard            0     0    10       0 Preproce~
5     7    28 brier_survival standard            0     0    10       0 Preproce~
```

- **Modèle final et évaluation après entraînement**

Nous avons construit ici un nouveau modèle, nommé “modèle final”, en utilisant les meilleurs ensembles de paramètres identifiés selon la métrique du C-index de Harrell.

```
# A tibble: 1 x 5
  .metric              .estimator .eval_time .estimate .config
  <chr>                <chr>           <dbl>     <dbl> <chr>
1 concordance_survival standard           NA     0.497 Preprocessor1_Model1
```

la valeur du C de Harrell estimé sur est relativement proche de 0.5, ce qui suggère que la performance du modèle est modeste en termes de capacité à classer correctement les paires d’individus selon leurs risque de survie.

Pour aller plus loin: comparaison de modèle de survie

Afin d’obtenir une vision plus détaillée et comparative des performances de la qualité des prédictions probabilistes de notre modèle sur les données de test et sur la période étudiée , nous avons explorer un modèle de Cox. Cette comparaison se base sur les valeurs de la métrique “score de Brier” obtenu au cours de ’la recherche par grille de l’entrainement.

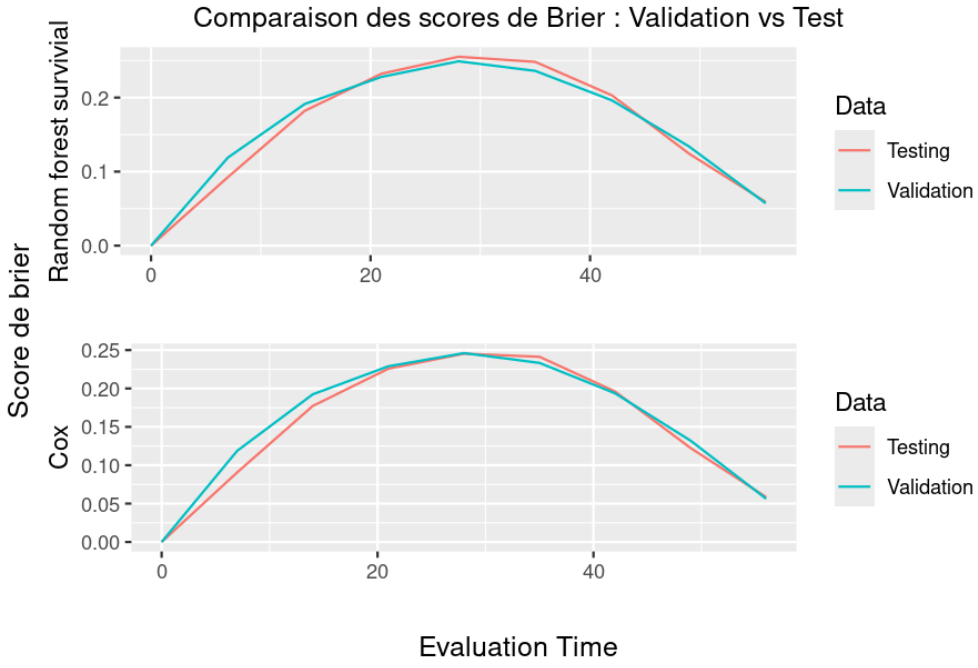
- Comparaison des deux modèles:

```
# A tibble: 2 x 3
  Model                    `C-index de Harrell` `Erreur de prédiction`
  <chr>                                   <dbl>                  <dbl>
1 Random forest survivial                 0.497                  0.503
2 Cox                                     0.557                  0.443
```

Le C-index de Harrell est une mesure de la capacité d’un modèle à classer correctement les paires d’individus en fonction de leur risque relatif. Ainsi, plus la valeur estimée se rapproche de 1, meilleure est la classification.

D’après les erreurs de prédictions obtenues sur l’échantillon test avec les modèles finaux, le modèle de Cox montre de meilleures performances par rapport au modèle Random Forest Survival.

- **Évaluation des performances de généralisation des modèles en visualisant les courbes de test et de validation à différents points temporels, en utilisant le Brier survival score:**



D'après l'analyse du score de Brier pour les échantillons de validation et de test, les deux modèles (Cox et Random Forest Survival) présentent globalement une qualité de prédiction similaire. Cependant, une différence notable apparaît sur certaines périodes de temps, où le modèle de Cox montre une meilleure performance prédictive par rapport au **Random Forest Survival**. Cette supériorité se traduit par une plus grande cohérence entre les courbes de validation et de test pour le modèle de Cox durant cette période, tandis que le Random Forest Survival présente des signes de surapprentissage, avec un écart plus marqué entre les courbes de validation et de test.

Conclusion

À travers cette étude, nous avons mené une analyse complète de survie sur nos données, ce qui nous a permis non seulement d'approfondir notre compréhension des outils d'étude de survie, mais aussi de mieux appréhender l'événement étudié, à savoir la survie infantojuvénile. Cette analyse nous a offert une vision globale de l'évolution de la survie infantojuvénile sur la durée de l'étude, révélant une diminution progressive de la probabilité de survie au fil du temps. Cette tendance a été mise en évidence grâce à l'utilisation de modèles non paramétriques tels que **Kaplan-Meier** et **Nelson-Aalen**, qui ont permis de décrire la fonction de survie sans faire d'hypothèses restrictives sur la distribution des données.

Nous avons également pu identifier des différences significatives dans la survie entre différents groupes, en fonction de variables explicatives qualitatives. Par exemple, en comparant les groupes définis par le sexe des nourrissons, nous avons observé une survie moins favorable pour les individus de sexe masculin, une différence confirmée par l'application du **test du log-rank**. Ces résultats illustrent l'importance des facteurs démographiques dans l'étude de la survie.

Ensuite, la mise en place d'un **modèle de forêt aléatoire de survie** nous a permis d'évaluer la capacité prédictive de notre approche.

Le **modèle de forêt aléatoire de survie** a montré des prédictions légèrement supérieures à celles d'une prédiction aléatoire. Cependant, nous avons finalement opté pour le modèle de **Cox**, en raison de sa meilleure performance prédictive par rapport à la forêt aléatoire de survie et de sa plus grande simplicité d'interprétation. Néanmoins, la qualité des prédictions de ce modèle varie en fonction de la période étudiée, une tendance également observée avec la forêt aléatoire. Nous avons constaté que le modèle de Cox prédit plus efficacement la survie après le premier mois de vie, suggérant une meilleure capacité à capturer les dynamiques de survie à long terme.

Cependant, ces résultats doivent être interprétés avec prudence, car ils sont basés sur des **données fictives**. Bien que ces données soient utiles pour explorer et valider des méthodes, leur comportement trop idéalisé limite la généralisation de nos conclusions à des situations réelles de survie infantojuvénile.

Cette étude nous a permis d'explorer les outils et les méthodes d'analyse de survie, tout en mettant en lumière les limites liées à l'utilisation de données simulées. Pour des applications concrètes, il sera essentiel de valider ces approches sur des données réelles, afin de mieux comprendre les facteurs influençant la survie infantojuvénile et d'améliorer les interventions visant à réduire la mortalité dans cette population.