Survival and longitudinal data analysis

Exercise 8.11 of Klein and Moeschberger

Etienne Gaucher

12/10/2021

Introduction

L'étude que nous allons mener consiste à savoir s'il existe ou non un lien entre l'allaitement des nouveau-nés et le risque de développer une pneumonie lors des 12 premiers mois de la vie. 3 470 nourrissons ont participé à l'enquête entre 1979 et 1986, et pour chacun, on a noté la durée entre la naissance et le développement d'une pneumonie. Puisque l'étude ne s'intéresse qu'aux 12 premiers mois, une censure à droite s'impose, car certains nourrissons ne seront pas atteints lors des 12 premiers mois. Parmi les covariables, on retrouve par exemple l'âge de la mère, sa consommation d'alcool et de tabac pendant la grossesse ou le nombre de frères et sœurs du nouveau-né.

Tout d'abord, on commence par importer les données nécessaires à cette étude. Les données sont stockées dans le package **KMsurv** sous le nom 'pneumon'. Les variables qualitatives font l'objet d'une attention particulière pour qu'elles soient considérées en tant que *factor* et non *integer* ou *numeric*. Cela aura un intérêt pour une éventuelle régression.

```
# import des librairies
library(KMsurv)
library(tidyverse)

# import des données et recodage
data(pneumon)
pneumon<-pneumon %>% dplyr::select(-agepn) %>% mutate(region = as.factor(region),
                                                      race = as.factor(race),
                                                      urban = as.factor(urban),
                                                      alcohol = as.factor(alcohol),
                                                      smoke = as.factor(smoke),
                                                      poverty = as.factor(poverty))
```

Question 1

Pour s'assurer que les données ont été correctement importées, on peut visualiser une partie des données, par exemple les premières lignes.

```
head(pneumon)
```

```
##   chldage hospital mthage urban alcohol smoke region poverty bweight race
## 1      12        0     22     1       0     0      1       1       1    1
## 2      12        0     20     1       1     0      1       1       0    1
## 3       3        0     24     1       3     0      1       1       0    1
## 4       2        0     22     1       2     2      1       1       0    1
## 5       4        0     21     1       1     2      1       1       1    1
```

```
## 6      12        0     20     1       0     0      1       1       0    1
##   education nsibs wmonth sfmonth
## 1        10     1      1       1
## 2        12     1      2       2
## 3        12     2      1       0
## 4         9     0      0       0
## 5        12     0      0       0
## 6        12     0      0       0
```

```
# type de chaque variable
sapply(pneumon, class)
```

```
##   chldage  hospital    mthage     urban   alcohol     smoke    region   poverty
## "numeric" "integer" "integer"  "factor"  "factor"  "factor"  "factor"  "factor"
##   bweight      race education     nsibs    wmonth   sfmonth
## "integer"  "factor" "integer" "integer" "integer" "integer"
```

Toutes les variables semblent être correctement importées.

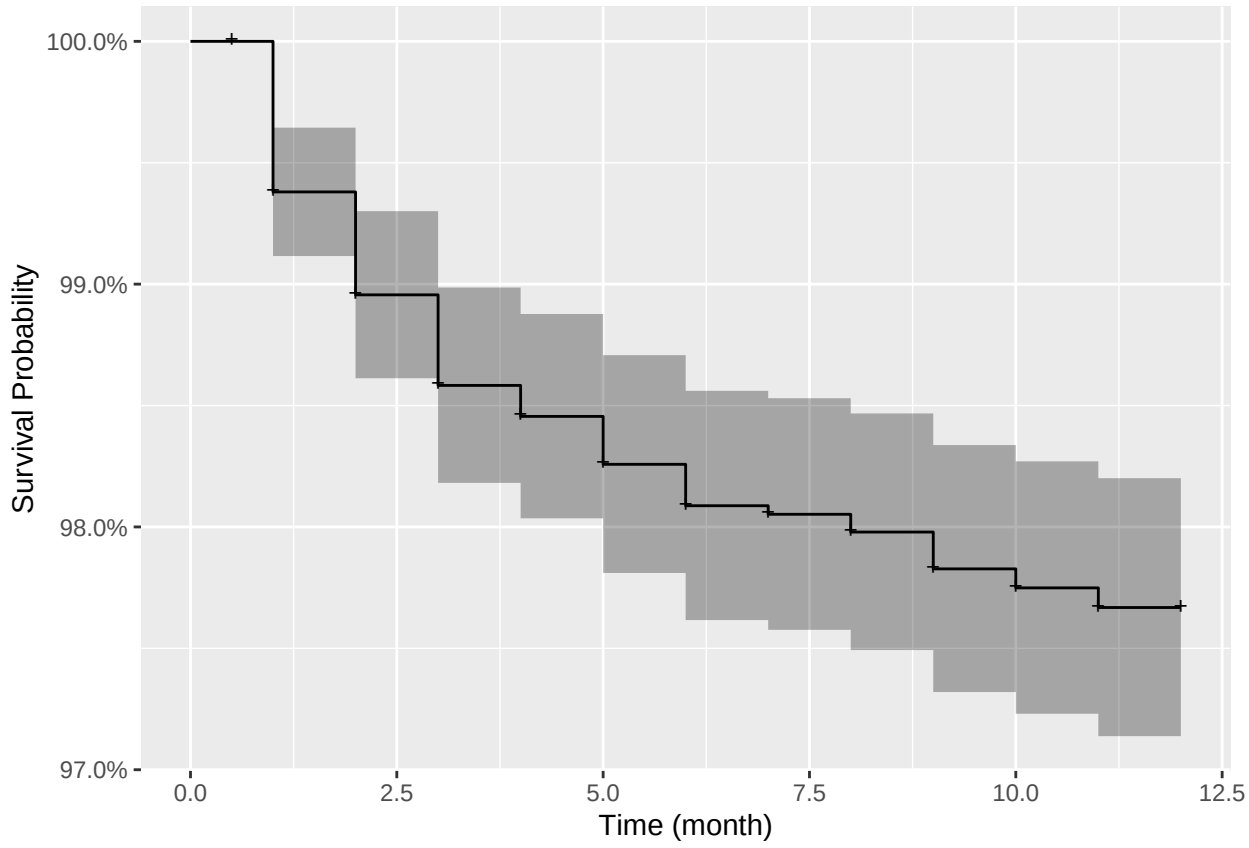
Question 2

Pour construire l'estimateur de Kaplan-Meier, on utilise la librairie 'survival'. L'estimateur de Kaplan-Meier est une estimation non-paramétrique de la fonction de survie.

```
# import des librairies
library(survival)
library(ggfortify)

# estimateur de Kaplan-Meier
km<-survfit(Surv(chldage, hospital) ~ 1, data = pneumon)

# graphique de l'estimateur de Kaplan-Meier
autoplot(km, xlab = "Time (month)", ylab = "Survival Probability")
```



L'intervalle de confiance de l'estimateur est représenté par la bande grisée sur le graphique. On souhaite connaître la probabilité de ne pas développer une pneumonie lors des 6 premiers mois pour un nouveau-né. Afin d'être plus précis qu'une simple lecture sur le graphique, on recherche les valeurs exactes.

```
summary(km)
```

```
## Call: survfit(formula = Surv(chldage, hospital) ~ 1, data = pneumon)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     1   3386      21    0.994 0.00135        0.991        0.996
##     2   3282      14    0.990 0.00176        0.986        0.993
##     3   3184      12    0.986 0.00205        0.982        0.990
##     4   3089       4    0.985 0.00215        0.980        0.989
##     5   2993       6    0.983 0.00229        0.978        0.987
##     6   2880       5    0.981 0.00241        0.976        0.986
##     7   2779       1    0.981 0.00243        0.976        0.985
##     8   2682       2    0.980 0.00249        0.975        0.985
##     9   2585       4    0.978 0.00260        0.973        0.983
##    10   2496       2    0.977 0.00265        0.972        0.983
##    11   2418       2    0.977 0.00271        0.971        0.982
```

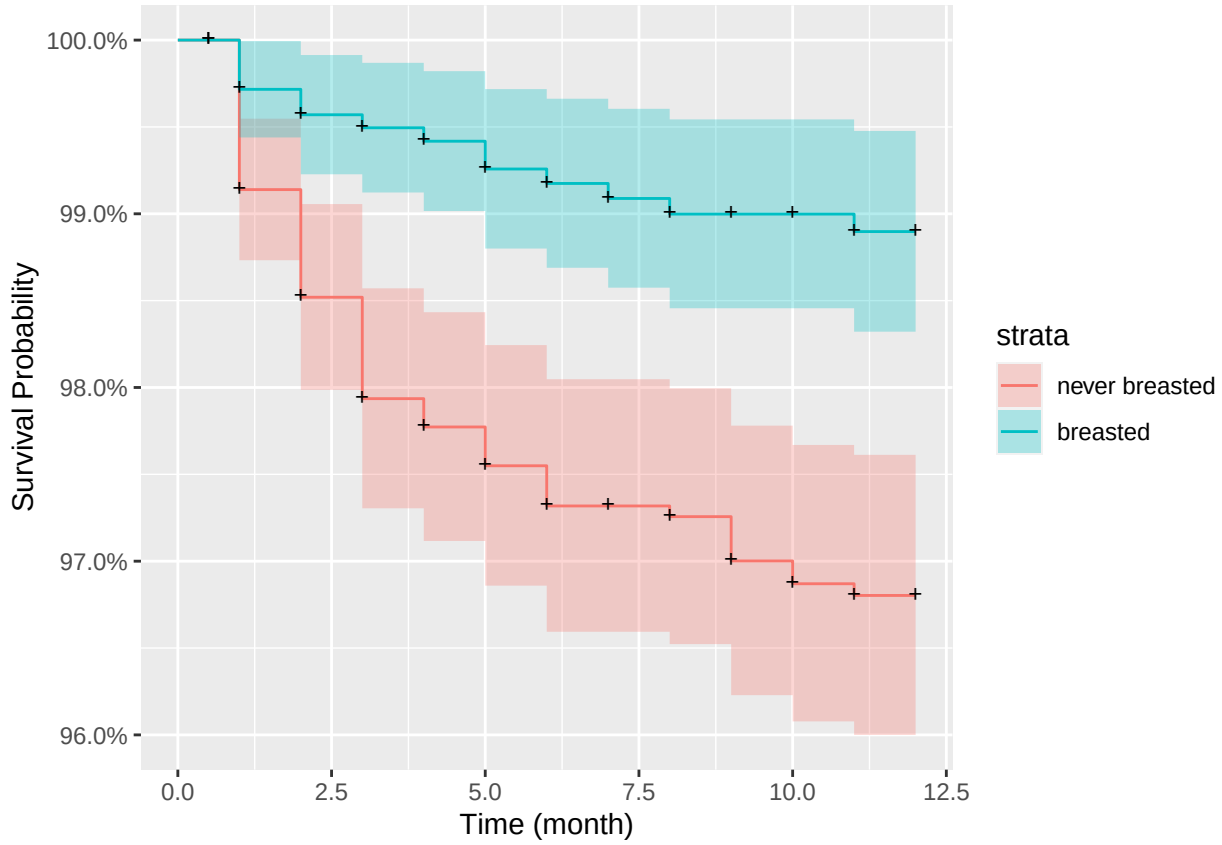
La probabilité pour un nouveau-né de ne pas développer une pneumonie lors des 6 premiers mois est 0,981. L'intervalle de confiance de cette probabilité est $[0.976, 0.986]$.

Question 3

L'allaitement par la mère est une variable binaire que l'on veut considérer. Cependant, le jeu de données ne donne que la durée entre la naissance et la fin de l'allaitement grâce à la variable *wmonth*. On doit donc créer une variable binaire nommée *BreastFed* égale à *breasted* si l'enfant a été allaité par la mère (donc *wmonth* strictement supérieur à 0), et *never breasted* si l'enfant n'a pas été allaité. On trace ensuite l'estimateur de Kaplan-Meier pour les 2 populations : enfants allaités et non allaités.

```
# création de la variable BreastFed
pneumon<-pneumon %>% mutate(BreastFed=recode(factor(wmonth>0),
                                             'FALSE'='never breasted',
                                             'TRUE'='breasted'))

# estimateur de Kaplan-Meier pour les deux populations
kmsurvival_BF <- survfit(Surv(chldage,hospital) ~ BreastFed, data = pneumon)
autoplot(kmsurvival_BF, xlab = "Time (month)", ylab = "Survival Probability")
```



Les deux fonctions de survie empiriques ne se croisent pas. En revanche, les intervalles de confiance se chevauchent. Par conséquent, on doit effectuer le test du log-rank pour savoir si l'allaitement a un impact sur le développement d'une pneumonie. Le test du log-rank compare les deux fonctions de survie.

```
# test du log-rank
survdiff(Surv(chldage,hospital) ~ BreastFed, data = pneumon)
```

```
## Call:
## survdiff(formula = Surv(chldage, hospital) ~ BreastFed, data = pneumon)
##
##                              N Observed Expected (O-E)^2/E (O-E)^2/V
```

```
## BreastFed=never breasted 2036       59     42.7      6.22        15
## BreastFed=breasted       1434       14     30.3      8.77        15
##
##  Chisq= 15  on 1 degrees of freedom, p= 1e-04
```

On obtient une p-value égale à 1e-04. La p-value est très faible, donc on rejette H_0 = il n'y a pas de différence pour l'âge où on développe une pneumonie entre les deux populations. On en conclut qu'il y a probablement une association entre l'allaitement et le développement d'une pneumonie. Le graphique laisse penser que les enfants allaités sont moins atteints que les enfants non-allaités.

Question 4

On cherche à tester l'hypothèse $H_0 : \beta^*_{breastFed} = 0$ en utilisant le test de Wald et le test du rapport de vraisemblance.

```
cox_model<-coxph(Surv(chldage,hospital) ~ BreastFed, data = pneumon)
summary(cox_model)
```

```
## Call:
## coxph(formula = Surv(chldage, hospital) ~ BreastFed, data = pneumon)
##
##   n= 3470, number of events= 73
##
##                       coef exp(coef) se(coef)     z Pr(>|z|)
## BreastFedbreasted -1.0970    0.3339   0.2973 -3.69 0.000224 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                   exp(coef) exp(-coef) lower .95 upper .95
## BreastFedbreasted    0.3339      2.995    0.1864    0.5979
##
## Concordance= 0.614  (se = 0.023 )
## Likelihood ratio test= 16.59  on 1 df,   p=5e-05
## Wald test            = 13.62  on 1 df,   p=2e-04
## Score (logrank) test = 15.04  on 1 df,   p=1e-04
```

La p-value associée au test de Wald est égale à 2e-04 $<$ 0,01. On rejette donc l'hypothèse que $\beta^*_{breastFed} = 0$. De même, la p-value associée au test du rapport de vraisemblance est égale à 5e-05 $<$ 0,01. Une nouvelle fois, on rejette l'hypothèse que $\beta^*_{breastFed} = 0$. Les fonctions de survie des deux populations semblent différentes.

La valeur estimée de $\beta^*_{breastFed}$ vaut $\hat{\beta}^*_{breastFed}$ = -1,097, avec un écart type de 0,2973.

Question 5

Les autres variables disponibles dans le jeu de données peuvent également être associées au développement d'une pneumonie. Ainsi, on souhaite tester si l'allaitement a réellement un impact sur le développement d'une pneumonie en ajoutant chaque variable dans des modèles séparés. Cela permettra de savoir si la durée entre la naissance et le développement d'une pneumonie est la même pour les enfants allaités et non-allaités lorsque l'on ajoute une variable dans le modèle initial. Cela revient à tester $H_0 : \beta^*_{breastFed} = 0$ pour chaque modèle. On utilise le test de Wald pour cette question.

```
# boucle for sur les variables du dataset
for (var in names(pneumon[3:14]))
{
  # formule du modèle du Cox
```

```
  formule<-paste(c("Surv(chldage, hospital) ~ BreastFed", var), collapse = '+')
  formule<-as.formula(formule)

  # séparation des résultats de chaque modèle
  cat(print(summary(coxph(formule, data=pneumon))),
      "\n",
      "-------------------------",
      "\n \n")
}
```

```
## Call:
## coxph(formula = formule, data = pneumon)
##
##   n= 3470, number of events= 73
##
##                       coef exp(coef) se(coef)      z Pr(>|z|)
## BreastFedbreasted -1.02651   0.35826  0.30096 -3.411 0.000648 ***
## mthage            -0.06776   0.93448  0.04521 -1.499 0.133908
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                   exp(coef) exp(-coef) lower .95 upper .95
## BreastFedbreasted    0.3583      2.791    0.1986    0.6462
## mthage               0.9345      1.070    0.8552    1.0211
##
## Concordance= 0.635  (se = 0.028 )
## Likelihood ratio test= 18.86  on 2 df,   p=8e-05
## Wald test            = 15.86  on 2 df,   p=4e-04
## Score (logrank) test = 17.29  on 2 df,   p=2e-04
##
##
##  -------------------------
##
## Call:
## coxph(formula = formule, data = pneumon)
##
##   n= 3470, number of events= 73
##
##                       coef exp(coef) se(coef)     z Pr(>|z|)
## BreastFedbreasted -1.0720    0.3423   0.2978 -3.60 0.000319 ***
## urban1            -0.3819    0.6826   0.2496 -1.53 0.125997
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                   exp(coef) exp(-coef) lower .95 upper .95
## BreastFedbreasted    0.3423      2.921    0.1910    0.6137
## urban1               0.6826      1.465    0.4185    1.1133
##
## Concordance= 0.638  (se = 0.029 )
## Likelihood ratio test= 18.82  on 2 df,   p=8e-05
## Wald test            = 16.01  on 2 df,   p=3e-04
## Score (logrank) test = 17.5  on 2 df,   p=2e-04
##
##
```

```
##  -------------------------
##
## Call:
## coxph(formula = formule, data = pneumon)
##
##   n= 3470, number of events= 73
##
##                      coef exp(coef) se(coef)      z Pr(>|z|)
## BreastFedbreasted -1.1108    0.3293   0.2989 -3.717 0.000202 ***
## alcohol1           0.2079    1.2311   0.3051  0.681 0.495597
## alcohol2          -0.1742    0.8402   0.4703 -0.370 0.711160
## alcohol3          -0.2152    0.8064   0.5952 -0.361 0.717729
## alcohol4          -0.0404    0.9604   0.5952 -0.068 0.945879
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                   exp(coef) exp(-coef) lower .95 upper .95
## BreastFedbreasted    0.3293     3.0367    0.1833    0.5915
## alcohol1             1.2311     0.8123    0.6770    2.2386
## alcohol2             0.8402     1.1903    0.3342    2.1121
## alcohol3             0.8064     1.2401    0.2512    2.5893
## alcohol4             0.9604     1.0412    0.2991    3.0839
##
## Concordance= 0.629  (se = 0.029 )
## Likelihood ratio test= 17.45  on 5 df,   p=0.004
## Wald test            = 14.44  on 5 df,   p=0.01
## Score (logrank) test = 15.85  on 5 df,   p=0.007
##
##
##  -------------------------
##
## Call:
## coxph(formula = formule, data = pneumon)
##
##   n= 3470, number of events= 73
##
##                      coef exp(coef) se(coef)      z Pr(>|z|)
## BreastFedbreasted -1.0514    0.3494   0.2978 -3.530 0.000415 ***
## smoke1             0.7644    2.1476   0.2554  2.993 0.002760 **
## smoke2             0.6821    1.9781   0.3474  1.963 0.049609 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                   exp(coef) exp(-coef) lower .95 upper .95
## BreastFedbreasted    0.3494     2.8617    0.1949    0.6264
## smoke1               2.1476     0.4656    1.3020    3.5426
## smoke2               1.9781     0.5055    1.0012    3.9084
##
## Concordance= 0.667  (se = 0.031 )
## Likelihood ratio test= 26.52  on 3 df,   p=7e-06
## Wald test            = 23.65  on 3 df,   p=3e-05
## Score (logrank) test = 25.69  on 3 df,   p=1e-05
##
##
```

```
##  -------------------------
##
## Call:
## coxph(formula = formule, data = pneumon)
##
##   n= 3470, number of events= 73
##
##                       coef exp(coef) se(coef)      z Pr(>|z|)
## BreastFedbreasted -1.0937    0.3350   0.3020 -3.621 0.000293 ***
## region2            0.1651    1.1795   0.3420  0.483 0.629197
## region3           -0.3849    0.6805   0.3401 -1.132 0.257744
## region4           -0.4401    0.6440   0.4367 -1.008 0.313572
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                   exp(coef) exp(-coef) lower .95 upper .95
## BreastFedbreasted    0.3350     2.9852    0.1853    0.6055
## region2              1.1795     0.8478    0.6034    2.3057
## region3              0.6805     1.4695    0.3494    1.3254
## region4              0.6440     1.5529    0.2736    1.5157
##
## Concordance= 0.649  (se = 0.029 )
## Likelihood ratio test= 21.47  on 4 df,   p=3e-04
## Wald test            = 18.5  on 4 df,   p=0.001
## Score (logrank) test = 20.07  on 4 df,   p=5e-04
##
##
##  -------------------------
##
## Call:
## coxph(formula = formule, data = pneumon)
##
##   n= 3470, number of events= 73
##
##                       coef exp(coef) se(coef)      z Pr(>|z|)
## BreastFedbreasted -1.0919    0.3356   0.2977 -3.668 0.000245 ***
## poverty1          -0.1331    0.8753   0.3981 -0.334 0.738039
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                   exp(coef) exp(-coef) lower .95 upper .95
## BreastFedbreasted    0.3356      2.980    0.1872    0.6015
## poverty1             0.8753      1.142    0.4012    1.9100
##
## Concordance= 0.616  (se = 0.024 )
## Likelihood ratio test= 16.69  on 2 df,   p=2e-04
## Wald test            = 13.73  on 2 df,   p=0.001
## Score (logrank) test = 15.16  on 2 df,   p=5e-04
##
##
##  -------------------------
##
## Call:
## coxph(formula = formule, data = pneumon)
```

```
##
##   n= 3470, number of events= 73
##
##                       coef exp(coef) se(coef)      z Pr(>|z|)
## BreastFedbreasted -1.0087    0.3647   0.3018 -3.342  0.00083 ***
## bweight            0.4203    1.5224   0.2376  1.768  0.07698 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                   exp(coef) exp(-coef) lower .95 upper .95
## BreastFedbreasted    0.3647     2.7420    0.2019    0.6589
## bweight              1.5224     0.6569    0.9555    2.4255
##
## Concordance= 0.643  (se = 0.029 )
## Likelihood ratio test= 19.7  on 2 df,   p=5e-05
## Wald test            = 16.83  on 2 df,   p=2e-04
## Score (logrank) test = 18.41  on 2 df,   p=1e-04
##
##
##  -------------------------
##
## Call:
## coxph(formula = formule, data = pneumon)
##
##   n= 3470, number of events= 73
##
##                        coef exp(coef) se(coef)      z Pr(>|z|)
## BreastFedbreasted -1.20623   0.29932  0.30291 -3.982 6.83e-05 ***
## race2             -0.46977   0.62515  0.28705 -1.637    0.102
## race3             -0.05003   0.95120  0.31772 -0.157    0.875
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                   exp(coef) exp(-coef) lower .95 upper .95
## BreastFedbreasted    0.2993      3.341    0.1653     0.542
## race2                0.6251      1.600    0.3562     1.097
## race3                0.9512      1.051    0.5103     1.773
##
## Concordance= 0.645  (se = 0.029 )
## Likelihood ratio test= 19.53  on 3 df,   p=2e-04
## Wald test            = 16.72  on 3 df,   p=8e-04
## Score (logrank) test = 18.28  on 3 df,   p=4e-04
##
##
##  -------------------------
##
## Call:
## coxph(formula = formule, data = pneumon)
##
##   n= 3470, number of events= 73
##
##                        coef exp(coef) se(coef)      z Pr(>|z|)
## BreastFedbreasted -0.97282   0.37802  0.30023 -3.240  0.00119 **
## education         -0.14935   0.86127  0.05377 -2.777  0.00548 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                   exp(coef) exp(-coef) lower .95 upper .95
## BreastFedbreasted    0.3780      2.645    0.2099    0.6809
## education            0.8613      1.161    0.7751    0.9570
##
## Concordance= 0.674  (se = 0.028 )
## Likelihood ratio test= 23.53  on 2 df,   p=8e-06
## Wald test            = 21.14  on 2 df,   p=3e-05
## Score (logrank) test = 22.22  on 2 df,   p=1e-05
##
##
##  -------------------------
##
## Call:
## coxph(formula = formule, data = pneumon)
##
##   n= 3470, number of events= 73
##
##                       coef exp(coef) se(coef)      z Pr(>|z|)
## BreastFedbreasted -1.0454    0.3516   0.2983 -3.505 0.000457 ***
## nsibs              0.2785    1.3212   0.1140  2.444 0.014545 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                   exp(coef) exp(-coef) lower .95 upper .95
## BreastFedbreasted    0.3516     2.8445    0.1959    0.6308
## nsibs                1.3212     0.7569    1.0567    1.6519
##
## Concordance= 0.656  (se = 0.027 )
## Likelihood ratio test= 21.91  on 2 df,   p=2e-05
## Wald test            = 19.76  on 2 df,   p=5e-05
## Score (logrank) test = 21.32  on 2 df,   p=2e-05
##
##
##  -------------------------
##
## Call:
## coxph(formula = formule, data = pneumon)
##
##   n= 3470, number of events= 73
##
##                       coef exp(coef) se(coef)      z Pr(>|z|)
## BreastFedbreasted -0.5174    0.5961   0.4154 -1.246    0.213
## wmonth            -0.1588    0.8532   0.1016 -1.563    0.118
##
##                   exp(coef) exp(-coef) lower .95 upper .95
## BreastFedbreasted    0.5961      1.678    0.2640     1.346
## wmonth               0.8532      1.172    0.6992     1.041
##
## Concordance= 0.623  (se = 0.022 )
## Likelihood ratio test= 20.11  on 2 df,   p=4e-05
## Wald test            = 13.11  on 2 df,   p=0.001
```

```
## Score (logrank) test = 16.23  on 2 df,   p=3e-04
##
##
##  ------------------------
##
## Call:
## coxph(formula = formule, data = pneumon)
##
##   n= 3470, number of events= 73
##
##                       coef exp(coef) se(coef)      z Pr(>|z|)
## BreastFedbreasted -0.5845    0.5574   0.4375 -1.336    0.182
## sfmonth           -0.2234    0.7998   0.1664 -1.342    0.179
##
##                   exp(coef) exp(-coef) lower .95 upper .95
## BreastFedbreasted    0.5574      1.794    0.2364     1.314
## sfmonth              0.7998      1.250    0.5772     1.108
##
## Concordance= 0.626  (se = 0.021 )
## Likelihood ratio test= 18.87  on 2 df,   p=8e-05
## Wald test            = 13.47  on 2 df,   p=0.001
## Score (logrank) test = 15.87  on 2 df,   p=4e-04
##
##
##  ------------------------
##
```

Lorsque le modèle contient séparément les variables *mthage*, *urban*, *alcohol*, *smoke*, *region*, *poverty*, *bweight*, *race*, *education* ou *nsibs*, la p-value associée au test de Wald pour la variable *BreastFed* est toujours inférieure à 0,05. Pour ces modèles, on en conclut que le temps lié au développement d'une pneumonie n'est pas la même pour les enfants allaités et non-allaités.

Pour les modèles avec les variables *wmonth* et *sfmonth*, la p-value est respectivement de 0,213 et 0,182. On rappelle que la variable *wmonth* correspond à la durée d'allaitement, et *sfmonth* la durée qu'il a fallu à l'enfant pour manger de la nourriture solide. Les résultats des deux tests concluent que le temps lié au développement d'une pneumonie est la même pour les enfants allaités et non-allaités. Cependant, il semble logique que ces 2 variables soient fortement corrélées à la variable *BreastFed*, ce qui fausse probablement le modèle et donc le test de Wald. Puisque l'on s'intéresse uniquement à la variable *BreastFed*, on ignore les variables *wmonth* et *sfmonth* dans la suite de l'exercice.

Question 6

On construit un modèle step-by-step pour obtenir un modèle précis avec les variables du dataset les plus significatives. Les variables *wmonth* et *sfmonth* ne sont pas prises en compte.

```r
# import de la librairie
library(MASS)

# construction du modèle step-by-step
model_all= coxph(Surv(chldage,hospital) ~ .-wmonth-sfmonth, data = pneumon)
model_selected = stepAIC(model_all, trace=F)
summary(model_selected)
```

```
## Call:
## coxph(formula = Surv(chldage, hospital) ~ mthage + smoke + nsibs +
```

```
##     BreastFed, data = pneumon)
##
##   n= 3470, number of events= 73
##
##                       coef exp(coef) se(coef)      z Pr(>|z|)
## mthage            -0.12102   0.88602  0.04989 -2.426  0.01529 *
## smoke1             0.74872   2.11429  0.25527  2.933  0.00336 **
## smoke2             0.63080   1.87911  0.34799  1.813  0.06988 .
## nsibs              0.38513   1.46980  0.12316  3.127  0.00177 **
## BreastFedbreasted -0.88129   0.41425  0.30241 -2.914  0.00357 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                   exp(coef) exp(-coef) lower .95 upper .95
## mthage               0.8860     1.1286    0.8035    0.9770
## smoke1               2.1143     0.4730    1.2820    3.4870
## smoke2               1.8791     0.5322    0.9500    3.7167
## nsibs                1.4698     0.6804    1.1546    1.8711
## BreastFedbreasted    0.4142     2.4140    0.2290    0.7493
##
## Concordance= 0.695  (se = 0.028 )
## Likelihood ratio test= 37.43  on 5 df,   p=5e-07
## Wald test            = 34.53  on 5 df,   p=2e-06
## Score (logrank) test = 36.67  on 5 df,   p=7e-07
```

Le modèle final utilise les variables *mthage* (âge de la mère), *smoke* (variable binaire; si la mère a fumé pendant la grossesse), *nsibs* (nombre de frères et sœurs du nouveau-né) et *BreastFed*. On implémente ensuite ce modèle.

```
cox_model_final<-coxph(Surv(chldage,hospital) ~ mthage + smoke + nsibs + BreastFed , data = pneumon)
summary(cox_model_final)
```

```
## Call:
## coxph(formula = Surv(chldage, hospital) ~ mthage + smoke + nsibs +
##     BreastFed, data = pneumon)
##
##   n= 3470, number of events= 73
##
##                       coef exp(coef) se(coef)      z Pr(>|z|)
## mthage            -0.12102   0.88602  0.04989 -2.426  0.01529 *
## smoke1             0.74872   2.11429  0.25527  2.933  0.00336 **
## smoke2             0.63080   1.87911  0.34799  1.813  0.06988 .
## nsibs              0.38513   1.46980  0.12316  3.127  0.00177 **
## BreastFedbreasted -0.88129   0.41425  0.30241 -2.914  0.00357 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                   exp(coef) exp(-coef) lower .95 upper .95
## mthage               0.8860     1.1286    0.8035    0.9770
## smoke1               2.1143     0.4730    1.2820    3.4870
## smoke2               1.8791     0.5322    0.9500    3.7167
## nsibs                1.4698     0.6804    1.1546    1.8711
## BreastFedbreasted    0.4142     2.4140    0.2290    0.7493
##
## Concordance= 0.695  (se = 0.028 )
```

```
## Likelihood ratio test= 37.43  on 5 df,   p=5e-07
## Wald test            = 34.53  on 5 df,   p=2e-06
## Score (logrank) test = 36.67  on 5 df,   p=7e-07
```

Les valeurs des coefficients $\exp(\beta_j^*)$ nous indiquent l'influence des variables sur le développement d'une pneumonie.

Le coefficient associé à *mthage* vaut 0.89, donc plus la mère est âgée et moins l'enfant risque de développer rapidement une pneumonie.

Les coefficients associés à *smoke* sont 2,11 et 1,88, donc si la mère a fumé pendant la grossesse, alors l'enfant risque de développer plus rapidement une pneumonie.

Le coefficient associé à *nsibs* vaut 1.47, donc plus l'enfant a de frères et sœurs et plus il risque de développer rapidement une pneumonie.

Le coefficient associé à *BreastFed* vaut 0.41, donc si l'enfant a été allaité, alors il a moins de risque de développer rapidement une pneumonie.

Question 7

On veut prédire la probabilité pour un nouveau-né de ne pas avoir développé de pneumonie à 6 mois. On connaît toutes les valeurs des variables, mais on utilise uniquement les variables du modèle. La mère du nouveau-né a 27 ans ($mthage = 27$), elle n'a pas fumé pendant la grossesse ($smoke = 0$) et n'a pas allaité l'enfant puisque $wmonth = 0$. Le nouveau-né a un frère ou une sœur ($nsibs = 1$).

```
# définition du nouveau-né avec ses variables
new_ind = data.frame(chldage = 6, hospital = 1, mthage = 27, smoke = 0, nsibs = 1,
                     BreastFed="never breasted") %>% mutate(smoke = as.factor(smoke))

# prédiction de la probabilité
print(exp(-predict(cox_model_final, newdata = new_ind, type = "expected")))
```

```
## [1] 0.9902698
```

La probabilité de ne pas avoir développé de pneumonie à 6 mois pour le nouveau-né est de 0,9903.