Survival Analysis - TP Lab 2

```
- REAL Philippe
```

Exercise 8.11 of Klein and Moeschberger

Import des packages

```
library(KMsurv)
library(tidyverse)
library(survival)
rm(list=ls())
```

Question 1

Vérification que les variables sont bien importées. Point d'attention particulier pour les variables de type facteurs.

```
data("pneumon")
head(pneumon)
```

	chldage <dbl>	hospital <int>	mthage <int>	urban <int>	alcohol <int>	smoke <int>	region <int>	poverty <int>	bweight <int>
1	12	0	22	1	0	0	1	1	1
2	12	0	20	1	1	0	1	1	0
3	3	0	24	1	3	0	1	1	0
4	2	0	22	1	2	2	1	1	0
5	4	0	21	1	1	2	1	1	1
6	12	0	20	1	0	0	1	1	0

6 rows | 1-10 of 16 columns

```
glimpse(pneumon)
```

```
## Observations: 3,470
## Variables: 15
## $ chldage   <dbl> 12.0, 12.0, 3.0, 2.0, 4.0, 12.0, 7.0, 3.0, 7.0, 12.0...
## $ hospital  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ mthage    <int> 22, 20, 24, 22, 21, 20, 24, 24, 26, 21, 24, 27, 20, ...
## $ urban     <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ alcohol   <int> 0, 1, 3, 2, 1, 0, 0, 3, 2, 1, 0, 0, 0, 2, 0, 4, 0, 2...
## $ smoke     <int> 0, 0, 0, 2, 2, 0, 0, 0, 2, 0, 0, 0, 0, 0, 1, 1, 1, 0...
## $ region    <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ poverty   <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ bweight   <int> 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0...
## $ race      <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ education <int> 10, 12, 12, 9, 12, 12, 12, 14, 12, 12, 16, 16, 12, 1...
## $ nsibs     <int> 1, 1, 2, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1...
## $ wmonth    <int> 1, 2, 1, 0, 0, 0, 0, 4, 1, 3, 0, 0, 0, 4, 0, 1, 9, 1...
## $ sfmonth   <int> 1, 2, 0, 0, 0, 0, 0, 2, 1, 2, 0, 0, 0, 2, 0, 1, 2, 3...
## $ agepn     <int> 1, 12, 3, 2, 4, 12, 7, 3, 6, 12, 12, 12, 3, 4, 2, 4,...
```

Traitements (fonction factor) des variables region et race sont des variables categorielles.

```
pneumon <- mutate(pneumon,race=as.factor(race))
pneumon <- mutate(pneumon,region=as.factor(region))
```

Question 2

On cherche à représenter la courbe de l'estimateur de Kaplan-Meyer pour la fonction de survie de l'âge de contraction de la pneumonie (the age at pneumonia). Pour définir l'estimateur de Kaplan-Meyer on se donne :

- Une durée T avec la fonction de survie associée $\bar{F}$
- Un temps de censure C, independant de T pour la fonction de survie
- Un n-échantillon i.i.d $(t_1^C, \delta_1), (t_2^C, \delta_2), \ldots, (t_n^C, \delta_n)$ avec $t_1^C < t_2^C < \ldots < t_n^C$ n-réalisations de la va $(T^C = min(T, C), \delta = \mathbb{I}_{T<C})$

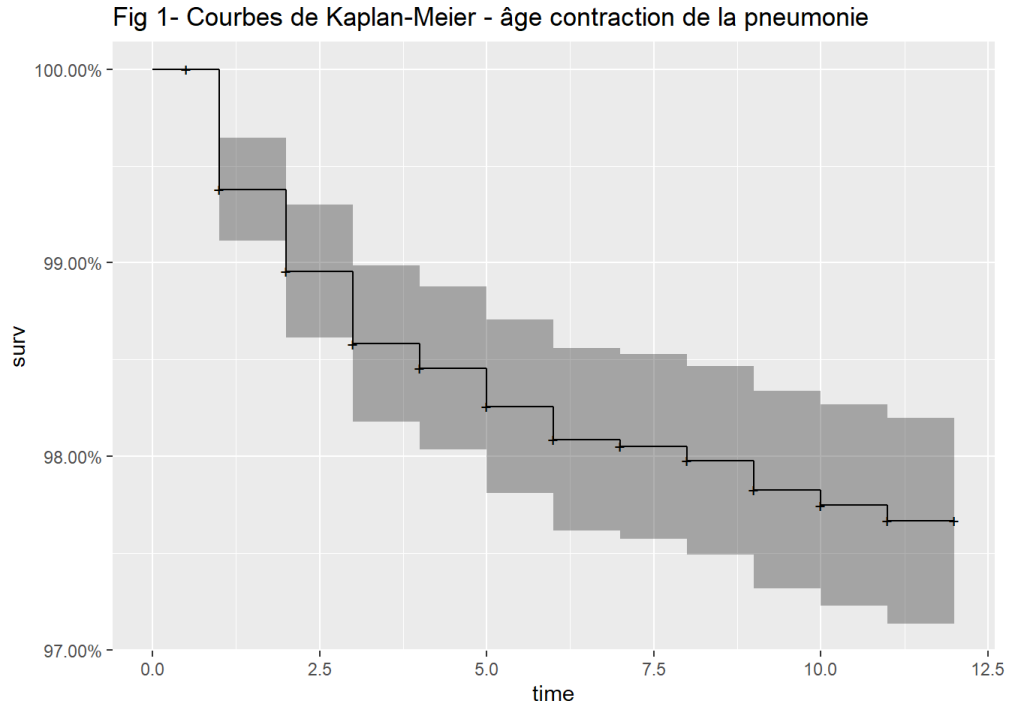
L'estimateur de Kaplan-Meyer est donnée par : $\hat{\bar{F}} = \prod_{i:t_i \leq t}(1 - \frac{\delta_i}{n-(i-1)}$ pour $t \geq t_1^C$ et 1 sinon $t < t_1^C$.

Pour calculer cet estimateur avec R on utilise la fonction survfit qui crée une courbe de survie à partir de l'estimateur Kaplan-Meier combiné à la fonction Surv qui renvoi un objet survie qui peut être passé en paramètre à survfit.

```
kmsurvival = survfit(Surv(chldage,hospital) ~ 1, data= pneumon)
```

On peut tracer la courbe de l'estimateur de Kapplan-Meyer.

```
library(ggfortify)
autoplot(kmsurvival,main="Fig 1- Courbes de Kaplan-Meier - âge contraction de la pneumonie")
```



La probabilité qu'un nouveau né n'ai pas développé une pneumonie à 6 mois peut être lu dans le résultat de la sortie ci dessous à partir de la colonne survival au time 6.

```
summary(kmsurvival)
```

```
## 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
```

On trouve que la probabilité pour un nouveau né de ne pas avoir contracté la pneumonie à 6 mois est de 0.981. Avec un intervalle de confiance de [0.976 , 0.986] pour le niveau de confiance 95% (valeur par défaut de survfit paramètre conf.int=.95).

Question 3

La variable dummy Z vaut 1 si les enfants sont allaités à la naissance et 0 sinon (ils ne le sont jamais). Pour cela on considère la variable : wmonth qui indique le mois à partir duquel l'enfant est sevré. On ajoute notre nouvelle variable Z au jeu de donnée initiale pneumon. Auparavant Z est transformée en varaiable categorielle (fonction de R : factor) par le biais de la fonction recode (qui transforme les valeurs 1 en Allaitement / 0 en Pas d'allaitement).

```
pneumon <- mutate(pneumon,Z=recode(factor(wmonth>0),"TRUE"="Allaitement","FALSE"="Pas d'allaitement"))
```

Comme à la question précédente on utilise la fonction survfit, mais cette fois ci sur nos deux groupes de population, obtenus par différentiation à partir de la variable Z.

```
KM.fit =survfit(Surv(chldage,hospital)~Z, data = pneumon)
summary(KM.fit)
```

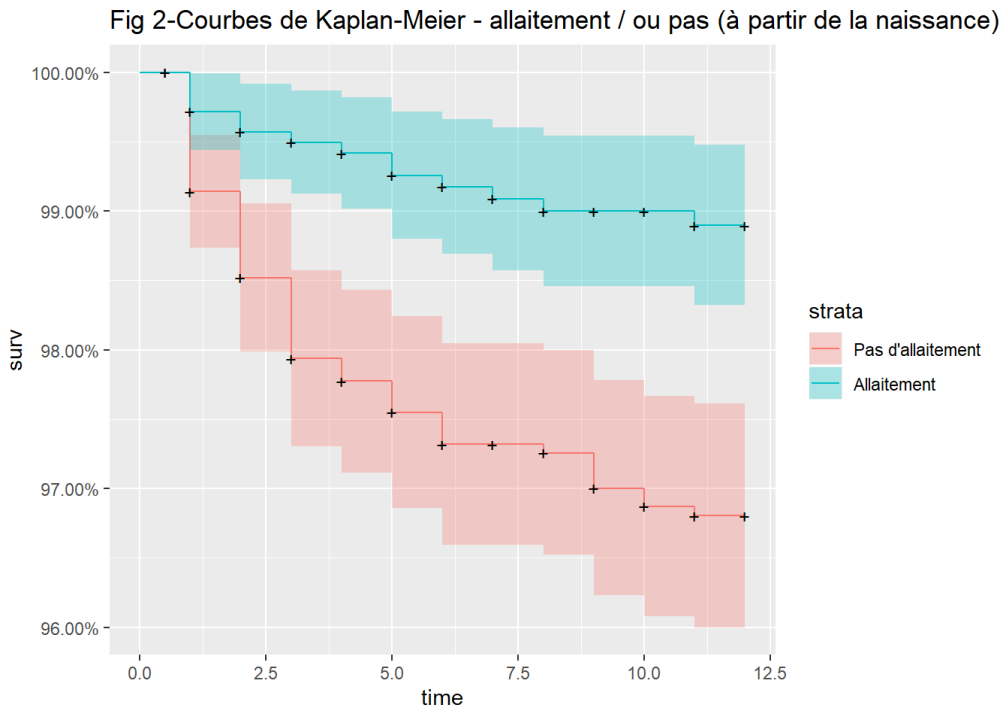
```
## Call: survfit(formula = Surv(chldage, hospital) ~ Z, data = pneumon)
##
##                Z=Pas d'allaitement
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     1   1975      17    0.991 0.00208        0.987        0.995
##     2   1920      12    0.985 0.00273        0.980        0.991
##     3   1856      11    0.979 0.00323        0.973        0.986
##     4   1803       3    0.978 0.00336        0.971        0.984
##     5   1750       4    0.975 0.00353        0.969        0.982
##     6   1687       4    0.973 0.00371        0.966        0.980
##     8   1580       1    0.973 0.00376        0.965        0.980
##     9   1526       4    0.970 0.00396        0.962        0.978
##    10   1478       2    0.969 0.00406        0.961        0.977
##    11   1436       1    0.968 0.00411        0.960        0.976
##
##                Z=Allaitement
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     1   1411       4    0.997 0.00142        0.994        1.000
##     2   1362       2    0.996 0.00175        0.992        0.999
##     3   1328       1    0.995 0.00190        0.991        0.999
##     4   1286       1    0.994 0.00205        0.990        0.998
##     5   1243       2    0.993 0.00234        0.988        0.997
##     6   1193       1    0.992 0.00248        0.987        0.997
##     7   1148       1    0.991 0.00263        0.986        0.996
##     8   1102       1    0.990 0.00277        0.985        0.995
##    11    982       1    0.989 0.00295        0.983        0.995
```

Pour avoir une idée de l'effet de l'allaitement sur le fait de contracter ou pas la pneumonie, on trace les courbes de survies des deux populations.

On remarque une différence assez nette des probabilités de survie. Lorsque l'enfant n'est pas allaité la probabilité estimé (par l'estimateur de Kaplan-Meyer) qu'il a de contacter une pneumonie est plus importante d'environ 2% au bout d'un an.

On le voit bien sur le graphique obtenu à partir de la fonction autoplot u package ggfortify. https://cran.r-project.org/web/packages/ggfortify/vignettes/plot_surv.html (https://cran.r-project.org/web/packages/ggfortify/vignettes/plot_surv.html)

```
kmsurvival.bf = survfit(Surv(chldage,hospital) ~ Z, data= pneumon)
library(ggfortify)
autoplot(kmsurvival.bf,main="Fig 2-Courbes de Kaplan-Meier - allaitement / ou pas (à partir de la naissance)")
```



L'ecartement des courbes tend à s'accentuer avec le temps toutes deux décroissante, sans jamais se recouper. La courbe de survie de la population qui n'allète pas à une décroissance assez rapidejusqu'au 3ème mois. La décroissance ralentit ensuite mais reste toujours plus importante que pour l'autre population.

La comparaison de la survie dans nos 2 groupes peut s'effectuer au moyen du test du log-rank ou du test de Wilcoxon. NB1 : Le test du log-rank est plus performant lorsque les deux courbes de survie ne se croisent pas. Ce qui est notre cas. NB2 : lorsque les taux de hasard instantanée sont proportionnels, le log-rank est le "meilleur" test que l'on puisse effectuer. Le test d'une différence de survie statistiquement significative entre plusieurs sous-groupes ou échantillons se fait dans le logiciel R au moyen de la fonction survdiff du package survival.

<Extrait de: http://iml.univ-mrs.fr/~reboul/R-survie.pdf (http://iml.univ-mrs.fr/~reboul/R-survie.pdf)>

On test donc l'hypothèse $H_0 : \beta^*_{breastf} = 0$ avec le test du log-rank

- test du log-rank

```
survdiff(Surv(chldage,hospital) ~ Z, data= pneumon)
```

```
## Call:
## survdiff(formula = Surv(chldage, hospital) ~ Z, data = pneumon)
##
##                       N Observed Expected (O-E)^2/E (O-E)^2/V
## Z=Pas d'allaitement 2036       59     42.7      6.22        15
## Z=Allaitement       1434       14     30.3      8.77        15
##
##  Chisq= 15  on 1 degrees of freedom, p= 1e-04
```

- test de Wilcoxon

```
survdiff(Surv(chldage,hospital) ~ Z, data= pneumon,rho=1)
```

```
## Call:
## survdiff(formula = Surv(chldage, hospital) ~ Z, data = pneumon,
##     rho = 1)
##
##                       N Observed Expected (O-E)^2/E (O-E)^2/V
## Z=Pas d'allaitement 2036     58.5     42.3      6.17      15.1
## Z=Allaitement       1434     13.9     30.0      8.70      15.1
##
##  Chisq= 15.1  on 1 degrees of freedom, p= 1e-04
```

Dans les 2 cas, on a une petite p-value= 1e-04. La différence entre les deux groupes d'enfants allaités ou pas est bien significative. Ce qui confirme la première inpression donnée par le graphique précédent (Fig-2.

Question 4

On test à nouveaux la même hypothèse $H_0 : \beta^*_{breastf} = 0$ mais cette fois la probabilité de survie est estimée à partir d'un modèle de Cox.

```
coxph.bf = coxph(Surv(chldage,hospital) ~ Z, data= pneumon)
summary(coxph.bf)
```

```
## Call:
## coxph(formula = Surv(chldage, hospital) ~ Z, data = pneumon)
##
##   n= 3470, number of events= 73
##
##                 coef exp(coef) se(coef)     z Pr(>|z|)
## ZAllaitement -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
## ZAllaitement    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
```

Tous les tests (wald, LRT, logrank) concluent à un effet significatif de la variable Z. Avec une division du risque par 3 (p-value Wald = $2 * 10^{-4}$ et p-value LRT = $5 * 10^{-5}$) pour les enfants allaités.

Cela signifie qu'il y a une association importante entre l'allaitement et le risque de pneumonie.

Il conviendrait de tester la validité des hypothèses du modèle de cox employé.

Question 5

Modèles de Cox à 2 facteurs, en prenant pour variable explicative la variable Z et successivement, une à une chacune des autres variables explicatives: mthage, urban, alcohol, smoke, region, poverty, bweight, race, education, nsibs, wmonth, sfmonth La variable esxplicative "agepn" est sortie.

"mthage" "urban" "alcohol" "smoke" "region" "poverty" "bweight"
"race" "education" "nsibs" "wmonth" "sfmonth" "agepn"

```
#### variable explicatives: Z + mthag
coxph.bf_mthage = coxph(Surv(chldage,hospital) ~ Z + mthage, data= pneumon)
summary(coxph.bf_mthage)
```

```
## Call:
## coxph(formula = Surv(chldage, hospital) ~ Z + mthage, data = pneumon)
##
##   n= 3470, number of events= 73
##
##                  coef exp(coef) se(coef)      z Pr(>|z|)
## ZAllaitement -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
## ZAllaitement    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
```

```
#### variable explicatives: Z + urban
coxph.bf_urban = coxph(Surv(chldage,hospital) ~ Z + urban, data= pneumon)
summary(coxph.bf_urban)
```

```
## Call:
## coxph(formula = Surv(chldage, hospital) ~ Z + urban, data = pneumon)
##
##   n= 3470, number of events= 73
##
##                 coef exp(coef) se(coef)     z Pr(>|z|)
## ZAllaitement -1.0720    0.3423   0.2978 -3.60 0.000319 ***
## urban        -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
## ZAllaitement    0.3423      2.921    0.1910    0.6137
## urban           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
```

```
#### variable explicatives: Z + alcohol
coxph.bf_alcohol = coxph(Surv(chldage,hospital) ~ Z + alcohol, data= pneumon)
summary(coxph.bf_alcohol)
```

```
## Call:
## coxph(formula = Surv(chldage, hospital) ~ Z + alcohol, data = pneumon)
##
##   n= 3470, number of events= 73
##
##                  coef exp(coef) se(coef)      z Pr(>|z|)
## ZAllaitement -1.09490   0.33457  0.29740 -3.682 0.000232 ***
## alcohol      -0.02865   0.97176  0.10892 -0.263 0.792560
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##              exp(coef) exp(-coef) lower .95 upper .95
## ZAllaitement    0.3346      2.989    0.1868    0.5993
## alcohol         0.9718      1.029    0.7850    1.2030
##
## Concordance= 0.613  (se = 0.026 )
## Likelihood ratio test= 16.66  on 2 df,   p=2e-04
## Wald test            = 13.69  on 2 df,   p=0.001
## Score (logrank) test = 15.11  on 2 df,   p=5e-04
```

```
#### variable explicatives: Z + alcohol
coxph.bf_smoke = coxph(Surv(chldage,hospital) ~ Z + smoke, data= pneumon)
summary(coxph.bf_smoke)
```

```
## Call:
## coxph(formula = Surv(chldage, hospital) ~ Z + smoke, data = pneumon)
##
##   n= 3470, number of events= 73
##
##                 coef exp(coef) se(coef)      z Pr(>|z|)
## ZAllaitement -1.0501    0.3499   0.2979 -3.525 0.000424 ***
## smoke         0.4224    1.5257   0.1510  2.797 0.005156 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##              exp(coef) exp(-coef) lower .95 upper .95
## ZAllaitement    0.3499     2.8579    0.1951    0.6274
## smoke           1.5257     0.6554    1.1348    2.0513
##
## Concordance= 0.665  (se = 0.031 )
## Likelihood ratio test= 23.83  on 2 df,   p=7e-06
## Wald test            = 21.62  on 2 df,   p=2e-05
## Score (logrank) test = 23.42  on 2 df,   p=8e-06
```

```
#### variable explicatives: Z + region
#variable explicatives: Z + region
coxph.bf_region = coxph(Surv(chldage,hospital) ~ Z + region, data= pneumon)
summary(coxph.bf_region)
```

```
## Call:
## coxph(formula = Surv(chldage, hospital) ~ Z + region, data = pneumon)
##
##   n= 3470, number of events= 73
##
##                 coef exp(coef) se(coef)      z Pr(>|z|)
## ZAllaitement -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
## ZAllaitement    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
```

```
#### variable explicatives: Z + poverty
coxph.bf_poverty = coxph(Surv(chldage,hospital) ~ Z + poverty, data= pneumon)
summary(coxph.bf_poverty)
```

```
## Call:
## coxph(formula = Surv(chldage, hospital) ~ Z + poverty, data = pneumon)
##
##   n= 3470, number of events= 73
##
##                 coef exp(coef) se(coef)      z Pr(>|z|)
## ZAllaitement -1.0919    0.3356   0.2977 -3.668 0.000245 ***
## poverty      -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
## ZAllaitement    0.3356      2.980    0.1872    0.6015
## poverty         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
```

```
#### variable explicatives: Z + bweightt
coxph.bf_bweight = coxph(Surv(chldage,hospital) ~ Z + bweight, data= pneumon)
summary(coxph.bf_bweight)
```

```
## Call:
## coxph(formula = Surv(chldage, hospital) ~ Z + bweight, data = pneumon)
##
##   n= 3470, number of events= 73
##
##                 coef exp(coef) se(coef)      z Pr(>|z|)
## ZAllaitement -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
## ZAllaitement    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
```

```
#### variable explicatives: Z + race
coxph.bf_race = coxph(Surv(chldage,hospital) ~ Z + race, data= pneumon)
summary(coxph.bf_race)
```

```
## Call:
## coxph(formula = Surv(chldage, hospital) ~ Z + race, data = pneumon)
##
##   n= 3470, number of events= 73
##
##                  coef exp(coef) se(coef)      z Pr(>|z|)
## ZAllaitement -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
## ZAllaitement    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
```

```
#### variable explicatives: Z + education
coxph.bf_education = coxph(Surv(chldage,hospital) ~ Z + education, data= pneumon)
summary(coxph.bf_education)
```

```
## Call:
## coxph(formula = Surv(chldage, hospital) ~ Z + education, data = pneumon)
##
##   n= 3470, number of events= 73
##
##                  coef exp(coef) se(coef)      z Pr(>|z|)
## ZAllaitement -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
## ZAllaitement    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
```

```
#### variable explicatives: Z + nsibs
coxph.bf_nsibs = coxph(Surv(chldage,hospital) ~ Z + nsibs, data= pneumon)
summary(coxph.bf_nsibs)
```

```
## Call:
## coxph(formula = Surv(chldage, hospital) ~ Z + nsibs, data = pneumon)
##
##   n= 3470, number of events= 73
##
##                 coef exp(coef) se(coef)      z Pr(>|z|)
## ZAllaitement -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
## ZAllaitement    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
```

```
#### variable explicatives: Z + wmonthh
coxph.bf_wmonth = coxph(Surv(chldage,hospital) ~ Z + wmonth, data= pneumon)
summary(coxph.bf_wmonth)
```

```
## Call:
## coxph(formula = Surv(chldage, hospital) ~ Z + wmonth, data = pneumon)
##
##   n= 3470, number of events= 73
##
##                 coef exp(coef) se(coef)      z Pr(>|z|)
## ZAllaitement -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
## ZAllaitement    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
```

```
#### variable explicatives: Z + sfmonth
coxph.bf_sfmonth = coxph(Surv(chldage,hospital) ~ Z + sfmonth, data= pneumon)
summary(coxph.bf_sfmonth)
```

```
## Call:
## coxph(formula = Surv(chldage, hospital) ~ Z + sfmonth, data = pneumon)
##
##   n= 3470, number of events= 73
##
##                 coef exp(coef) se(coef)      z Pr(>|z|)
## ZAllaitement -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
## ZAllaitement    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
```

Pour chacun de ces modèles, l'indice de concordance est plutôt bon, varie entre ($0.613 $) et ($0.67$).

L'hypothèse $H_0 : ^_1 = ^_2 $ est rejetée par le LRT (p-value<$3e-04$) pour tous les modèles. Dans le cas du test de Wald on retrouve aussi une petite p-value.

La p-value = 0.001 est la p-value la plus grande. Elle est rencontrée pour les modèles incluant les variables Z+alcohool / Z+powerty / Z+wmounth

On reamrque que les p-values des tests de Wald univariés associées au coefficient des variables nsibs (nombre d'enfant dans la famille) et education dans les 2 modèles correspondants sont inférieures à 5% et 1% respectivement.
Dans tous les cas la différence entre les 2 groupes est significative. Le fait d'allaiter contribue de manière significative à la survie (ne pas tombre malade et se retouver hospitalisé)

Question 6

On fait un modèle de Cox multivarié en incluant toutes les variables. on sort la variable wmonth qui est utilisée pour construire la variable "dummy" que l'on a ajouté, et donc déjà présente.

```
fit.complet=coxph(Surv(chldage,hospital) ~ . -wmonth ,data=pneumon)
summary(fit.complet)
```

```
## Call:
## coxph(formula = Surv(chldage, hospital) ~ . - wmonth, data = pneumon)
##
##   n= 3470, number of events= 73
##
##                    coef exp(coef)  se(coef)      z Pr(>|z|)
## mthage       -0.104484  0.900789  0.057707 -1.811   0.0702 .
## urban        -0.335452  0.715015  0.267167 -1.256   0.2093
## alcohol      -0.078737  0.924283  0.112848 -0.698   0.4853
## smoke         0.322854  1.381064  0.168541  1.916   0.0554 .
## region2       0.056449  1.058073  0.349724  0.161   0.8718
## region3      -0.429256  0.650994  0.354085 -1.212   0.2254
## region4      -0.499080  0.607089  0.446205 -1.118   0.2634
## poverty      -0.028873  0.971540  0.400869 -0.072   0.9426
## bweight       0.193981  1.214074  0.260837  0.744   0.4571
## race2        -0.377878  0.685314  0.319462 -1.183   0.2369
## race3         0.200445  1.221947  0.366819  0.546   0.5848
## education    -0.033302  0.967246  0.073865 -0.451   0.6521
## nsibs         0.338471  1.402801  0.138457  2.445   0.0145 *
## sfmonth      -0.179804  0.835434  0.162149 -1.109   0.2675
## agepn         0.004359  1.004369  0.028748  0.152   0.8795
## ZAllaitement -0.488073  0.613808  0.437598 -1.115   0.2647
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##              exp(coef) exp(-coef) lower .95 upper .95
## mthage          0.9008     1.1101    0.8045     1.009
## urban           0.7150     1.3986    0.4235     1.207
## alcohol         0.9243     1.0819    0.7409     1.153
## smoke           1.3811     0.7241    0.9925     1.922
## region2         1.0581     0.9451    0.5331     2.100
## region3         0.6510     1.5361    0.3252     1.303
## region4         0.6071     1.6472    0.2532     1.456
## poverty         0.9715     1.0293    0.4428     2.131
## bweight         1.2141     0.8237    0.7281     2.024
## race2           0.6853     1.4592    0.3664     1.282
## race3           1.2219     0.8184    0.5954     2.508
## education       0.9672     1.0339    0.8369     1.118
## nsibs           1.4028     0.7129    1.0694     1.840
## sfmonth         0.8354     1.1970    0.6080     1.148
## agepn           1.0044     0.9957    0.9493     1.063
## ZAllaitement    0.6138     1.6292    0.2603     1.447
##
## Concordance= 0.726  (se = 0.028 )
## Likelihood ratio test= 46.16  on 16 df,   p=9e-05
## Wald test            = 41.84  on 16 df,   p=4e-04
## Score (logrank) test = 45.38  on 16 df,   p=1e-04
```

On reamrque que les p-values des tests de Wald univariés associées aux coefficients des variables nsibs (nombre d'enfant dans la famille) et smoke sont inférieures à 5%. Et inférieur à 10% pour region.

On peut essayer de faire un choix de variables de manière automatique en utilisant la méthode stepAIC. Procédure backward basée sur les tests de Wald :

```
library(MASS)
cox.complet = coxph(Surv(chldage,hospital) ~ . -wmonth ,data=pneumon)
modele.final<-stepAIC(cox.complet,trace = F,direction = "both")
summary(modele.final)
```

```
## Call:
## coxph(formula = Surv(chldage, hospital) ~ mthage + smoke + nsibs +
##     sfmonth, data = pneumon)
##
##   n= 3470, number of events= 73
##
##             coef exp(coef) se(coef)      z Pr(>|z|)
## mthage  -0.12386   0.88351  0.04965 -2.495   0.0126 *
## smoke    0.38743   1.47318  0.15091  2.567   0.0103 *
## nsibs    0.38437   1.46869  0.12182  3.155   0.0016 **
## sfmonth -0.32331   0.72375  0.12615 -2.563   0.0104 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##         exp(coef) exp(-coef) lower .95 upper .95
## mthage     0.8835     1.1319    0.8016    0.9738
## smoke      1.4732     0.6788    1.0960    1.9802
## nsibs      1.4687     0.6809    1.1567    1.8648
## sfmonth    0.7238     1.3817    0.5652    0.9268
##
## Concordance= 0.701  (se = 0.028 )
## Likelihood ratio test= 35.16  on 4 df,   p=4e-07
## Wald test            = 29.71  on 4 df,   p=6e-06
## Score (logrank) test = 31.91  on 4 df,   p=2e-06
```

On obtient un modèle à 5 variables explicatives: mthage: Age de la mère smoke: 1 si la mère à fumée pendant la grossesse. region: 1=Nord-est, 2=Nord, 3=sud, 4=Ouest nsibs: Nombre d'enfants dans la famille sfmonth: Age à partir duquel l'enfant prends de la nouriture "solide".

Question 7

A partir du modèle obtenu à la question précédente, on va prédire la probabilité qu'un nouveau né a de contracter la pneumonie à 6 mois.
Les nouvelles variables en entrée sont les suivantes : mthage = 27, urban=1, alcohol=3, smoke=0, region=2, poverty=1, bweight=0, race=1, education=12, nsibs=1, wmonth=0, sfmonth=0, agepn=4

Pour cet enfant, la mère n'a pas fumée pendant la grossesse , ils habitent dans le nord, la mère est âgée de 27ans il n'y a qu'un seul autre enfant dans la famille, que celui-ci ne prend pas encore de la nouriture "solide" et qu'il n'a été allaité dès la naissance.

Pour les autre variables qui n'entrent pas dans le modèle, on sait qu'ils sont citadin, d'un milieu pauvre, la mère a été peu scolarisée, l'enfant à la naissance avait un poids inférieur à la normale.

```
cox.Result = coxph(Surv(chldage, hospital) ~ mthage + smoke + region + nsibs + sfmonth, data = pneumon)
glimpse(pneumon)
```

```
## Observations: 3,470
## Variables: 16
## $ chldage   <dbl> 12.0, 12.0, 3.0, 2.0, 4.0, 12.0, 7.0, 3.0, 7.0, 12.0...
## $ hospital  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ mthage    <int> 22, 20, 24, 22, 21, 20, 24, 24, 26, 21, 24, 27, 20, ...
## $ urban     <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ alcohol   <int> 0, 1, 3, 2, 1, 0, 0, 3, 2, 1, 0, 0, 0, 2, 0, 4, 0, 2...
## $ smoke     <int> 0, 0, 0, 2, 2, 0, 0, 0, 2, 0, 0, 0, 0, 0, 1, 1, 1, 0...
## $ region    <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ poverty   <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ bweight   <int> 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0...
## $ race      <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ education <int> 10, 12, 12, 9, 12, 12, 12, 14, 12, 12, 16, 16, 12, 1...
## $ nsibs     <int> 1, 1, 2, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1...
## $ wmonth    <int> 1, 2, 1, 0, 0, 0, 0, 4, 1, 3, 0, 0, 0, 4, 0, 1, 9, 1...
## $ sfmonth   <int> 1, 2, 0, 0, 0, 0, 0, 2, 1, 2, 0, 0, 0, 2, 0, 1, 2, 3...
## $ agepn     <int> 1, 12, 3, 2, 4, 12, 7, 3, 6, 12, 12, 12, 3, 4, 2, 4,...
## $ Z         <fct> Allaitement, Allaitement, Allaitement, Pas d'allaite...
```

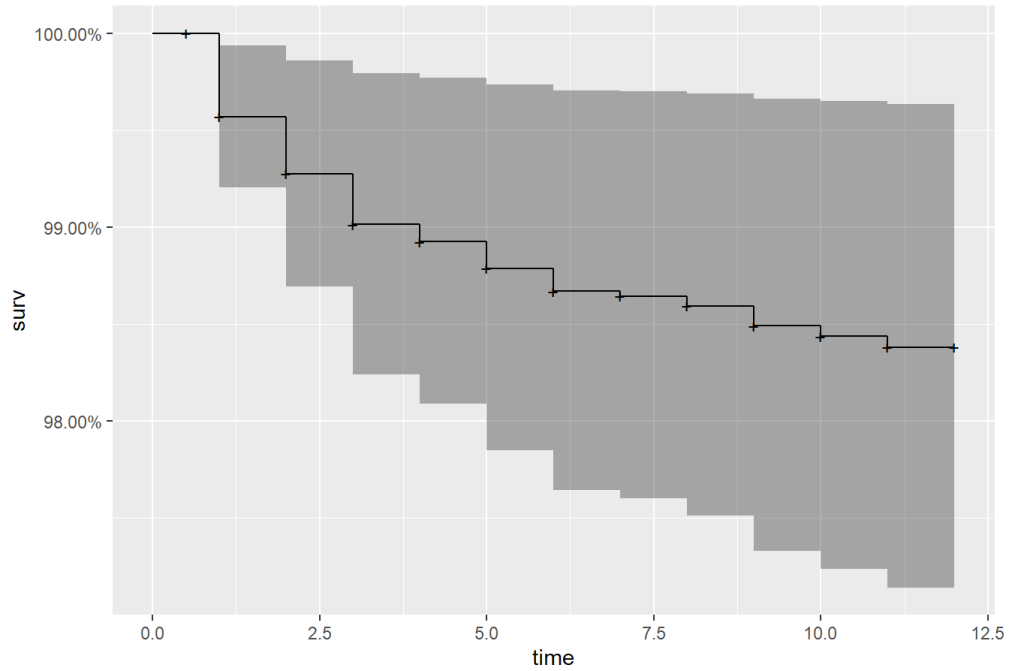
```
DataPredict = data.frame( mthage = 27 ,smoke =  0,region=2, nsibs = 1, sfmonth = 0)
DataPredict = mutate(DataPredict , region = factor(region,levels = c(1,2,3,4)))
predict(cox.Result , newdata = DataPredict, type ="risk")
```

```
##         1
## 0.9551334
```

```
kmsurvival.final<-survfit(cox.Result , newdata = DataPredict)
library(ggfortify)

autoplot(kmsurvival.final,main="Fig 3 - Courbes estiamteur KM - proba-de survie pour l'individu")
```

Fig 3 - Courbes estiamteur KM - proba-de survie pour l'individu



Notre individu dont la probabilité de survie (ne pas contracter la pneumonie) simulé à partir de notre modèle (courbe de survie fig-3) se trouve bien dans l'Intervalle de confiance du groupe des enfants allaités (cf. Question 3 - fig-2 courbe verte) mais plutôt dans la "fourchette" basse à la limite de la borne inférieur de l'IC de niveau de confiance 95% (valeur par défaut de survfit paramètre conf.int=.95).

```
kmsurvival.final$time
```

```
##  [1]  0.5  1.0  2.0  3.0  4.0  5.0  6.0  7.0  8.0  9.0 10.0 11.0 12.0
```

```
kmsurvival.final$surv
```

```
##  [1] 1.0000000 0.9957002 0.9927551 0.9901496 0.9892586 0.9878825 0.9866959
##  [8] 0.9864519 0.9859484 0.9849098 0.9843723 0.9838211 0.9838211
```

La probabilité correspond au 7ème enregistrement de la sortie soit P-Survie = 0.9866959
D'après le modèle le nouveau né à quasiment 99% de chance de ne pas développer une pneumonie.

```
1-kmsurvival.final$surv
```

```
##  [1] 0.000000000 0.004299770 0.007244904 0.009850443 0.010741355
##  [6] 0.012117516 0.013304100 0.013548070 0.014051556 0.015090171
## [11] 0.015627711 0.016178869 0.016178869
```

C'est à dire, un peu plus de 1.3% (0.013304100) de chance de contracter la pneumonie.

Notre individu est à la limite de la borne inférieure de l'intervalle de confiance à 95% obtenu. Ceci pouvant peut-être s'expliquer par certains critères qui n'ont pas été pris en compte dans notre modèle et qui, on a pu le remarquer semblaient être assez significatifs. Comme l'education (ici un niveau faible) et l'appartenance à un milieu pauvre en zone urbaine dans le nord des Etats-Unis.