Exercise 8.11 of Klein and Moeschberger 2005

Devoir maison obligatoire (Dauphine)

Paul Hardouin

December 11, 2019

Contents

We study data gathered from annual personal interviews conducted for the National Longitudinal Survey of Youth (NLSY) from 1979 through 1986. This data was used to study whether or not the mother's feeding choice protected the infant against hospitalized pneumonia in the first year of life. Ages of young children at the time they were hospitalized with pneumonia were recorded as well as the observed ages of those infants that were not hospitalized with pneumonia during the study period. The data pneumon is available in R package KMsurv. Use the discrete method for handling ties in the following.

1. Load data

==

Check if the variables have been correcly imported, especially the factors.

==

We load librairies.

```r
library(survival)
library(fitdistrplus)
library(tidyverse)
library(KMsurv)
library(ggfortify)
```

We load data.

```r
data("pneumon")
```

We suppress useless attribute.

```r
pneumon = pneumon[,!colnames(pneumon) %in% c("agepn")]
```

We encode categorial data as factor.

```r
pneumon$urban = as.factor(pneumon$urban)
pneumon$alcohol = as.factor(pneumon$alcohol)
pneumon$smoke = as.factor(pneumon$smoke)
pneumon$region = as.factor(pneumon$region)
pneumon$poverty = as.factor(pneumon$poverty)
pneumon$bweight = as.factor(pneumon$bweight)
pneumon$race = as.factor(pneumon$race)
```

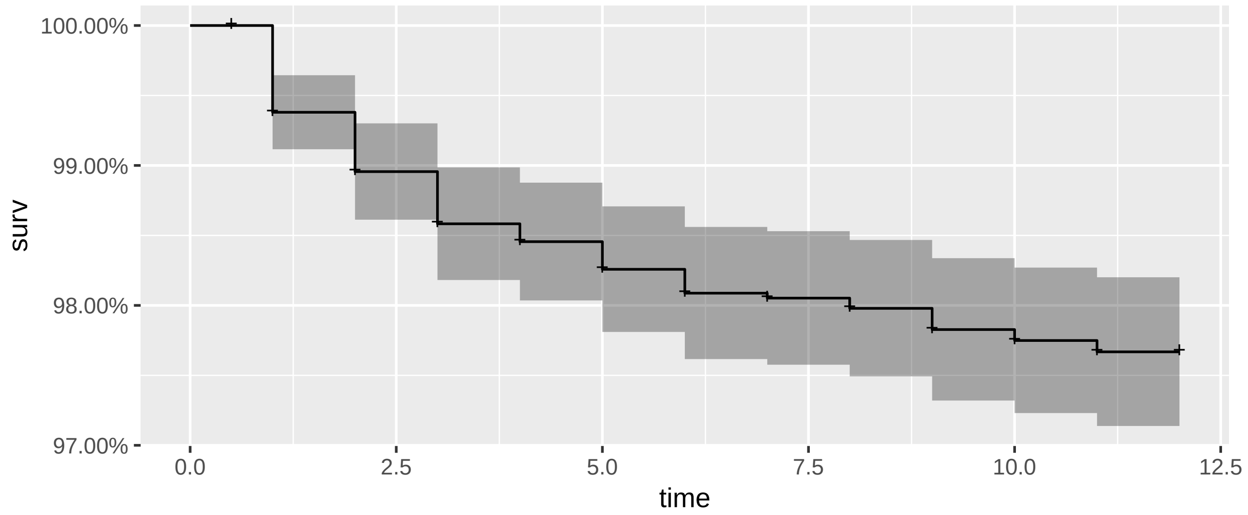
2. Kaplan-Meier estimator

===

Plot the Kaplan-Meier estimator for the survival function of the age at pneumonia. Give an estimation and a confidence interval for the probability for a newborn of not having developed pneumonia at 6 months.

===

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



```
summary(KMfit)
```

```
## 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
```

At 6 months, the probability for a newborn of not having developed pneumonia is **98.1%** with the confidence interval [**97.6% - 98.6%**].

3. Breast fedding / not fedding distributions

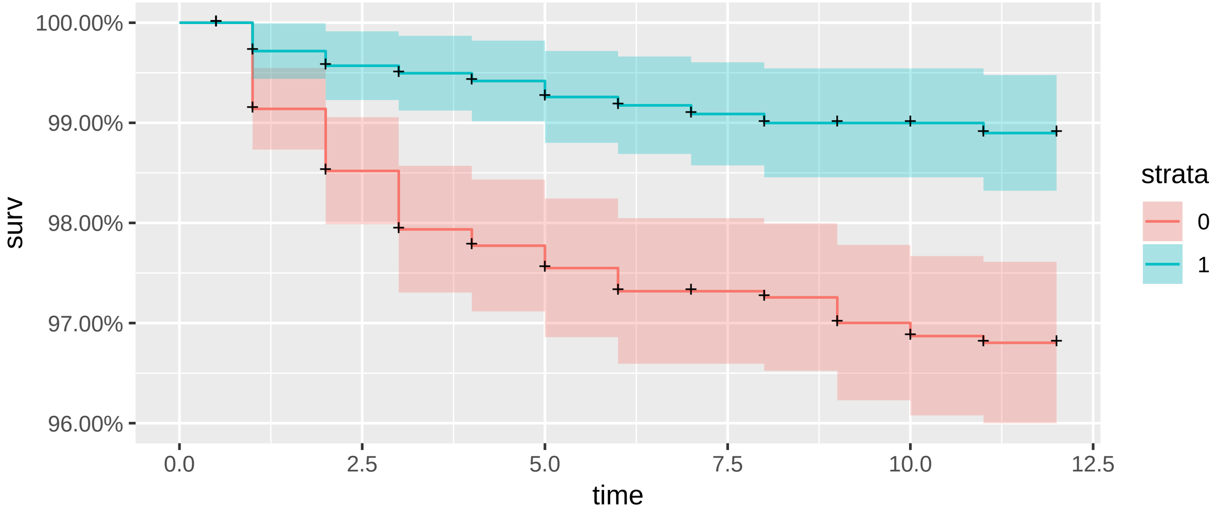
==

Construct a dummy variable Z = 1 if infants were breast fed at birth, 0 if infants were never breast fed, and test the hypothesis H_0: there is not difference in distributions of age at first pneumonia between child whether were breast fed or not.

==

We build a dummy variable Z to describe the breast fedding or not.

```
pneumon$Z = as.integer(pneumon$wmonth>0)
autoplot(survfit(Surv(chldage,hospital)~Z, data = pneumon))
```



On the previous figure, we can oberve 2 distinct curves. It makes us assume that the distributions will be different between the 2 groups, and that the hypothesis H_0 will be rejected. Let's make a log-rank test to validate this observation.

```
# LOG-RANK test
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=0 2036       59     42.7      6.22        15
## Z=1 1434       14     30.3      8.77        15
##
##  Chisq= 15  on 1 degrees of freedom, p= 1e-04
```

We got a very low p-value = 1e-4, and the hypothesis H_0 is effectively rejected.

4. Breast fedding / not fedding hazard ratio

==
Test the hypothesis H_0: $\beta^*_{breast} = 0$, i.e., the survival functions for the two types of breast feeding are equal, using the likehood ratio, and the Wald tests. Find the estimate of β^*_{breast}, $\hat{\beta}^*_{breast}$, its standard error, and its relative risk.
==

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

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

The LRT test p-value = 5e-5, so the hypothesis H_0: $\beta^*_{breast} = 0$ is rejected according this test.
The wald test p-value = 2e-4, so the hypothesis H_0: $\beta^*_{breast} = 0$ is rejected according this test.

The estimate of β^*_{breast} is $\hat{\beta}^*_{breast} = -1.0970 = log(0.3339)$. The relative risk of modality **Breast fedding** is multiplied by 0.3339 compared to the relative risk of modality **Breast not fedding**. The standard error value is 0.023.

5. Bivariate Cox Model

==

Also available in the data set is information on other factors that may be associated with the timing of hospitalized pneumonia. These factors are age of the mother at the infant's birth, rural-urban environment of the mother, use of alcohol by the mother (no drinks, less than one drink, 1-2 drinks, 3-4 drinks, or more than 4 drinks per month), mother's cigarette use (none, less than 1 pack/day, 1 or more pack/day), region of country (northeast, north central, south, or west), birthweight of infant (less the 5.5 lbs or 5.5 lbs or more), poverty status of mother (yes/no), race of mother (white, black, or other), or number of siblings of infant. Test the hypothesis that the times to hospitalized pneumonia are the same for the two feeding groups adjusting for each of these factors in a separate model using the Wald test.

==

5.1. Breast Feeding + age of the mother at birth

```r
summary(coxph(Surv(chldage,hospital)~Z+mthage, data = pneumon))
```

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

About **breast feeding (Z)**, we have a low p-value = 6e-4, so $\beta^*_{breast} = 0$ hypothesis is rejected. As $exp(coef_{breast}) = 0.3583 \neq 1$, it involves the the times to hospitalized pneumonia are different for the two feeding groups.

About the **age of the mother**, we got a high p-value = 0.133, so this factor doesn't impact significantly the time to hospitalized pneumonia.

5.2. Breast Feeding + urban environnement of mother

```r
summary(coxph(Surv(chldage,hospital)~Z+urban, data = pneumon))
```

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

About **breast feeding (Z)**, we have a low p-value = 3e-4, so $\beta^*_{breast} = 0$ hypothesis is rejected. As $exp(coef_{breast}) = 0.3423 \neq 1$, it involves the the times to hospitalized pneumonia are different for the two feeding groups.

About the **rural-urban environment of the mother**, we got a high p-value = 0.125, so this factor doesn't impact significantly the time to hospitalized pneumonia.

5.3. Breast Feeding + alcohol consumption of mother

```
summary(coxph(Surv(chldage,hospital)~Z+alcohol, data = pneumon))
```

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

About **breast feeding (Z)**, we have a low p-value = 2e-4, so $\beta^*_{breast} = 0$ hypothesis is rejected. As $exp(coef_{breast}) = 0.3293 \neq 1$, it involves the the times to hospitalized pneumonia are different for the two feeding groups.

About the **alcohol consumption of mother**, we got high p-values = [0.495 0.711 0.717 0.945], so this factor doesn't impact significantly the time to hospitalized pneumonia.

5.4. Breast Feeding + cigarette consumption of mother

```r
summary(coxph(Surv(chldage,hospital)~Z+smoke, data = pneumon))
```

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

About **breast feeding (Z)**, we have a low p-value = 4e-4, so $\beta^*_{breast} = 0$ hypothesis is rejected. As $exp(coef_{breast}) = 0.3494 \neq 1$, it involves the the times to hospitalized pneumonia are different for the two feeding groups.

About the **cigarette consumption of mother**, we got high p-values = [0.002 0.049]. As intuitively expected, the hazard rate is increased by the smoking mother modalities.

- the risk of modality **less than one pack/day** is multiplied by 2.14 compared to the risk of modality **not smoking**
- the risk of modality **more than one pack/day** is multiplied by 1.97 compared to the risk of modality **not smoking**

Paradoxically, risk of **more than one pack/day** is multiplied by $\frac{1.97}{2.17} = 0.92 < 1$ compared to risk of **less than one pack/day**. But 0.92 is very close to 1, so this last value is not so significant.

5.5. Breast Feeding + region

```
summary(coxph(Surv(chldage,hospital)~Z+region, data = pneumon))
```

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

About **breast feeding (Z)**, we have a low p-value = 3e-4, so $\beta^*_{breast} = 0$ hypothesis is rejected. As $exp(coef_{breast}) = 0.3350 \neq 1$, it involves the the times to hospitalized pneumonia are different for the two feeding groups.

About the **region**, we got high p-values = [0.629 0.257 0.313], so this factor doesn't impact significantly the time to hospitalized pneumonia.

5.6. Breast Feeding + poverty status of mother

```
summary(coxph(Surv(chldage,hospital)~Z+poverty, data = pneumon))
```

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

About **breast feeding (Z)**, we have a low p-value = 2e-4, so $\beta^*_{breast} = 0$ hypothesis is rejected. As $exp(coef_{breast}) = 0.3356 \neq 1$, it involves the the times to hospitalized pneumonia are different for the two feeding groups.

About the **poverty status of mother**, we got a high p-value = 0.738, so this factor doesn't impact significantly the time to hospitalized pneumonia.

5.7. Breast Feeding + birth weight

```
summary(coxph(Surv(chldage,hospital)~Z+bweight, data = pneumon))
```

```
## Call:
## coxph(formula = Surv(chldage, hospital) ~ Z + bweight, data = pneumon)
##
##   n= 3470, number of events= 73
##
##             coef exp(coef) se(coef)      z Pr(>|z|)
## Z        -1.0087    0.3647   0.3018 -3.342  0.00083 ***
## bweight1  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
## Z           0.3647     2.7420    0.2019    0.6589
## bweight1    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
```

About **breast feeding (Z)**, we have a low p-value = 8e-4, so $\beta^*_{breast} = 0$ hypothesis is rejected. As $exp(coef_{breast}) = 0.3647 \neq 1$, it involves the the times to hospitalized pneumonia are different for the two feeding groups.

About the **birth weight**, we got a quite low p-value = 0.076. The hazard rate is increased by the fact to have a weight higher than 5.5 lbs at birth. The risk of modality **more than 5.5 lbs at birth** is multiplied by 1.52 compared to the risk of modality **less than 5.5 lbs at birth**. This result is not intuitive, because a high weight at birth is normally a sign of good health.

- maybe the definition of TRUE value is inverted ?
- maybe this factor is not so signifiant : p-value is higher than 5% ?

5.8. Breast Feeding + race of mother

```
summary(coxph(Surv(chldage,hospital)~Z+race, data = pneumon))
```

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

About **breast feeding (Z)**, we have a low p-value = 6e-5, so $\beta^*_{breast} = 0$ hypothesis is rejected. As $exp(coef_{breast}) = 2993 \neq 1$, it involves the the times to hospitalized pneumonia are different for the two feeding groups.

About the **race of mother**, we got high p-values = [0.102 0.773], so this factor doesn't impact significantly the time to hospitalized pneumonia.

5.9. Breast Feeding + number of siblings

```
summary(coxph(Surv(chldage,hospital)~Z+nsibs, data = pneumon))
```

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

About **breast feeding (Z)**, we have a low p-value = 4e-4, so $\beta^*_{breast} = 0$ hypothesis is rejected. As $exp(coef_{breast}) = 0.3516 \neq 1$, it involves the the times to hospitalized pneumonia are different for the two feeding groups.

About the **number of siblings**, we got a low p-values = 0.014, so this factor impacts significantly the time to hospitalized pneumonia. The hazard rate increases when the number of sibling increase. Having a supplementary child multiplies the relative risk by 1.3212. It was intuitively expectable :

- when parents have more children to take care of, they have less time by child to take care of them.
- when a child have siblings, they are a supplementary vector of disease contamination.

5.10. Conclusions

Looking the last bivariate cox models, we can select a list of pertinents factors that impact the times to hospitalized pneumonia are different for the two feeding groups.

- Z
- smoke
- (birth weight) : to be confirmed/unconfirmed
- number of siblings

6. Multivariate Cox Model

==
Since one is primarily interested in comparing the two types of breast feeding, interest will center upon building a model with the view of testing the particular comparison of interest adjusting for the other non controllable fixed covariates in question 4. Build such a model using the Wald test.
==

6.1. Cox Model

We build a cox model with all the factors. The p-values of the Wald tests are lower than 10% for **Z**, **mthage**, **smoke** and **nsibs**. Finally, the doubts about influence of **birth weight** factor (question 5) are confirmed and we disqualify it. However, **mthage** is finally added.

```
fit_total = coxph(Surv(chldage,hospital)~ Z + mthage + urban + alcohol + smoke + region + poverty + bwe
summary(fit_total)
```

```
## Call:
## coxph(formula = Surv(chldage, hospital) ~ Z + mthage + urban +
##     alcohol + smoke + region + poverty + bweight + race + education +
##     nsibs, data = pneumon)
##
##   n= 3470, number of events= 73
##
##               coef exp(coef) se(coef)      z Pr(>|z|)
## Z         -0.89048   0.41046  0.31334 -2.842  0.00448 **
## mthage    -0.10715   0.89839  0.05693 -1.882  0.05982 .
## urban1    -0.33203   0.71746  0.26764 -1.241  0.21477
## alcohol1   0.14368   1.15451  0.31790  0.452  0.65130
## alcohol2  -0.33574   0.71481  0.47982 -0.700  0.48410
## alcohol3  -0.35642   0.70018  0.60397 -0.590  0.55511
## alcohol4  -0.28827   0.74956  0.60482 -0.477  0.63363
## smoke1     0.70437   2.02257  0.26850  2.623  0.00871 **
## smoke2     0.48116   1.61795  0.38015  1.266  0.20561
## region2    0.06947   1.07194  0.34977  0.199  0.84256
## region3   -0.38319   0.68168  0.35516 -1.079  0.28062
## region4   -0.51868   0.59531  0.44667 -1.161  0.24556
## poverty1  -0.01816   0.98200  0.40104 -0.045  0.96388
## bweight1   0.18697   1.20559  0.26022  0.719  0.47244
## race2     -0.39048   0.67673  0.32122 -1.216  0.22413
## race3      0.22603   1.25361  0.36638  0.617  0.53729
## education -0.04469   0.95629  0.07451 -0.600  0.54868
## nsibs      0.34800   1.41623  0.13947  2.495  0.01259 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## Z            0.4105     2.4363    0.2221    0.7586
## mthage       0.8984     1.1131    0.8035    1.0044
## urban1       0.7175     1.3938    0.4246    1.2123
## alcohol1     1.1545     0.8662    0.6192    2.1528
## alcohol2     0.7148     1.3990    0.2791    1.8307
## alcohol3     0.7002     1.4282    0.2143    2.2872
## alcohol4     0.7496     1.3341    0.2291    2.4526
## smoke1       2.0226     0.4944    1.1950    3.4234
## smoke2       1.6180     0.6181    0.7680    3.4084
## region2      1.0719     0.9329    0.5401    2.1276
## region3      0.6817     1.4670    0.3398    1.3674
## region4      0.5953     1.6798    0.2480    1.4287
## poverty1     0.9820     1.0183    0.4475    2.1552
## bweight1     1.2056     0.8295    0.7239    2.0077
## race2        0.6767     1.4777    0.3606    1.2701
## race3        1.2536     0.7977    0.6114    2.5706
## education    0.9563     1.0457    0.8264    1.1067
## nsibs        1.4162     0.7061    1.0775    1.8615
##
## Concordance= 0.73  (se = 0.029 )
## Likelihood ratio test= 48.65  on 18 df,   p=1e-04
## Wald test            = 45.51  on 18 df,   p=3e-04
## Score (logrank) test = 48.36  on 18 df,   p=1e-04
```

6.2. StepAIC

We decide to run a model selection with AIC research.
This selection confirmed conclusion of question 6.

```
stepAIC(fit_total,trace = F)
```

```
## Call:
## coxph(formula = Surv(chldage, hospital) ~ Z + mthage + smoke +
##     nsibs, data = pneumon)
##
##             coef exp(coef) se(coef)      z       p
## Z      -0.88129   0.41425  0.30241 -2.914 0.00357
## 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
##
## Likelihood ratio test=37.43  on 5 df, p=4.9e-07
## n= 3470, number of events= 73
```

7. Prediction

==

In the final model, predict the probability of not having developed pneumonia at 6 months for a newborn whith covariates.

Factors	Values
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

==

We build final model with variables selection.

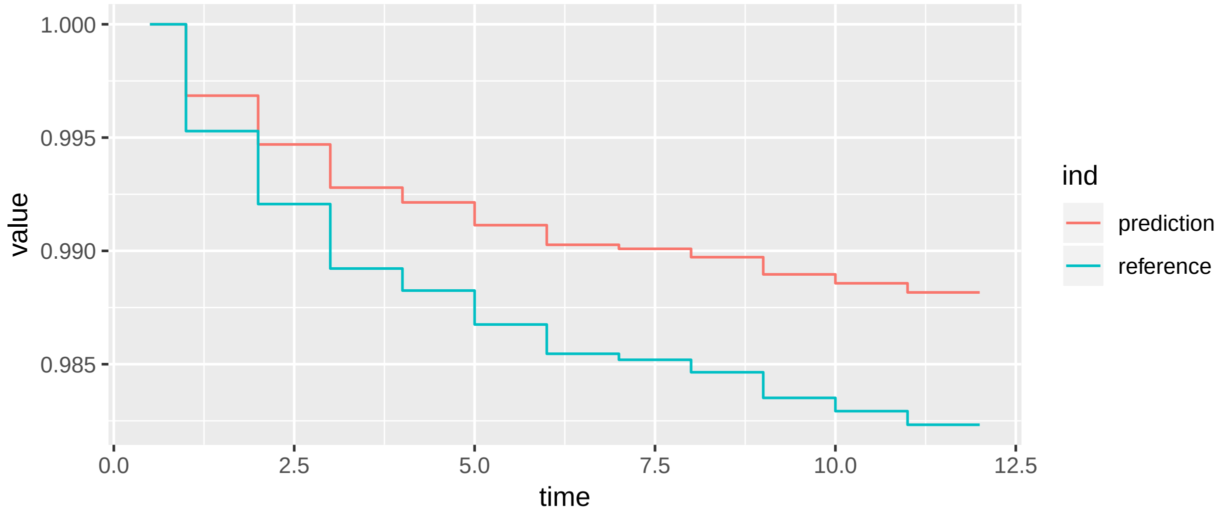
```
fit_final = coxph(Surv(chldage,hospital)~ Z + mthage + smoke + nsibs, data = pneumon)
```

We create a data frame with given covariates.

```
newDF = data.frame(Z=0, mthage=27, urban=1, alcohol=3, smoke=0, region=2, poverty=1, bweight=0, race=1,
newDF$urban = as.factor(newDF$urban)
newDF$alcohol = as.factor(newDF$alcohol)
newDF$smoke = as.factor(newDF$smoke)
newDF$region = as.factor(newDF$region)
newDF$poverty = as.factor(newDF$poverty)
newDF$bweight = as.factor(newDF$bweight)
newDF$race = as.factor(newDF$race)
```

We display the prediction

```
prediction_model = survfit(fit_final)
marqueurs = predict(fit_final,newDF)
time = prediction_model$time
reference = prediction_model$surv
prediction =  exp(-prediction_model$cumhaz*exp(marqueurs))
pred = tibble(time,reference, prediction) %>% gather("ind","value",2:3)
ggplot(pred,aes(x=time,y=value,color=ind)) + geom_step()
```



```
knitr::kable(data.frame(Time=time,Prediction=prediction))
```

Time	Prediction
0.5	1.0000000
1.0	0.9968508
2.0	0.9946966
3.0	0.9927914
4.0	0.9921398
5.0	0.9911359
6.0	0.9902698
7.0	0.9900912
8.0	0.9897226
9.0	0.9889630
10.0	0.9885704
11.0	0.9881671
12.0	0.9881671

With our final model, for the newborn we specified, **the probability of not having developed pneumonia at 6 months is 0.9902698**.