Stroke data

Cyril Matthey-Doret

26 décembre 2016

Introduction

In this report, I assess the effect of covariables on the survival of stroke victims.

Data

I used the 'stroke' dataset, available in the ISwR package.

```r
library(ISwR)
data(stroke)
```

This dataset contains 10 variables observed on 829 individuals who were followed for up to 5 years after a stroke.

The variables in the dataset are:

- sex: Indicates the gender of the subject. 2-levels factor: `Male`, `Female`.
- died: Date at which the subject died. Date in format YYYY-MM-DD.
- dstr: Date of the stroke. Date in format YYYY-MM-DD.
- age: Age at stroke in years. Discrete numeric values.
- dgn: Diagnosis of the patient. 4-levels factor: `INF` (infarction, ischaemic), `ICH` (intracranial haemorrhage), `ID` (unidentified), `SAH` (subarchnoid haemorrhage)
- coma: Indicates whether patient was in a coma after the stroke. 2-levels factor: `Yes`, `No`
- diab: History of diabetes. 2-levels factor: `Yes`, `No`
- minf: History of myocardial infarction. 2-levels factor: `Yes`, `No`
- han: History of hypertension. 2-levels factor: `Yes`, `No`
- obsmonths: Observation times in months. Set to 0.1 for patients dying on the same day as the stroke. Continuous numeric values.
- dead: Whether or not the patient died during the study. Boolean value.

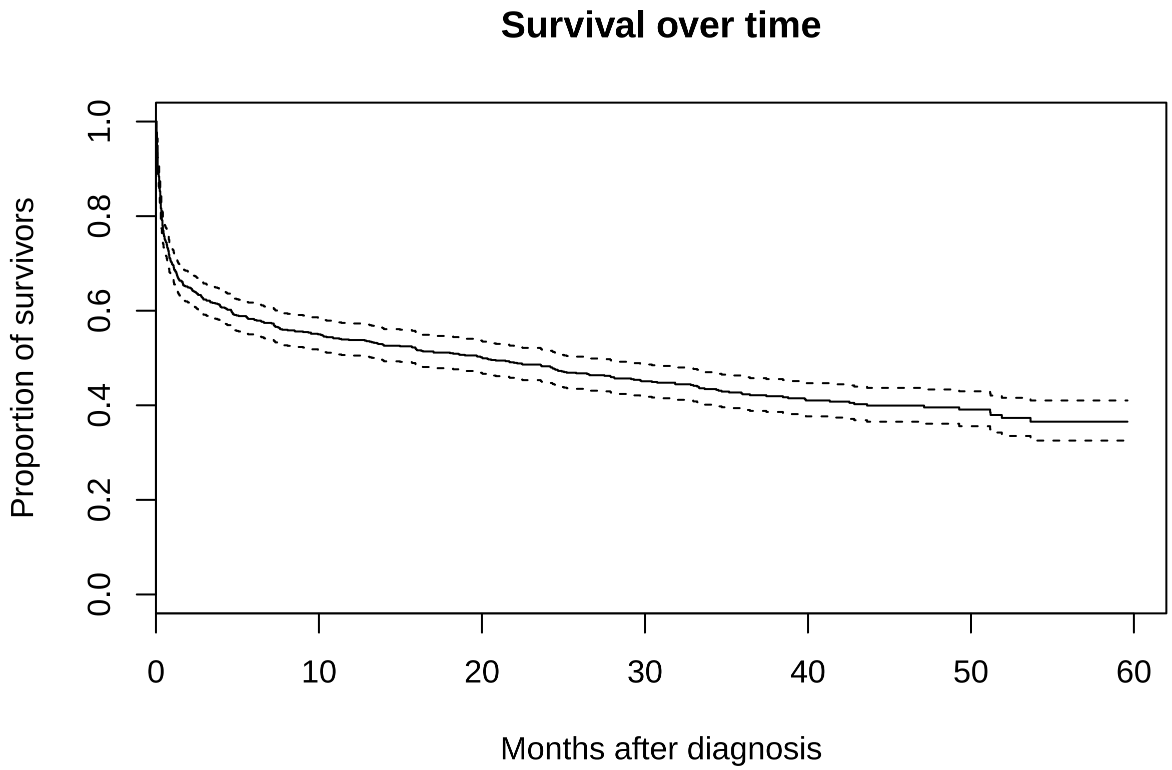
I am especially interested in the difference observed in the results when assessing the effect of `sex` alone, or in conjunction with `age`.

Analyses

I use tools built in the `survival` package to analyze this dataset.

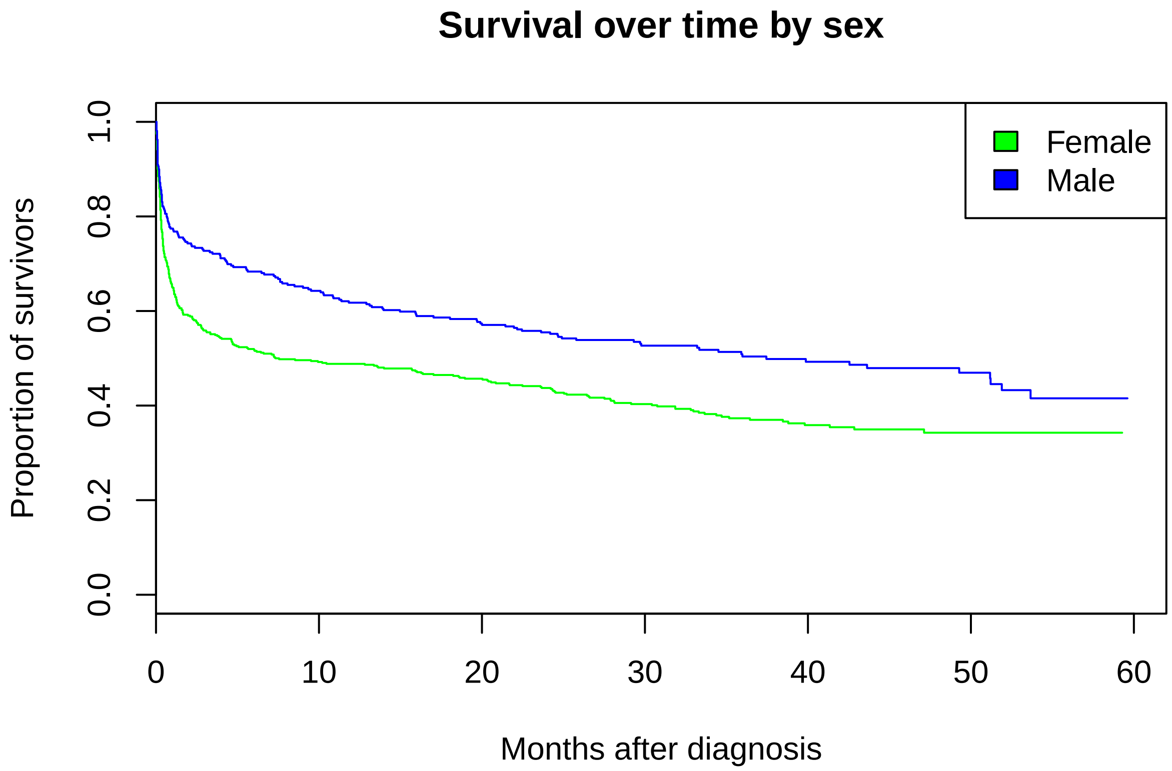
```r
library(survival)
stroke.surv.fit <- Surv(stroke$obsmonths, stroke$dead=="TRUE")
stroke.surv <- survfit(stroke.surv.fit~1)
```

Surv generates a survival object and survfit estimates the survival function using Kaplan-Meier estimate. When plotting the model, we can visualize the proportion of survivors decreasing over time.



Splitting the population by sex allows to visualize the proportion of survival for each sex independently, revealing that males seem to survive strokes better than women.

```
stroke.surv.bysex <- survfit(stroke.surv.fit~stroke$sex)
```



We can test if this difference between genders is statistically significant using a log-rank test:

```
stroke.surv.diff <- survdiff(stroke.surv.fit~stroke$sex)
stroke.surv.diff
```

```
## Call:
## survdiff(formula = stroke.surv.fit ~ stroke$sex)
##
##                        N Observed Expected (O-E)^2/E (O-E)^2/V
## stroke$sex=Female 510      321      280      6.11      14.6
## stroke$sex=Male   319      164      205      8.32      14.6
##
##  Chisq= 14.6  on 1 degrees of freedom, p= 0.000132
```

According to the log-rank test, males survive significantly better to strokes than females (p<0.001) in this study.

We can also add the `age` variable to see if age has an effect on survival, and to see if genders still survive differently after correcting for age. In order to assess the influence of a numerical variable (age), we need to use a Cox model of proportional hazards.

```
stroke.coxph.sex.age <- coxph(stroke.surv.fit~stroke$sex+stroke$age)
summary(stroke.coxph.sex.age)
```

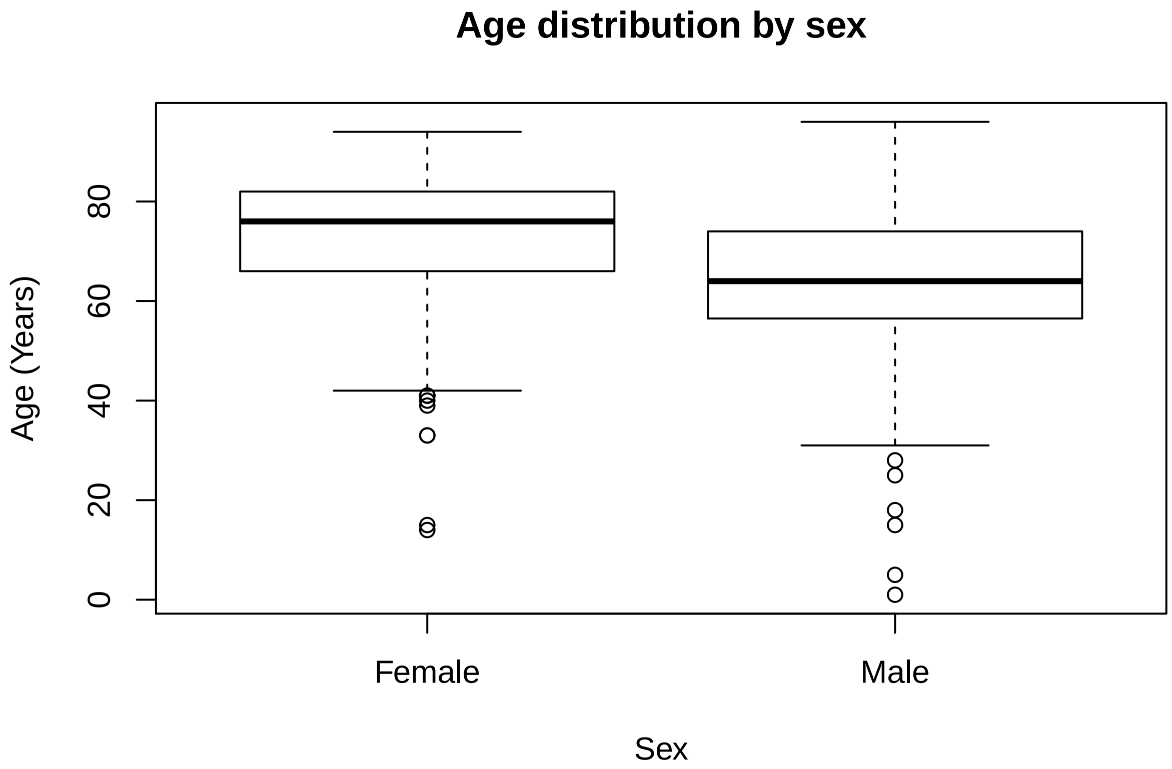
```
## Call:
## coxph(formula = stroke.surv.fit ~ stroke$sex + stroke$age)
```

```
## 
##   n= 829, number of events= 485 
## 
##                     coef exp(coef) se(coef)      z Pr(>|z|)    
## stroke$sexMale 0.022332  1.022584 0.100717  0.222    0.825    
## stroke$age     0.049289  1.050524 0.004374 11.268   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
##                exp(coef) exp(-coef) lower .95 upper .95
## stroke$sexMale     1.023     0.9779    0.8394     1.246
## stroke$age         1.051     0.9519    1.0416     1.060
## 
## Concordance= 0.66  (se = 0.014 )
## Rsquare= 0.174   (max possible= 0.999 )
## Likelihood ratio test= 158.6  on 2 df,   p=0
## Wald test            = 137.7  on 2 df,   p=0
## Score (logrank) test = 136.3  on 2 df,   p=0
```

In the summary of the Cox model, the `exp(coef)` column shows the estimated hazard ratio. The value for age is above 1 (exp(coef)=1.051), indicating age and risk of death are positively related: each year of age increases the risk of death by 5.1%. This value is statistically significant ($p<0.001$), therefore we can deduce that age significantly increases risk of death. The value for Male is also above 1 (exp(coef)=1.023), indicating that males have an increased risk of death, now that we corrected for age. However, the associated p-value is not significant ($p=0.825$),therefore the gender doesn't affect risk of death when age is incorporated in the model.

Looking at the data might explain why the effect of sex disappeared when correcting for age.

```
boxplot(stroke$age~stroke$sex, main="Age distribution by sex", ylab="Age (Years)", xlab="Sex")
```



Visualizing the distribution of ages separately for males and females reveals that females tend to be older. Therefore, the lower mortality of males is likely a consequence of their younger age and not of the sex itself. Indeed, when correcting by sex, we show that females of the same age as males do not have a higher risks mortality after a stroke. It is therefore important to know how the predictions of the model may be affected by covariates.

To have an idea of which variables in the dataset predict best the survival rate, we can include all predictors in the Cox model:

```
stroke.coxph.all <- coxph(stroke.surv.fit~stroke$sex+stroke$age+stroke$dgn+stroke$coma+stroke$diab+strol
summary(stroke.coxph.all)
```

```
## Call:
## coxph(formula = stroke.surv.fit ~ stroke$sex + stroke$age + stroke$dgn +
##     stroke$coma + stroke$diab + stroke$minf + stroke$han)
##
##   n= 814, number of events= 471
##    (15 observations deleted due to missingness)
##
##                     coef exp(coef)  se(coef)      z Pr(>|z|)
## stroke$sexMale  0.040303  1.041126  0.104167  0.387 0.698828
## stroke$age      0.050392  1.051683  0.004826 10.441  < 2e-16 ***
## stroke$dgnID   -0.294154  0.745162  0.169896 -1.731 0.083384 .
## stroke$dgnINF  -0.661077  0.516295  0.159111 -4.155 3.26e-05 ***
## stroke$dgnSAH  -0.396942  0.672373  0.259040 -1.532 0.125434
## stroke$comaYes  2.531158 12.568056  0.150927 16.771  < 2e-16 ***
## stroke$diabYes  0.451163  1.570137  0.139725  3.229 0.001243 **
## stroke$minfYes  0.476096  1.609777  0.129802  3.668 0.000245 ***
## stroke$hanYes  -0.041723  0.959135  0.095228 -0.438 0.661284
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                exp(coef) exp(-coef) lower .95 upper .95
## stroke$sexMale    1.0411    0.96050    0.8489    1.2769
## stroke$age        1.0517    0.95086    1.0418    1.0617
## stroke$dgnID      0.7452    1.34199    0.5341    1.0396
## stroke$dgnINF     0.5163    1.93688    0.3780    0.7052
## stroke$dgnSAH     0.6724    1.48727    0.4047    1.1171
## stroke$comaYes   12.5681    0.07957    9.3497   16.8942
## stroke$diabYes    1.5701    0.63689    1.1940    2.0648
## stroke$minfYes    1.6098    0.62120    1.2482    2.0761
## stroke$hanYes     0.9591    1.04261    0.7958    1.1559
##
## Concordance= 0.753  (se = 0.014 )
## Rsquare= 0.404   (max possible= 0.999 )
## Likelihood ratio test= 421.1  on 9 df,   p=0
## Wald test            = 480.1  on 9 df,   p=0
## Score (logrank) test = 720.9  on 9 df,   p=0
```

The full model confirms that age is significantly increasing mortality rate and that sex has no effect. According to this model, the strongest predictor is `coma`, with `exp(coef)=12.57`, indicating that people who have been in coma are much less likely to survive. Other predictors that significantly increase the risk of mortality are `diab` and `minf`, indicating that patients who have a diabetes or myocardial infection history have also a significantly increased risk of death. On the other hand, the `INF` levels of factor `dgn` is a significant predictor with a `exp(coef)<1`, indicating patients diagnosed with infarction have a lower mortality risk.

Conclusion

In conclusion, it is important to look for covariates among predictors and include them in the model (i.e. sex and age) to avoid drawing erroneous conclusions. Here, the conclusion to draw from the full model are that older patients, those who have a history of diabetes or myocardial infection and those who were in a coma after the stroke have less survival chance, whereas patients diagnosed with ischaemic infarction have higher survival chances.