Exercise 4: Multivariable analysis in R part 2: Cox proportional hazard model

At the end of this exercise you should be able to:

a. Use the Cox proportional hazard model
b. Test the assumption for proportionality and if violated, carry out a stratified analysis

Often the results of the logistic regression are the culminating final summary of your analysis. In our specific setting, the logistic regression was an intermediate step to an adjusted form of a survival analysis. You have learned the principle of using a Kaplan-Meier survival analysis in an earlier Exercise. We extend on this approach and look at factors that determine the binary outcome in the treatment of multidrug-resistant tuberculosis. A Kaplan-Meier analysis would get us quite far, but we would face the problem of adjustment and the problem of dealing with continuous predictor variables such as age. While categorizing age is an option, it might be arbitrary and not the most satisfactory solution. We could end up with multiple small groups if values from more than one variable are made strata of a combining new variable. Almost inevitably, differences might become meaningless with the loss of power.

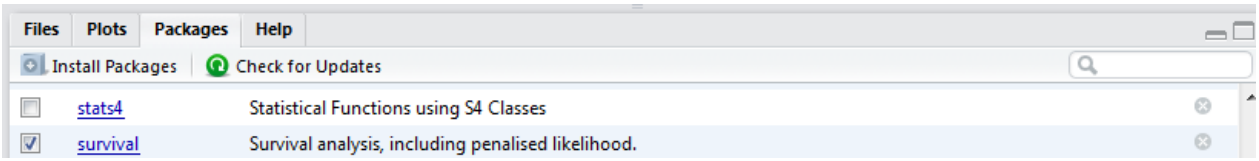
There are several techniques of multivariate analysis. Principally all do the same thing – simultaneous adjustment for multiple variables and providing independent effect estimates for each exposure variable in the model on the outcome. Depending on the type of outcome variable, different techniques are used.

If the outcome is a continuous variable, we use linear regression. If the outcome is categorical, we use logistic regression. If the outcome is 'time to event', we use a Cox proportional hazard model. If the outcome is 'number of events' (discrete numeric), then we use Poisson regression. Here, the outcome measure is 'time to event', and hence we use Kaplan-Meier analysis for univariate analysis and the Cox proportional hazard model for multivariate analysis.

In our approach to the analysis of the dataset on multidrug-resistant tuberculosis we combine the two techniques of logistic regression modeling and the Cox proportional hazard model in a way that is quite common: logistic regression is used first to evaluate and determine which variables have to be considered. There are alternative approaches, including determining the factors within the Cox model itself. In our case, we had isolated three factors, initial fluoroquinolone resistance, age, and sex. After fitting the Cox proportional hazard model including these three variables, we test whether any of the variables grossly violates the assumption of proportionality of hazards (which must be met). If there is such a variable, it must be removed from the adjustment and instead stratification by this variable is indicated. Then each of the strata are adjusted for the remaining variables and it is checked again whether anything remains that violates the proportionality assumption, and so on, until the final model emerges. This is the procedure we are going to apply.

The R survival package

The base package of R does not include survival analysis, and the package "survival" must thus be installed (see lower right quadrant in RStudio):



The "survival" package was written by Terry Therneau from the Mayo Clinic. The procedure is the same as we used before for the "foreign" package. Open a new file in the Source editor and save it as `e_ex04.r` file. Analogous to calling "foreign" from the library, we also need to be calling "survival" from the library. We are going to use the `e_ex02_02.dat` as our starting dataset. Thus, make sure that it is attached before we start doing anything.

```
* Exercise 4: Multivariable analysis in R part 2: Cox proportional hazard
  model
library(survival)
rm(list=ls())
e_ex04 <- read.table("e_ex02_02.dat")
names(e_ex04a)
```

The last line gives us the variables:

```
[1] "age"        "fq04"       "sex"        "totobstime" "agequart"   "agemed"     "outcome07"
[8] "outcome02"  "pza02"      "kmy02"      "pth02"      "cxr02"      "fq02"
```

We need only `AGE`, `FQ04`, `SEX`, `TOTOBSTIME`, and `OUTCOME02`, thus:

```
e_ex04b <- e_ex04a[c(1:4, 8)]
```

(It is not just to repeat what was learnt before, there is also some reason to reduce datasets: like EpiData Analysis, R works in the memory which makes it a relatively "slow processor"). We then make the same modification we discussed and did in Exercise 3, attach the file and check the names:

```
attach(e_ex04b)
e_ex04b$fq03[fq04 == "Susceptible"]           <- "1-Susceptible"
e_ex04b$fq03[fq04 == "Low-level resistance"]  <- "2-Low-level resistance"
e_ex04b$fq03[fq04 == "High-level resistance"] <- "3-High-level resistance"
e_ex04b$out02[outcome02 == "Success"] <- 0
e_ex04b$out02[outcome02 == "Failure"] <- 1
detach(e_ex04b)
names(e_ex04b)
e_ex04c <- e_ex04b[c(1, 3:4, 6:7)]
attach(e_ex04c)
names(e_ex04c)
```

The last line shows:

```
[1] "age"        "sex"        "totobstime" "fq03"       "out02"
```

We have thus reduced our dataset to the bare essential minimum with which we can work.

Kaplan-Meier survival analysis

Let's first compare notes, i.e. check whether we get identical survival probabilities using the Kaplan-Meier method in EpiData Analysis and in R. In EpiData Analysis, we would use:

```
lifetable out02 totobstime /by=fq03 /i=b30 /adj \
          /ymin=0.30 /ymax=1 /NG /e3 \
          /ti="KM successful outcome probability" \
          /sub="by intitial fluoroquinolone susceptibility" \
          /t
```

We get:

Survival probability for `susceptible`: 0.865
Survival probability for `low-level resistance`: 0.909
Survival probability for `high-level resistance`: 0.480

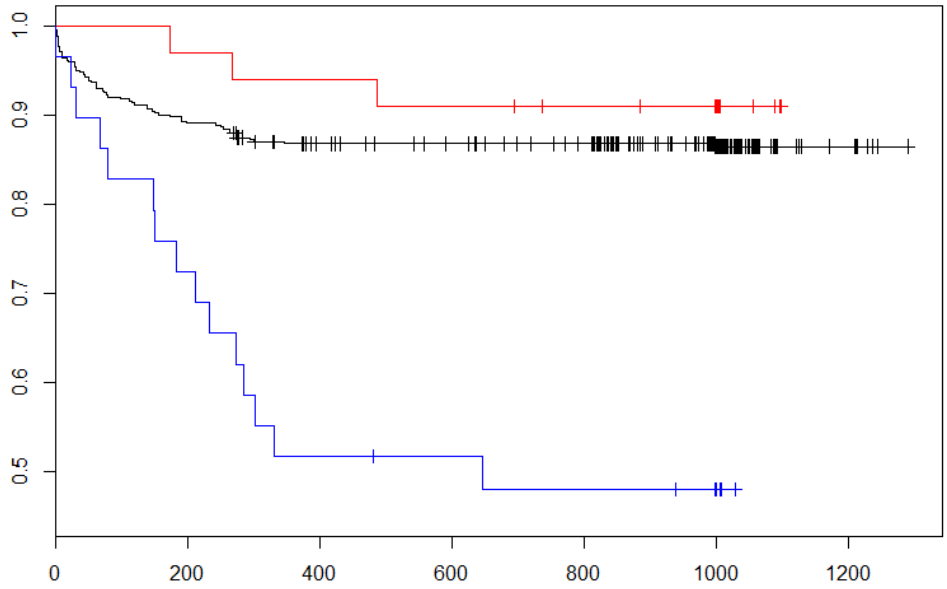
The model commands in R:

```
fit1 <- survfit(formula=Surv(totobstime, out02) ~ fq03, data=e_ex04c)
```

We want to add a simple plot:

```
plot(fit1, col = c(1, 2, 4), ymin=0.45)
# colors: 1=black; 2=red, 4=blue
```

which gives us the plot in the lower-right quadrant:



Remember that initially we defined R as a "language and environment for statistical computing and graphing". The graphics capabilities of R are enormous but it will take time to learn and acquire them (see more information in the text by Thomas Lumley). As it is not our primary purpose to evaluate the graphics capabilities of R, we leave that for the time being. We must see the statistical output, however, so we add a third command line:

```
fit1 <- survfit(formula=Surv(totobstime, out02) ~ fq03, data=e_ex04c)
plot(fit1, col = c(1, 2, 4), ymin=0.45)
# colors: 1=black; 2=red, 4=blue
summary(fit1, times = seq(0, 2000, 30))
```

We get exactly the same survival probabilities as with the `lifetable` in EpiData Analysis. This should strengthen our confidence in both software and our skills. We note, however, that the confidence intervals are different. In EpiData Analysis, events change the survival probability but censored observations do not. This is identical in R and it is what we expect. In EpiData analysis, the 95% confidence interval, however, continues to widen as observations with the passage of time become censored, while this is not the case in R. EpiData Analysis uses the philosophy that smaller numbers lead to larger uncertainty, while R focuses on the importance of uncertainty at the point of the last event. Comparison shows that Stata and R get the same results. EpiData Analysis has adapted the approach proposed by Altman, and this will be reviewed during the re-writing of the EpiData Analysis module. In any case, we can reproduce the survival probability in the Kaplan-Meier approach.

The Cox proportional hazard model

The proportional hazards model allows the analysis of survival data by regression modeling. Linearity is assumed on the log scale of the hazard. Relative to a referent, say the rate of death among a control group, the rate of death among the experimental group might be half that of the control group and the hazard ratio is thus 0.5. In contrast to relative risks which are cumulative over observation time, hazard ratios reflect an instantaneous risk over the study period or a subset of the period. Hazard ratios suffer therefore somewhat less from possible selection bias introduced by endpoints. Under the Cox proportional hazard model, the hazard ratio is constant. The Cox model thus assumes an underlying hazard function with a corresponding survival curve. In a stratified analysis, there will be one such curve for each stratum.

The command lines for the Cox model are:

```
mdrcox  <- coxph(Surv(totobstime, out02) ~ factor(fq03) + factor(sex) + age,
  data=e_ex04c)
summary(mdrcox)
```

where `totobstime` is the variable for the total observation time from treatment start until the event occurs or the observation time is censored (be it e.g. to loss from follow-up after treatment cessation or reaching the end of the observation time). Note that "`Surv`" is capitalized (R is case-sensitive). The second line gives the output:

```
Call:
coxph(formula = Surv(totobstime, out02) ~ factor(fq03) + factor(sex) +
    age, data = e_ex04c)

  n= 501, number of events= 77

                                      coef exp(coef)  se(coef)      z Pr(>|z|)
factor(fq03)2-Low-level resistance  -0.329337  0.719400  0.593274 -0.555 0.578814
factor(fq03)3-High-level resistance  1.506031  4.508801  0.293728  5.127 2.94e-07 ***
factor(sex)Male                     -0.643091  0.525665  0.256959 -2.503 0.012325 *
age                                  0.031887  1.032400  0.008783  3.630 0.000283 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
                                  exp(coef) exp(-coef) lower .95 upper .95
factor(fq03)2-Low-level resistance    0.7194     1.3900    0.2249    2.3013
factor(fq03)3-High-level resistance   4.5088     0.2218    2.5354    8.0183
factor(sex)Male                       0.5257     1.9024    0.3177    0.8698
age                                   1.0324     0.9686    1.0148    1.0503

Concordance= 0.687  (se = 0.033 )
Rsquare= 0.068   (max possible= 0.848 )
Likelihood ratio test= 35.06  on 4 df,   p=4.521e-07
Wald test            = 43.95  on 4 df,   p=6.565e-09
Score (logrank) test = 49.31  on 4 df,   p=5.026e-10
```

As it should be, the three variables are significantly associated. The `exp(coef)`, i.e. the exponent of the coefficient is the hazard ratio.

As mentioned above, the Cox proportional hazard model requires that the assumption of proportionality is met, that is the survival function for different factors are required to change proportionately and do not, for instance cross each other.

The test diagnostic to evaluate whether the assumption of proportionality is met is:

```
cox.zph(mdrcox)
```

gives:

```
                                    rho chisq        p
factor(fq03)2-Low-level resistance  0.248  4.72 0.029826
factor(fq03)3-High-level resistance 0.257  5.16 0.023055
factor(sex)Male                    -0.105  0.82 0.365224
age                                -0.241  4.18 0.040909
GLOBAL                                 NA 18.51 0.000982
```

The global chi-square test is highly significant, that is the assumption of proportionality is violated. The most important culprit is seemingly the variable `fq03`. `AGE` is also significant but not that grossly and `sex` not at all. The way to resolve it is stratification, that is `fq03` must be taken out and the model must be stratified by `fq03`. R makes it easy for us to do that with the following modification:

```
mdrcox  <- coxph(Surv(totobstime, out02) ~ strata(fq03) + factor(sex) + age,
  data=e_ex04c)
summary(mdrcox)
```

This gives:

```
                     coef exp(coef)  se(coef)      z Pr(>|z|)
factor(sex)Male -0.643315  0.525547  0.256857 -2.505 0.012260 *
age              0.031399  1.031898  0.008758  3.585 0.000337 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                exp(coef) exp(-coef) lower .95 upper .95
factor(sex)Male    0.5255     1.9028    0.3177    0.8695
age                1.0319     0.9691    1.0143    1.0498

Concordance= 0.647  (se = 0.039 )
Rsquare= 0.028   (max possible= 0.806 )
Likelihood ratio test= 14.26  on 2 df,   p=0.0008024
Wald test            = 14.38  on 2 df,   p=0.0007543
Score (logrank) test = 14.73  on 2 df,   p=0.0006345
```

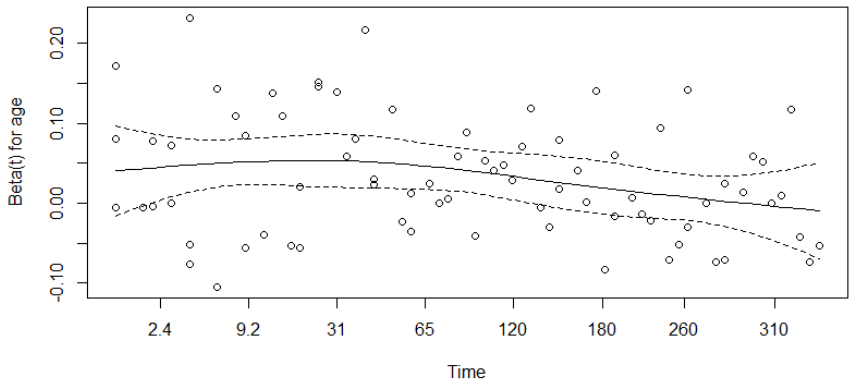
We have now the hazard of `sex` and `age` but not anymore of fluoroquinolone resistance because the latter was not a constant hazard and was thus taken out by stratification.

Testing the assumption for proportionality and plotting the test result:

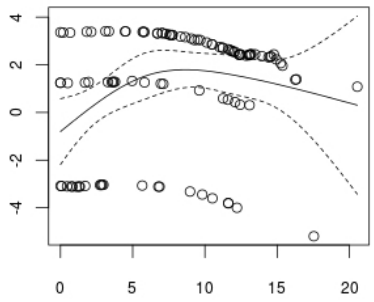
```
cox.zph(mdrcox)
plot(cox.zph(mdrcox))
```

```
                    rho chisq      p
factor(sex)Male -0.112 0.926 0.3358
age             -0.245 4.288 0.0384
GLOBAL              NA 7.864 0.0196
```

and



This indicates that the model is not ideal for the variable age, but with a bit lenience and the more or less regular graphical test we will allow it to pass without further more complex stratification. The interpretation of the graph in its most simplest way is how curved it is: if it is fairly flat (as we think it is here), the assumption of proportionality is not (much) violated. If it is decidedly different from flat, then the assumption is violated as in this example:



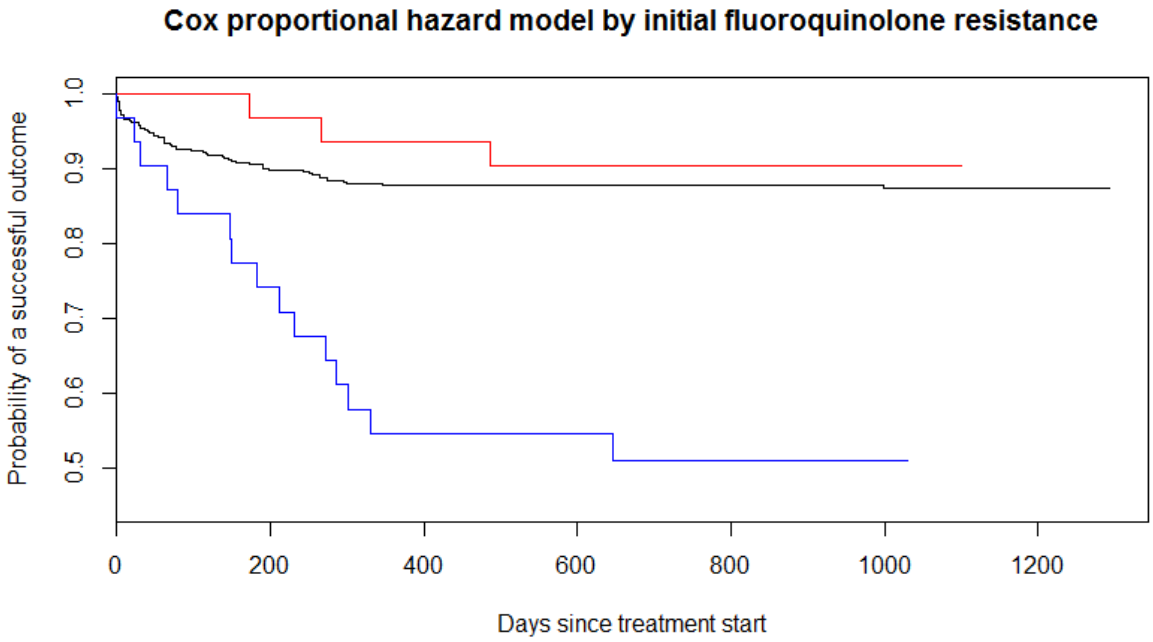
Copied from:

http://stats.stackexchange.com/questions/15114/how-to-understand-the-plotting-of-the-cox-zph-function-in-r

```
help(plot.survfit)
```

gives us information on how to plot it. It can get elaborate, of course, but here just a bare-bone sequence of commands:

```
p <- plot(survfit(mdrcox), ylim=c(.45, 1), xlab="Days since treatment start",
  mark.time=F, ylab="Probability of a successful outcome",  col=c(1, 2, 4),
  main="Cox  proportional  hazard  model  by  initial  fluoroquinolone
  resistance")
```



Our main interest is in the numeric output quantification of survival probability:

```
print(p)
```

```
$x
[1] 1292 1099 1030

$y
[1] 0.8744596 0.9045559 0.5096336
```

Thus, our main result, the survival probabilities among patients with initial fluoroquinolone susceptibility, low-, and high-level resistance, adjusted for age and sex are:

Survival probability `susceptible`: `0.874`
Survival probability `low-level resistant`: `0.905`
Survival probability `high-level resistant`: `0.510`

To obtain the detailed data:

```
mdrcox2 <- survfit(mdrcox)
summary(mdrcox2, times = seq(0, 3000, 30))
```

gives (just the first 150 days for susceptible, low-level resistant, and high-level resistant):

```
                fq03=1-Susceptible
time n.risk n.event survival std.err lower 95% CI upper 95% CI
   0    439       2    0.996 0.00292        0.990        1.000
  30    419      19    0.956 0.00954        0.938        0.975
  60    411       7    0.942 0.01103        0.920        0.963
  90    404       7    0.927 0.01235        0.903        0.951
 120    400       4    0.918 0.01304        0.893        0.944
 150    397       4    0.910 0.01369        0.883        0.937
```

```
                fq03=2-Low-level resistance
time n.risk n.event survival std.err lower 95% CI upper 95% CI
   0     33       0    1.000  0.0000        1.000            1
  30     33       0    1.000  0.0000        1.000            1
  60     33       0    1.000  0.0000        1.000            1
  90     33       0    1.000  0.0000        1.000            1
 120     33       0    1.000  0.0000        1.000            1
 150     33       0    1.000  0.0000        1.000            1

                fq03=3-High-level resistance
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    0     29       1    0.969  0.0309        0.910        1.000
   30     27       2    0.905  0.0526        0.807        1.000
   60     26       0    0.905  0.0526        0.807        1.000
   90     24       2    0.839  0.0667        0.718        0.981
  120     24       0    0.839  0.0667        0.718        0.981
  150     23       2    0.775  0.0764        0.638        0.940
```

Here, we show it for 30-day intervals, out for up to 3000 days but we could also just get it for the first 5 days, but daily:

```
summary(mdrcox2, times = seq(0, 5, 1))
```

```
                fq03=1-Susceptible
time n.risk n.event survival std.err lower 95% CI upper 95% CI
   0    439       2    0.996 0.00292        0.990        1.000
   1    437       1    0.994 0.00358        0.987        1.000
   2    436       2    0.990 0.00463        0.981        0.999
   3    434       2    0.986 0.00548        0.975        0.996
   4    432       3    0.979 0.00656        0.967        0.992
   5    429       2    0.975 0.00719        0.961        0.989

                fq03=2-Low-level resistance
time n.risk n.event survival std.err lower 95% CI upper 95% CI
   0     33       0        1       0            1            1
   1     33       0        1       0            1            1
   2     33       0        1       0            1            1
   3     33       0        1       0            1            1
   4     33       0        1       0            1            1
   5     33       0        1       0            1            1

                fq03=3-High-level resistance
time n.risk n.event survival std.err lower 95% CI upper 95% CI
   0     29       1    0.969  0.0309         0.91            1
   1     28       0    0.969  0.0309         0.91            1
   2     28       0    0.969  0.0309         0.91            1
   3     28       0    0.969  0.0309         0.91            1
   4     28       0    0.969  0.0309         0.91            1
   5     28       0    0.969  0.0309         0.91            1
```

R is indeed powerful through flexibility.

Task:

This analysis shows that low-level fluoroquinolone resistance has seemingly no influence on the outcome while in contrast high-level fluoroquinolone resistance is a powerful predictor for an adverse outcome. Nevertheless, even with high-level fluoroquinolone resistance, a remarkable 51% still had a successful outcome. Unsuccessful here also included death and default. What is of key interest with resistance to the core drug fluoroquinolone is whether the outcome is bacteriologically favorable or unfavorable.

- ***The first task is to identify factors that are predictors for an unsuccessful bacteriological outcome (failure or relapse) versus a bacteriologically favorable outcome (completion and relapse-free cure) among cases with any type of fluoroquinolone resistance (low-level or high-level) and who did neither die nor default.***
- ***Summarize your analysis in a table showing numbers, odds ratios and 95% confidence intervals both univariate and adjusted multivariate for the factors identified in your regression analysis.***