W12 TP

Lucile Favero

12/6/2021

Contents

1 Aim

Report how bone marrow transplant survival times relates to graft versus host disease (GHVD)

2 Preprocessing

2.1 Load library and data

```
library(GGally)       # for ggpairs
```

```
## Loading required package: ggplot2
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(ggfortify)    # for autoplot
library(ggplot2)      # for ggplot
library('MASS')       # for the glm model selection
library(glmnet)
```

```
## Loading required package: Matrix
```

```
## Loaded glmnet 4.1-1
```

```
library(patchwork)
```

```
##
## Attaching package: 'patchwork'
```

```
## The following object is masked from 'package:MASS':
##
##     area
```

```
library(ISwR)
library(survival)
```

```
##
## Attaching package: 'survival'

## The following object is masked from 'package:ISwR':
##
##     lung
```

```
library(survminer)
```

```
## Loading required package: ggpubr
```

```
d<- graft.vs.host
```

2.2 Understand datas

> "The gvhd data frame has 37 rows and 7 columns. It contains data from patients receiving a nondepleted >allogenic bone marrow transplant with the purpose of finding variables associated with the development of >acute graft-versus-host disease."

```
str(d)
```

```
## 'data.frame':    37 obs. of  9 variables:
##  $ pnr   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ rcpage: int  27 13 19 21 28 22 19 20 33 18 ...
##  $ donage: int  23 18 19 22 38 20 19 23 36 19 ...
##  $ type  : int  2 2 1 2 2 2 2 2 1 1 ...
##  $ preg  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ index : num  0.27 0.31 0.39 0.48 0.49 0.5 0.81 0.82 0.86 0.92 ...
##  $ gvhd  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ time  : int  95 1385 465 810 1497 1181 993 138 266 579 ...
##  $ dead  : int  1 0 1 1 0 1 0 1 1 0 ...
```

```
summary(d)
```

```
##       pnr          rcpage          donage           type            preg
##  Min.   : 1   Min.   :13.00   Min.   :14.00   Min.   :1.000   Min.   :0.0000
##  1st Qu.:10   1st Qu.:20.00   1st Qu.:20.00   1st Qu.:1.000   1st Qu.:0.0000
##  Median :19   Median :23.00   Median :23.00   Median :2.000   Median :0.0000
##  Mean   :19   Mean   :25.43   Mean   :25.81   Mean   :1.973   Mean   :0.2703
##  3rd Qu.:28   3rd Qu.:29.00   3rd Qu.:34.00   3rd Qu.:3.000   3rd Qu.:1.0000
##  Max.   :37   Max.   :43.00   Max.   :43.00   Max.   :3.000   Max.   :1.0000
##      index             gvhd             time            dead
##  Min.   : 0.270   Min.   :0.0000   Min.   :  41.0   Min.   :0.0000
##  1st Qu.: 0.920   1st Qu.:0.0000   1st Qu.: 177.0   1st Qu.:0.0000
##  Median : 2.010   Median :0.0000   Median : 667.0   Median :0.0000
##  Mean   : 2.556   Mean   :0.4595   Mean   : 669.8   Mean   :0.4865
##  3rd Qu.: 3.730   3rd Qu.:1.0000   3rd Qu.:1105.0   3rd Qu.:1.0000
##  Max.   :10.110   Max.   :1.0000   Max.   :1504.0   Max.   :1.0000
```

Transform into factor the variables : type, preg, gvhd, dead

```
d$type <-as.factor(d$type)#type of leukaemia coded 1: AML, 2: ALL, 3: CML for acute myeloid, acute lymp
d$preg<-as.factor(d$preg)# indicating whether donor has been pregnant. 0: no, 1: yes.``
levels(d$preg)<-c("no","yes")
d$gvhd<-as.factor(d$gvhd)# graft-versus-host disease, 0: no, 1: yes
levels(d$gvhd)<-c("no","yes")
d$dead <- as.factor(d$dead) # a numeric vector code, 0: no (censored), 1: yes
levels(d$dead)<-c("no","yes")
```

```r
str(d)
```

```
## 'data.frame':    37 obs. of  9 variables:
##  $ pnr   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ rcpage: int  27 13 19 21 28 22 19 20 33 18 ...
##  $ donage: int  23 18 19 22 38 20 19 23 36 19 ...
##  $ type  : Factor w/ 3 levels "1","2","3": 2 2 1 2 2 2 2 2 1 1 ...
##  $ preg  : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ index : num  0.27 0.31 0.39 0.48 0.49 0.5 0.81 0.82 0.86 0.92 ...
##  $ gvhd  : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ time  : int  95 1385 465 810 1497 1181 993 138 266 579 ...
##  $ dead  : Factor w/ 2 levels "no","yes": 2 1 2 2 1 2 1 2 2 1 ...
```

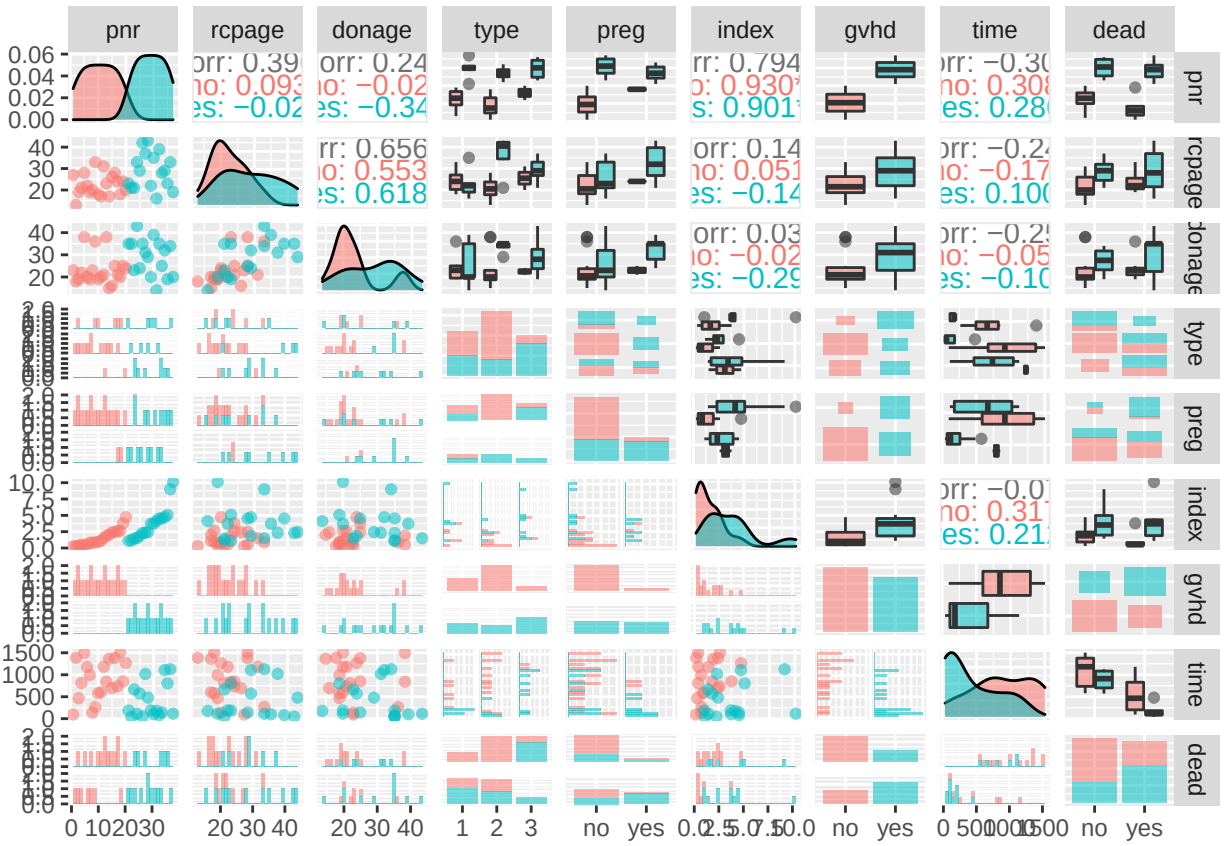
```r
summary(d)
```

```
##       pnr         rcpage          donage      type    preg         index       
##  Min.   : 1   Min.   :13.00   Min.   :14.00   1:11   no :27   Min.   : 0.270  
##  1st Qu.:10   1st Qu.:20.00   1st Qu.:20.00   2:16   yes:10   1st Qu.: 0.920  
##  Median :19   Median :23.00   Median :23.00   3:10            Median : 2.010  
##  Mean   :19   Mean   :25.43   Mean   :25.81                   Mean   : 2.556  
##  3rd Qu.:28   3rd Qu.:29.00   3rd Qu.:34.00                   3rd Qu.: 3.730  
##  Max.   :37   Max.   :43.00   Max.   :43.00                   Max.   :10.110  
##   gvhd         time         dead   
##  no :20   Min.   :  41.0   no :19  
##  yes:17   1st Qu.: 177.0   yes:18  
##           Median : 667.0           
##           Mean   : 669.8           
##           3rd Qu.:1105.0           
##           Max.   :1504.0           
```

Plot the ggpairs:

```r
ggpairs(d, aes(color=gvhd, alpha = 0.3))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



3 Modelization

```
surv<-Surv(d$time,d$dead=="yes")
```

3.1 A simple model

We set a model depending only on gvhd

```
m1<-coxph(data=d,surv~ gvhd)
summary(m1)
```

```
## Call:
## coxph(formula = surv ~ gvhd, data = d)
##
##   n= 37, number of events= 18
##
##            coef exp(coef) se(coef)     z Pr(>|z|)
## gvhdyes 1.2419    3.4620   0.5144 2.414   0.0158 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##         exp(coef) exp(-coef) lower .95 upper .95
## gvhdyes     3.462     0.2888     1.263     9.489
##
## Concordance= 0.671  (se = 0.052 )
## Likelihood ratio test= 6.19  on 1 df,   p=0.01
```

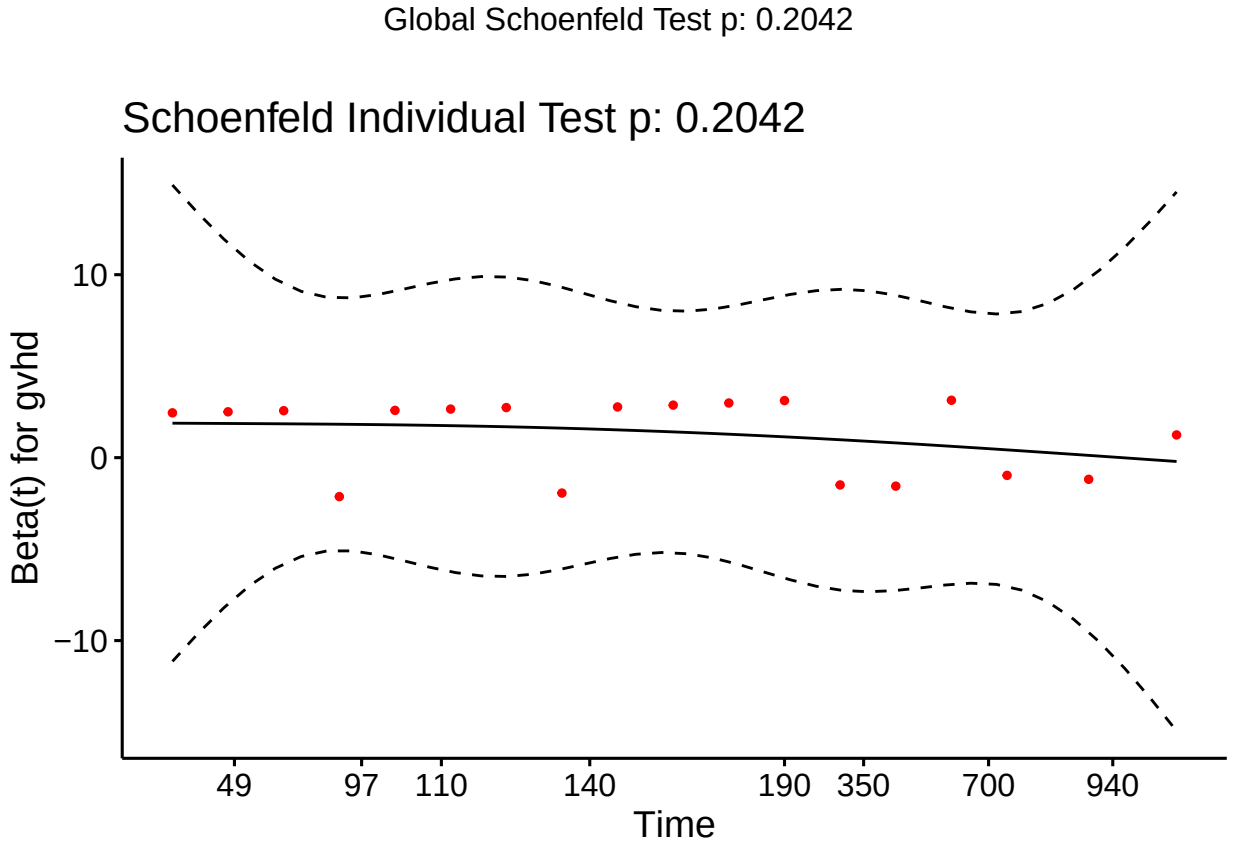
```
## Wald test            = 5.83  on 1 df,   p=0.02
## Score (logrank) test = 6.55  on 1 df,   p=0.01
```

The variable is significant for this model.

```
m1.diag<-cox.zph(m1)
m1.diag
```

```
##        chisq df   p
## gvhd    1.61  1 0.2
## GLOBAL  1.61  1 0.2
```

```
p1<-ggsurv(survfit(m1))+ylim(0,1)
p2<-ggcoxdiagnostics(m1,hline=FALSE)+ geom_smooth()
ggcoxzph(m1.diag)
```




```
(p1 | p2)
```

```
## `geom_smooth()` using formula 'y ~ x'
```

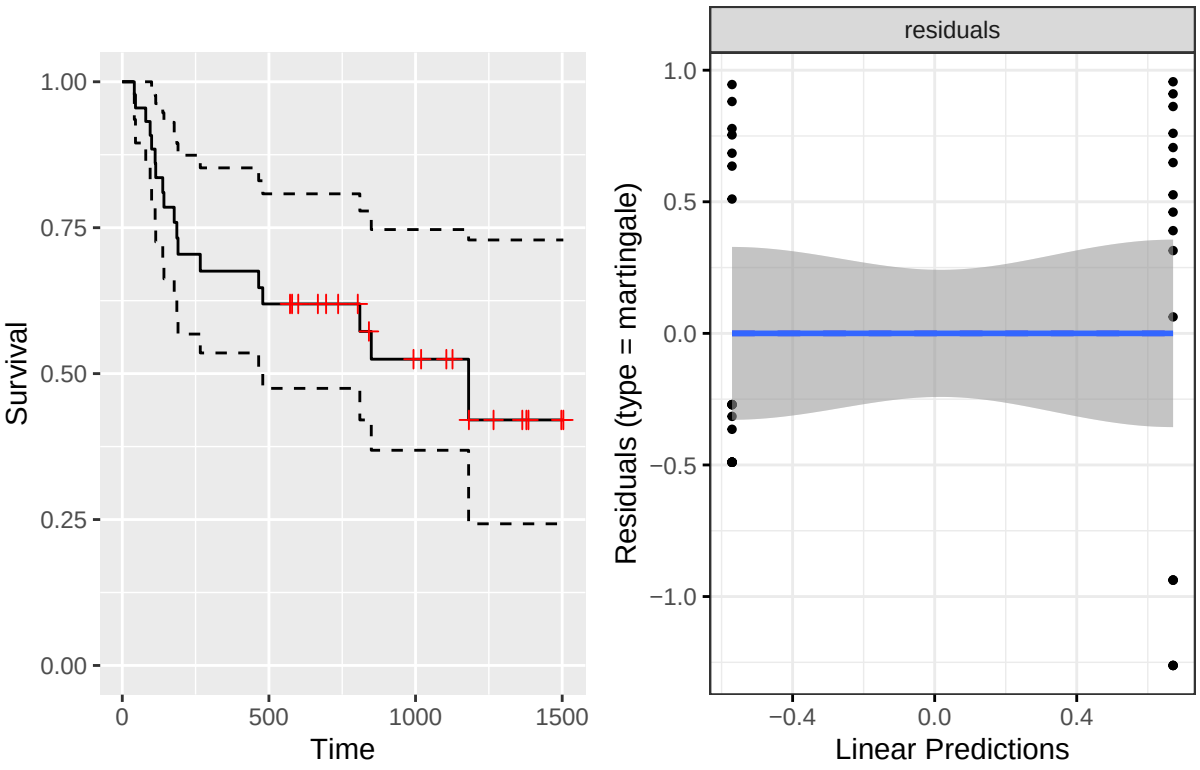
```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : pseudoinverse used at -0.57679
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : neighborhood radius 1.2481
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : reciprocal condition number 0
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : There are other near singularities as well. 1.5577
```

```
## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : pseudoinverse used at
## -0.57679

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : neighborhood radius
## 1.2481

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : reciprocal condition
## number 0

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : There are other near
## singularities as well. 1.5577

## `geom_smooth()` using method = 'loess' and formula 'y ~ x'

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : pseudoinverse used at -0.57679

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : neighborhood radius 1.2481

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : reciprocal condition number 0

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : There are other near singularities as well. 1.5577

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : pseudoinverse used at
## -0.57679

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : neighborhood radius
## 1.2481

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : reciprocal condition
## number 0

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : There are other near
## singularities as well. 1.5577
```



The p-value is larger than 5%, so the null hypothesis of proportional hazards is rejected. However, the second plot has a correct regression line and the in the third, all the points are in the CI.

3.2 Add more variables

```
m2<-coxph(data=d,surv~ gvhd+rcpage+ donage+ type+ preg +index )
summary(m2)
```

```
## Call:
## coxph(formula = surv ~ gvhd + rcpage + donage + type + preg +
##     index, data = d)
##
##   n= 37, number of events= 18
##
##               coef  exp(coef)   se(coef)      z Pr(>|z|)
## gvhdyes  2.0942285  8.1191750  0.7775552  2.693  0.00707 **
## rcpage  -0.0161510  0.9839787  0.0437183 -0.369  0.71180
## donage   0.0441502  1.0451393  0.0393515  1.122  0.26189
## type2   -0.0331175  0.9674249  0.6712286 -0.049  0.96065
## type3   -2.4211375  0.0888205  0.9523310 -2.542  0.01101 *
## pregyes  0.3439450  1.4105011  0.7410858  0.464  0.64257
## index    0.0004593  1.0004594  0.1456073  0.003  0.99748
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##         exp(coef) exp(-coef) lower .95 upper .95
## gvhdyes   8.11917     0.1232   1.76871   37.2707
## rcpage    0.98398     1.0163   0.90318    1.0720
## donage    1.04514     0.9568   0.96756    1.1289
## type2     0.96742     1.0337   0.25958    3.6055
```

```
## type3     0.08882    11.2587   0.01374    0.5743
## pregyes   1.41050     0.7090   0.33004    6.0282
## index     1.00046     0.9995   0.75207    1.3309
##
## Concordance= 0.801  (se = 0.052 )
## Likelihood ratio test= 22.62  on 7 df,   p=0.002
## Wald test            = 21.36  on 7 df,   p=0.003
## Score (logrank) test = 25.65  on 7 df,   p=6e-04
```

The two variables significant for this model are gvhd and type.

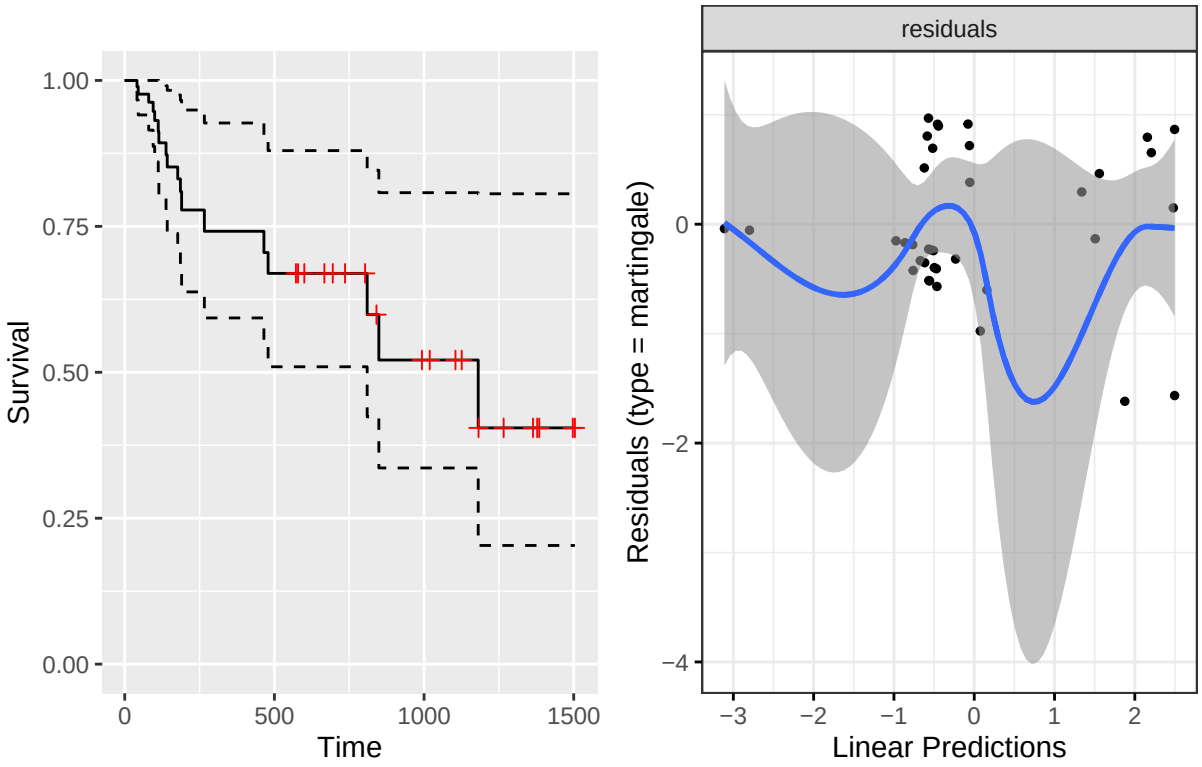
```
m2.diag<-cox.zph(m2)
m2.diag
```

```
##          chisq df      p
## gvhd    0.1596  1 0.6896
## rcpage  0.7716  1 0.3797
## donage  2.0623  1 0.1510
## type    9.9676  2 0.0068
## preg    0.0479  1 0.8267
## index   0.0224  1 0.8811
## GLOBAL 12.6576  7 0.0809
```

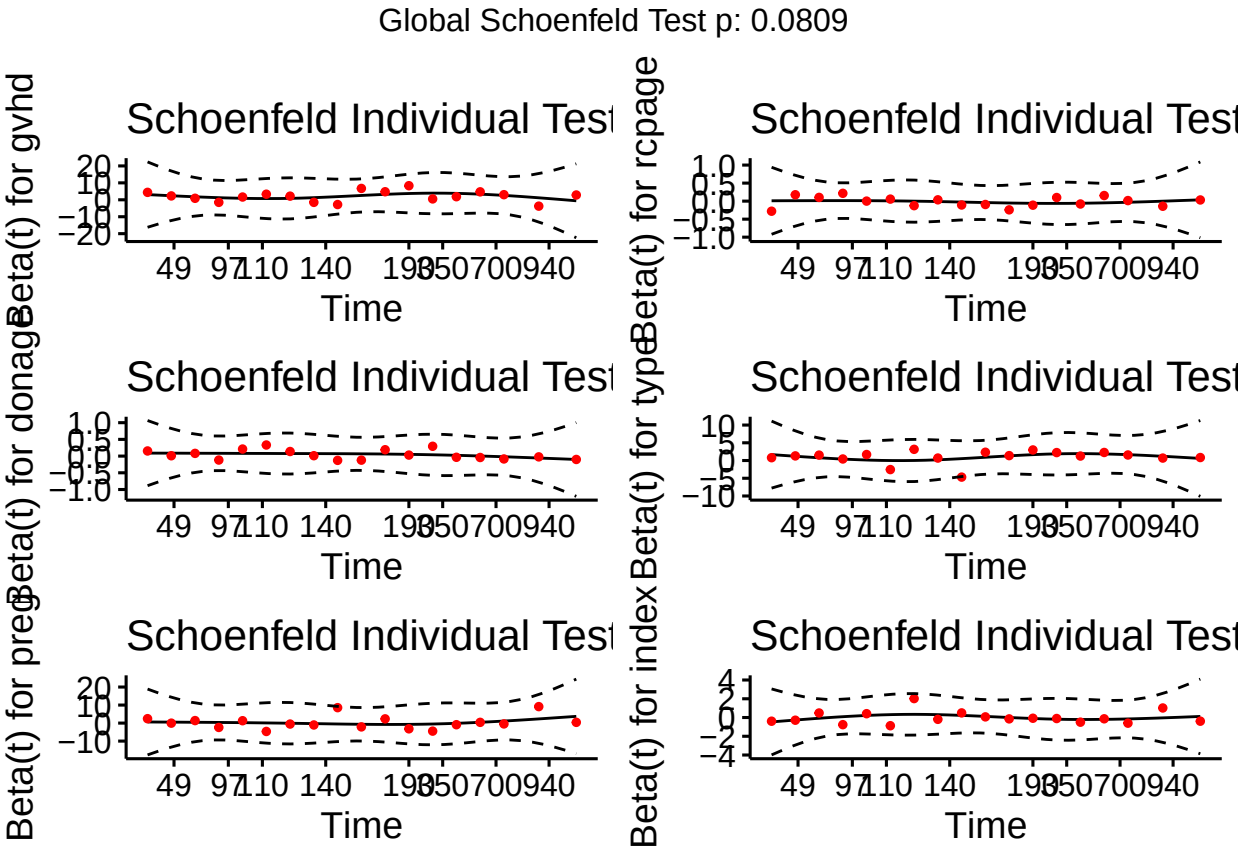
```
p1<-ggsurv(survfit(m2))+ylim(0,1)
p2<-ggcoxdiagnostics(m2,hline=FALSE)+ geom_smooth()
p3<-ggcoxzph(m2.diag)
p1 | p2
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



```
p3
```



From the second plot, we see that the linear prediction is not respected. The p-value is larger than 5%, so the null hypothesis of proportional hazards is rejected. In the last plot all the points are in the CI. ## only gvhd and type

```
m3<-coxph(data=d,surv~ gvhd+type )
summary(m3)
```

```
## Call:
## coxph(formula = surv ~ gvhd + type, data = d)
##
##   n= 37, number of events= 18
##
##             coef exp(coef) se(coef)      z Pr(>|z|)
## gvhdyes  2.29593   9.93370  0.60242  3.811 0.000138 ***
## type2   -0.04537   0.95565  0.55440 -0.082 0.934783
## type3   -2.52955   0.07969  0.84987 -2.976 0.002916 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##         exp(coef) exp(-coef) lower .95 upper .95
## gvhdyes   9.93370     0.1007   3.05022   32.3512
## type2     0.95565     1.0464   0.32240    2.8327
## type3     0.07969    12.5479   0.01507    0.4215
##
## Concordance= 0.745  (se = 0.059 )
## Likelihood ratio test= 20.37  on 3 df,   p=1e-04
```

```
## Wald test            = 18.81  on 3 df,   p=3e-04
## Score (logrank) test = 21.72  on 3 df,   p=7e-05
```

The two variables are significant.

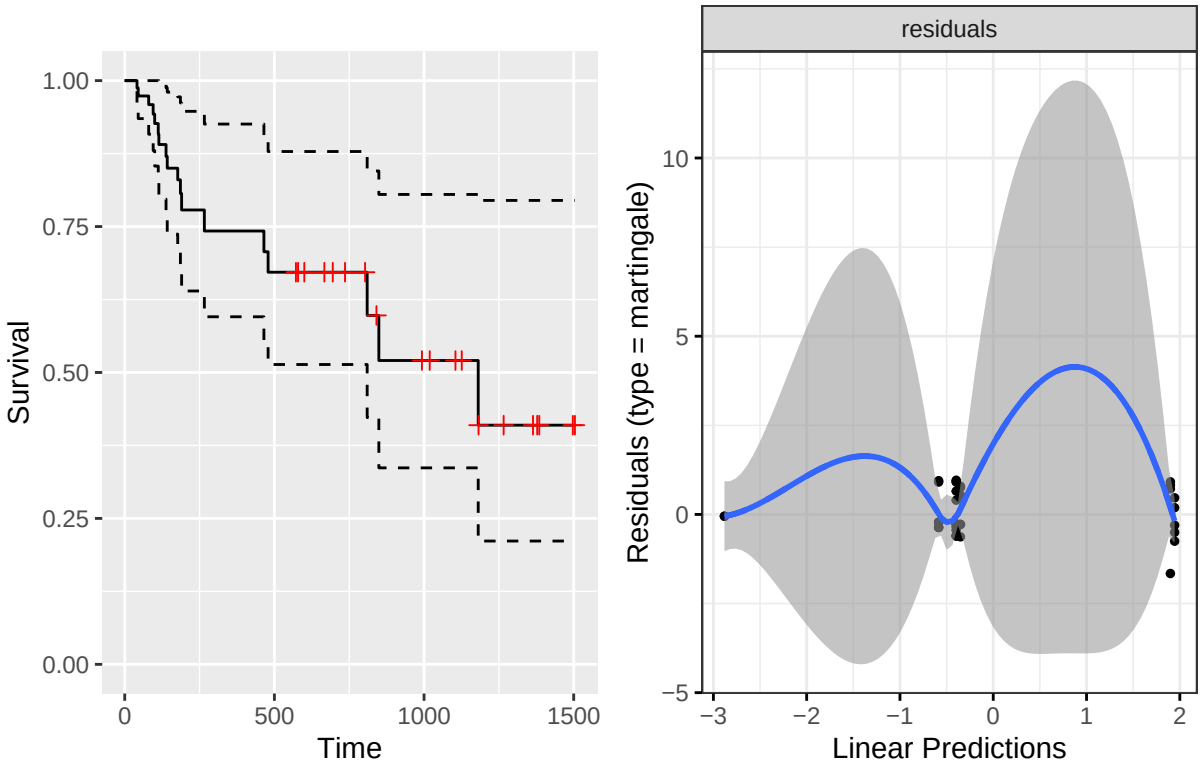
```
m3.diag<-cox.zph(m3)
m3.diag
```

```
##        chisq df      p
## gvhd   0.198  1 0.6561
## type   9.396  2 0.0091
## GLOBAL 9.426  3 0.0241
```

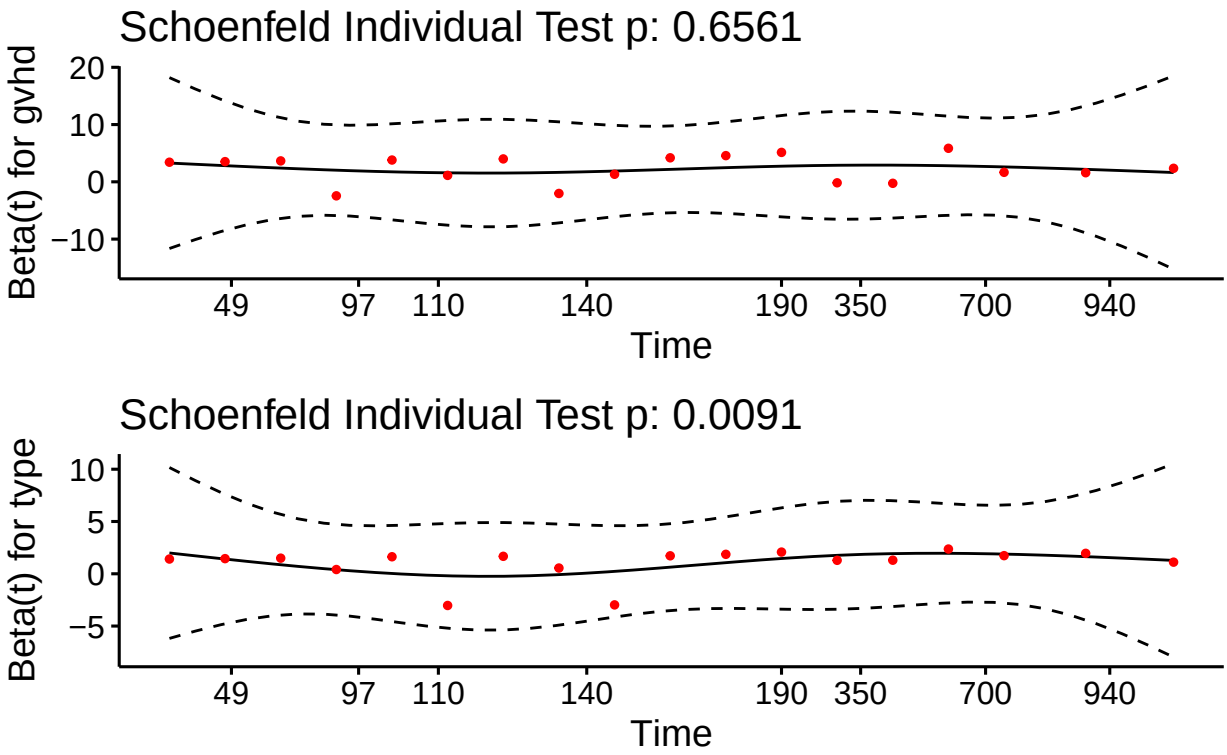
```
p1<-ggsurv(survfit(m3))+ylim(0,1)
p2<-ggcoxdiagnostics(m3,hline=FALSE)+ geom_smooth()
p3<-ggcoxzph(m3.diag)
p1 | p2
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
p3
```



From the second plot, we see that the linear prediction is not respected. The p-value is lower than 5%, so the null hypothesis of proportional hazards is not rejected. In the last plot all the points are in the CI. # Version of R used

```
## R version 4.0.1 (2020-06-06)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Pop!_OS 21.04
##
## Matrix products: default
## BLAS:   /usr/lib/x86_64-linux-gnu/openblas-pthread/libblas.so.3
## LAPACK: /usr/lib/x86_64-linux-gnu/openblas-pthread/libopenblasp-r0.3.13.so
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=fr_CH.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=fr_CH.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=fr_CH.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=fr_CH.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] survminer_0.4.9  ggpubr_0.4.0     survival_3.1-12  ISwR_2.0-8
##  [5] patchwork_1.1.1  glmnet_4.1-1     Matrix_1.2-18    MASS_7.3-51.6
##  [9] ggfortify_0.4.11 GGally_2.1.1     ggplot2_3.3.3
##
```

```
## loaded via a namespace (and not attached):
##  [1] tidyr_1.1.3        splines_4.0.1      foreach_1.5.1      carData_3.0-4
##  [5] assertthat_0.2.1   highr_0.9          cellranger_1.1.0   yaml_2.2.1
##  [9] pillar_1.6.1       backports_1.2.1    lattice_0.20-41    glue_1.4.2
## [13] digest_0.6.27      RColorBrewer_1.1-2 ggsignif_0.6.1     colorspace_2.0-1
## [17] htmltools_0.5.1.1  plyr_1.8.6         pkgconfig_2.0.3    broom_0.7.10
## [21] haven_2.4.1        purrr_0.3.4        xtable_1.8-4       scales_1.1.1
## [25] km.ci_0.5-2        openxlsx_4.2.3     rio_0.5.26         KMsurv_0.1-5
## [29] tibble_3.1.2       mgcv_1.8-31        farver_2.1.0       generics_0.1.0
## [33] car_3.0-10         ellipsis_0.3.2     withr_2.4.2        magrittr_2.0.1
## [37] crayon_1.4.1       readxl_1.3.1       evaluate_0.14      fansi_0.5.0
## [41] nlme_3.1-148       rstatix_0.7.0      forcats_0.5.1      foreign_0.8-80
## [45] tools_4.0.1        data.table_1.14.0  hms_1.1.0          lifecycle_1.0.0
## [49] stringr_1.4.0      munsell_0.5.0      zip_2.2.0          compiler_4.0.1
## [53] rlang_0.4.11       grid_4.0.1         iterators_1.0.13   labeling_0.4.2
## [57] rmarkdown_2.8      gtable_0.3.0       codetools_0.2-16   abind_1.4-5
## [61] DBI_1.1.1          reshape_0.8.8      curl_4.3.1         R6_2.5.0
## [65] zoo_1.8-9          gridExtra_2.3      knitr_1.33         dplyr_1.0.6
## [69] survMisc_0.5.5     utf8_1.2.1         shape_1.4.6        stringi_1.6.2
## [73] Rcpp_1.0.6         vctrs_0.3.8        tidyselect_1.1.1   xfun_0.23
```