Week11

Daniel Granja

08/12/2021

Aim

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

#Preprocessing ##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(ISwR)         # for the dataset

d<-graft.vs.host

#Select variables of interest
d[c(1,2,6)]<-NULL
```

##Data explanation > This data aims to find the variables that explain bone marrow transplant survival times in relation with graft versus host disease (GHVD).

```
str(d)
```

```
## 'data.frame':    37 obs. of  6 variables:
##  $ 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 ...
##  $ 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)
```

```
##      donage           type            preg             gvhd       
##  Min.   :14.00   Min.   :1.000   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:20.00   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:0.0000  
##  Median :23.00   Median :2.000   Median :0.0000   Median :0.0000  
##  Mean   :25.81   Mean   :1.973   Mean   :0.2703   Mean   :0.4595  
##  3rd Qu.:34.00   3rd Qu.:3.000   3rd Qu.:1.0000   3rd Qu.:1.0000  
##  Max.   :43.00   Max.   :3.000   Max.   :1.0000   Max.   :1.0000  
##       time             dead       
##  Min.   :  41.0   Min.   :0.0000  
##  1st Qu.: 177.0   1st Qu.:0.0000  
##  Median : 667.0   Median :0.0000  
##  Mean   : 669.8   Mean   :0.4865  
##  3rd Qu.:1105.0   3rd Qu.:1.0000  
##  Max.   :1504.0   Max.   :1.0000
```

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

```r
d$type <-as.factor(d$type) #type of leukaemia coded 1: AML, 2: ALL, 3: CML for acute myeloid, acute lym
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")
str(d)
```

```
## 'data.frame':    37 obs. of  6 variables:
##  $ 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 ...
##  $ 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 ...
```

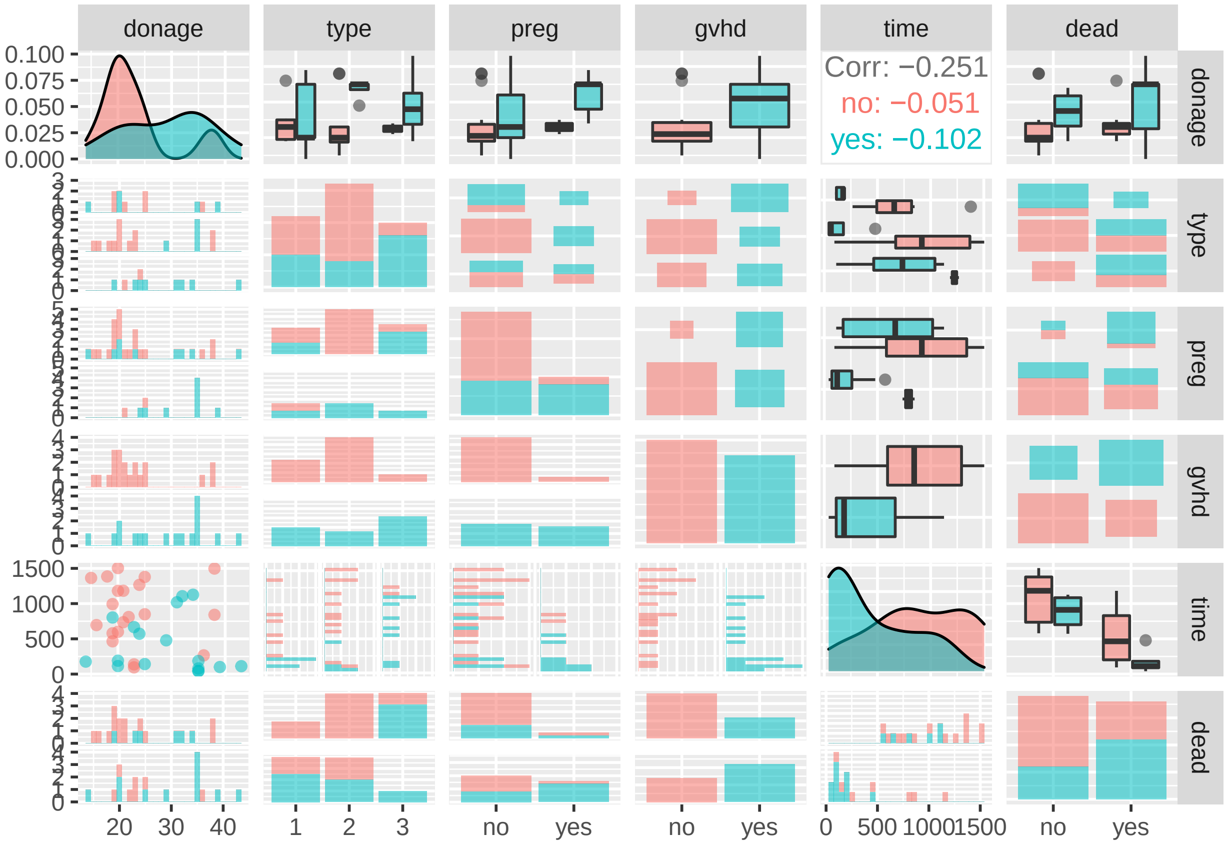
```r
summary(d)
```

```
##      donage       type    preg     gvhd          time          dead   
##  Min.   :14.00   1:11   no :27   no :20   Min.   :  41.0   no :19  
##  1st Qu.:20.00   2:16   yes:10   yes:17   1st Qu.: 177.0   yes:18  
##  Median :23.00   3:10                     Median : 667.0           
##  Mean   :25.81                            Mean   : 669.8           
##  3rd Qu.:34.00                            3rd Qu.:1105.0           
##  Max.   :43.00                            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`.
```



#Modelization

```
m1<-lm(time~ gvhd,data=d)
summary(m1)
```

```
##
## Call:
## lm(formula = time ~ gvhd, data = d)
##
## Residuals:
##     Min      1Q Median      3Q     Max
## -796.2 -297.3  -81.2  374.8  716.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   891.20      94.79   9.402 4.15e-11 ***
## gvhdyes      -481.91     139.84  -3.446   0.0015 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 423.9 on 35 degrees of freedom
## Multiple R-squared:  0.2533, Adjusted R-squared:  0.232
## F-statistic: 11.88 on 1 and 35 DF,  p-value: 0.001496
```

The p-value is significant for this model

#Adding other variables

```
m2<-lm( time~ gvhd+preg, data=d)
summary(m2)
```

```
##
## Call:
## lm(formula = time ~ gvhd + preg, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -826.47 -321.47  -63.05  322.95  582.53
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   921.47      93.36   9.870 1.63e-11 ***
## gvhdyes      -369.74     149.00  -2.482   0.0182 *
## pregyes      -302.68     167.20  -1.810   0.0791 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 410.8 on 34 degrees of freedom
## Multiple R-squared:  0.319,  Adjusted R-squared:  0.2789
## F-statistic: 7.963 on 2 and 34 DF,  p-value: 0.001458
```

```
anova(m1,m2)
```

```
## Analysis of Variance Table
##
## Model 1: time ~ gvhd
## Model 2: time ~ gvhd + preg
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1     35 6289317
## 2     34 5736398  1    552919 3.2772 0.0791 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
m3<-lm( time~ gvhd+type, data=d)
summary(m3)
```

```
##
## Call:
## lm(formula = time ~ gvhd + type, data = d)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -798.4 -198.4  -10.9  222.3  630.4
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    747.6      130.6   5.724 2.18e-06 ***
## gvhdyes       -636.7      139.8  -4.554 6.82e-05 ***
## type2          145.8      151.0   0.966   0.3410
## type3          561.2      172.3   3.258   0.0026 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 378.4 on 33 degrees of freedom
## Multiple R-squared:  0.4389, Adjusted R-squared:  0.3879
## F-statistic: 8.605 on 3 and 33 DF,  p-value: 0.0002323
```

```
anova(m1,m3)
```

```
## Analysis of Variance Table
##
## Model 1: time ~ gvhd
## Model 2: time ~ gvhd + type
##   Res.Df     RSS Df Sum of Sq      F   Pr(>F)
## 1     35 6289317
## 2     33 4726155  2   1563161 5.4573 0.008962 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
m4<-lm( time~ gvhd+donage, data=d)
summary(m4)
```

```
##
## Call:
## lm(formula = time ~ gvhd + donage, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -796.63 -303.69  -85.89  351.03  738.04
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  989.781    247.161   4.005  0.00032 ***
## gvhdyes     -456.727    152.987  -2.985  0.00522 **
## donage        -4.268      9.861  -0.433  0.66792
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 428.9 on 34 degrees of freedom
## Multiple R-squared:  0.2574, Adjusted R-squared:  0.2138
## F-statistic: 5.894 on 2 and 34 DF,  p-value: 0.006345
```

```
anova(m1,m4)
```

```
## Analysis of Variance Table
##
## Model 1: time ~ gvhd
## Model 2: time ~ gvhd + donage
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1     35 6289317
## 2     34 6254863  1     34454 0.1873 0.6679
```

Only the addition of the type IV brings something more to the model as per the anova tests' significance.

#Adding interaction

```
m31<-lm( time~ gvhd*type, data=d)
summary(m31)
```

```
##
```

```
## Call:
## lm(formula = time ~ gvhd * type, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -830.25 -121.25  -26.25  255.75  665.83
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     712.17     158.00   4.508 8.76e-05 ***
## gvhdyes        -558.77     234.34  -2.384   0.0234 *
## type2           213.08     193.50   1.101   0.2793
## type3           511.83     315.99   1.620   0.1154
## gvhdyes:type2  -205.23     323.79  -0.634   0.5308
## gvhdyes:type3    28.02     385.39   0.073   0.9425
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 387 on 31 degrees of freedom
## Multiple R-squared:  0.4488, Adjusted R-squared:  0.3599
## F-statistic: 5.048 on 5 and 31 DF,  p-value: 0.001691
```

```
anova(m3,m31)
```

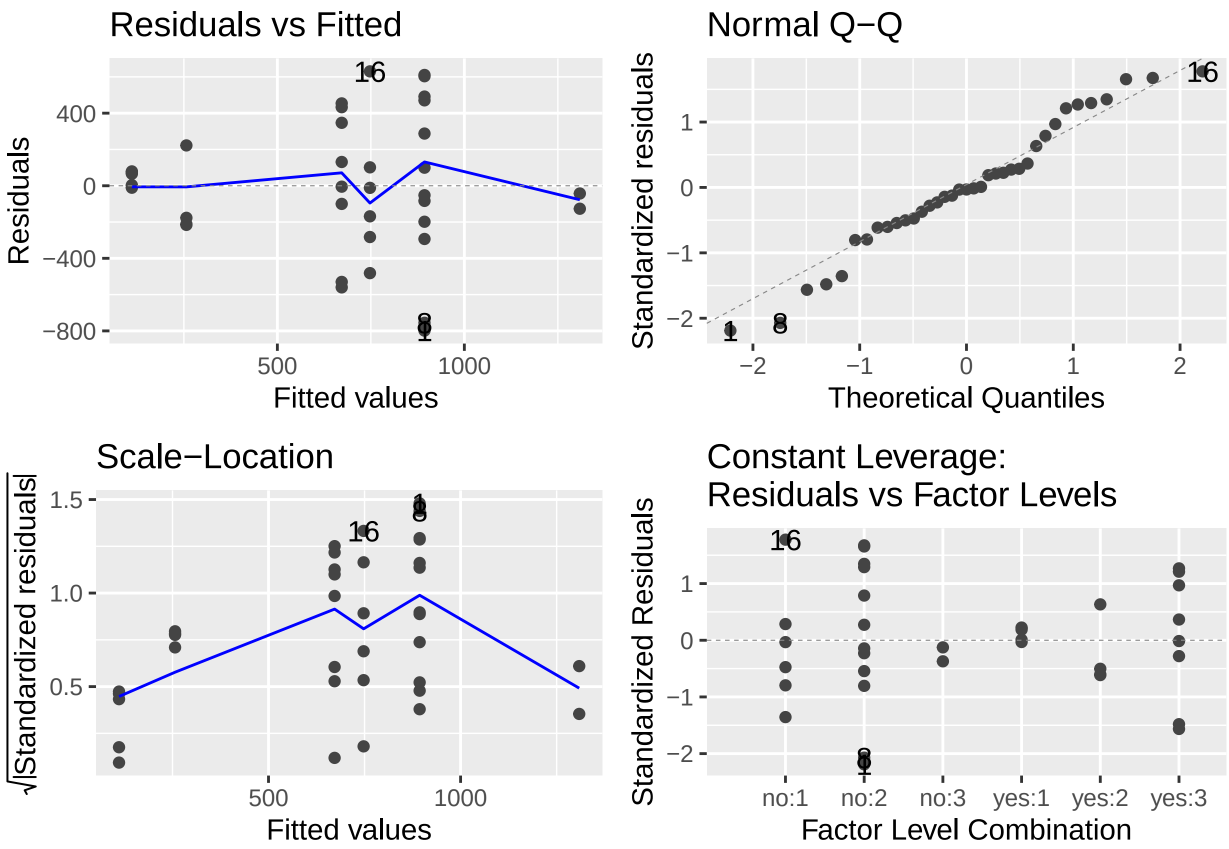
```
## Analysis of Variance Table
##
## Model 1: time ~ gvhd + type
## Model 2: time ~ gvhd * type
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1     33 4726155
## 2     31 4643015  2     83141 0.2776 0.7595
```

The interaction doesn't bring much (p =< 0.75). The best model is m3: " Y = 747.6 + (-636.7)*gvhdyes + 145.8*type2 + 561.2*type3 ".

#Controling the postulates

```
autoplot(m3)
```

```
## Warning: `arrange_()` was deprecated in dplyr 0.7.0.
## Please use `arrange()` instead.
## See vignette('programming') for more help
```

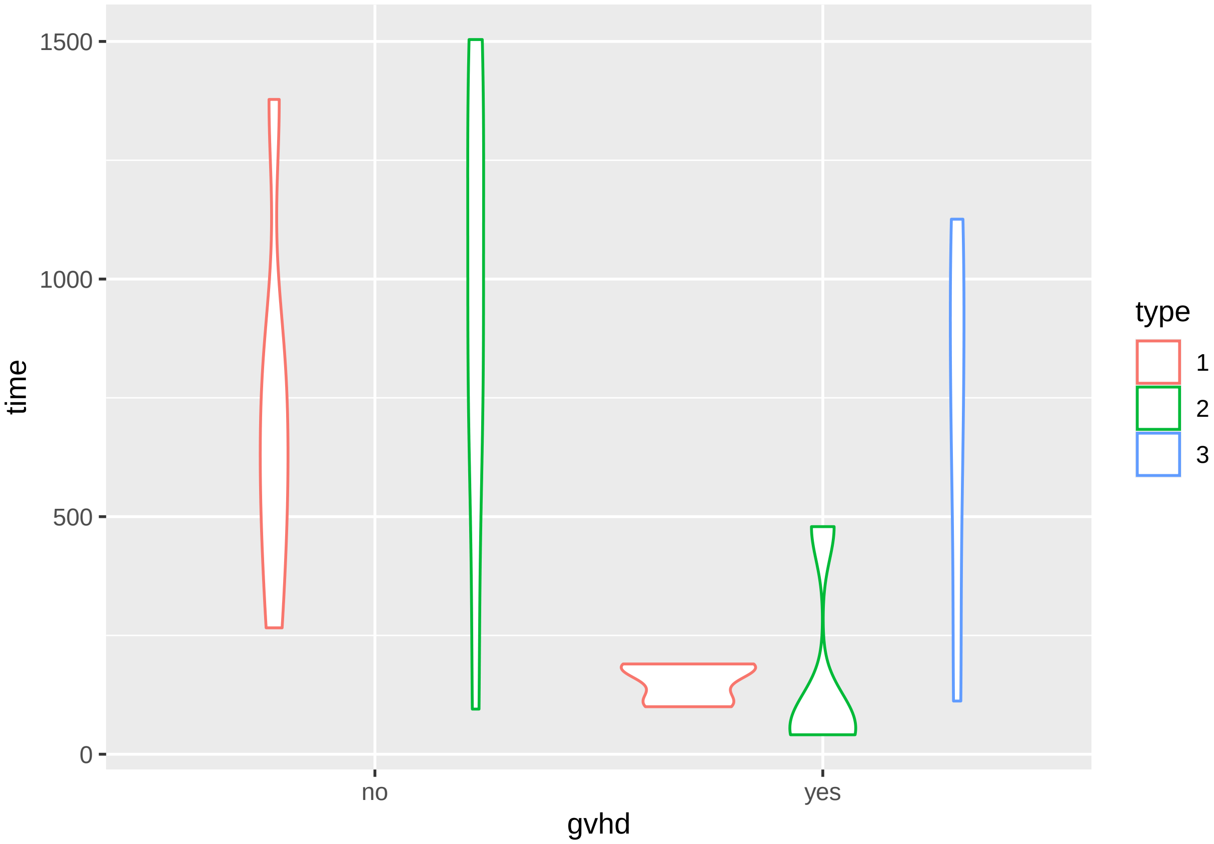


>

The postulates seem ok, even if we can discuss some bortherline observations.

#Representation

```
ggplot(data=d, aes(x=gvhd,y=time, colour=type))+geom_violin()
```



In conclusion, we can predict that the graft-versus-host desease reduce survival time when it's present, and the acute lymphatic leukaemia (type 2) and the chronic myeloid leukaemia (type 3) are associated with a better survival time than acute myeloid leukaemia.

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] ISwR_2.0-8       ggfortify_0.4.11 GGally_2.1.1     ggplot2_3.3.3
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.6         highr_0.9          pillar_1.6.1       compiler_4.0.1
##  [5] RColorBrewer_1.1-2 plyr_1.8.6         tools_4.0.1        digest_0.6.27
##  [9] evaluate_0.14      lifecycle_1.0.0    tibble_3.1.2       gtable_0.3.0
## [13] pkgconfig_2.0.3    rlang_0.4.11       DBI_1.1.1          yaml_2.2.1
## [17] xfun_0.23          gridExtra_2.3      withr_2.4.2        stringr_1.4.0
## [21] dplyr_1.0.6        knitr_1.33         generics_0.1.0     vctrs_0.3.8
## [25] grid_4.0.1         tidyselect_1.1.1   reshape_0.8.8      glue_1.4.2
## [29] R6_2.5.0           fansi_0.5.0        rmarkdown_2.8      farver_2.1.0
## [33] tidyr_1.1.3        purrr_0.3.4        magrittr_2.0.1     scales_1.1.1
## [37] ellipsis_0.3.2     htmltools_0.5.1.1  assertthat_0.2.1   colorspace_2.0-1
## [41] labeling_0.4.2     utf8_1.2.1         stringi_1.6.2      munsell_0.5.0
## [45] crayon_1.4.1
```