Survival Analysis in Pediatric Patients with Acute Lymphoblastic Leukemia in Argentina- Assumptions-Statistical testing

Assessment of Assumptions and Statistical Testing

```r
#Load required packages
library(readxl)
library(ggplot2)
library(car)
library(dplyr)
```

```r
    #Load data set
data_lla <- read_excel("data_lla.xlsx")
#Verify dimensions and first rows
dim(data_lla)
```

```
## [1] 2096   17
```

```r
head(data_lla,3)
```

```
## # A tibble: 3 x 17
##     Sexo  Down Blancos Blastos    MO   SNC Ploidia Estirpe RTA_PRED CAT_ERM  Edad
##    <dbl> <dbl>   <dbl> <chr>   <dbl> <dbl>   <dbl> <chr>      <dbl>   <dbl> <dbl>
## 1      1     0       3 2          90     1       1 B              1      NA     2
## 2      1     0      36 48         50     1       1 T              0       3    14
## 3      0     0      10 95         98     1       2 B              0       3     4
## # i 6 more variables: MLL <dbl>, TIEMPOSG <dbl>, SGSTATUS <dbl>,
## #   Edad_cat <dbl>, Blancos_cat <dbl>, TEL <dbl>
```

```r
#Rename variables
data_lla <- data_lla %>%
  rename(
    Sex = Sexo,
    Age_cat = Edad_cat,
    WBC_cat = Blancos_cat,
    DownSyndrome = Down,
    CNS = SNC,
    Lineage = Estirpe,
    TEL_gene = TEL,
    MLL_gene = MLL,
    Blasts_PB = Blastos,
    Blasts_BM = MO,
    Ploidy = Ploidia,
    Pred_Response = RTA_PRED,
```

```
    MRD_cat = CAT_ERM,
    OS_status = SGSTATUS
  )


#Categorical variables
data_lla <- data_lla %>%
  mutate_at(c("Sex","DownSyndrome","MLL_gene","Pred_Response",
              "Lineage","TEL_gene","CNS","Age_cat","WBC_cat",
              "Ploidy","MRD_cat"), factor)


#Continuous variables
data_lla <- data_lla %>%
  mutate(
    Blasts_PB = as.numeric(Blasts_PB),
    Blasts_BM = as.numeric(Blasts_BM),
    Edad = as.numeric(Edad)
  )


#Save data set in csv file
write.csv(data_lla, "data_lla.csv", row.names = FALSE)


# Categorical variables


# MRD_cat vs DownSyndrome

table1 <- table(data_lla$MRD_cat, data_lla$DownSyndrome)
prop.table(table1)


##
##              0          1
##   1 0.370264064 0.006314581
##   2 0.455797933 0.006888634
##   3 0.156716418 0.004018370

plot(table1, col=c("red","blue"), main="MRD categories vs Down syndrome")
```
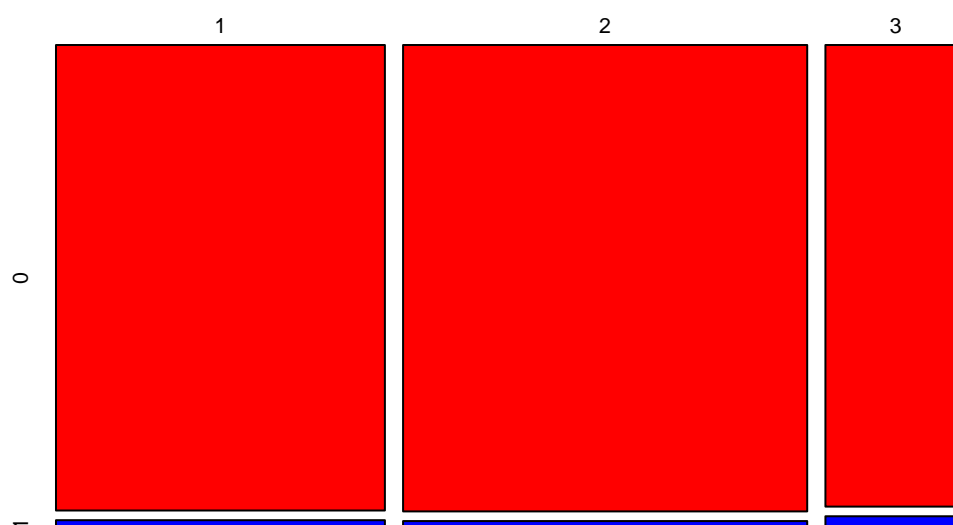
MRD categories vs Down syndrome


```r
# Chi-square test

chisq.test(table1)
```

```
## Warning in chisq.test(table1): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  table1
## X-squared = 1.2682, df = 2, p-value = 0.5304
```

```r
# If assumptions are not met, Fisher's exact test

fisher.test(table1)
```
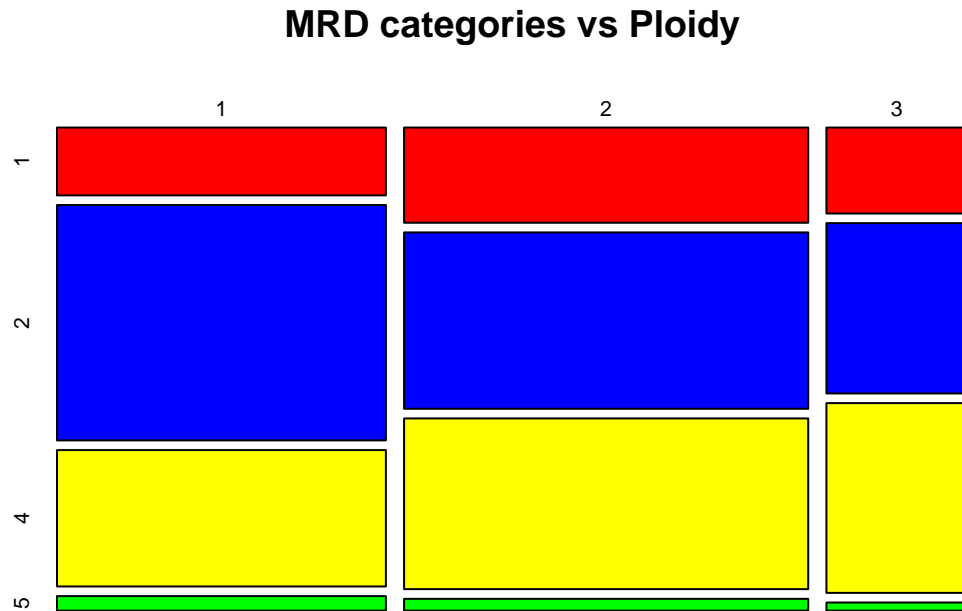
```
##
##  Fisher's Exact Test for Count Data
##
## data:  table1
## p-value = 0.479
## alternative hypothesis: two.sided
```

```
# MRD_cat vs Ploidy

table2 <- table(data_lla$MRD_cat, data_lla$Ploidy)
plot(table2, col=c("red","blue","yellow","green"), main="MRD categories vs Ploidy")
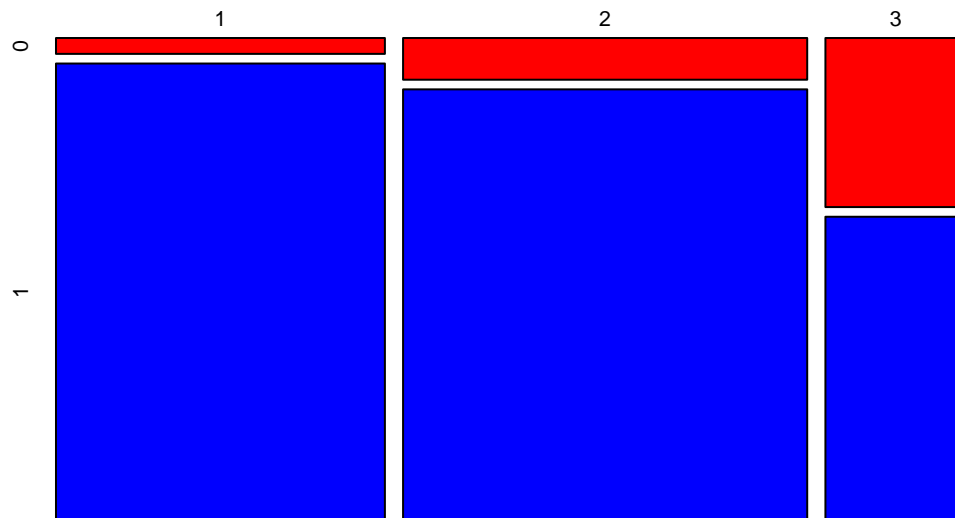```



**MRD categories vs Ploidy**

```
chisq.test(table2)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table2
## X-squared = 35.184, df = 6, p-value = 3.97e-06
```

```
# MRD_cat vs Pred_Response

table3 <- table(data_lla$MRD_cat, data_lla$Pred_Response)
plot(table3, col=c("red","blue","yellow"), main="MRD categories vs Prednisone response")
```

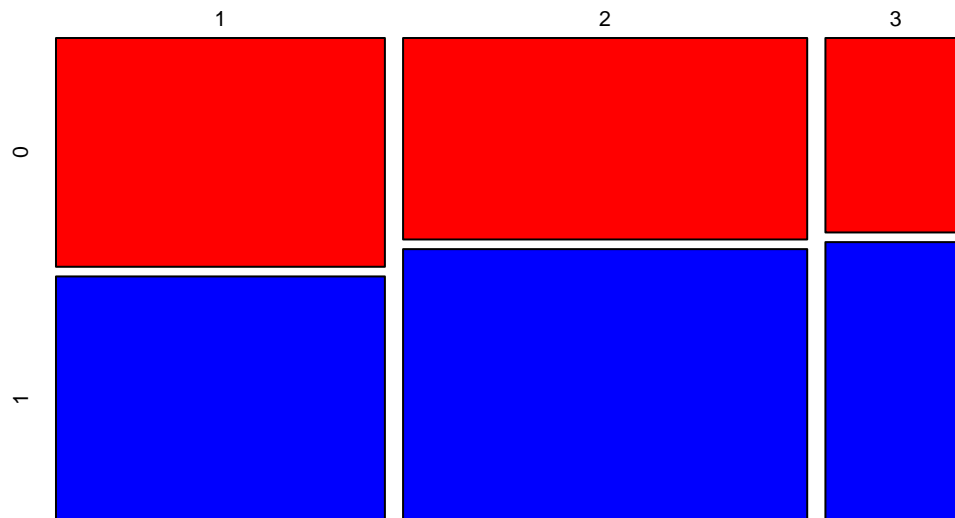## MRD categories vs Prednisone response


```r
chisq.test(table3)
```

```
## 
##  Pearson's Chi-squared test
## 
## data:  table3
## X-squared = 216.44, df = 2, p-value < 2.2e-16
```

```r
# MRD_cat vs Sex

table4 <- table(data_lla$MRD_cat, data_lla$Sex)
plot(table4, col=c("red","blue"), main="MRD categories vs Sex")
```
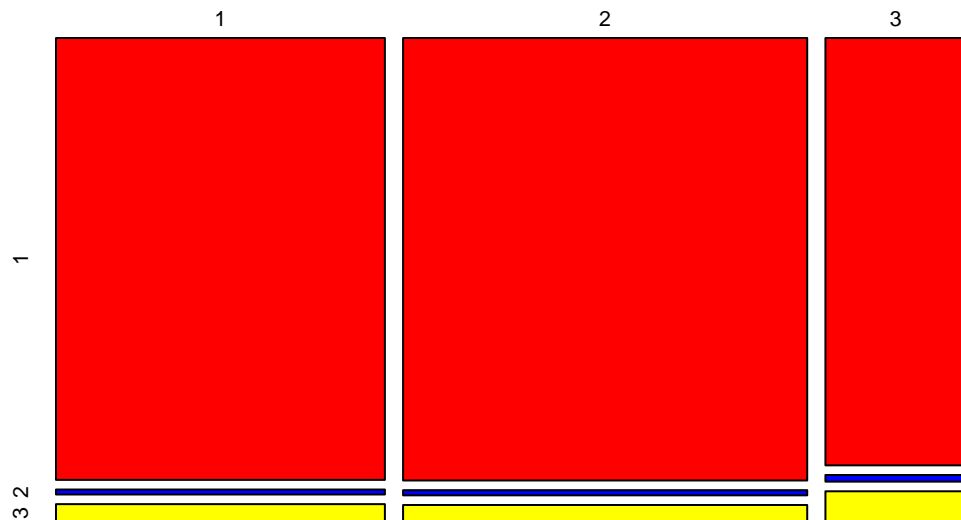
MRD categories vs Sex


```r
chisq.test(table4)
```

```
## 
##  Pearson's Chi-squared test
## 
## data:  table4
## X-squared = 6.4488, df = 2, p-value = 0.03978
```

```r
# MRD_cat vs CNS

table5 <- table(data_lla$MRD_cat, data_lla$CNS)
plot(table5, col=c("red","blue","yellow"), main="MRD categories vs CNS involvement")
```
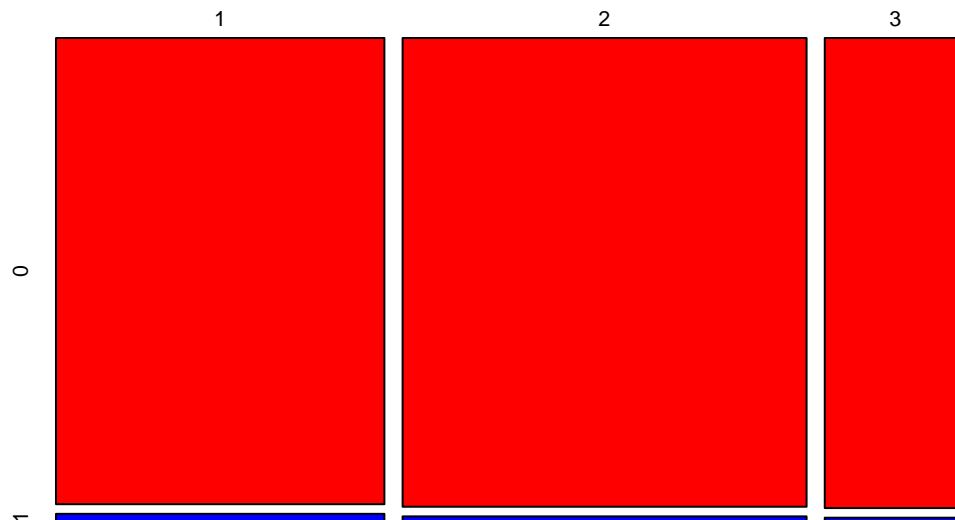
MRD categories vs CNS involvement


```r
fisher.test(table5)  # if assumptions are not met
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table5
## p-value = 0.2638
## alternative hypothesis: two.sided
```

```r
# MRD_cat vs MLL_gene

table6 <- table(data_lla$MRD_cat, data_lla$MLL_gene)
plot(table6, col=c("red","blue"), main="MRD categories vs MLL gene")
```

MRD categories vs MLL gene


```r
chisq.test(table6)
```

```
## Warning in chisq.test(table6): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  table6
## X-squared = 1.4718, df = 2, p-value = 0.4791
```
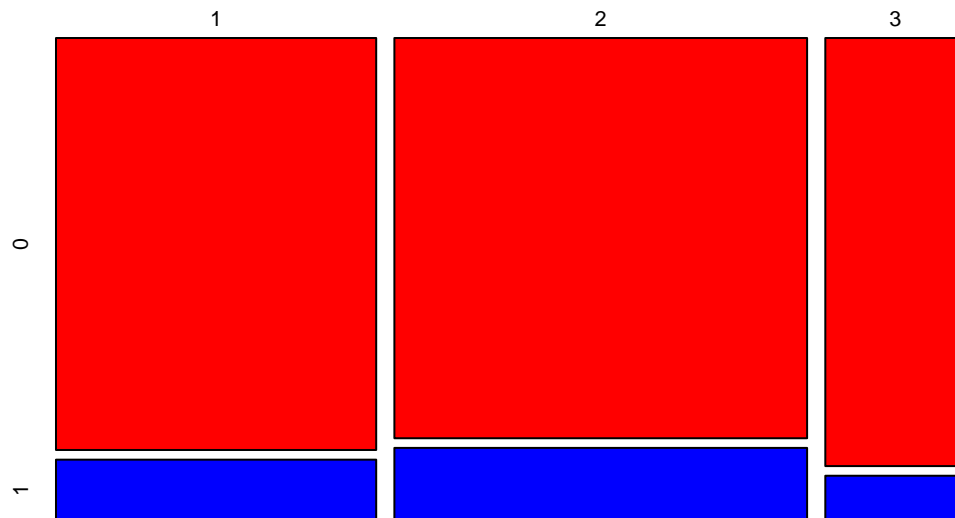
```r
fisher.test(table6)  # if assumptions are not met
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table6
## p-value = 0.5936
## alternative hypothesis: two.sided
```

```r
# MRD_cat vs TEL_gene

table7 <- table(data_lla$MRD_cat, data_lla$TEL_gene)
plot(table7, col=c("red","blue"), main="MRD categories vs TEL gene")
```
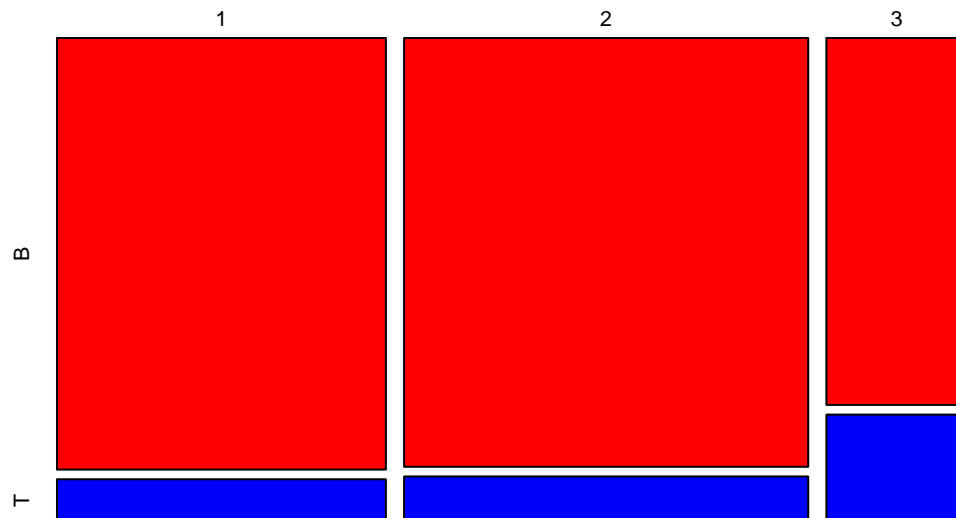
MRD categories vs TEL gene


```r
chisq.test(table7)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table7
## X-squared = 5.5501, df = 2, p-value = 0.06235
```

```r
# MRD_cat vs Lineage

table8 <- table(data_lla$MRD_cat, data_lla$Lineage)
plot(table8, col=c("red","blue"), main="MRD categories vs Lineage")
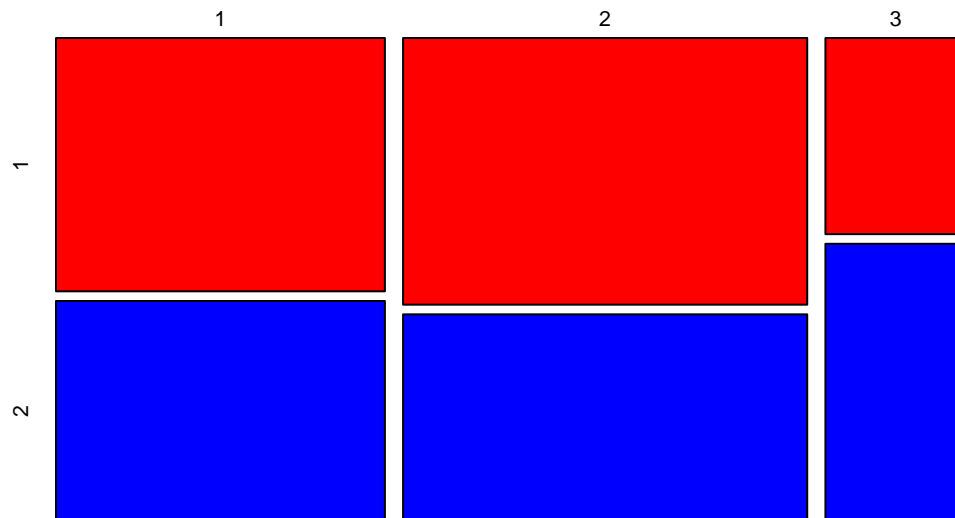```

# MRD categories vs Lineage



```r
chisq.test(table8)
```

```
## 
##  Pearson's Chi-squared test
## 
## data:  table8
## X-squared = 41.784, df = 2, p-value = 8.448e-10
```

```r
# MRD_cat vs Age_cat

table9 <- table(data_lla$MRD_cat, data_lla$Age_cat)
plot(table9, col=c("red","blue"), main="MRD categories vs Age category")
```
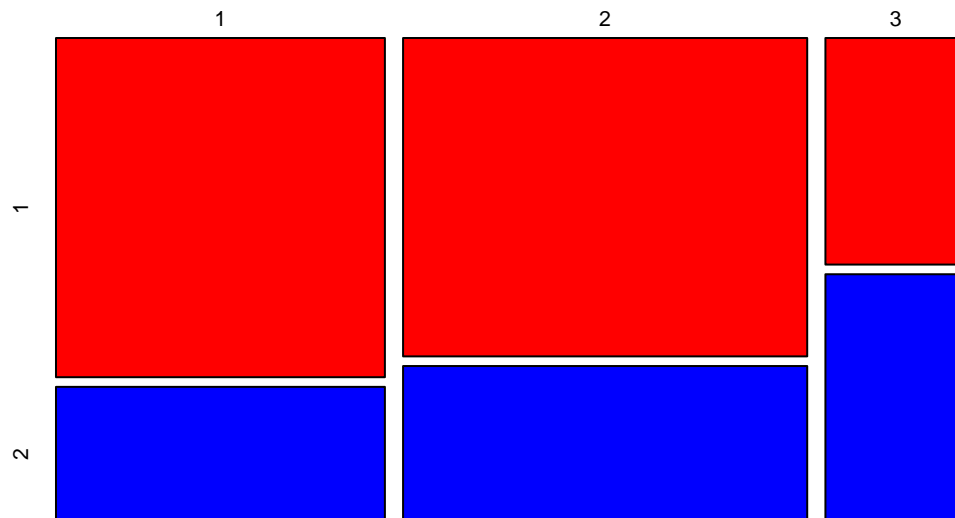
MRD categories vs Age category


```r
chisq.test(table9)
```

```
## 
##  Pearson's Chi-squared test
## 
## data:  table9
## X-squared = 18.683, df = 2, p-value = 8.769e-05
```

```r
# MRD_cat vs WBC_cat

table10 <- table(data_lla$MRD_cat, data_lla$WBC_cat)
plot(table10, col=c("red","blue"), main="MRD categories vs WBC category")
```

MRD categories vs WBC category
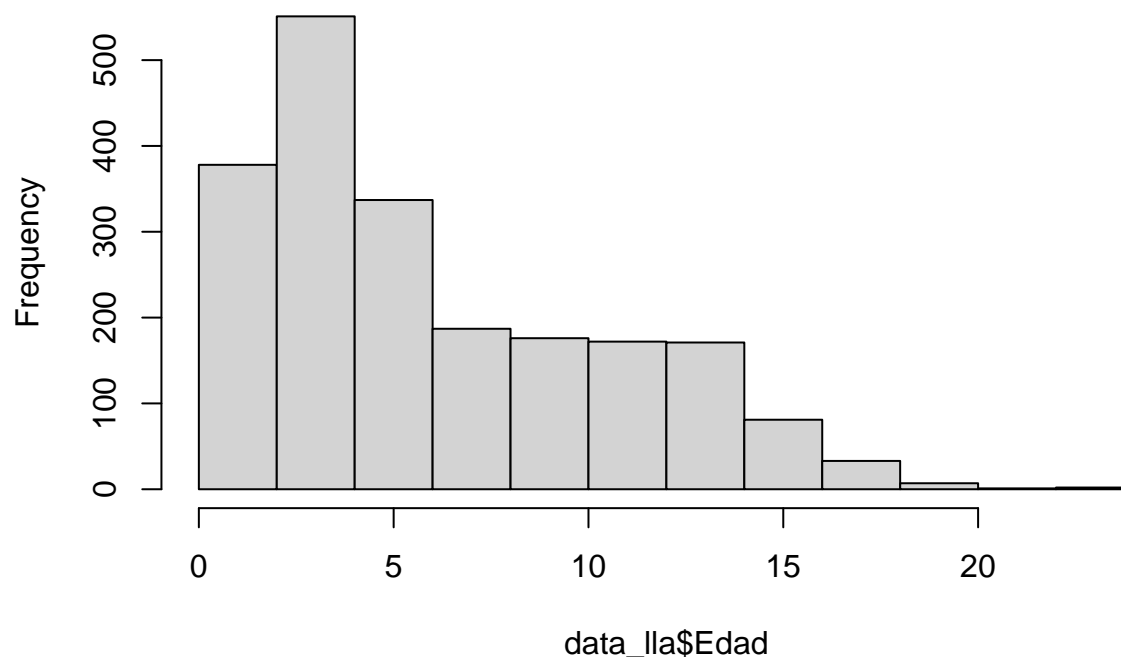


```r
chisq.test(table10)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table10
## X-squared = 50.76, df = 2, p-value = 9.495e-12
```

```r
# Continuous variables

# Age

hist(data_lla$Edad)
```

Histogram of data_lla$Edad
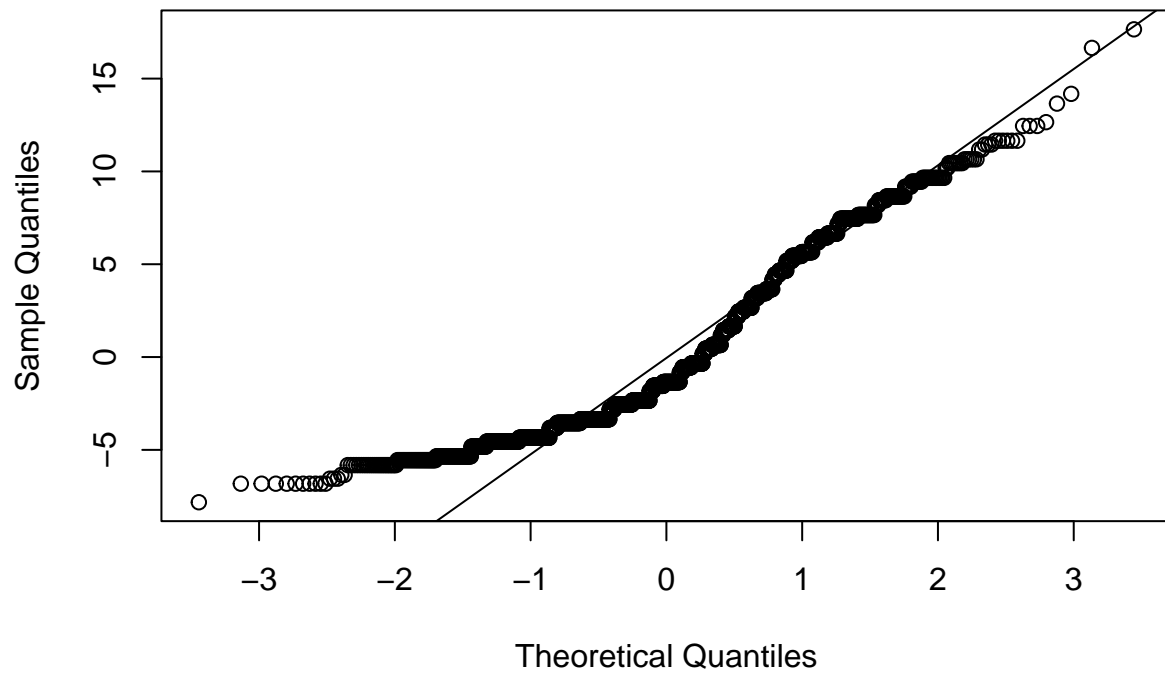


```r
shapiro.test(data_lla$Edad)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data_lla$Edad
## W = 0.90726, p-value < 2.2e-16
```

```r
mod_age <- lm(Edad ~ MRD_cat, data = data_lla)
resid_age <- resid(mod_age)
rstd_age <- rstandard(mod_age)
pred_age <- fitted(mod_age)
resid_data_age <- data.frame(resid_age, rstd_age, pred_age)

shapiro.test(resid_age)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid_age
## W = 0.91802, p-value < 2.2e-16
```
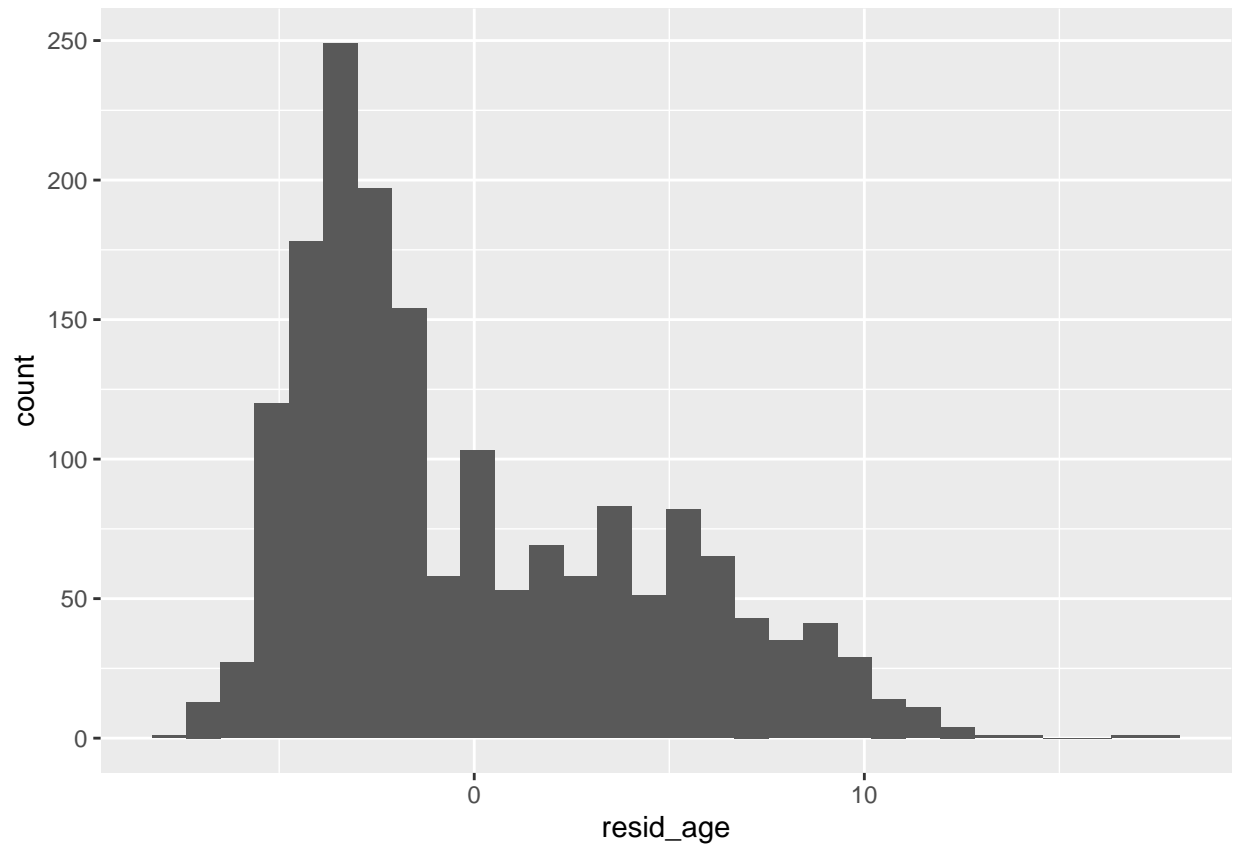
```r
qqnorm(resid_age); qqline(resid_age)
```
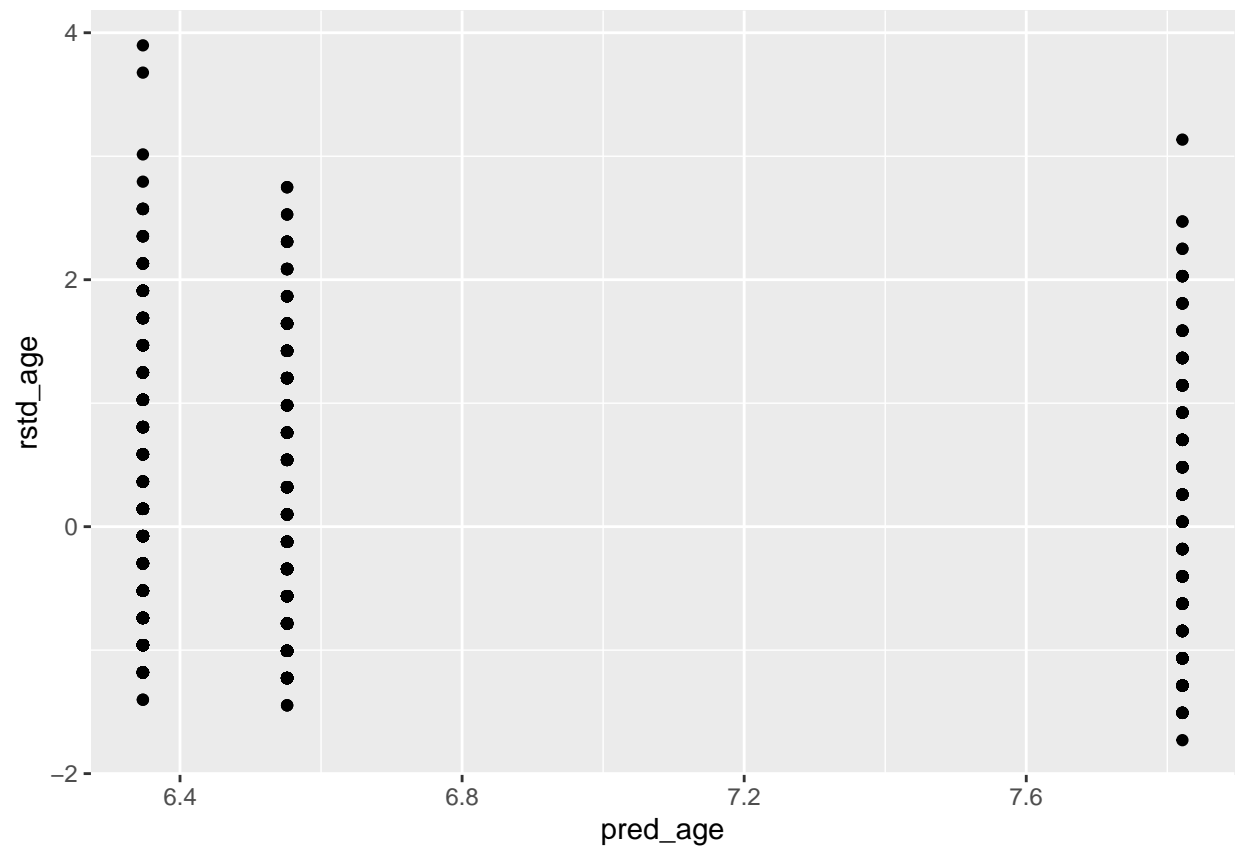
Normal Q–Q Plot


```r
ggplot(resid_data_age, aes(x=resid_age)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value `binwidth`.
```
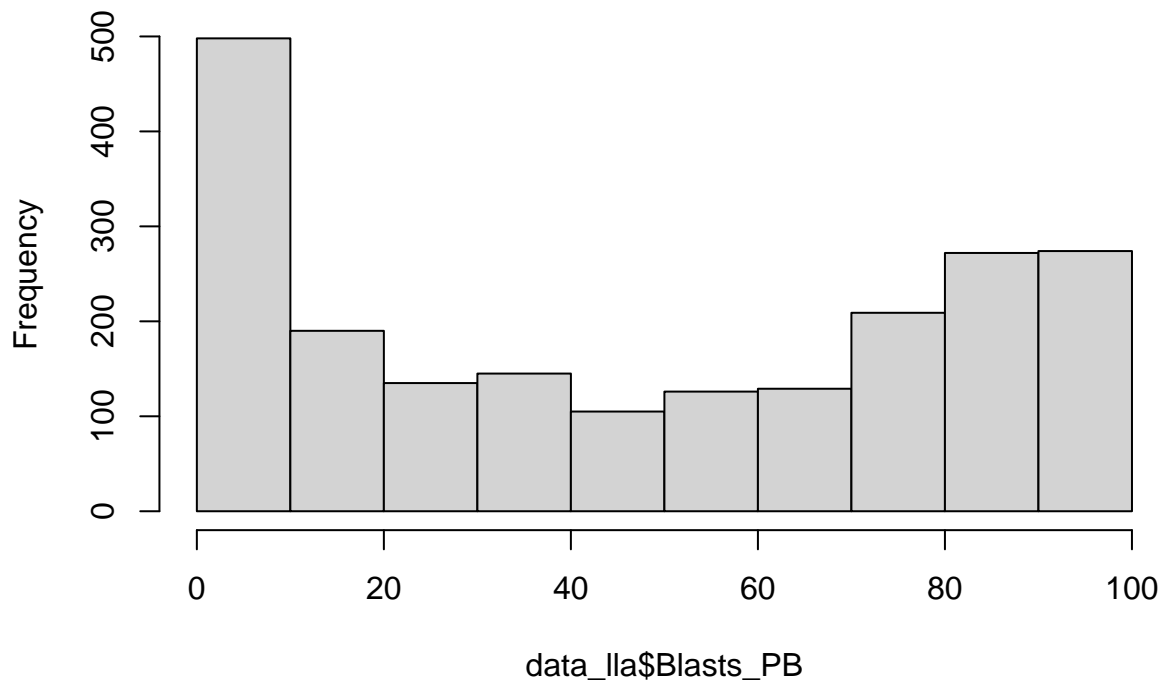
```r
ggplot(resid_data_age, aes(x=pred_age, y=rstd_age)) + geom_point()
```

```
# Blasts_PB

hist(data_lla$Blasts_PB)
```

Histogram of data_lla$Blasts_PB


```r
shapiro.test(data_lla$Blasts_PB)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data_lla$Blasts_PB
## W = 0.89253, p-value < 2.2e-16
```

```r
mod_blastsPB <- lm(Blasts_PB ~ MRD_cat, data = data_lla)
resid_blastsPB <- resid(mod_blastsPB)
shapiro.test(resid_blastsPB)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid_blastsPB
## W = 0.91901, p-value < 2.2e-16
```

```r
# Blasts_BM

hist(data_lla$Blasts_BM)
```

Histogram of data_lla$Blasts_BM


```r
shapiro.test(data_lla$Blasts_BM)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data_lla$Blasts_BM
## W = 0.74845, p-value < 2.2e-16
```
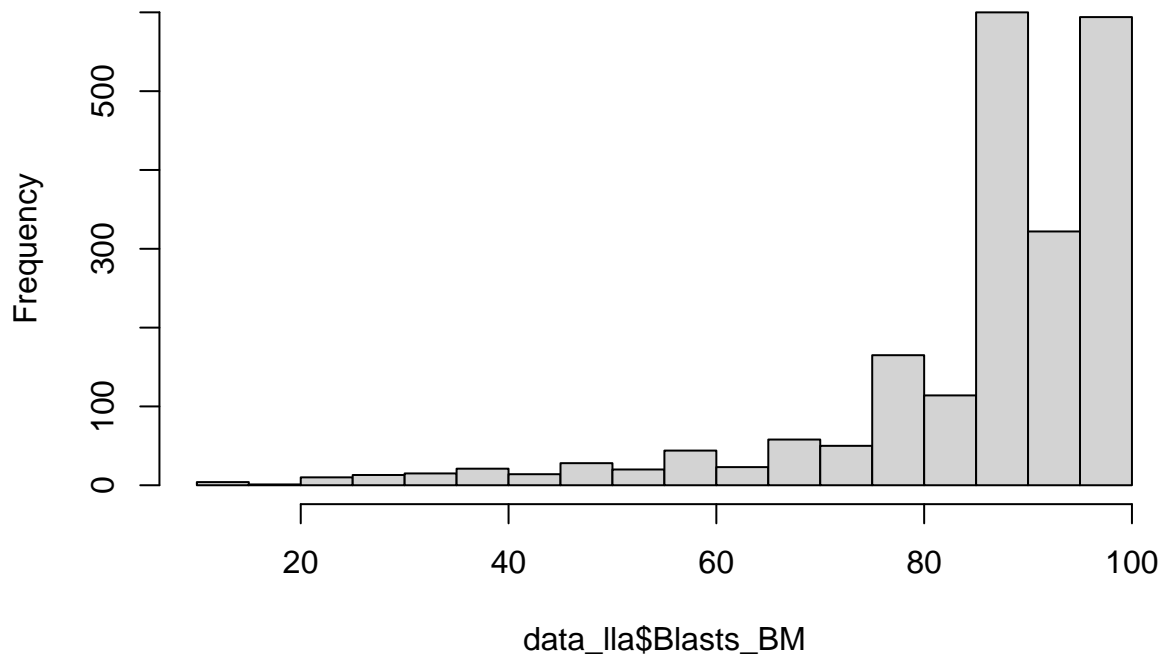
```r
mod_blastsBM <- lm(Blasts_BM ~ MRD_cat, data = data_lla)
resid_blastsBM <- resid(mod_blastsBM)
shapiro.test(resid_blastsBM)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid_blastsBM
## W = 0.75815, p-value < 2.2e-16
```

```r
## Homoscedasticity

leveneTest(Edad ~ MRD_cat, data = data_lla)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value  Pr(>F)
```

```
## group     2  4.1252 0.01632 *
##        1739
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
leveneTest(Blasts_PB ~ MRD_cat, data = data_lla)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value Pr(>F)
## group    2  0.0868 0.9168
##        1726
```

```
leveneTest(Blasts_BM ~ MRD_cat, data = data_lla)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value  Pr(>F)
## group    2  3.4411 0.03225 *
##        1739
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In this case of the following variables: Blasts in Peripheral Blood (Blasts_PB): does not meet normality or homoscedasticity assumptions. Therefore, a Median test was applied. Blasts in Bone Marrow (MO): does not meet the assumption of normality but meets homoscedasticity. Therefore, a Wilcoxon/Mann-Whitney test is applied.

```
library(coin)
```

```
## Warning: package 'coin' was built under R version 4.4.2
```

```
## Loading required package: survival
```

```
data_lla$OS_status <- factor(data_lla$OS_status)

#Median test for Blasts_PB
median_test(Blasts_PB ~ OS_status, data = data_lla)
```

```
##
##  Asymptotic Two-Sample Brown-Mood Median Test
##
## data:  Blasts_PB by OS_status (0, 1)
## Z = -4.0017, p-value = 6.29e-05
## alternative hypothesis: true mu is not equal to 0
```

```
#Wilcoxon test for MO
wilcox.test(Blasts_BM ~ OS_status, data = data_lla)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Blasts_BM by OS_status
## W = 369200, p-value = 0.3045
## alternative hypothesis: true location shift is not equal to 0
```