final project

Chrissy Chen

2024-11-20

1. Data cleaning

```
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(skimr)
library(DataExplorer)
library(survival)
library(survminer)
```

```
## Loading required package: ggpubr
##
## Attaching package: 'survminer'
##
## The following object is masked from 'package:survival':
##
##     myeloma
```

```r
library(MASS)
```

```
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:dplyr':
##
##     select
```

```r
library(knitr)
cirrhosis <- read_csv("cirrhosis.csv")
```

```
## Rows: 418 Columns: 20
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (7): Status, Drug, Sex, Ascites, Hepatomegaly, Spiders, Edema
## dbl (13): ID, N_Days, Age, Bilirubin, Cholesterol, Albumin, Copper, Alk_Phos...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
# Recode the Status variable
cirrhosis$Surv_Status <- ifelse(cirrhosis$Status == "D", 1, 0)
```

1. EDA

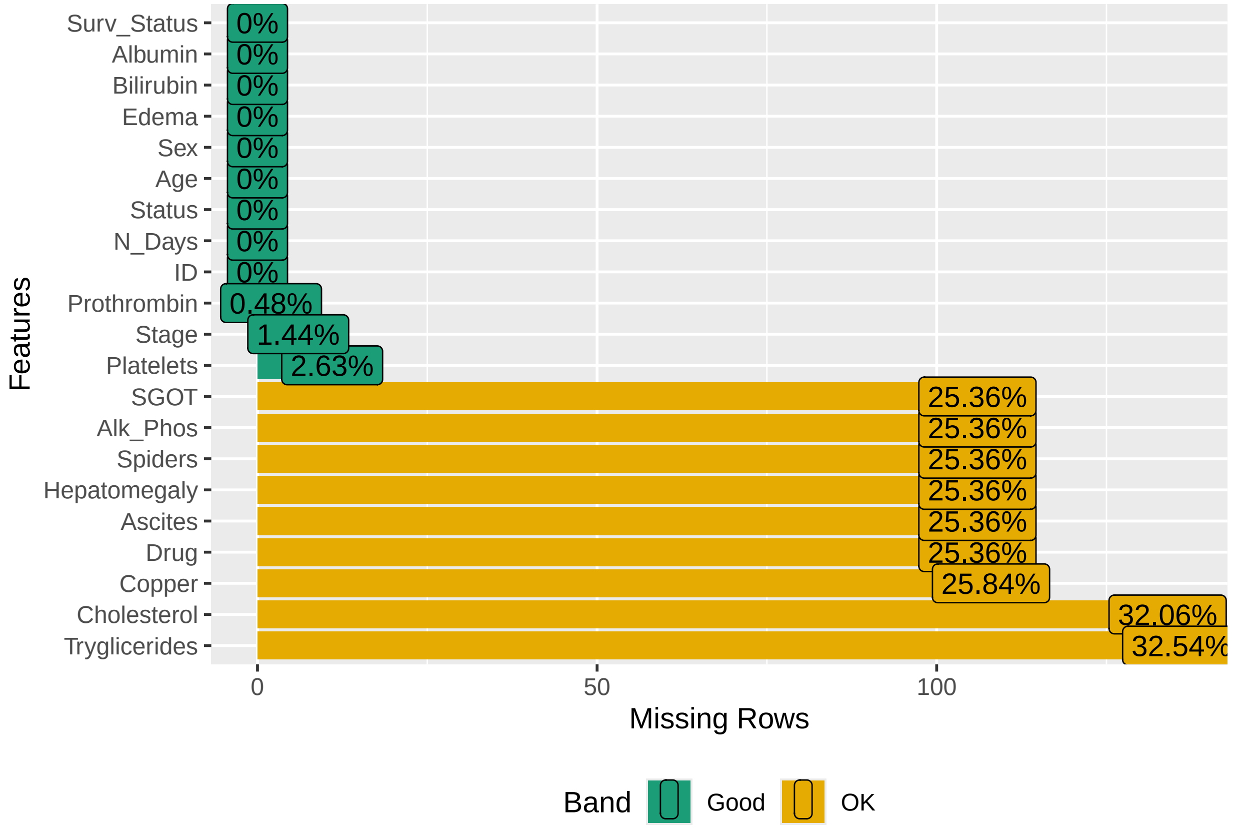
```r
# Summary statistics of each column
summary(cirrhosis)
```

```
##        ID            N_Days        Status              Drug
##  Min.   :  1.0   Min.   :  41   Length:418         Length:418
##  1st Qu.:105.2   1st Qu.:1093   Class :character   Class :character
##  Median :209.5   Median :1730   Mode  :character   Mode  :character
##  Mean   :209.5   Mean   :1918
##  3rd Qu.:313.8   3rd Qu.:2614
##  Max.   :418.0   Max.   :4795
##
##       Age            Sex              Ascites          Hepatomegaly
##  Min.   : 9598   Length:418         Length:418         Length:418
##  1st Qu.:15644   Class :character   Class :character   Class :character
##  Median :18628   Mode  :character   Mode  :character   Mode  :character
##  Mean   :18533
##  3rd Qu.:21272
##  Max.   :28650
##
##    Spiders             Edema              Bilirubin       Cholesterol
##  Length:418         Length:418         Min.   : 0.300   Min.   : 120.0
##  Class :character   Class :character   1st Qu.: 0.800   1st Qu.: 249.5
##  Mode  :character   Mode  :character   Median : 1.400   Median : 309.5
##                                        Mean   : 3.221   Mean   : 369.5
##                                        3rd Qu.: 3.400   3rd Qu.: 400.0
##                                         Max.   :28.000   Max.   :1775.0  
##                                                          NA's   :134     
##     Albumin          Copper          Alk_Phos            SGOT       
##  Min.   :1.960   Min.   :  4.00   Min.   :  289.0   Min.   : 26.35  
##  1st Qu.:3.243   1st Qu.: 41.25   1st Qu.:  871.5   1st Qu.: 80.60  
##  Median :3.530   Median : 73.00   Median : 1259.0   Median :114.70  
##  Mean   :3.497   Mean   : 97.65   Mean   : 1982.7   Mean   :122.56  
##  3rd Qu.:3.770   3rd Qu.:123.00   3rd Qu.: 1980.0   3rd Qu.:151.90  
##  Max.   :4.640   Max.   :588.00   Max.   :13862.4   Max.   :457.25  
##                  NA's   :108      NA's   :106       NA's   :106     
##  Tryglicerides      Platelets      Prothrombin        Stage      
##  Min.   : 33.00   Min.   : 62.0   Min.   : 9.00   Min.   :1.000  
##  1st Qu.: 84.25   1st Qu.:188.5   1st Qu.:10.00   1st Qu.:2.000  
##  Median :108.00   Median :251.0   Median :10.60   Median :3.000  
##  Mean   :124.70   Mean   :257.0   Mean   :10.73   Mean   :3.024  
##  3rd Qu.:151.00   3rd Qu.:318.0   3rd Qu.:11.10   3rd Qu.:4.000  
##  Max.   :598.00   Max.   :721.0   Max.   :18.00   Max.   :4.000  
##  NA's   :136      NA's   :11      NA's   :2       NA's   :6      
##   Surv_Status    
##  Min.   :0.0000  
##  1st Qu.:0.0000  
##  Median :0.0000  
##  Mean   :0.3852  
##  3rd Qu.:1.0000  
##  Max.   :1.0000  
## 
```

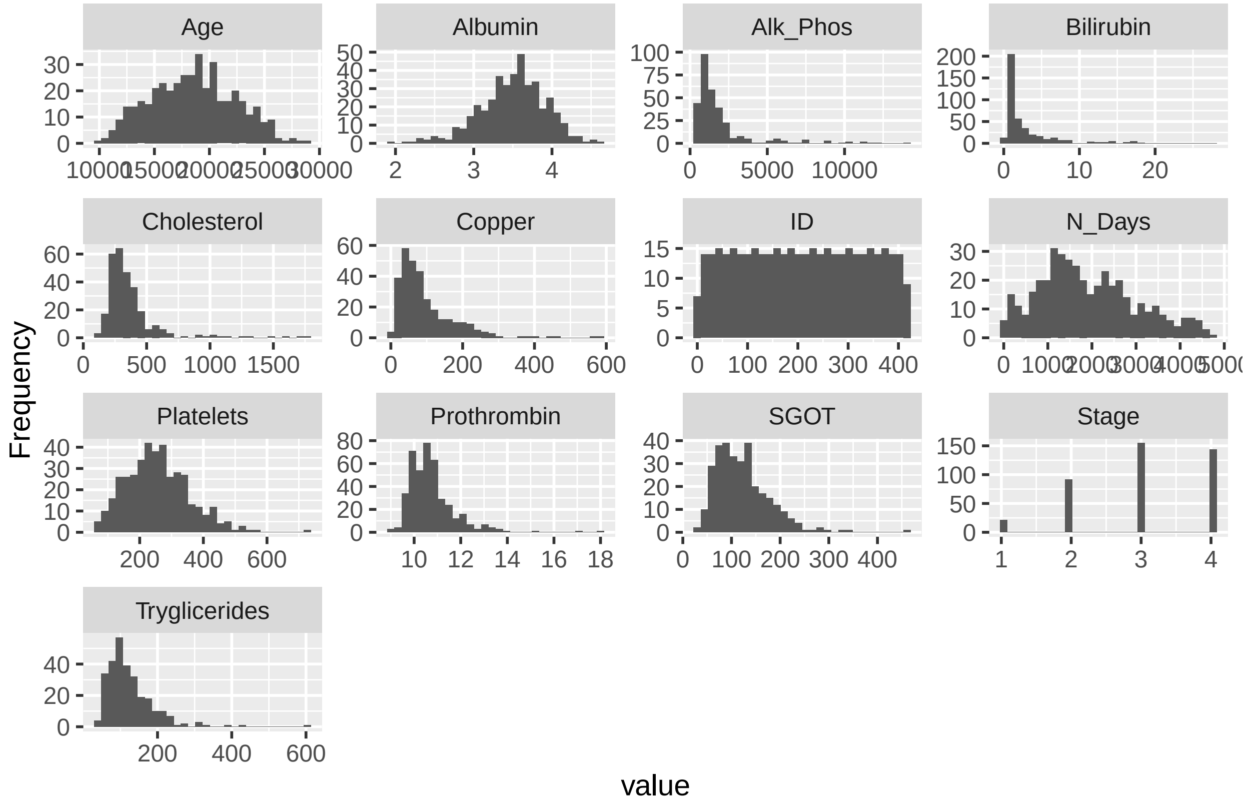
```r
# Identify missing values
colSums(is.na(cirrhosis))
```

```
##            ID        N_Days        Status          Drug           Age 
##             0             0             0           106             0 
##           Sex       Ascites  Hepatomegaly       Spiders         Edema 
##             0           106           106           106             0 
##     Bilirubin   Cholesterol       Albumin        Copper      Alk_Phos 
##             0           134             0           108           106 
##          SGOT Tryglicerides     Platelets   Prothrombin         Stage 
##           106           136            11             2             6 
##   Surv_Status 
##             0 
```

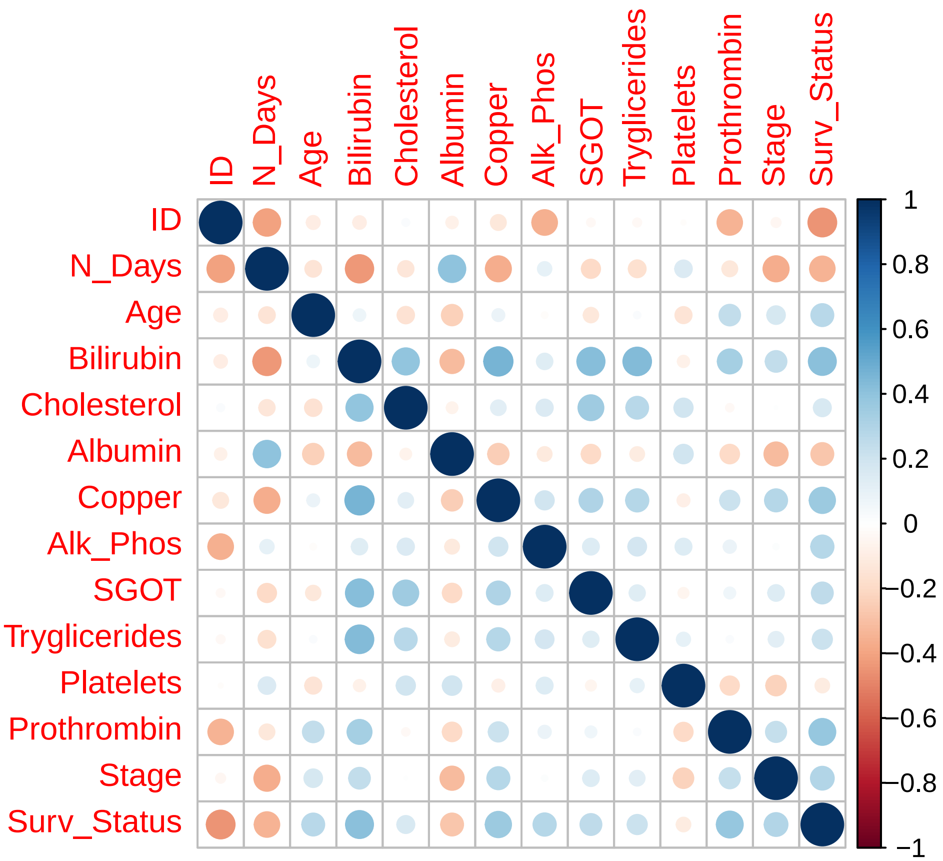
```r
# Visualize missing data
plot_missing(cirrhosis)
```



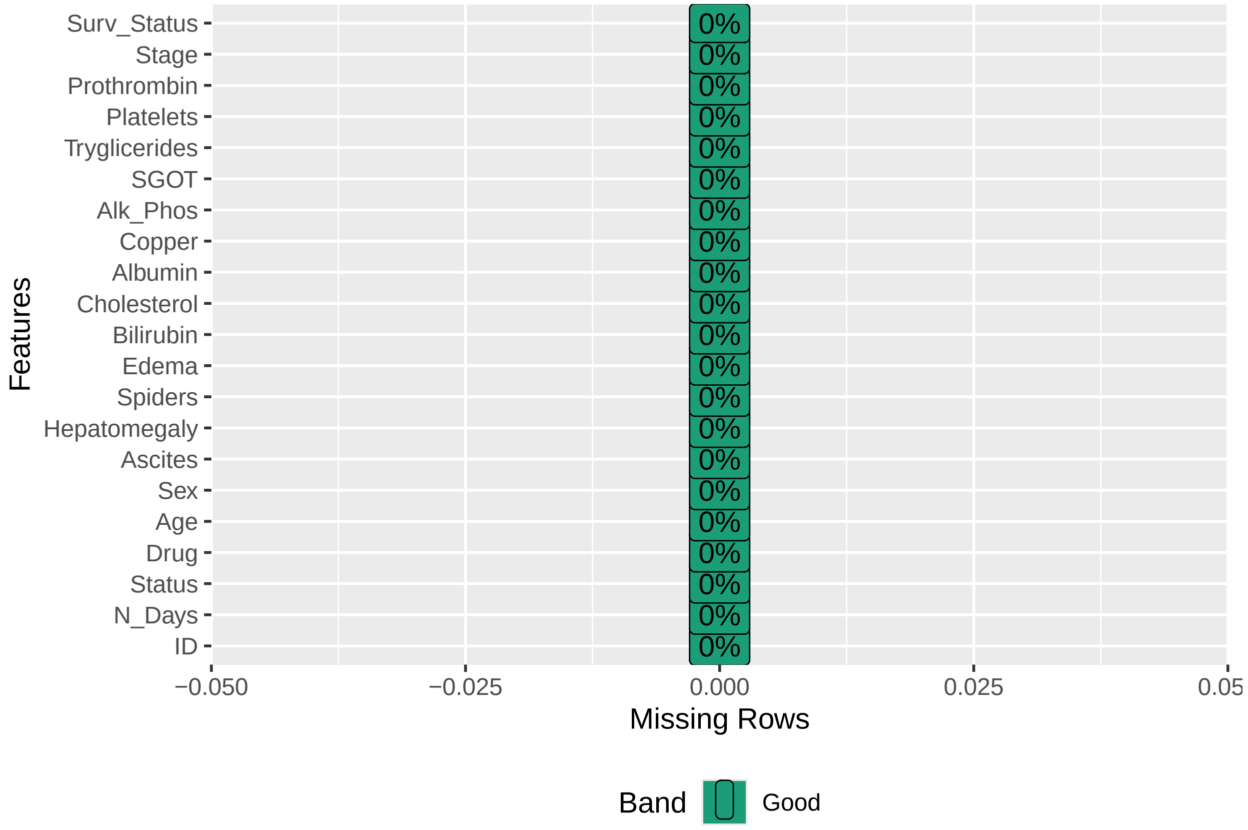
```
# Visualize distributions of numerical columns
plot_histogram(cirrhosis)
```



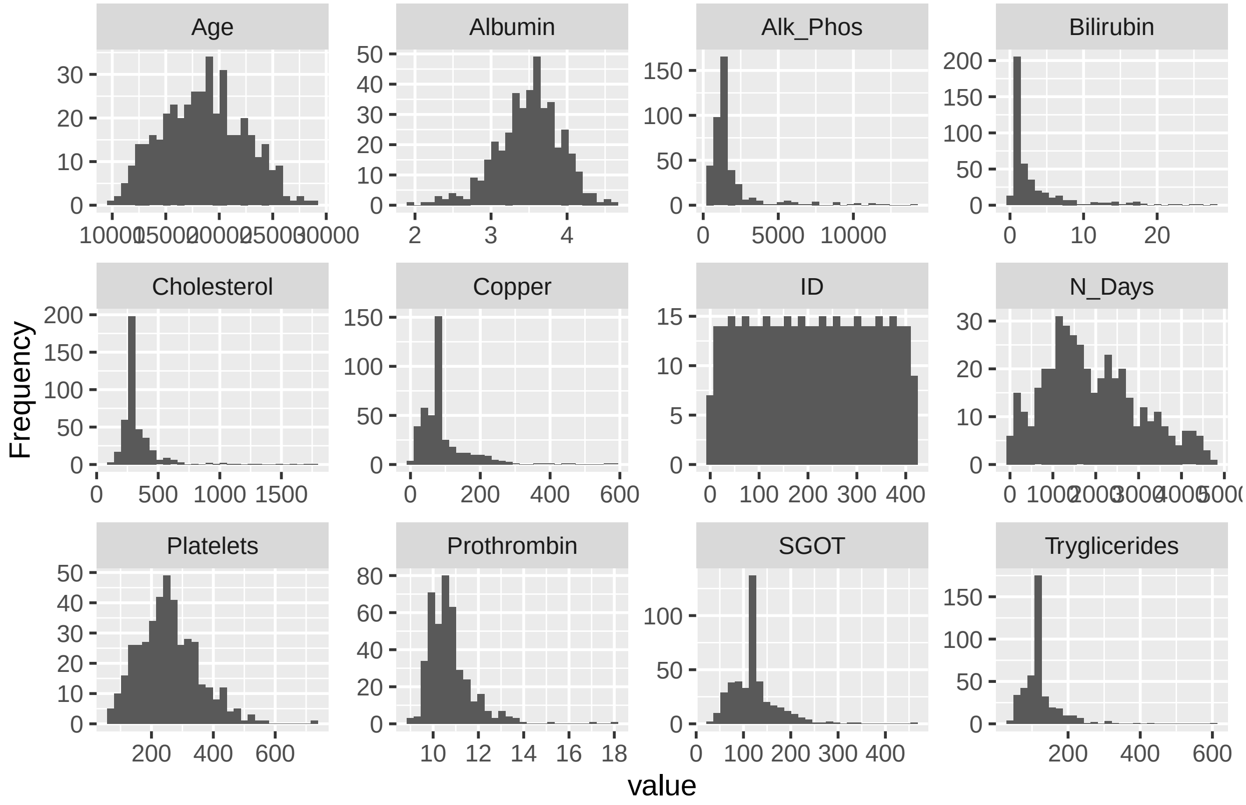
```r
# Correlation matrix for numerical variables
correlation <- cor(select_if(cirrhosis, is.numeric), use = "complete.obs")
corrplot::corrplot(correlation, method = "circle")
```



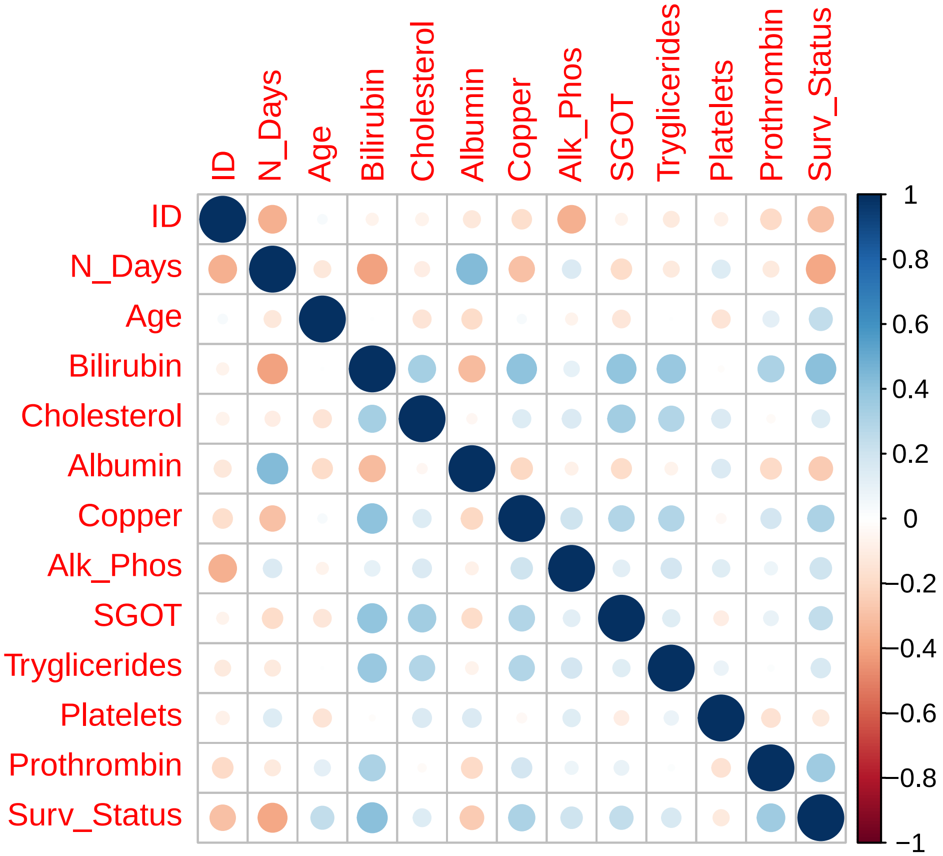
```
# Dealing with missing values:
# For continuous variables: Replace missing values with median
# For categorical variables: Replace with the most frequent category (mode).
cirrhosis$Prothrombin[is.na(cirrhosis$Prothrombin)] <- median(cirrhosis$Prothrombin, na.rm = TRUE)
cirrhosis$SGOT[is.na(cirrhosis$SGOT)] <- median(cirrhosis$SGOT, na.rm = TRUE)
cirrhosis$Tryglicerides[is.na(cirrhosis$Tryglicerides)] <- median(cirrhosis$Tryglicerides, na.rm = TRUE
cirrhosis$Cholesterol[is.na(cirrhosis$Cholesterol)] <- median(cirrhosis$Cholesterol, na.rm = TRUE)
cirrhosis$Copper[is.na(cirrhosis$Copper)] <- median(cirrhosis$Copper, na.rm = TRUE)
cirrhosis$Platelets[is.na(cirrhosis$Platelets)] <- median(cirrhosis$Platelets, na.rm = TRUE)
cirrhosis$Alk_Phos[is.na(cirrhosis$Alk_Phos)] <- median(cirrhosis$Alk_Phos, na.rm = TRUE)

mode_impute <- function(x) {
  x[is.na(x)] <- as.character(names(which.max(table(x, useNA = "no"))))
  return(x)
}
cirrhosis$Stage <- mode_impute(cirrhosis$Stage)

# Function to recode NA values based on percentages of known values
recode_na_by_percentage <- function(data, columns) {
  for (column in columns) {
    counts <- table(data[[column]], useNA = "no")
    percentages <- counts / sum(counts)

    # Replace NA values based on the probabilities
    data[[column]] <- sapply(data[[column]], function(x) {
      if (is.na(x)) {
        sample(names(percentages), size = 1, prob = percentages)
      } else {
        x
      }
    })
  }
  return(data)
}

columns_to_recode <- c("Ascites", "Hepatomegaly", "Spiders", "Drug")

# Apply the function to recode NA values
cirrhosis <- recode_na_by_percentage(cirrhosis, columns_to_recode)

# Visualize missing data
plot_missing(cirrhosis)
```



```
# Visualize distributions of numerical columns
plot_histogram(cirrhosis)
```



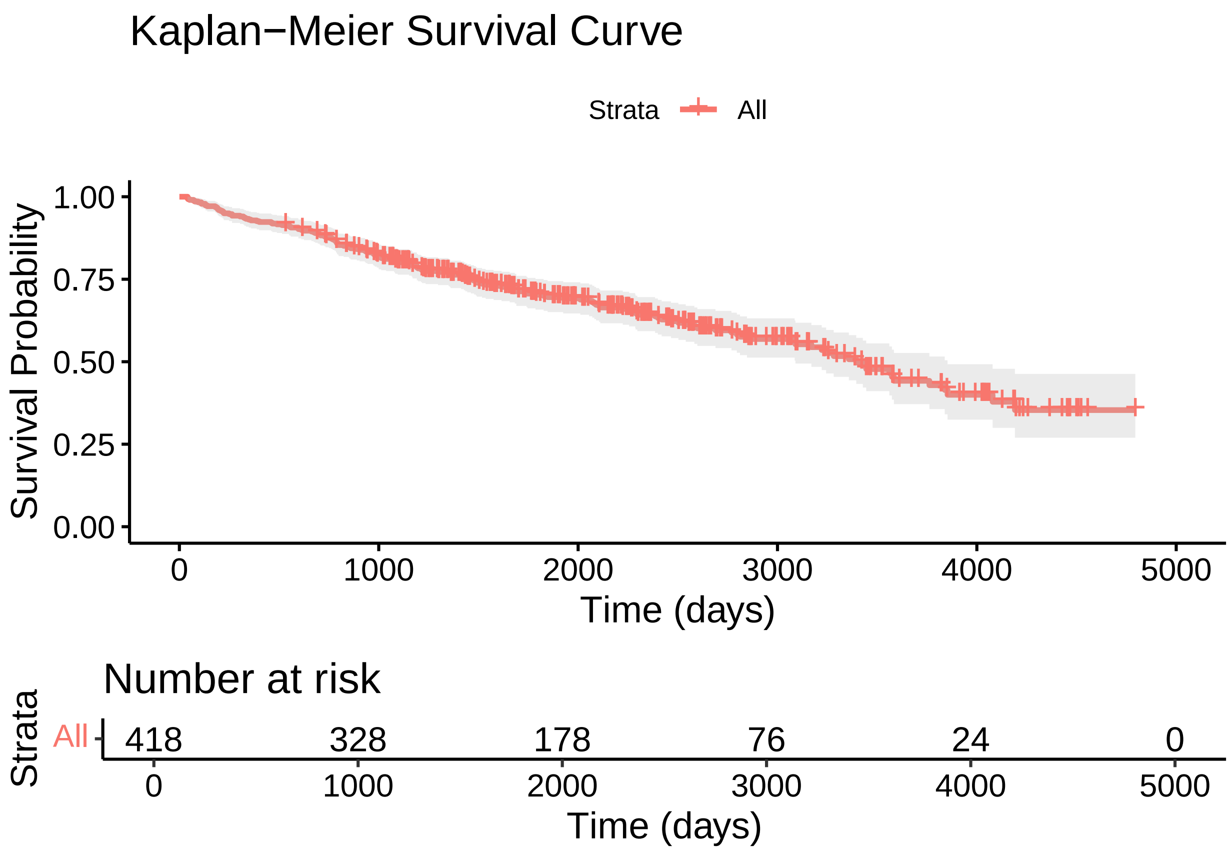
```
# Correlation matrix for numerical variables
correlation <- cor(select_if(cirrhosis, is.numeric), use = "complete.obs")
corrplot::corrplot(correlation, method = "circle")
```



2. Non-parametric Methods: Kaplan-Meier Estimator

```
# Create a survival object
surv_obj <- Surv(time = cirrhosis$N_Days, event = (cirrhosis$Status == "D"))

# Kaplan-Meier survival curve
km_fit <- survfit(surv_obj ~ 1, data = cirrhosis)

# Plot the survival curve
ggsurvplot(km_fit,
           conf.int = TRUE,
           pval = TRUE,
           risk.table = TRUE,
           title = "Kaplan-Meier Survival Curve",
           xlab = "Time (days)",
           ylab = "Survival Probability")
```

```
## Warning in .pvalue(fit, data = data, method = method, pval = pval, pval.coord = pval.coord, : There a
##  This is a null model.
```



```r
summary(km_fit, times = c(365, 1095, 1825, 2555, 3285, 3650))  # Survival at 1, 3, 5, 7, 9, 10 years
```

```
## Call: survfit(formula = surv_obj ~ 1, data = cirrhosis)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   365    388      30    0.928  0.0126        0.904        0.953
##  1095    313      52    0.801  0.0197        0.764        0.841
##  1825    197      33    0.703  0.0236        0.658        0.751
##  2555    116      21    0.612  0.0278        0.560        0.669
##  3285     56      13    0.517  0.0342        0.454        0.589
##  3650     35       7    0.442  0.0394        0.371        0.527
```

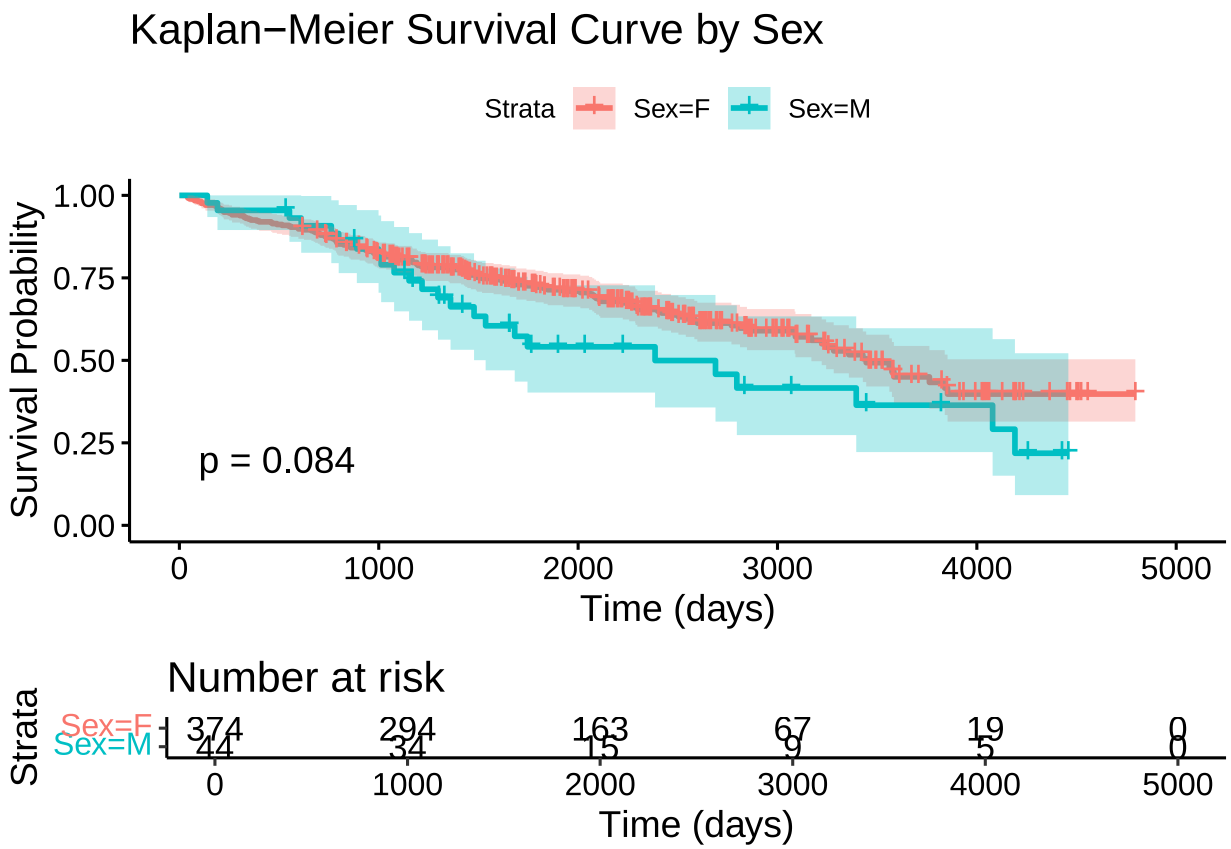
```r
# Log-rank test
survdiff(Surv(N_Days, Status == "D") ~ Sex, data = cirrhosis)
```

```
## Call:
## survdiff(formula = Surv(N_Days, Status == "D") ~ Sex, data = cirrhosis)
##
##         N Observed Expected (O-E)^2/E (O-E)^2/V
## Sex=F 374      137    143.7     0.317      2.98
## Sex=M  44       24     17.3     2.640      2.98
##
##  Chisq= 3  on 1 degrees of freedom, p= 0.08
```

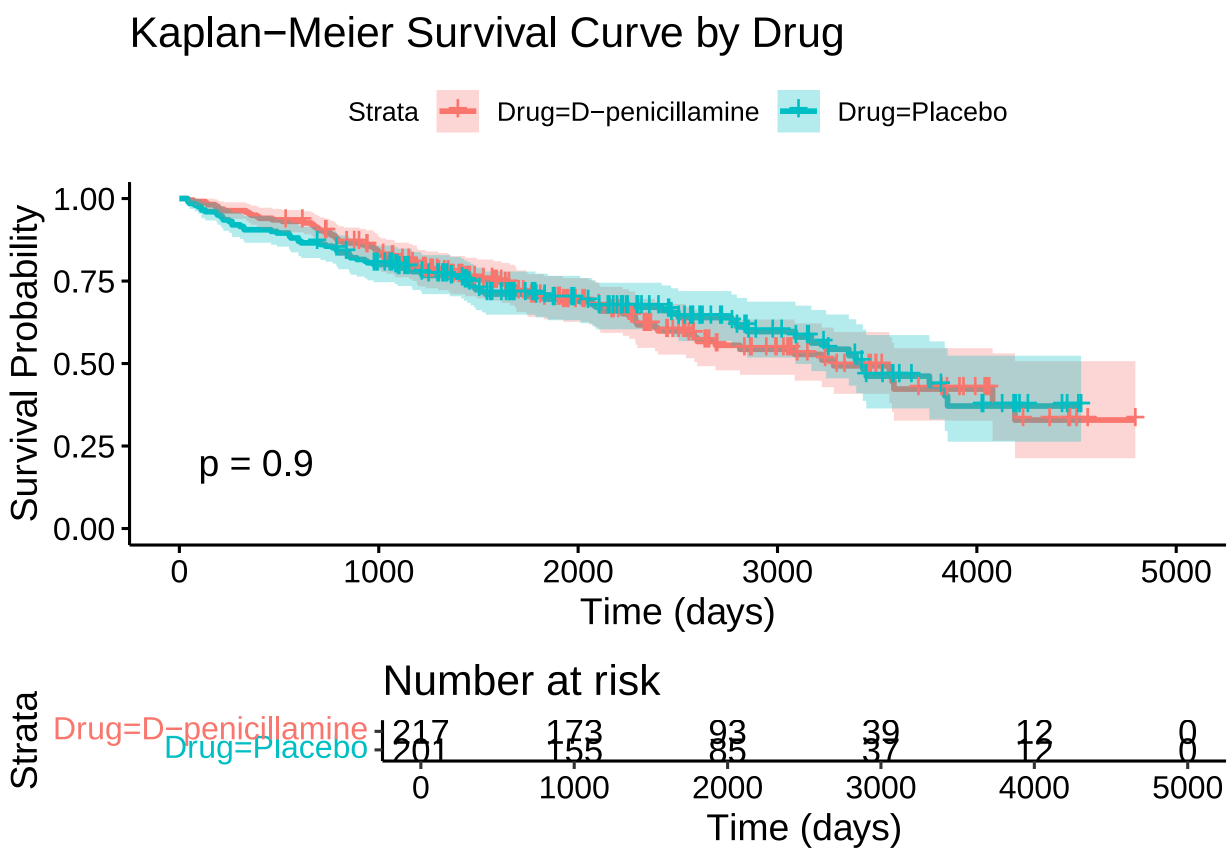
```r
survdiff(Surv(N_Days, Status == "D") ~ Drug, data = cirrhosis)
```

```
## Call:
## survdiff(formula = Surv(N_Days, Status == "D") ~ Drug, data = cirrhosis)
##
##                           N Observed Expected (O-E)^2/E (O-E)^2/V
## Drug=D-penicillamine 217       85     84.2   0.00758    0.0159
## Drug=Placebo         201       76     76.8   0.00831    0.0159
##
##  Chisq= 0  on 1 degrees of freedom, p= 0.9
```

```r
# draw KM curve stratified by sex and drug
km_fit_sex <- survfit(Surv(N_Days, Status == "D") ~ Sex, data = cirrhosis)
ggsurvplot(km_fit_sex,
           conf.int = TRUE,
           pval = TRUE,
           risk.table = TRUE,
           title = "Kaplan-Meier Survival Curve by Sex",
           xlab = "Time (days)",
           ylab = "Survival Probability")
```



```r
km_fit_drug <- survfit(Surv(N_Days, Status == "D") ~ Drug, data = cirrhosis)
ggsurvplot(km_fit_drug,
           conf.int = TRUE,
           pval = TRUE,
           risk.table = TRUE,
           title = "Kaplan-Meier Survival Curve by Drug",
           xlab = "Time (days)",
           ylab = "Survival Probability")
```



3. Semi-parametric Methods: Cox Proportional Hazards Model 3.1 Model selection

```
# # 1)  Fit a univariate model for each covariate, and identify the predictors significant at some leve
# uni_Drug <- coxph(surv_obj ~ Drug, data = cirrhosis)
# summary(uni_Drug)
#
# uni_Age <- coxph(surv_obj ~ Age, data = cirrhosis)
# summary(uni_Age) **
#
# uni_Sex <- coxph(surv_obj ~ Sex, data = cirrhosis)
# summary(uni_Sex) **
#
# uni_Ascites <- coxph(surv_obj ~ Ascites, data = cirrhosis)
# summary(uni_Ascites) **
#
# uni_Hepatomegaly <- coxph(surv_obj ~ Hepatomegaly, data = cirrhosis)
# summary(uni_Hepatomegaly) **
#
# uni_Spiders <- coxph(surv_obj ~ Spiders, data = cirrhosis)
# summary(uni_Spiders) **
#
# uni_Bilirubin <- coxph(surv_obj ~ Bilirubin, data = cirrhosis)
# summary(uni_Bilirubin) **
#
# uni_Albumin <- coxph(surv_obj ~ Albumin, data = cirrhosis)
# summary(uni_Albumin) **
#
# uni_Edema <- coxph(surv_obj ~ Edema, data = cirrhosis)
# summary(uni_Edema) **
#
# uni_Copper <- coxph(surv_obj ~ Copper, data = cirrhosis)
# summary(uni_Copper) **
#
# uni_Alk_Phos <- coxph(surv_obj ~ Alk_Phos, data = cirrhosis)
# summary(uni_Alk_Phos) **
#
# uni_SGOT <- coxph(surv_obj ~ SGOT, data = cirrhosis)
# summary(uni_SGOT) **
#
# uni_Tryglicerides <- coxph(surv_obj ~ Tryglicerides, data = cirrhosis)
# summary(uni_Tryglicerides) **
#
# uni_Prothrombin <- coxph(surv_obj ~ Prothrombin, data = cirrhosis)
# summary(uni_Prothrombin) **
#
# uni_Cholesterol <- coxph(surv_obj ~ Cholesterol, data = cirrhosis)
# summary(uni_Cholesterol) **
#
# uni_Platelets <- coxph(surv_obj ~ Platelets, data = cirrhosis)
# summary(uni_Platelets) **
# # drop Drug
#
# # 2)  Fit a multivariate model with all significant univariate predictors, and use backward selection
# model_2 <- coxph(surv_obj ~ Age + Ascites + Hepatomegaly + Spiders + Bilirubin + Albumin + Edema + Cop
#                      data = cirrhosis)
# model_backward <- stepAIC(model_2, direction = "backward")
# summary(model_backward)
# # drop Ascites, Spiders, Alk_Phos, Tryglicerides, Cholesterol, Platelets
#
# # 3) Starting with final step (2) model, consider each of the non-significant variables from step (1)
# model_3_ascites <- coxph(surv_obj ~ Ascites + Age + Hepatomegaly + Bilirubin + Albumin + Edema + Copp
# model_forward_ascites <- stepAIC(model_3_ascites, direction = "forward")
# summary(model_forward_ascites)
#
# model_3_spiders <- coxph(surv_obj ~ Spiders + Age + Hepatomegaly + Bilirubin + Albumin + Edema + Copp
# model_forward_spiders <- stepAIC(model_3_spiders, direction = "forward")
# summary(model_forward_spiders)
#
# model_3_alk <- coxph(surv_obj ~ Alk_Phos + Age + Hepatomegaly + Bilirubin + Albumin + Edema + Copper
# model_forward_alk <- stepAIC(model_3_alk, direction = "forward")
# summary(model_forward_alk)
#
```

```r
# model_3_tryglicerides <- coxph(surv_obj ~ Tryglicerides + Age + Hepatomegaly + Bilirubin + Albumin + 
# model_forward_tryglicerides <- stepAIC(model_3_tryglicerides, direction = "forward")
# summary(model_forward_tryglicerides)
#
# model_3_cholesterol <- coxph(surv_obj ~ Cholesterol + Age + Hepatomegaly + Bilirubin + Albumin + Edem
# model_forward_cholesterol <- stepAIC(model_3_cholesterol, direction = "forward")
# summary(model_forward_cholesterol)
#
# model_3_platelets <- coxph(surv_obj ~ Platelets + Age + Hepatomegaly + Bilirubin + Albumin + Edema + 
# model_forward_platelets <- stepAIC(model_3_platelets, direction = "forward")
# summary(model_forward_platelets)
# # no new variables added
#
# # 4)  Do final pruning of main-effects model (omit variables that are non-significant, add any that a
# model_4 <- coxph(surv_obj ~ Age + Hepatomegaly + Bilirubin + Albumin + Edema + Copper + SGOT + Prothr
#                      data = cirrhosis)
# model_final<- stepAIC(model_4, direction = "both")
# summary(model_final)

# Overall stepwise model selection
model_all <- coxph(surv_obj ~ Drug + Sex + Age + Ascites + Hepatomegaly + Spiders + Bilirubin + Albumin
                    data = cirrhosis)
model_stepwise <- stepAIC(model_all, direction = "both")
```

```
## Start:  AIC=1561.13
## surv_obj ~ Drug + Sex + Age + Ascites + Hepatomegaly + Spiders + 
##     Bilirubin + Albumin + Edema + Copper + Alk_Phos + SGOT + 
##     Tryglicerides + Prothrombin + Cholesterol + Platelets + Stage
## 
##                 Df    AIC
## - Spiders        1 1559.1
## - Platelets      1 1559.1
## - Alk_Phos       1 1559.2
## - Drug           1 1559.2
## - Sex            1 1559.4
## - Cholesterol    1 1559.6
## - Ascites        1 1559.8
## - Tryglicerides  1 1560.4
## <none>             1561.1
## - Copper         1 1562.2
## - Hepatomegaly   1 1562.6
## - SGOT           1 1562.7
## - Edema          2 1563.1
## - Prothrombin    1 1565.6
## - Stage          3 1566.3
## - Albumin        1 1566.6
## - Age            1 1572.3
## - Bilirubin      1 1582.2
## 
## Step:  AIC=1559.14
## surv_obj ~ Drug + Sex + Age + Ascites + Hepatomegaly + Bilirubin + 
##     Albumin + Edema + Copper + Alk_Phos + SGOT + Tryglicerides + 
##     Prothrombin + Cholesterol + Platelets + Stage
```

```
##
##                 Df    AIC
## - Platelets      1 1557.2
## - Alk_Phos       1 1557.2
## - Drug           1 1557.2
## - Sex            1 1557.4
## - Cholesterol    1 1557.6
## - Ascites        1 1557.8
## - Tryglicerides  1 1558.4
## <none>             1559.1
## - Copper         1 1560.4
## - Hepatomegaly   1 1560.7
## - SGOT           1 1560.7
## + Spiders        1 1561.1
## - Edema          2 1561.2
## - Prothrombin    1 1563.7
## - Albumin        1 1564.7
## - Stage          3 1564.9
## - Age            1 1570.3
## - Bilirubin      1 1580.5
##
## Step:  AIC=1557.15
## surv_obj ~ Drug + Sex + Age + Ascites + Hepatomegaly + Bilirubin +
##     Albumin + Edema + Copper + Alk_Phos + SGOT + Tryglicerides +
##     Prothrombin + Cholesterol + Stage
##
##                 Df    AIC
## - Alk_Phos       1 1555.2
## - Drug           1 1555.3
## - Sex            1 1555.4
## - Cholesterol    1 1555.7
## - Ascites        1 1555.8
## - Tryglicerides  1 1556.4
## <none>             1557.2
## - Copper         1 1558.4
## - Hepatomegaly   1 1558.7
## - SGOT           1 1559.0
## + Platelets      1 1559.1
## + Spiders        1 1559.1
## - Edema          2 1559.3
## - Prothrombin    1 1561.7
## - Albumin        1 1562.7
## - Stage          3 1563.0
## - Age            1 1568.3
## - Bilirubin      1 1578.6
##
## Step:  AIC=1555.23
## surv_obj ~ Drug + Sex + Age + Ascites + Hepatomegaly + Bilirubin +
##     Albumin + Edema + Copper + SGOT + Tryglicerides + Prothrombin +
##     Cholesterol + Stage
##
##                 Df    AIC
## - Drug           1 1553.3
## - Sex            1 1553.5
## - Cholesterol    1 1553.7
## - Ascites        1 1553.9
## - Tryglicerides  1 1554.5
## <none>             1555.2
## - Copper         1 1556.5
## - Hepatomegaly   1 1556.8
## - SGOT           1 1557.1
## + Alk_Phos       1 1557.2
## + Spiders        1 1557.2
## + Platelets      1 1557.2
## - Edema          2 1557.3
## - Prothrombin    1 1559.8
## - Albumin        1 1560.7
## - Stage          3 1561.2
## - Age            1 1566.9
## - Bilirubin      1 1576.7
##
## Step:  AIC=1553.34
## surv_obj ~ Sex + Age + Ascites + Hepatomegaly + Bilirubin + Albumin +
##     Edema + Copper + SGOT + Tryglicerides + Prothrombin + Cholesterol +
##     Stage
##
##                 Df    AIC
## - Sex            1 1551.6
## - Cholesterol    1 1551.9
## - Ascites        1 1552.0
## - Tryglicerides  1 1552.8
## <none>             1553.3
## - Copper         1 1554.7
## - Hepatomegaly   1 1554.8
## - SGOT           1 1555.2
## + Drug           1 1555.2
## + Alk_Phos       1 1555.3
## + Spiders        1 1555.3
## + Platelets      1 1555.3
## - Edema          2 1555.6
## - Prothrombin    1 1557.8
## - Albumin        1 1558.7
## - Stage          3 1559.6
## - Age            1 1565.0
## - Bilirubin      1 1575.1
##
## Step:  AIC=1551.64
## surv_obj ~ Age + Ascites + Hepatomegaly + Bilirubin + Albumin +
##     Edema + Copper + SGOT + Tryglicerides + Prothrombin + Cholesterol +
##     Stage
##
##                 Df    AIC
## - Cholesterol    1 1550.2
## - Ascites        1 1550.4
## - Tryglicerides  1 1551.1
## <none>             1551.6
## - Hepatomegaly   1 1553.3
## + Sex            1 1553.3
## + Drug           1 1553.5
## + Alk_Phos       1 1553.5
## - Edema          2 1553.6
## + Spiders        1 1553.6
## + Platelets      1 1553.6
## - SGOT           1 1553.6
## - Copper         1 1553.8
## - Prothrombin    1 1556.2
## - Albumin        1 1556.8
## - Stage          3 1557.7
## - Age            1 1565.5
## - Bilirubin      1 1573.2
##
## Step:  AIC=1550.17
## surv_obj ~ Age + Ascites + Hepatomegaly + Bilirubin + Albumin +
##     Edema + Copper + SGOT + Tryglicerides + Prothrombin + Stage
##
##                 Df    AIC
## - Ascites        1 1548.9
## - Tryglicerides  1 1549.4
## <none>             1550.2
## + Cholesterol    1 1551.6
## - Edema          2 1551.7
## - Hepatomegaly   1 1551.8
## + Sex            1 1551.9
## + Drug           1 1552.0
## - Copper         1 1552.1
## + Alk_Phos       1 1552.1
## + Platelets      1 1552.2
## + Spiders        1 1552.2
## - SGOT           1 1552.9
## - Prothrombin    1 1554.6
## - Albumin        1 1555.5
## - Stage          3 1556.0
## - Age            1 1563.6
## - Bilirubin      1 1574.8
##
## Step:  AIC=1548.9
## surv_obj ~ Age + Hepatomegaly + Bilirubin + Albumin + Edema +
##     Copper + SGOT + Tryglicerides + Prothrombin + Stage
##
##                 Df    AIC
## - Tryglicerides  1 1547.6
## <none>             1548.9
## + Ascites        1 1550.2
## + Cholesterol    1 1550.4
## + Sex            1 1550.5
## - Hepatomegaly   1 1550.6
## + Alk_Phos       1 1550.8
## + Drug           1 1550.8
## + Spiders        1 1550.9
## + Platelets      1 1550.9
## - SGOT           1 1551.3
## - Edema          2 1551.5
## - Copper         1 1551.8
## - Prothrombin    1 1553.6
## - Stage          3 1554.8
## - Albumin        1 1555.7
## - Age            1 1564.0
## - Bilirubin      1 1573.6
##
## Step:  AIC=1547.64
## surv_obj ~ Age + Hepatomegaly + Bilirubin + Albumin + Edema +
##     Copper + SGOT + Prothrombin + Stage
##
##                  Df    AIC
## <none>              1547.6
## + Tryglicerides  1 1548.9
## + Sex            1 1549.2
## + Cholesterol    1 1549.3
## + Ascites        1 1549.4
## + Drug           1 1549.4
## + Alk_Phos       1 1549.5
## - Hepatomegaly   1 1549.5
## + Platelets      1 1549.6
## + Spiders        1 1549.6
## - Copper         1 1550.0
## - SGOT           1 1550.5
## - Edema          2 1551.3
## - Prothrombin    1 1552.5
## - Stage          3 1553.1
## - Albumin        1 1554.2
## - Age            1 1563.6
## - Bilirubin      1 1573.4
```

```
summary(model_stepwise)
```

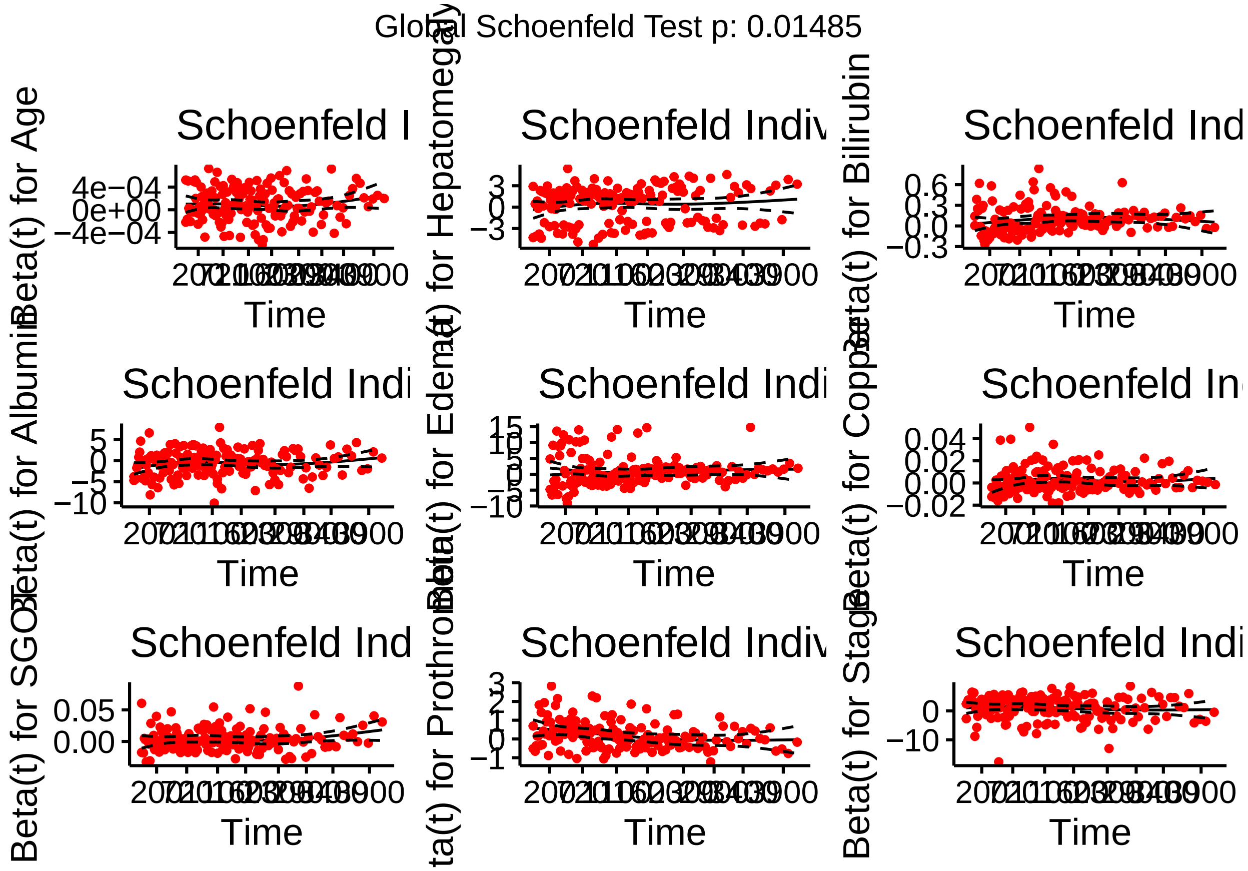
```
## Call:
## coxph(formula = surv_obj ~ Age + Hepatomegaly + Bilirubin + Albumin +
##     Edema + Copper + SGOT + Prothrombin + Stage, data = cirrhosis)
##
##   n= 418, number of events= 161
##
##                      coef  exp(coef)   se(coef)      z Pr(>|z|)
## Age             9.650e-05  1.000e+00  2.301e-05  4.194 2.74e-05 ***
## HepatomegalyY   3.708e-01  1.449e+00  1.899e-01  1.953  0.05086 .
## Bilirubin       9.222e-02  1.097e+00  1.595e-02  5.782 7.37e-09 ***
## Albumin        -6.552e-01  5.193e-01  2.198e-01 -2.981  0.00288 **
## EdemaS          2.068e-01  1.230e+00  2.332e-01  0.887  0.37519
## EdemaY          8.679e-01  2.382e+00  2.992e-01  2.901  0.00372 **
## Copper          1.956e-03  1.002e+00  9.015e-04  2.170  0.03002 *
## SGOT            3.700e-03  1.004e+00  1.592e-03  2.323  0.02016 *
## Prothrombin     2.033e-01  1.225e+00  7.123e-02  2.854  0.00432 **
## Stage2          7.112e-01  2.036e+00  7.452e-01  0.954  0.33986
## Stage3          1.024e+00  2.784e+00  7.340e-01  1.395  0.16299
## Stage4          1.431e+00  4.183e+00  7.367e-01  1.942  0.05208 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
##
##                 exp(coef) exp(-coef) lower .95 upper .95
## Age                1.0001     0.9999    1.0001     1.000
## HepatomegalyY      1.4489     0.6902    0.9986     2.102
## Bilirubin          1.0966     0.9119    1.0629     1.131
## Albumin            0.5193     1.9255    0.3376     0.799
## EdemaS             1.2297     0.8132    0.7786     1.942
## EdemaY             2.3818     0.4199    1.3251     4.281
## Copper             1.0020     0.9980    1.0002     1.004
## SGOT               1.0037     0.9963    1.0006     1.007
## Prothrombin        1.2254     0.8161    1.0657     1.409
## Stage2             2.0365     0.4910    0.4727     8.773
## Stage3             2.7844     0.3591    0.6606    11.736
## Stage4             4.1827     0.2391    0.9872    17.723
##
## Concordance= 0.836  (se = 0.016 )
## Likelihood ratio test= 223.3  on 12 df,   p=<2e-16
## Wald test            = 244.7  on 12 df,   p=<2e-16
## Score (logrank) test = 359.8  on 12 df,   p=<2e-16
```

```
# Test proportional hazards assumption
cox_zph <- cox.zph(model_stepwise)
cox_zph
```

```
##                 chisq df      p
## Age            0.0531  1 0.8177
## Hepatomegaly   1.2362  1 0.2662
## Bilirubin      6.8079  1 0.0091
## Albumin        2.1082  1 0.1465
## Edema          3.8781  2 0.1438
## Copper         1.5195  1 0.2177
## SGOT           5.7431  1 0.0166
## Prothrombin    6.9178  1 0.0085
## Stage          5.6052  3 0.1325
## GLOBAL        24.9950 12 0.0148
```

```
# Plot Schoenfeld residuals to check proportional hazards
ggcoxzph(cox_zph)
```



```
# Bilirubin & Prothrombin violate the PH assumption

# Fit the extended Cox model with time-dependent terms
cox_model_td <- coxph(surv_obj ~ Age + Hepatomegaly + Albumin + Bilirubin + Bilirubin*N_Days + Edema +
                      Prothrombin + Prothrombin*N_Days + SGOT + SGOT*N_Days + Stage, data = cirrhosis)
```

```
## Warning in coxph.fit(X, Y, istrat, offset, init, control, weights = weights, :
## Ran out of iterations and did not converge
```

```
## Warning in coxph.fit(X, Y, istrat, offset, init, control, weights = weights, :
## one or more coefficients may be infinite
```

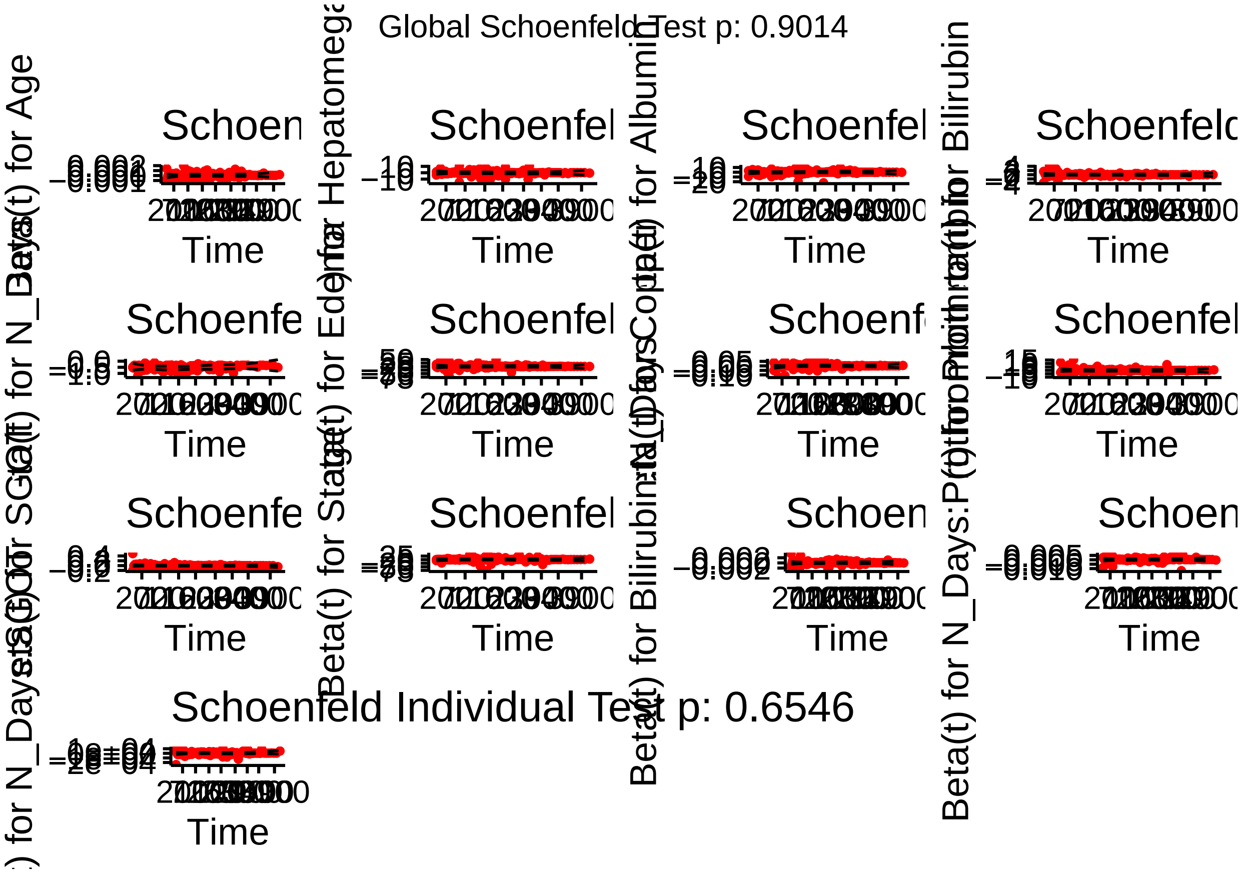
```
summary(cox_model_td)
```

```
## Call:
## coxph(formula = surv_obj ~ Age + Hepatomegaly + Albumin + Bilirubin +
##     Bilirubin * N_Days + Edema + Copper + Prothrombin + Prothrombin *
##     N_Days + SGOT + SGOT * N_Days + Stage, data = cirrhosis)
##
##   n= 418, number of events= 161
##
##                           coef  exp(coef)   se(coef)      z Pr(>|z|)
## Age                  4.033e-05  1.000e+00  5.064e-05  0.797   0.4257
## HepatomegalyY        1.304e-01  1.139e+00  3.616e-01  0.361   0.7185
## Albumin             4.712e-02  1.048e+00  4.014e-01  0.117   0.9066    
## Bilirubin          -9.630e-03  9.904e-01  3.041e-02 -0.317   0.7515    
## N_Days             -3.293e-01  7.194e-01  4.366e-02 -7.542 4.65e-14 ***
## EdemaS             -1.895e-01  8.274e-01  5.015e-01 -0.378   0.7056    
## EdemaY              3.281e-01  1.388e+00  5.602e-01  0.586   0.5581    
## Copper              2.931e-05  1.000e+00  1.792e-03  0.016   0.9870    
## Prothrombin        -9.678e-02  9.078e-01  1.742e-01 -0.556   0.5785    
## SGOT               -1.659e-03  9.983e-01  3.057e-03 -0.543   0.5875    
## Stage2              1.017e+00  2.766e+00  4.739e-01  2.147   0.0318 *  
## Stage3              1.418e+00  4.129e+00  3.403e-01  4.167 3.09e-05 ***
## Stage4              1.392e+00  4.021e+00  3.302e-01  4.214 2.50e-05 ***
## Bilirubin:N_Days    5.224e-05  1.000e+00  3.216e-05  1.624   0.1043    
## N_Days:Prothrombin  2.894e-04  1.000e+00  1.385e-04  2.089   0.0367 *  
## N_Days:SGOT         4.071e-06  1.000e+00  2.214e-06  1.839   0.0660 .  
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
##                    exp(coef) exp(-coef) lower .95 upper .95
## Age                   1.0000     1.0000    0.9999    1.0001
## HepatomegalyY         1.1392     0.8778    0.5608    2.3141
## Albumin               1.0482     0.9540    0.4773    2.3023
## Bilirubin             0.9904     1.0097    0.9331    1.0512
## N_Days                0.7194     1.3900    0.6604    0.7837
## EdemaS                0.8274     1.2086    0.3096    2.2109
## EdemaY                1.3883     0.7203    0.4631    4.1622
## Copper                1.0000     1.0000    0.9965    1.0035
## Prothrombin           0.9078     1.1016    0.6452    1.2771
## SGOT                  0.9983     1.0017    0.9924    1.0043
## Stage2                2.7661     0.3615    1.0927    7.0025
## Stage3                4.1286     0.2422    2.1190    8.0439
## Stage4                4.0213     0.2487    2.1052    7.6811
## Bilirubin:N_Days      1.0001     0.9999    1.0000    1.0001
## N_Days:Prothrombin    1.0003     0.9997    1.0000    1.0006
## N_Days:SGOT           1.0000     1.0000    1.0000    1.0000
## 
## Concordance= 1  (se = 0 )
## Likelihood ratio test= 1681  on 16 df,   p=<2e-16
## Wald test            = 109  on 16 df,   p=7e-16
## Score (logrank) test = 674.6  on 16 df,   p=<2e-16
```

```
# Plot Schoenfeld residuals to check proportional hazards
cox_zph_td <- cox.zph(cox_model_td)
cox_zph_td
```

```
##                     chisq df    p
## Age                0.2703  1 0.60
## Hepatomegaly       0.1364  1 0.71
## Albumin            0.6071  1 0.44
## Bilirubin          0.6039  1 0.44
## N_Days             0.6953  1 0.40
## Edema              0.1954  2 0.91
## Copper             1.4673  1 0.23
## Prothrombin        0.4853  1 0.49
## SGOT               0.0381  1 0.85
## Stage                0.4495  3 0.93
## Bilirubin:N_Days     0.1753  1 0.68
## N_Days:Prothrombin 0.1251  1 0.72
## N_Days:SGOT          0.2001  1 0.65
## GLOBAL               9.2807 16 0.90
```

```
ggcoxzph(cox_zph_td)
```



```
# PH assumption: The hazard ratio for a given covariate is constant over time.
# global p-value = 1, do not reject the null hypothesis, so the PH assumtion holds.

# compare AIC
# model with interaction
AIC(cox_model_td)
```

```
## [1] 97.83188
```

```
# model without interaction
AIC(model_stepwise)
```

```
## [1] 1547.637
```

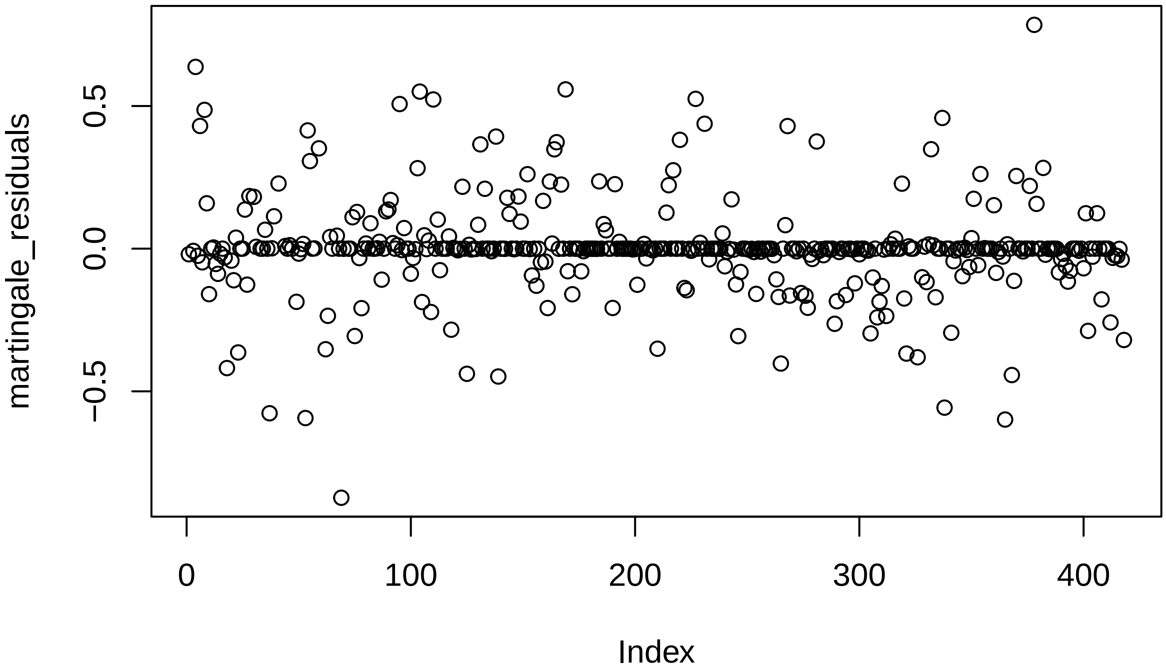
```r
# model with interaction has lower AIC


anova(model_stepwise, cox_model_td, test = "LRT")
```

```
## Analysis of Deviance Table
##  Cox model: response is  surv_obj
##  Model 1: ~ Age + Hepatomegaly + Bilirubin + Albumin + Edema + Copper + SGOT + Prothrombin + Stage
##  Model 2: ~ Age + Hepatomegaly + Albumin + Bilirubin + Bilirubin * N_Days + Edema + Copper + Prothro
##    loglik  Chisq Df Pr(>|Chi|)
## 1 -761.82
## 2  -32.92 1457.8  4  < 2.2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

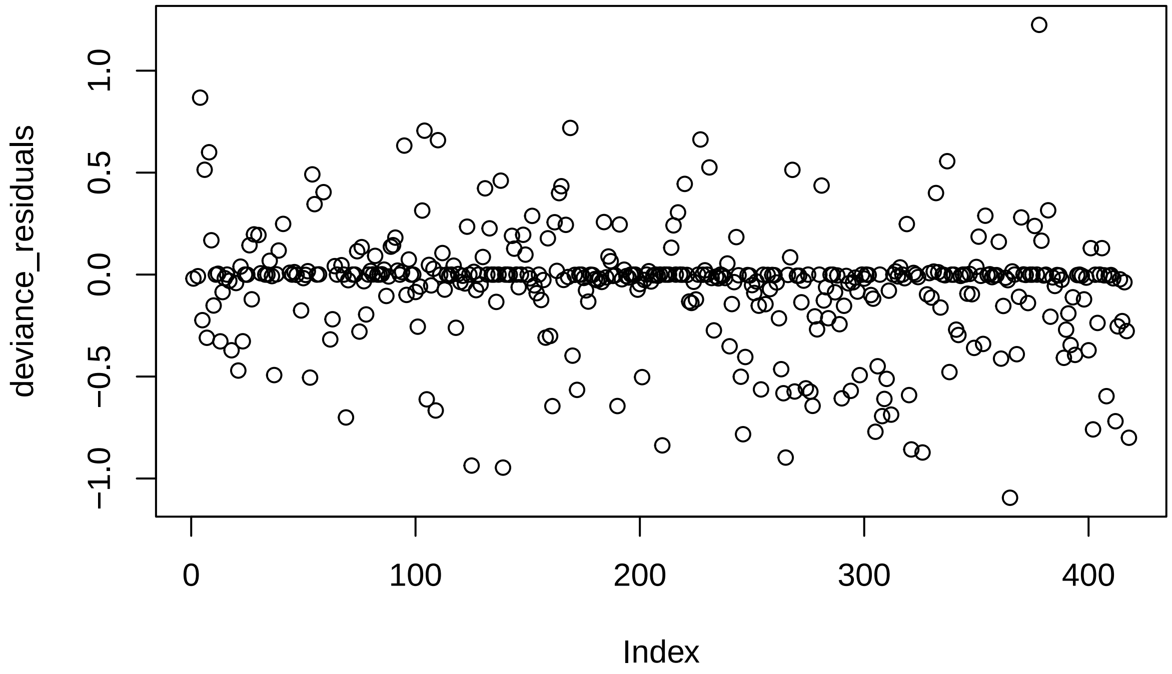
```r
# The time-dependent model (Model 2) is a significantly better fit compared to the simpler model (Model

# so the final model is:
# coxph(surv_obj ~ Age + Hepatomegaly + Albumin + Bilirubin + Bilirubin*N_Days + Edema + Copper +
# Prothrombin + Prothrombin*N_Days + SGOT + SGOT*N_Days + Stage, data = cirrhosis)
```

3.2 Residual analysis

```r
# Martingale Residuals
martingale_residuals <- residuals(cox_model_td, type = "martingale")
plot(martingale_residuals)
```



```
# Deviance Residuals
deviance_residuals <- residuals(cox_model_td, type = "deviance")
plot(deviance_residuals)
```



```r
# Residual analysis suggest a generally good fit for the model, but a few observations with larger resi
# Addressing or further investigating these outliers may improve interpretability and robustness of the

# Identify observations with deviance residuals > |0.5|
influential_obs <- which(abs(deviance_residuals) > 0.5)
print(influential_obs)
```

```
##   4   6   8  53  69  71  95 104 105 109 110 125 139 161 169 172 190 201 210 227
##   4   6   8  53  69  71  95 104 105 109 110 125 139 161 169 172 190 201 210 227
## 231 245 246 254 264 265 268 269 274 276 277 290 294 305 308 309 310 312 320 321
## 231 245 246 254 264 265 268 269 274 276 277 290 294 305 308 309 310 312 320 321
## 326 337 365 378 402 408 412 418
## 326 337 365 378 402 408 412 418
```