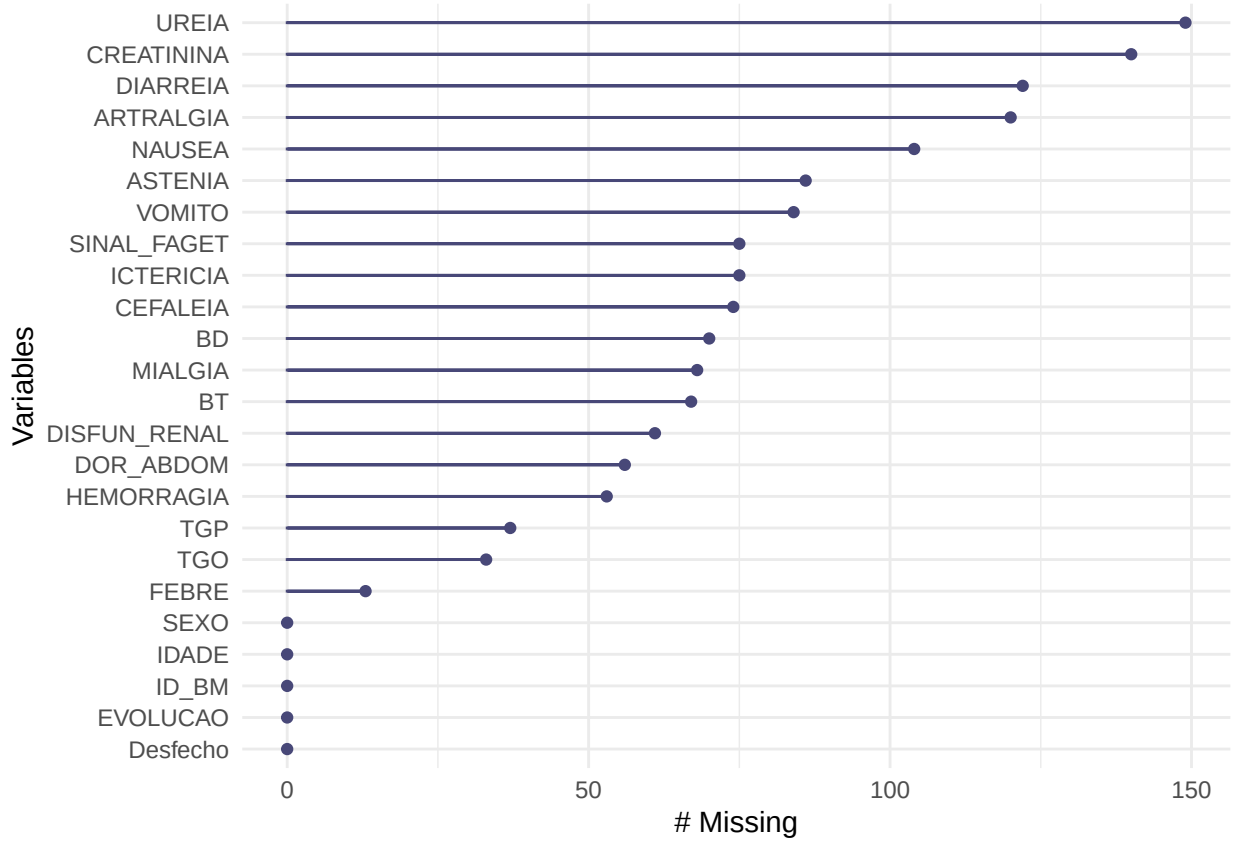
Relatório

Análise dos dados pontuais

```
dados_adicionais <- read_excel('Casos_Hum2017_tese.xlsx') %>% filter(CLASSIF_FINAL =='CONFIRMADO' ) %>%
  select(ID_BM,EVOLUCAO,TGO,TGP,CREATININA,BD,BT,UREIA,FEBRE:DIARREIA,IDADE,SEXO) %>% filter(!is.na(EVO
  mutate(Desfecho = ifelse(EVOLUCAO=='CURA',0,1)) %>% mutate_at(.vars = c('BD','BT'), .funs = as.numeri
```

```
summary_missing <- dados_adicionais %>%ungroup %>%
  miss_var_summary()
not_missing <-summary_missing %>% filter(n_miss==0) %>% select(variable) %>% pull()
gg_miss_var(dados_adicionais)
```



Existe uma porcentagem de valores faltantes altíssima para algumas variáveis, a modelagem com essa parcela dos dados será um pouco complicada.

Uma saída possível é a utilização das variáveis com a menor proporção de missing (TGO, HEMORRAGIA, BT e BD).

Uma linha de raciocínio interessante é a comparação desse modelo com poucas variáveis com o modelo longitudinal buscando uma possível argumentação sobre a importância de alguns resultados de exames que discriminam bem o óbito ou não.

```
ajuste_logistica <-glm(Desfecho ~ log2(TGO)+BT+ICTERICIA+IDADE+SEXO,
  data=dados_adicionais, family=binomial(link='logit'))
```

```
ajuste_logistica %>% summary
```

```
##
## Call:
## glm(formula = Desfecho ~ log2(TGO) + BT + ICTERICIA + IDADE +
##     SEXO, family = binomial(link = "logit"), data = dados_adicionais)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9726  -0.8386  -0.4310   0.8765   2.2464
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -5.16535    1.22203  -4.227 2.37e-05 ***
## log2(TGO)     0.18596    0.08047   2.311  0.02085 *
## BT            0.10834    0.04745   2.283  0.02243 *
## ICTERICIASIM  1.22675    0.42588   2.880  0.00397 **
## IDADE         0.02622    0.01274   2.058  0.03961 *
## SEXOM         0.48172    0.52875   0.911  0.36227
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 207.61  on 153  degrees of freedom
## Residual deviance: 161.05  on 148  degrees of freedom
##   (121 observations deleted due to missingness)
## AIC: 173.05
##
## Number of Fisher Scoring iterations: 4
```

Análise da predição

```
prob=predict(ajuste_logistica,type=c("response"))

dados_roc <- dados_adicionais %>%ungroup %>%
  select(TGO,BT,ICTERICIA,Desfecho) %>% drop_na() %>% mutate(p= prob)
require(pROC)
g <- auc(Desfecho ~ p, data = dados_roc)
g
```

```
## Area under the curve: 0.8008
```

```
dados_roc <- dados_roc %>% ungroup %>% mutate(predicao = ifelse(p>0.4,1,0))

confusionMatrix(dados_roc$predicao %>% as.factor,dados_roc$Desfecho%>% as.factor,positive='1')
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 68 19
##          1 24 43
##
```

```
##                Accuracy : 0.7208
##                  95% CI : (0.6429, 0.79)
##     No Information Rate : 0.5974
##     P-Value [Acc > NIR] : 0.0009635
##
##                   Kappa : 0.4271
##
##  Mcnemar's Test P-Value : 0.5418656
##
##             Sensitivity : 0.6935
##             Specificity : 0.7391
##          Pos Pred Value : 0.6418
##          Neg Pred Value : 0.7816
##              Prevalence : 0.4026
##          Detection Rate : 0.2792
##    Detection Prevalence : 0.4351
##       Balanced Accuracy : 0.7163
##
##        'Positive' Class : 1
##
```

Análise dos dados longitudinais

Limpeza dos dados

Os pacientes 773, 489 e 212 aparentam apresentar erros no registro da data de evolução.

```
data <- readRDS('dados_att.rds') %>%group_by(pac,exame) %>%
mutate(DT_EVOL=replace(DT_EVOL, pac==773, '2017-05-26')) %>%
  mutate(DT_EVOL=replace(DT_EVOL, pac==489, '2017-04-14')) %>%
  mutate(DT_EVOL=replace(DT_EVOL, pac==212, '2017-02-16'))%>%
  mutate(DATA_FINAL =difftime(DT_EVOL,DT_IS,units='days') ) %>%
  mutate(dia_doenca=DIAS,dt_inic =lag(DIAS) ) %>% select(-DIAS) %>%
  mutate(dt_ate_inter =difftime(DT_INTERN,DT_IS,units='days') %>% as.numeric )
```

Estruturação dos dados

```
dados <- data %>%
  select(pac,EVOLUCAO,exame,valor,SEXO,DT_NASC,FEBRE:DIARREIA,
         dt_inic,dia_doenca,DATA_FINAL,OBITO,dt_ate_inter) %>%
  mutate(Idade = year(as.Date('2019-12-30'))-year(DT_NASC)) %>%
  select(-DT_NASC) %>%
  spread(exame,valor) %>% mutate(Desfecho = ifelse(OBITO =='OBITO',1,0)) %>%
  select(-OBITO,-EVOLUCAO)

####### Seleção de variáveis
data_inic <- dados %>% group_by(pac) %>% filter(is.na(dt_inic)) %>% ungroup(pac) %>%
  select(Idade,SEXO,TGO,TGP,HEMORRAGIA,DOR_ABDOM,MIALGIA,Leucocitos,
         Bastoes,Creatinina,Linfocitos,Ureia,DATA_FINAL,Desfecho)
```

```
dados <- data %>%
  select(pac,EVOLUCAO,exame,valor,SEXO,DT_NASC,FEBRE:DIARREIA,
         dt_inic,dia_doenca,DATA_FINAL,OBITO,dt_ate_inter) %>%
```

```
  mutate(Idade = year(as.Date('2019-12-30'))-year(DT_NASC)) %>%
  select(-DT_NASC) %>%
  spread(exame,valor) %>% mutate(Desfecho = ifelse(OBITO =='OBITO',1,0)) %>%
  select(-OBITO,-EVOLUCAO)
```

Organização do banco

Nesse cenário, todas as observações de um indivíduo que foi a óbito foram consideradas como tendo "Óbito" como variável resposta". Essa aplicação se mostra necessária pois, obviamente, é preciso ajustar o modelo para prever o óbito do paciente antes do evento de interesse. Sendo assim,

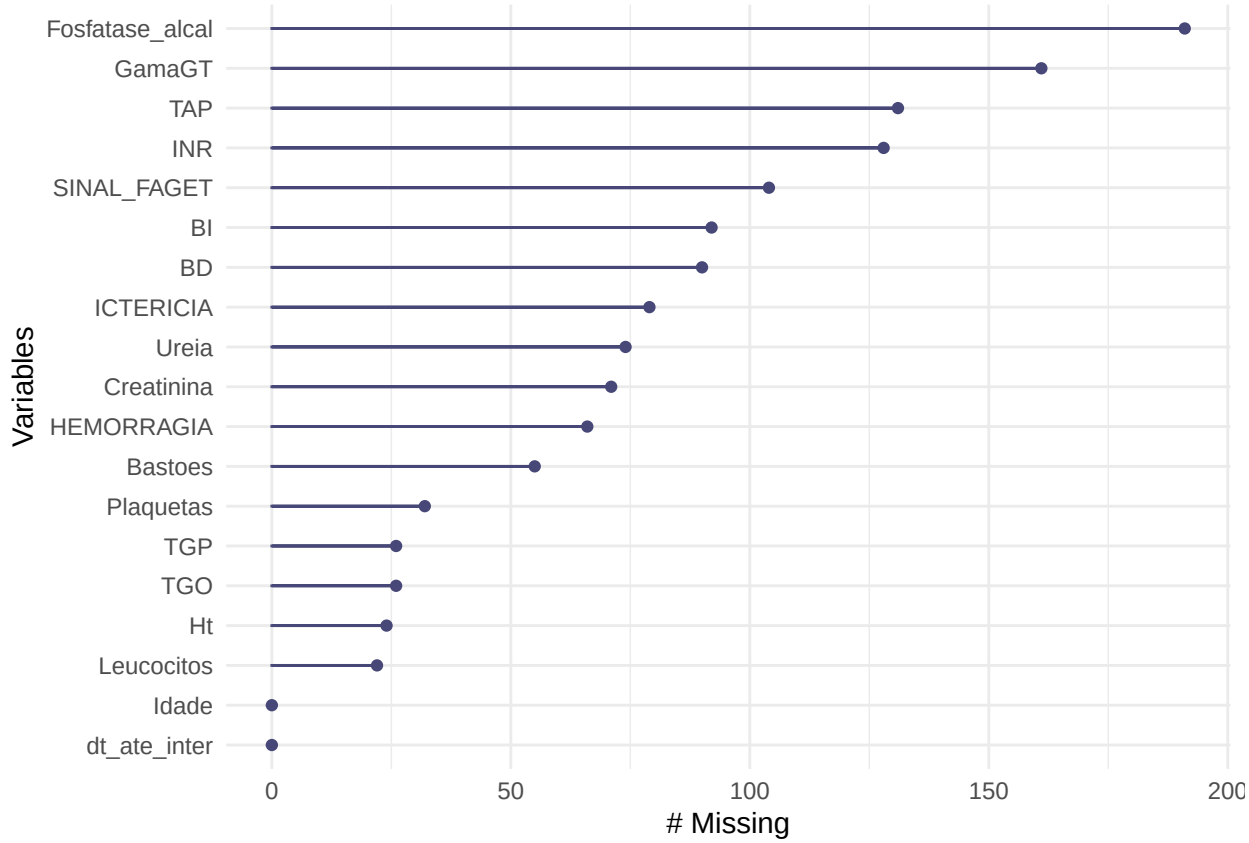
```
dados2 <- dados %>%
  select(pac,dt_inic,dia_doenca,DATA_FINAL,TGO,TGP,Desfecho,Idade,Ht,BD,BI,TAP,INR,
         Leucocitos,Bastoes,Creatinina,ICTERICIA,HEMORRAGIA,SEXO,dt_ate_inter
         ,Ureia,GamaGT,Fosfatase_alcal,Plaquetas,SINAL_FAGET) %>%
  mutate(DATA_FINAL = ifelse(DATA_FINAL<0,max(dia_doenca),DATA_FINAL))  %>%
filter(dia_doenca <=15)

dados_class <- dados2 %>% group_by(pac) %>% mutate(Desfecho=max(Desfecho))%>%
  arrange(pac,dia_doenca) %>%
  filter(dia_doenca <=15)

# dados_completos <- dados2 %>% group_by(pac)%>% filter(any(DATA_FINAL == dia_doenca))
#
# dados_faltantes <- dados2 %>% filter(!(pac %in% (dados_completos$pac %>% unique)))
#
# dados_faltantes_2 <- dados_faltantes %>%
#   group_by(pac) %>%
#   summarise(dt_inic = max(dia_doenca),dia_doenca =max(DATA_FINAL)) %>%
#   bind_rows(dados_faltantes,.) %>% arrange(pac)
#
# dados_final <- bind_rows(dados_completos,dados_faltantes_2)  %>%
#   group_by(pac) %>% mutate(Desfecho = ifelse(dia_doenca==max(dia_doenca),max(Desfecho,na.rm = TRUE),0
#
# dados_input <- dados_final %>% group_by(pac) %>%
#   mutate_at(.vars=vars(DATA_FINAL,TGO,TGP,Ht,Leucocitos,Creatinina,Ht,Idade,SEXO),
#             .funs=list(~na.locf(., na.rm = FALSE))) %>% filter(dia_doenca !=0) %>%
#   mutate_at(.vars=vars(DATA_FINAL,TGO,TGP,Ht,Leucocitos,Creatinina,Ht,Idade,SEXO),
#             .funs=list(~na.locf(., na.rm = FALSE,fromLast=TRUE)))
#
# dados_input <- dados_input%>%
#   group_by(pac) %>% mutate(Desfecho = ifelse(dia_doenca==max(dia_doenca),
#       max(Desfecho,na.rm = TRUE),0)) %>% ungroup %>%  mutate(dt_inic = ifelse(is.na(dt_inic),0,dt_ini
```

Análise de dados Faltantes

```
summary_missing <- dados_class %>%ungroup %>%
  miss_var_summary()
not_missing <-summary_missing %>% filter(n_miss==0) %>% select(variable) %>% pull()
gg_miss_var(dados_class %>% ungroup %>%
    select(-c(pac,dt_inic,dia_doenca,DATA_FINAL,Desfecho,SEXO)))
```



Devemos evitar utilizar variáveis com muitos valores faltantes

##Inputação de valores utilizando substituindo o valor faltante pelo registro do dia anterior.

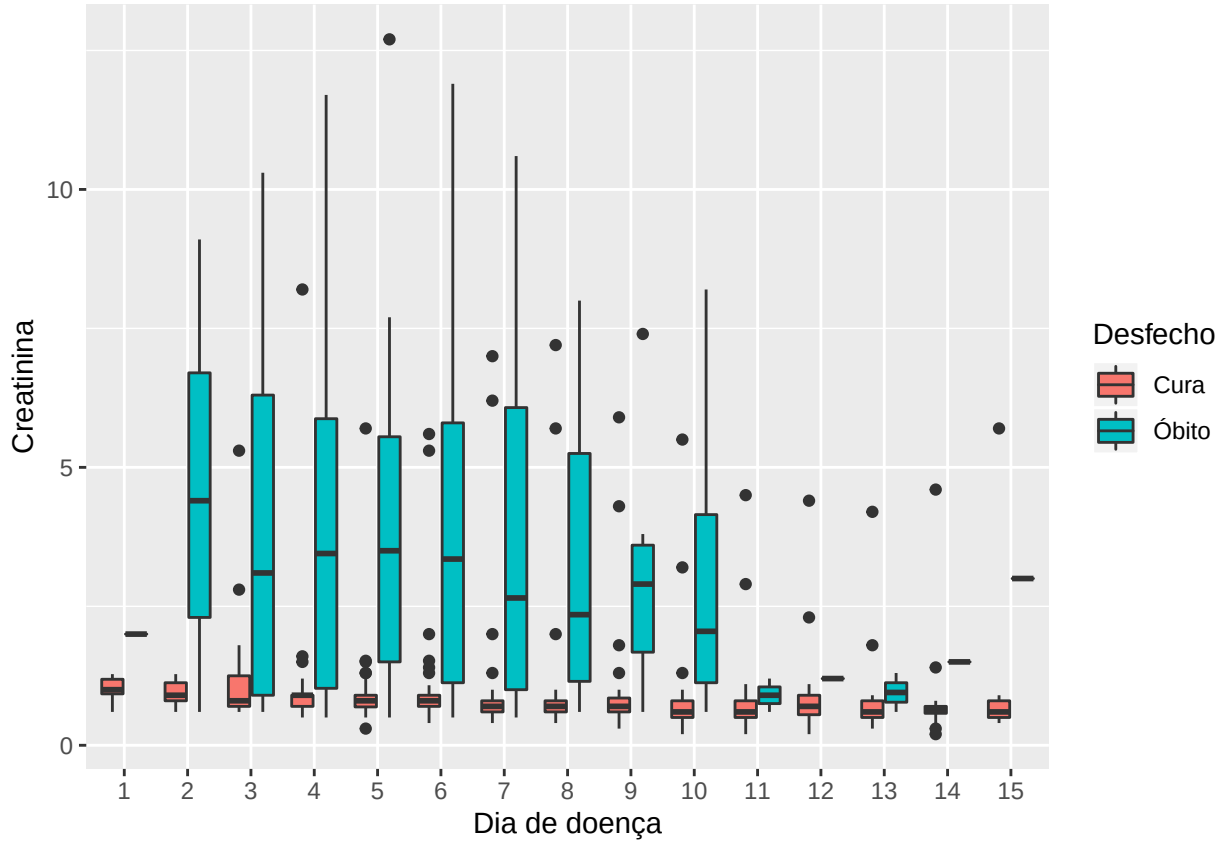
Caso o registro do dia anterior também esteja faltante, o registro do dia posterior é utilizado.

```
dados_class_fill <- dados_class %>%group_by(pac) %>%
  fill(c("Plaquetas","Ht",'TGP','Ureia','TGO','Creatinina','ICTERICIA','BD',
         'BI','TAP','INR','Leucocitos'),.direction="updown")
```

Estatísticas resumo após imputação

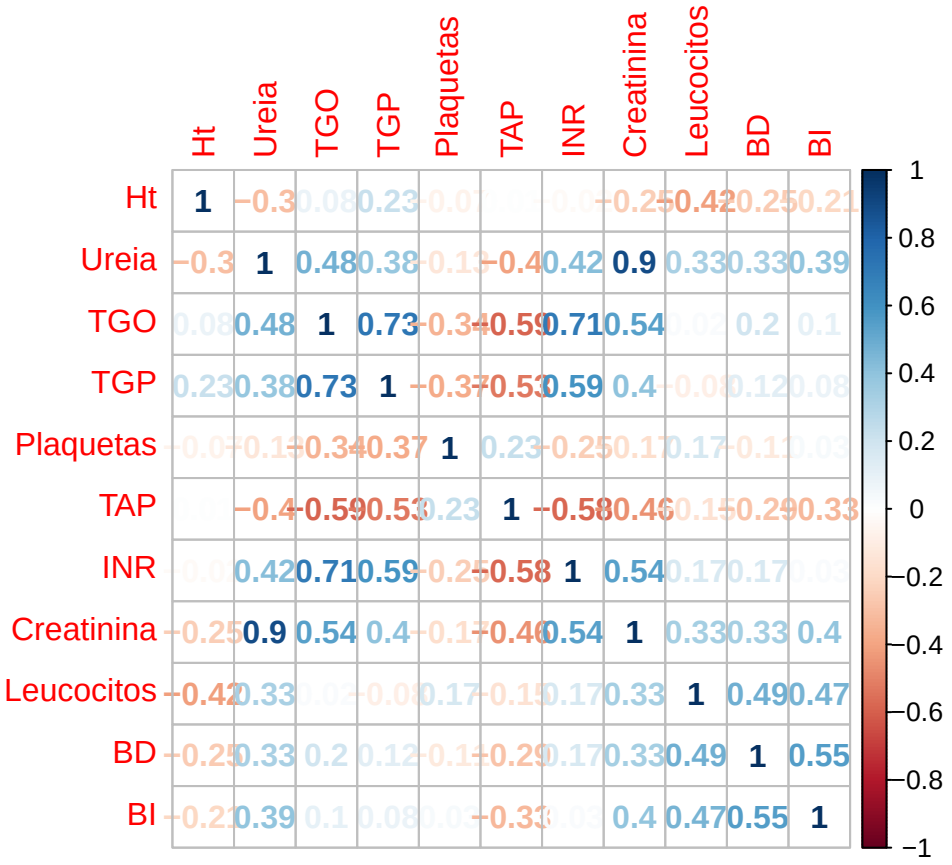
Gráfico Creatinina

```
ggplot(data=dados_plot %>% filter(dia_doenca>0),aes(y=Creatinina,fill=Desfecho,x=as.factor(dia_doenca)))
  geom_boxplot()+xlab("Dia de doença")
```



Correlações

```
require(corrplot)
M <- cor(dados_class_fill %>% ungroup%>%select(Ht,Ureia,TGO,TGP,Plaquetas,TAP,INR,
                                               Creatinina,Leucocitos,BD,BI)%>% drop_na()  )
corrplot(M, method = "number")
```



Ainda existem dados faltantes de ICTERICIA que não podem ser inputados pois não existe o regristro anível de paciente.

Aplicação de equações de estimação generalizadas (GEE)

O método de equações de estimação generalizadas (GEE) é uma alternativa para a modelagem de dados longitudinais.

Essa técnica de modelagem requer a indicação de uma estrutura de correlação entre o tempo e estima o efeito médio populacional das variáveis explicativas levando em consideração a correlação interna de cada paciente.

O algorítmo obtem a estimativa dos coeficientes β através da maximização da equação abaixo em β:

$$U(\tilde{\beta}) = \sum_{i=1}^{n} \frac{\partial \mu_{ij}}{\partial \beta_p} V_i^{-1} \{Y_i - \mu_i(\tilde{\beta})\}$$

Onde:

- μ_{ij} é o valor da variável explicativa do indivíduo i no tempo j.
- $\beta_p,\ p = 1, \ldots, k$ são os coeficientes das variáveis explicativas
- V_i é a estrutura de correlação do tempo.

Nessa aplicação as GEE nos fornencem uma estimativa da probabilidade de óbito para um indivíduo infectado com o vírus da febre amarela, independe do dia de doença em que ele se encontrar.

```
require(geepack)
dados_scaled <- dados2  %>% mutate(TGO=log(TGO)) %>% select(-dt_inic) %>%
  ungroup %>% mutate(pac = as.factor(pac)) %>%
  select(TGO,ICTERICIA,pac,dia_doenca,Desfecho,SEXO,Idade) %>%
  group_by(pac) %>% mutate(Idade= ifelse(Idade <=40,'<=40 ','>40')) %>%
  drop_na()

ajuste_gee<- geeglm(Desfecho~TGO+ICTERICIA+Idade+SEXO,id=pac, data=dados_scaled,
family=binomial(link='logit'),corstr='ar1')


summary(ajuste_gee)
```

```
##
## Call:
## geeglm(formula = Desfecho ~ TGO + ICTERICIA + Idade + SEXO, family = binomial(link = "logit"),
##     data = dados_scaled, id = pac, corstr = "ar1")
##
##  Coefficients:
##              Estimate  Std.err   Wald Pr(>|W|)
## (Intercept)  -1.36081  0.81745  2.771  0.09597 .
## TGO           0.06853  0.01458 22.080 2.62e-06 ***
## ICTERICIASIM  1.76956  0.55062 10.328  0.00131 **
## Idade>40     -0.71296  0.53625  1.768  0.18367
## SEXOM        -0.38385  0.69623  0.304  0.58141
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Estimated Scale Parameters:
##             Estimate Std.err
## (Intercept)   0.7917  0.1229
##
## Correlation: Structure = ar1  Link = identity
##
```

```
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha   0.9722 0.03714
## Number of clusters:   83   Maximum cluster size: 13
```

```
anova(ajuste_gee)
```

```
## Analysis of 'Wald statistic' Table
## Model: binomial, link: logit
## Response: Desfecho
## Terms added sequentially (first to last)
##
##           Df    X2 P(>|Chi|)
## TGO        1 15.64   7.7e-05 ***
## ICTERICIA  1  9.82    0.0017 **
## Idade      1  1.64    0.2001
## SEXO       1  0.30    0.5814
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Explicação do modelo

Foi-se escolhido um modelo com estrutura de correlações baseadas em um processo auto regressivo de ordem um que resultou em um valor α para a matriz de correlações de 0.972. Optou-se por utilizar o logarítmo natural da variável TGO afim de suavizar o comportamento dessa variável. Além disso, a variável de idade foi dicotomizada afim de melhor sua significância estatística.

Interpretação das variávies

Estimativas Negativas indicam diminuição na razão de chances de óbito e estimativas prositivas indicam aumento, mesma interpretação da regressão logística.

- O aumento de uma unidade de HT aumenta a razão de chances de cura em = vezes.
- Um aumento de uma unidade no logratímo do TGO aumenta razão de chances de óbito em exp(0.0641)=1.07 vezes
- A presença do sintoma de Icterícia aumenta a razão de chances de óbito em exp(1.6710)=5.32 vezes.

As variáveis de SEXO e idade não se mostraram significativas mas foram incluídas.

Análise da predição

```
prob=predict(ajuste_gee,type=c("response"))

dados_roc <- dados_scaled %>%ungroup %>%
  select(Desfecho,TGO,ICTERICIA,pac,dia_doenca) %>%
mutate(p= prob)
require(pROC)
g <- auc(Desfecho ~ p, data = dados_roc)
g
```

```
## Area under the curve: 0.801
```

```
dados_roc <- dados_roc %>% ungroup %>% mutate(predicao = ifelse(p>0.45,1,0))

confusionMatrix(dados_roc$predicao %>% as.factor,dados_roc$Desfecho%>% as.factor,positive='1')
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 243  27
##          1  91  87
##
##                Accuracy : 0.737
##                  95% CI : (0.693, 0.777)
##     No Information Rate : 0.746
##     P-Value [Acc > NIR] : 0.69
##
##                   Kappa : 0.414
##
##  Mcnemar's Test P-Value : 6.65e-09
##
##             Sensitivity : 0.763
##             Specificity : 0.728
##          Pos Pred Value : 0.489
##          Neg Pred Value : 0.900
##              Prevalence : 0.254
##          Detection Rate : 0.194
##    Detection Prevalence : 0.397
##       Balanced Accuracy : 0.745
##
##        'Positive' Class : 1
##
```

```
algum_acerto <- dados_roc %>% select(pac,p,Desfecho,dia_doenca) %>%group_by(pac) %>%  filter(Desfecho ==
```

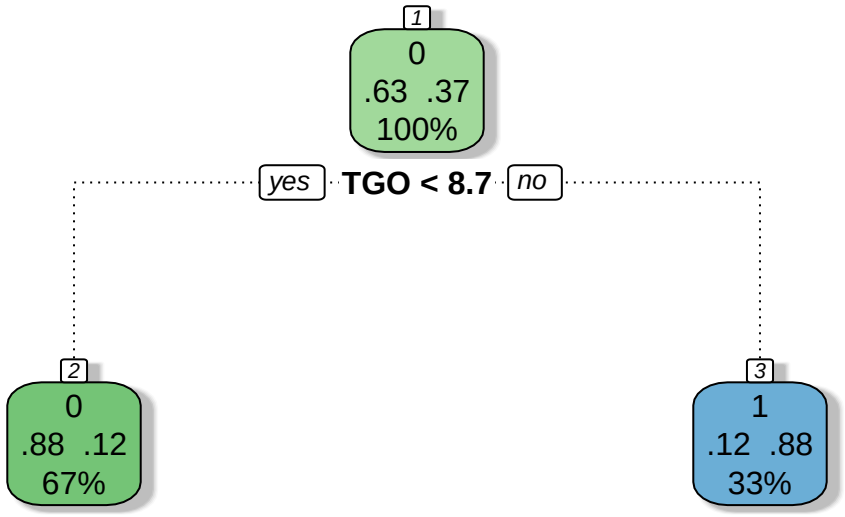
É importante notar que apesar da acurácia não ser extremamente alta, estamos falando da acurária de predição para todos os registros observados. Em alguns períodos de doença a diferenciação de indivíduos pode ser mais complexa.

Árvore de Decisão

Podemos nos basear no primeiro registro observado para construir uma árvore de decisão para tentar prever o óbito.

```
require(rpart)
require(rattle)
dados_primeira_obs=dados2  %>% mutate(Plaquetas=log2(Plaquetas),Leucocitos=log2(Leucocitos),
                                      TGO=log(TGO),TGP=log2(TGP)) %>% select(-dt_inic) %>% group_by(pa
  filter(dia_doenca==min(dia_doenca))

Arvore = rpart(Desfecho~ Idade+SEXO+TGO+ICTERICIA,data=dados_primeira_obs,maxdepth=6,method='class')
fancyRpartPlot(Arvore)
```



Rattle 2019−dez−22 19:35:12 gusta

```
t_pred <- predict(Arvore,dados_primeira_obs)[,1]


confusionMatrix(ifelse(t_pred<.5,1,0) %>% as.factor,dados_primeira_obs$Desfecho %>% as.factor,positive=
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 58  8
##          1  4 28
##
##                Accuracy : 0.878
##                  95% CI : (0.796, 0.935)
##     No Information Rate : 0.633
##     P-Value [Acc > NIR] : 5.07e-08
##
##                   Kappa : 0.73
##
##  Mcnemar's Test P-Value : 0.386
##
##             Sensitivity : 0.778
##             Specificity : 0.935
##          Pos Pred Value : 0.875
##          Neg Pred Value : 0.879
##              Prevalence : 0.367
```

```
##          Detection Rate : 0.286
##    Detection Prevalence : 0.327
##       Balanced Accuracy : 0.857
##
##        'Positive' Class : 1
##
```

O resultado baseando-se apenas em uma variável já tem uma discriminação aceitável, tem em vista que 88% dos casos que não vão a óbito tem o valor do TGO inferior a $\exp(8.7) = 6003$ e 88% dos casos de óbito te valor do TGO superior a 6003