Análise de sobrevivência com dados de diálise - Segunda Avaliação

Universidade Federal da Paraíba - CCEN

Gabriel de Jesus Pereira

20 de outubro de 2024

```r
library(flexsurv)
library(survminer)
library(discSurv)
library(survival)
library(tidyverse)
library(vroom)
library(ggsurvfit)
library(mice)

df <- read_delim(
    "sobrevivencia/segunda_avaliacao/includes/dialcompete.txt",
    delim = " ") |>
  mutate(
    intervalo = cut(
      tempo,
      breaks = 1:44,
      labels = paste0("[", 1:43, ",",  2:44, ")"),
      right = FALSE
    ),
    doenca = factor(doenca),
  ) |>
  select(-intervalo)
```

Questão 1

Com seu banco de dados utilizado na primeira prova, escolhido em: http://sobrevida.fiocruz.br/dados.html. A partir do banco de dados escolhido por você, faça o que se pede a seguir.

(a) Ajuste o modelo de Cox aos dados. Interprete os coeficientes do modelo escolhido. Quais critérios você considerou para obter os modelos?

A base de dados utilizada para o relatório possuí poucas variáveis, o que torna difícil a tarefa de melhorar o modelo ou até mesmo fazer uma análise mais elaborada. Por esse motivo, uma das variáveis que não apresentou significância estatística para o modelo (doenca), foi mantida para que pudesse analisar o seu impacto nos pacientes. Assim, foram mantidas duas variáveis, a variável de idade que o paciente iniciou a diálise e de pacientes que tinham doenças específicas. O critério para seleção das variáveis foi pelo teste t, que permitiu verificar que apenas a variável de idade era significante. O teste considerou um nível de significância de 5%.

```
fit_cox <- coxph(
    Surv(tempo, status) ~ idade + doenca,
    data = df,
    method = "breslow")
summary(fit_cox)
```

```
Call:
coxph(formula = Surv(tempo, status) ~ idade + doenca, data = df,
    method = "breslow")

  n= 2453, number of events= 686

                coef exp(coef)  se(coef)      z Pr(>|z|)
idade       0.011220  1.011283  0.002845  3.943 8.04e-05 ***
doencadiab  0.344346  1.411066  0.249773  1.379    0.168
```

```
doencahiper -0.162741  0.849812  0.237105 -0.686    0.492
doencaoutr  -0.222810  0.800267  0.257236 -0.866    0.386
doencarim   -0.297050  0.743007  0.243107 -1.222    0.222
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

            exp(coef) exp(-coef) lower .95 upper .95
idade          1.0113     0.9888    1.0057     1.017
doencadiab     1.4111     0.7087    0.8648     2.302
doencahiper    0.8498     1.1767    0.5339     1.353
doencaoutr     0.8003     1.2496    0.4834     1.325
doencarim      0.7430     1.3459    0.4614     1.197

Concordance= 0.582  (se = 0.012 )
Likelihood ratio test= 53.49  on 5 df,   p=3e-10
Wald test            = 57.97  on 5 df,   p=3e-11
Score (logrank) test = 59.61  on 5 df,   p=1e-11
```

Ao analisar primeiro os pacientes com diabetes, observa-se que a razão de risco é aproximadamente 1,41, indicando que eles têm um risco 41% maior de sofrer o evento (óbito, transplante, etc.) em comparação com o grupo de referência. No entanto, o valor-p sugere que esse resultado não é estatisticamente significativo ao nível de 5%. Para os pacientes com hipertensão, a razão de risco é aproximadamente 85% em relação aos outros grupos de doenças. Além disso, para cada ano de idade ao iniciar a diálise do paciente, contribui para o aumento do risco de morte em 1,13%.

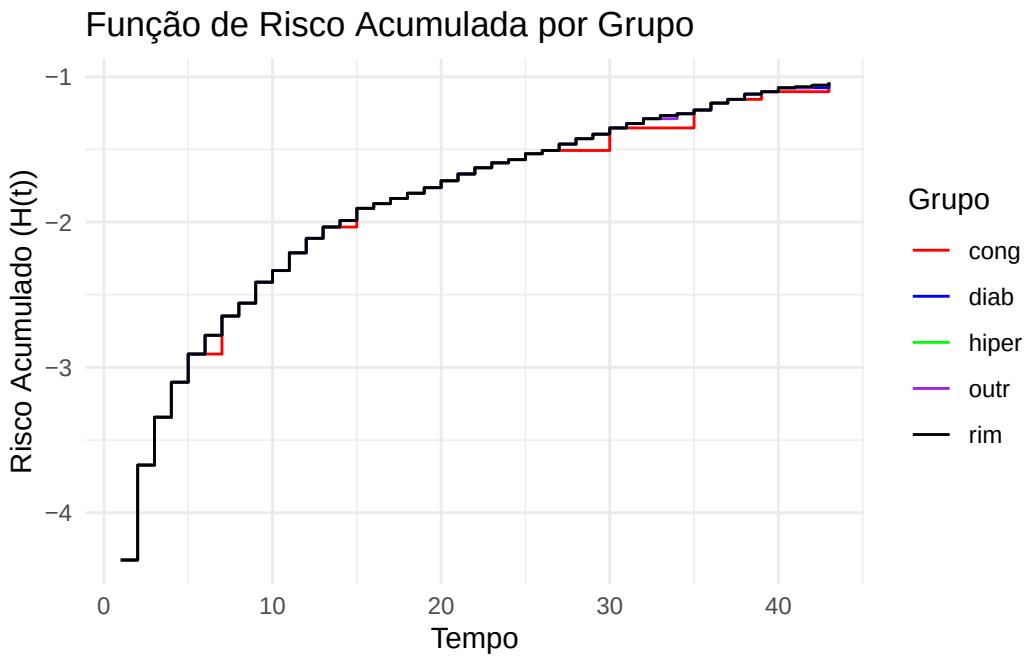
Para a doença renal, o risco é apenas de 74,30% em relação aos pacientes pertencentes aos demais grupos de doença. Para os que estão no grupo de outras doenças, o risco é de 80,03% em relação aos demais.

(b) Verifique a proporcionalidade dos riscos de acordo com o método que você desejar

A proporcionalidade de risco foi verificada a partir de três métodos. O primeiro deles foi pelo método gráfico, como pode ser visto na figura abaixo:

```
risco_acumulado <- basehaz(fit_cox, centered = FALSE) |>
    mutate(hazard = log(hazard))

dados <- merge(
```

```
    df |>
        rename(time = "tempo"),
    risco_acumulado,
    by="time")

ggplot(dados, aes(x = time, y = hazard, color = doenca)) +
  geom_step() +
  scale_color_manual(values = c("red", "blue", "green", "purple", "black")) +
  labs(title = "Função de Risco Acumulada por Grupo",
       x = "Tempo",
       y = "Risco Acumulado (H(t))",
       color = "Grupo") +
  theme_minimal()
```



Através do método gráfico, não foram encontrados indícios de violação da suposição de riscos proporcionais. O segundo método utilizado foi o teste de hipótese para verificar essa suposição, que confirmou os resultados observados graficamente. Ao nível de significância de 5%, tanto a variável idade quanto a variável doença, assim como o teste global, não indicaram qualquer violação da suposição de riscos proporcionais.

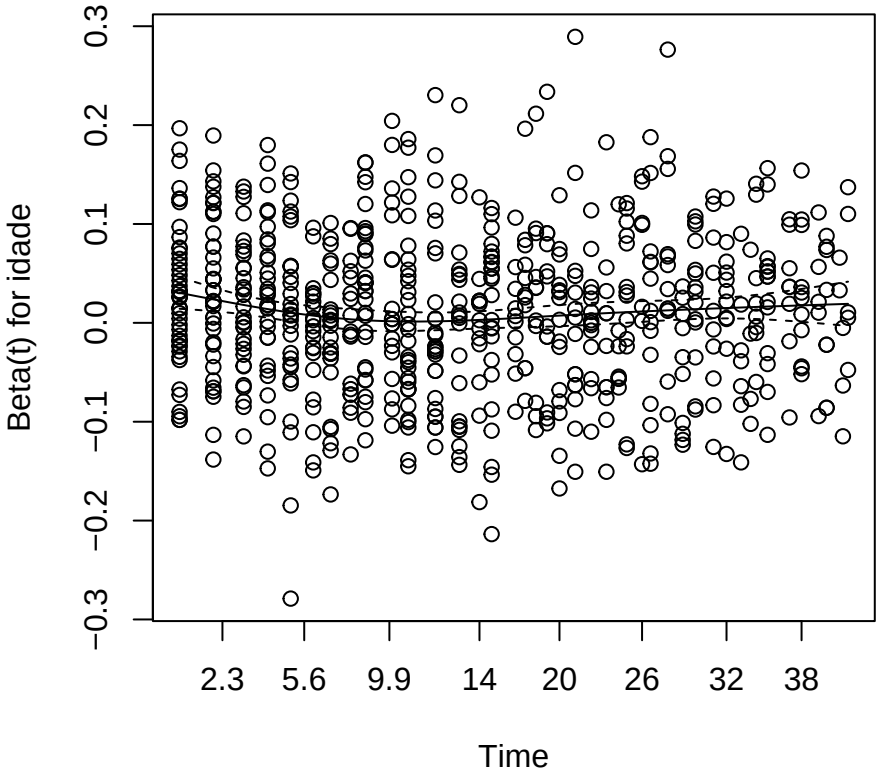
```
cox.zph(fit_cox, transform = "identity")
```

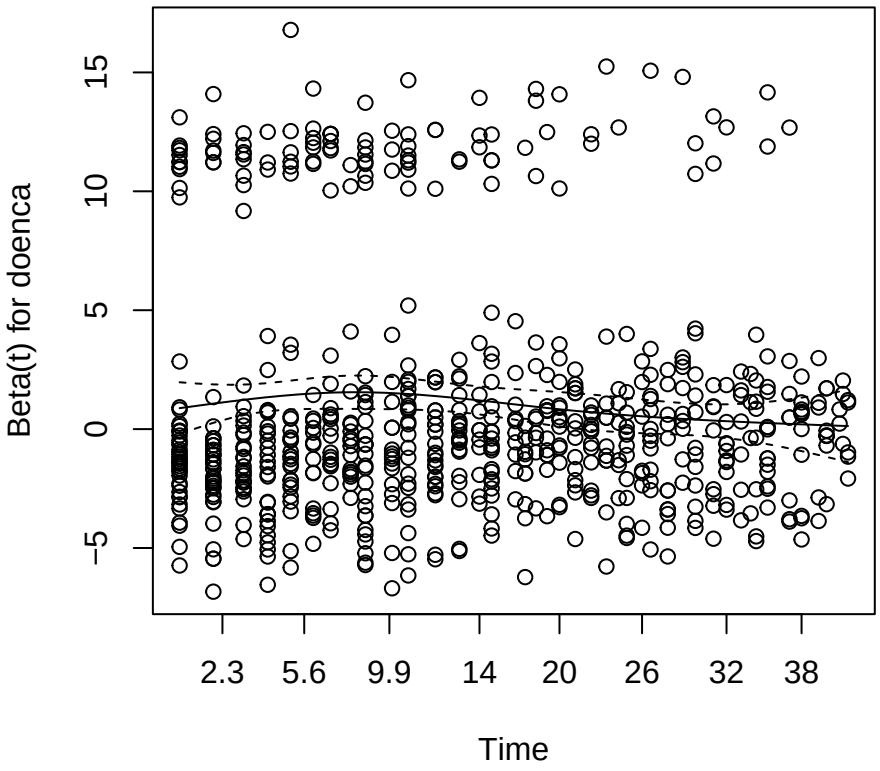
```
       chisq df    p
idade  0.443  1 0.51
```

```
doenca 6.806  4 0.15
GLOBAL 6.824  5 0.23
```

Por fim, foram utilizados os resíduos padronizados de Schoenfeld para analisar a suposição de riscos proporcionais, conforme ilustrado nos gráficos abaixo. A partir do gráfico de resíduos padronizados de Schoenfeld para a variável idade, utilizada na construção do modelo, não há evidências que sugiram a rejeição da hipótese de riscos proporcionais, uma vez que não são observadas tendências ao longo do tempo. O mesmo resultado se aplica à variável doença, pois também não se identificam tendências evidentes ao longo do tempo.

```
plot(cox.zph(fit_cox))
```





(c) Faça uma análise de resíduos do modelo que você escolheu. Qual ou quais resíduos você escolheu? Interprete.

Os resíduos escolhidos foram o resíduo de martingal e o de deviance. O resíduo de martingal é uma modificação dos resíduos de Cox-Snell e é visto como uma estimativa do número de falhas em excesso observadas na amostra. Ele é definido da seguinte forma:

$$m_i = \delta_i - e_i$$

em que δ_i é a variável indicadora de falha e e_i os resíduos de Cox-Snell.

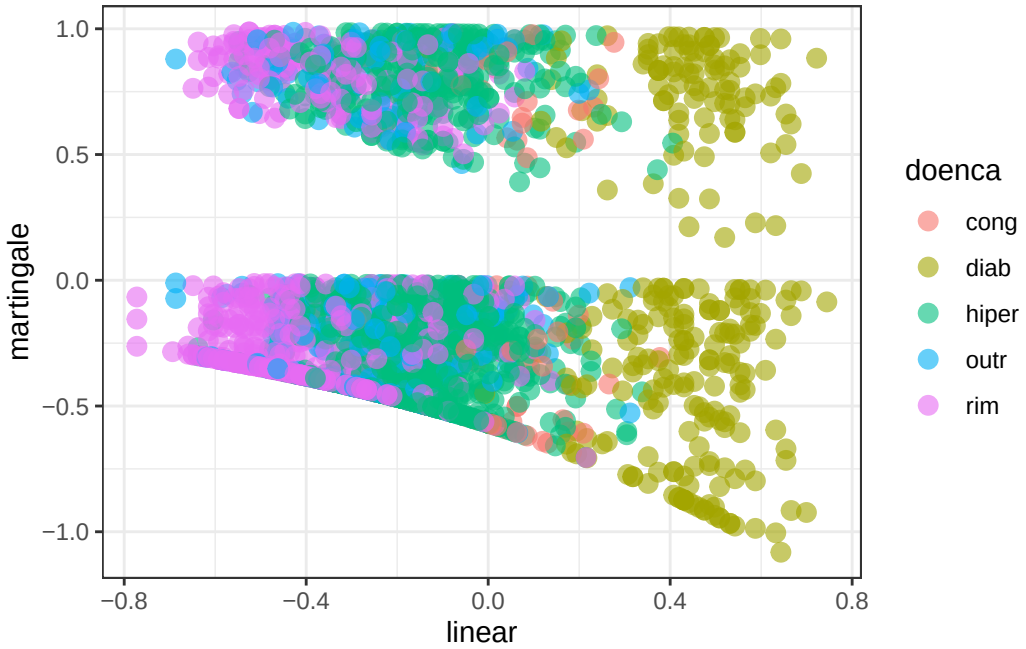
O gráfico abaixo mostra os resíduos de martingal em relação ao preditor linear do

modelo de Cox ajustado. O gráfico não sugere a presença de outliers. O mesmo pode ser verificado no gráfico dos resíduos de deviance. No entanto, ambos os gráficos apresentam problemas de adequação do modelo, uma vez que os resíduos não exibem um comportamento aleatório em torno de 0. Os resíduos de deviance são definidos da seguinte forma:

$$d_i = \text{sinal}\,(m_i)\,\{-2\,[m_i + \log\,(\delta_i - m_i)]\}^{1/2}$$

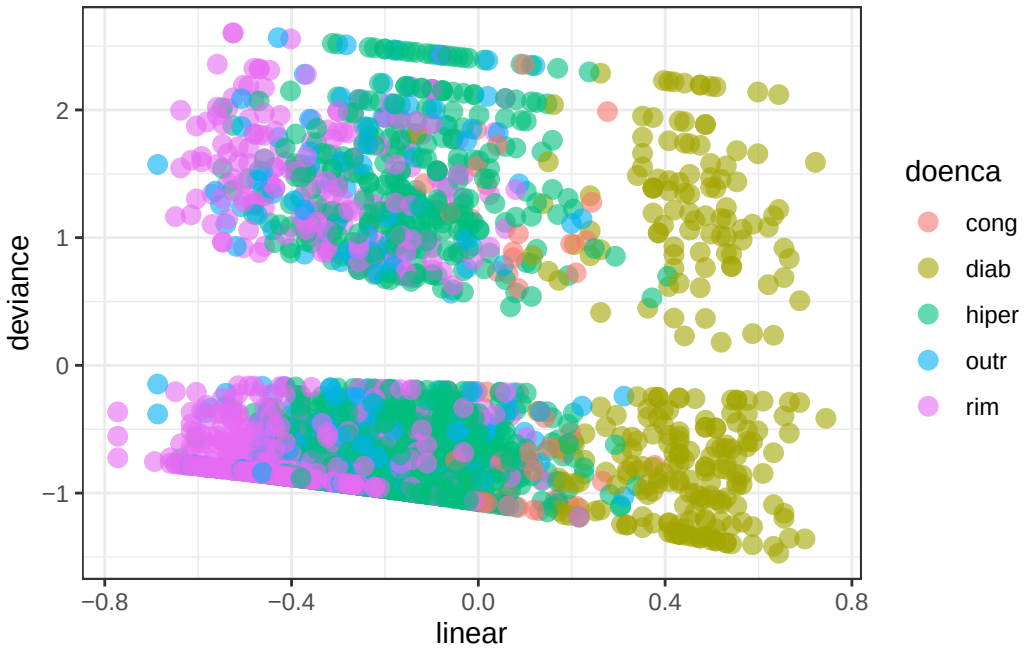
```
mart = df |>
    mutate(
        linear = fit_cox$linear.predictors,
        martingale = resid(fit_cox, type = "martingale"),
        deviance = resid(fit_cox, type = "deviance")
    ) |>
    ggplot(aes(x = linear, y = martingale, color = doenca)) +
    geom_point(size = 3, alpha = .6) +
    theme_bw()
```

```
mart
```




```
dev = df |>
    mutate(
        linear = fit_cox$linear.predictors,
```

```
        martingale = resid(fit_cox, type = "martingale"),
        deviance = resid(fit_cox, type = "deviance")
    ) |>
    ggplot(aes(x = linear, y = deviance, color = doenca)) +
    geom_point(size = 3, alpha = .6) +
    theme_bw()
dev
```



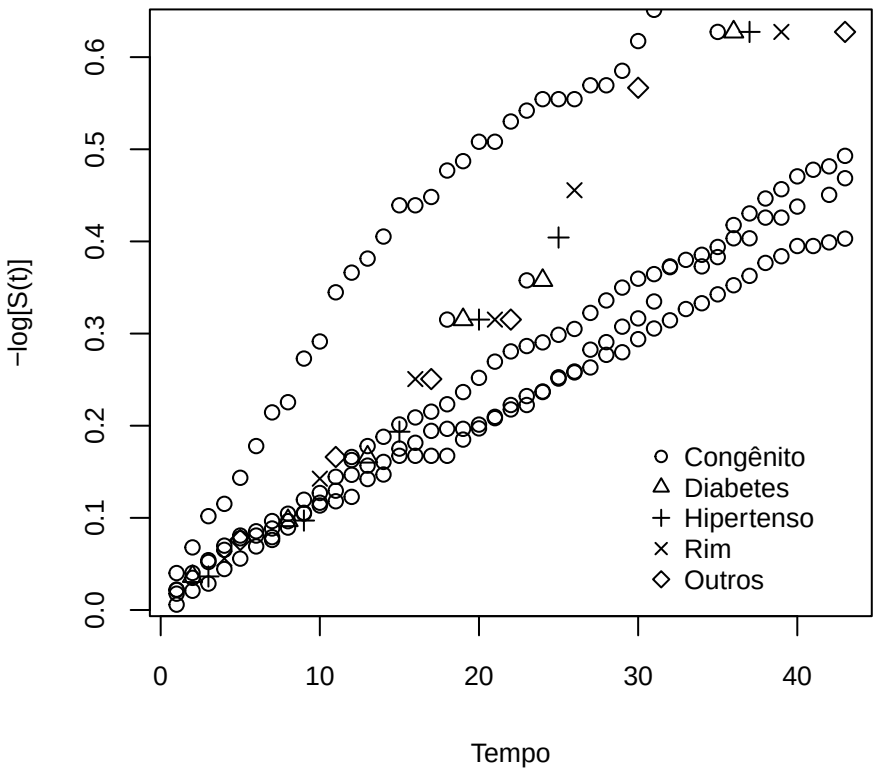
(d)

Ajuste seu modelo utilizando os modelos paramétricos de regressão Exponencial e Weibull.

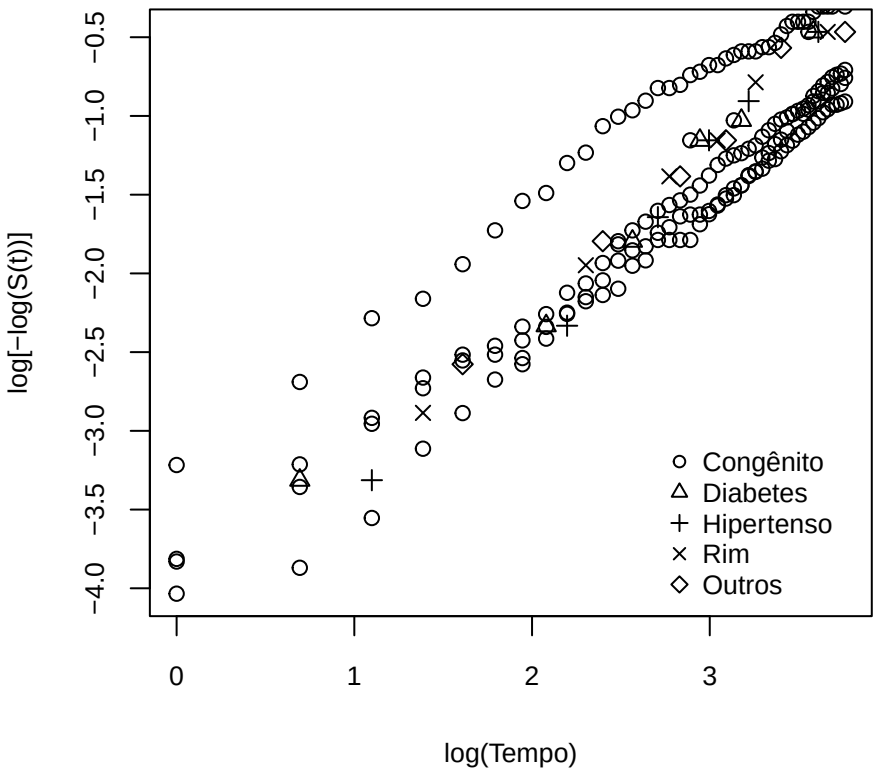
Primeiramente, foi realizada uma análise gráfica, separada por grupos, para identificar os melhores candidatos à modelagem dos dados em estudo. Os gráficos abaixo ilustram essa análise. O primeiro corresponde ao modelo exponencial, e o segundo ao modelo Weibull. No gráfico do modelo exponencial, observa-se que alguns grupos seguem aproximadamente uma linha reta, mas há desvios consideráveis na parte superior. Já no gráfico da distribuição Weibull, a maioria dos grupos apresenta um bom ajuste aos dados, sugerindo que este modelo pode ser uma boa opção para a análise.

```
ekm <- survfit(
    Surv(tempo, status) ~ doenca,
```

```
    data = df, type = "kaplan-meier")

st1<-ekm[1]$surv
time1<-ekm[1]$time

st2<-ekm[2]$surv
time2<-ekm[2]$time

st3<-ekm[3]$surv
time3<-ekm[3]$time

st4<-ekm[4]$surv
time4<-ekm[4]$time

st5<-ekm[5]$surv
time5<-ekm[5]$time


plot(time1, -log(st1), pch=1:5, xlab="Tempo", ylab="-log[S(t)]",
     cex.lab=.8, cex.axis=.8)
points(time2, -log(st2))
points(time3, -log(st3))
points(time4, -log(st4))
points(time5, -log(st5))
legend(30, 0.2, pch=1:5,
    c("Congênito","Diabetes", "Hipertenso", "Rim", "Outros"),
    bty="n", cex=.8)
```




```
plot(log(time1), log(-log(st1)), pch=1:5, xlab="log(Tempo)",
     ylab="log[-log(S(t))]", cex.lab=.8, cex.axis=.8)
points(log(time2), log(-log(st2)))
points(log(time3), log(-log(st3)))
points(log(time4), log(-log(st4)))
points(log(time5), log(-log(st5)))
legend(2.7, -3.0, pch=1:5,
    c("Congênito","Diabetes", "Hipertenso", "Rim", "Outros"),
     bty="n", cex=.8)
```



O coeficiente estimado para a idade é 0,01203, com um intervalo de confiança de 95% entre 0,00645 e 0,01761. Isso indica que a idade ao iniciar a diálise apresenta um aumento no risco do evento.

O coeficiente estimado para a categoria de diabetes é 0,36694, com um intervalo de confiança de 95% entre -0,12249 e 0,85637. A razão de tempo mediano correspondente é 1,44331, indicando que os pacientes com diabetes tem um tempo mediano até a ocorrência do evento 1,44 vezes maior que os que não pertencem ao grupo de diabetes.

No caso de hipertensão (doencahiper), o coeficiente estimado é 0,44802, com uma razão de tempo mediano de 1,53266. O tempo mediano de vida até a ocorrência do evento é 1,53 maior para os pacientes que pertencem ao grupo de hipertensão do que aqueles que não pertencem a esse grupo.

Para outras doenças (doencaoutr), o coeficiente estimado é 0,13290, com uma razão de tempo mediano de 0,80272, também sem significância estatística, já que o intervalo

de confiança de 95% inclui 0. Os pertencentes ao grupo de outras doenças apresentam uma redução no risco dos eventos de óbitos considerados (transplante, óbito por causa renal, ...).

Em relação à doença renal (doencarim), o coeficiente estimado é 0,28170, com uma razão de tempo mediano de 0,73191. Os pacientes pertencentes às doenças de rim apresentam uma redução no risco dos eventos. No entanto, é importante dizer que apenas a variável de idade possuí significância, pois é o único que não contém o valor 0 em seu intervalo de confiança.

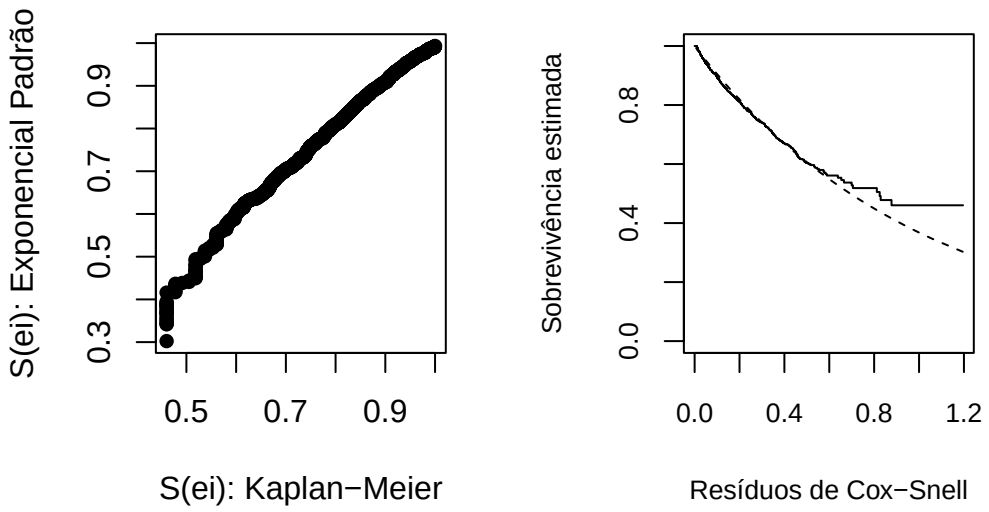
```
fit_exp <- flexsurvreg(
    Surv(tempo, status) ~ idade + doenca,
    data = df,
    dist='exponential')
fit_exp
```

```
Call:
flexsurvreg(formula = Surv(tempo, status) ~ idade + doenca, data = df,
    dist = "exponential")

Estimates:
             data mean  est       L95%      U95%      se         exp(est)
rate               NA    0.00841   0.00495   0.01426   0.00227        NA
idade         42.34570   0.01203   0.00645   0.01761   0.00285   1.01210
doencadiab     0.11415   0.36694  -0.12249   0.85637   0.24971   1.44331
doencahiper    0.44802  -0.16519  -0.62986   0.29948   0.23708   0.84773
doencaoutr     0.13290  -0.21975  -0.72394   0.28444   0.25724   0.80272
doencarim      0.28170  -0.31210  -0.78836   0.16415   0.24299   0.73191
             L95%      U95%
rate               NA        NA
idade          1.00647   1.01776
doencadiab     0.88471   2.35460
doencahiper    0.53266   1.34915
doencaoutr     0.48484   1.32901
doencarim      0.45459   1.17839

N = 2453,  Events: 686,  Censored: 1767
Total time at risk: 57312
Log-likelihood = -3691.498, df = 6
AIC = 7394.995
```

O gráfico abaixo mostra que o modelo exponencial parece apresentar um bom ajuste.

No entanto, no gráfico a direita, não parece apresentar um bom ajuste, pois há um desvio para valores maiores que 0,6 nos resíduos de Cox-Snell.

```
res_exp <- resid(fit_exp, type = "coxsnell")
ekm1 <- survfit(
    Surv(res_exp, status) ~ 1,
    data = df, type = "kaplan-meier")
t <- ekm1$time
st <- ekm1$surv
sexp <- exp(-t)

par(mfrow=c(1,2))

plot(st, sexp, xlab = "S(ei): Kaplan-Meier",
  ylab = "S(ei): Exponencial Padrão", pch = 16)

plot(ekm1, conf.int=F, lty=c(1,1), xlab="Resíduos de Cox-Snell",
     ylab="Sobrevivência estimada", cex.lab=.8, cex.axis=.8)
res_exp<-sort(res_exp)
exp1<-exp(-res_exp)
lines(res_exp, exp1, lty=2)
legend(1.3, 0.9, lty=c(1,2), c("Kaplan-Meier","Exponencial padrão"),
       lwd=1, bty="n", cex=0.8)
```



Agora analisando os resultados do modelo de Weibull. O coeficiente estimado para a idade é -0,01226, com um intervalo de confiança de 95% entre -0,01806 e -0,00646. Isso indica que a idade tem um efeito negativo sobre o tempo de sobrevivência, ou seja, à medida que a idade aumenta, o tempo de sobrevivência diminui.

O coeficiente estimado para diabetes é -0,37534, com um intervalo de confiança de 95% entre -0,88369 e 0,13300, indicando que a diabetes pode estar associada a um risco menor de falha, embora o intervalo de confiança inclua 0. A razão de tempo mediano é 0,68705, mas, como o intervalo de confiança inclui 0, o efeito não é estatisticamente significativo ao nível de 5%. Além disso, o tempo mediano até a ocorrência do evento de pacientes sem diabetes é 1,455 vezes menor do que aqueles pacientes que tiveram diabetes.

O coeficiente estimado para hipertensão é 0,17018, com um intervalo de confiança entre -0,31233 e 0,65269. A razão de risco estimada é 1,18552, sugerindo um risco de tempo mediano levemente maior para pacientes com hipertensão. No entanto, o efeito também não é estatisticamente significativo, pois o intervalo de confiança de 95% inclui 0. O tempo mediano até a ocorrência do evento para os pacientes com hipertensão é quase 1,2 vezes maior que dos pacientes que não tem hipertensão.

Para outras doenças (doencaoutr), o coeficiente estimado é 0,22870 (razão de tempo mediano = 1,25697), mas o intervalo de confiança também inclui 0, sugerindo ausência de significância estatística. Para o grupo de pacientes que pertence a outros doenças, o tempo mediano até a ocorrência do evento é 1,25 vezes maior do que os pacientes que não pertecem a esse grupo.

No caso de doença renal (doencarim), o coeficiente estimado é 0,31916 (razão de tempo mediano = 1,37598), mas, da mesma forma, o efeito não apresenta significância estatística, já que o intervalo de confiança de 95% inclui 0. O tempo mediano até a ocorrência do evento para os pacientes pertencentes ao grupo de doença renal é 1,37 vezes maior do que os pacientes que não pertencem.

```
fit_wei <- flexsurvreg(
    Surv(tempo, status) ~ idade + doenca,
    data = df,
    dist='weibull')
fit_wei
```

```
Call:
flexsurvreg(formula = Surv(tempo, status) ~ idade + doenca, data = df,
    dist = "weibull")
```

```
Estimates:
             data mean  est        L95%       U95%       se          exp(est)
shape               NA    0.96319    0.90221    1.02829    0.03214         NA
scale               NA  124.48342   71.48726  216.76758   35.22775         NA
idade         42.34570   -0.01226   -0.01806   -0.00646    0.00296    0.98782
doencadiab     0.11415   -0.37534   -0.88369    0.13300    0.25937    0.68705
doencahiper    0.44802    0.17018   -0.31233    0.65269    0.24618    1.18552
doencaoutr     0.13290    0.22870   -0.29498    0.75239    0.26719    1.25697
doencarim      0.28170    0.31916   -0.17546    0.81379    0.25236    1.37598
             L95%       U95%
shape               NA         NA
scale               NA         NA
idade          0.98211    0.99356
doencadiab     0.41325    1.14225
doencahiper    0.73174    1.92070
doencaoutr     0.74455    2.12206
doencarim      0.83907    2.25644

N = 2453,  Events: 686,  Censored: 1767
Total time at risk: 57312
Log-likelihood = -3690.856, df = 7
AIC = 7395.713
```

Analisando o ajuste do modelo Weibull, ele apresenta um ajuste muito semelhante com o modelo exponencial.

```
res_wei <- resid(fit_wei, type = "coxsnell")
ekm2 <- survfit(
    Surv(res_wei, status) ~ 1,
    data = df, type = "kaplan-meier")
t_wei <- ekm2$time
st_wei <- ekm2$surv
sexp_wei <- exp(-t_wei)

par(mfrow=c(1,2))

plot(st_wei, sexp_wei, xlab = "S(ei): Kaplan-Meier",
  ylab = "S(ei): Weibull Padrão", pch = 16)

plot(ekm2, conf.int=F, lty=c(1,1), xlab="Resíduos de Cox-Snell",
     ylab="Sobrevivência Estimada", cex.lab=1.4, cex.axis=1.3)
```

```
res_wei <- sort(res_wei)
exp2 <- exp(-res_wei)
lines(res_wei, exp2, lty=2)
legend(1.3, 0.9, lty=c(1,2), c("Kaplan-Meier", "Weibull Padrão"),
       lwd=1, bty="n", cex=0.8)
```