Universidade de Brasília - UnB
Instituto de Ciências Exatas - IE
Departamento de Estatística - EST

Análise de Sobrevivência de pacientes com Insuficiência Renal

Guilherme Mendes, Rafael Ribeiro e Tonny Barbosa

Brasília
04 de dezembro de 2018

Introdução

A análise de sobrevivência consiste, de modo geral, em analisar os tempos de vida dos indivíduos desde o seu momento de entrada no estudo, até o momento em que ocorre o acontecimento de interesse, que é definido à partida. Este acontecimento é geralmente denominado como uma **falha**, sendo assim, podemos estimar a função de sobrevivência, que é dada pela probabilidade de que um item falhe até um determinado tempo t. Neste trabalho, a falha do estudo é o óbito do paciente.

Indivíduos que continuam vivos após o término do estudo ou que abandonam o tratamento são exemplos de indivíduos com tempo de vida censurado, ou seja, quando algum paciente não chega a ocorrer o acontecimento de interesse durante o período de observação, isto é, o paciente parou de participar do estudo. Esse caso particular temos um cenário denominado como **censura**.

O tempo de vida dos indivíduos é afetado por variáveis que são observadas no estudo. Este tipo de dados sugere uma análise de regressão para avaliar quais variáveis são mais significativas para explicar o modelo, porém não podemos fazer uma regressão habitual devido as particularidades que este tipo de dados apresenta. Isso ocorre devido as observações censuradas. Assim a análise de sobrevivência é uma ótima escolha para modelar esse tipo de dados.

Nesse estudo, iremos analisar pacientes submetidos à hemodialise no Rio de Janeiro. Os dados foram coletados num período de 44 meses e nos informa o tempo que um paciente sobreviveu desde o início do estudo, bem como a causa da sua insuficiência renal. O banco de dados será explicado com mais propriedades ao longo do trabalho. Primeiro, iremos abordar conceitos e metodologia que será utilizada na análise.

Conceitos Básicos

O tempo de vida de um determinado indivíduo, de uma população homogênea, é representado por uma variável aleatória T, não negativa e absolutamente contínua. Podemos então definir a **Função de Sobrevivência** desse indivíduo, denotada por $S(t)$, é definida como a probabilidade de um indivíduo sobreviver a um tempo t. Ela é expressa por:

$$S(t) = P(T \geq t) = \int_t^\infty f(x)dx,$$

sendo que $S(t)$ é uma função monótona decrescente e contínua. Em consequência a função de distribuição acumulada é definida como a probabilidade de uma observação não sobreviver ao tempo t, isto é, $F(t) = 1 - S(t)$.

A **função de risco** pode ser expressa em termos da função densidade de probabilidade e da função de sobrevivência, isto é,

$$h(T) = \frac{f(t)}{S(t)}$$

A **função de risco acumulado** $H(t)$ também pode ser utilizada para representar o tempo de sobrevivência. Sua relação com a função de sobrevivência é dada por:

$$H(t) = -[\log S(t)]$$

O gráfico da função de taxa de falha da variável T pode assumir várias formas e é importante utilizar uma metodologia para identificar o modelo mais apropriado para esta variável. Gráfico do **tempo total em teste** também conhecido como curva TTT, é um desses gráficos que podem ser utilizados. A curva TTT é obtida construindo um gráfico de

$$G(r/n) = \frac{[(\sum_{i=1}^{r} T_{i:n}) + (n-r)T_{r:n}]}{\sum_{i=1}^{n} T_i}$$

sendo que as estatisticas de ordem da amostra são r/n, onde $r = 1, ..., n$ e $T_{i:n}$, com $i = 1, ..., n$. A interpretação de $\hat{H}(t)$ é o inversa da curva TTT.

Estimação

Estimador de Kaplan-Meier

Para estimar a função de sobrevivência na presença de censura geralmente é utilizado alguns estimadores. Eles são, Kaplan-Meier, Nelson Aalen e Tabela de vida.

Utilizamos o estimador de Kaplan-Meier, que é definido por:

$$\hat{S}(t) = \prod_{j:t_j<t} \left(1 - \frac{d_j}{n_j}\right)$$

- $t_1 < t_2 < \cdots < t_k$ tempos distintos e ordenados de falha;
- d_j o número de falhas em $t_j, j = 1, \ldots, k,$;
- n_j o número de individuos sob risco em t_j ou seja, os indivíduos que não falharam e não foram censurados até o instante imediatamente anterior a t_j.

As propriedades do estimador de Kaplan-Meier são:

- $\hat{S}(t)$ é um estimador não viciado para amostras grandes;
- $\hat{S}(t)$ é fracamente consistente;
- Kaplan e Meier é o estimador de máxima verossimilhança de $\hat{S}(t)$.

Teste de log-Rank

Para comparar curvas de sobrevivência de variáveis categóricas, fazemos o uso de testes não-paramétricos como por exemplo o teste de logRank e o teste de Wilcoxon.

O teste de logRank é apropriado quando a razão das funções de risco dos grupos a serem comparados é aproximadamente constante, ou seja, temos a suposição de riscos proporcionais, que pode ser verificada pelo comportamento das curvas de sobrevivência de cada grupo. O objetivo é testar se as curvas são iguais, para isso temos as seguintes hipóteses:

$H_0 : S_1(t) = S_2(t)$ Não existe diferença entre as curvas de sobrevivência,

$H_1 : S_1(t) \neq S_2(t)$ Existe diferença entre as curvas de sobrevivência.

O teste de Wilcoxon é mais adequado do que o teste de logRank em situações que não temos suposição dos riscos proporcionais. Esse teste utiliza peso igual ao número de indivíduos sob risco, ou seja, o impacto do teste é maior nos tempos iniciais.

Método de Máxima Verossimilhança

A função de verossimilhança, considerando tempos de falha e censura, é expressa por:

$$L(\theta) = \prod_{i=1}^{r} f(t_i, \theta) \prod_{i=r+1}^{n} S(t_i, \theta)$$

nesse caso,

- r é o número de falhas,
- $n - r$ é o número de censuras,
- $f(t_i, \theta)$ é a $f.d.p$ para os tempos de falha,
- $S(t_i, \theta)$ é a função de sobrevivência para os tempos de censura.

Modelos Probabilísticos

As distribuições mais utilizadas para modelar o tempo de sobrevivência são a Exponencial, Weibull, Log-logístico, Log-normal e Gama.

Normalmente escolhemos o modelo probabilístico baseado no comportamento da curva TTT, entretanto essa curva não leva em consideração as censuras. Como nossos dados possui muitas censuras, não é muito apropriado selecionar o modelo com base na

curva TTT, como alternativa podemos utilizar como referência a função de risco acumulado.

Os critérios de informação Akaike e Bayesiano nos permitem comparar modelos definidos por diferentes distribuições, em que menores valores de AIC e BIC nos indicam um melhor modelo. Dessa forma, será comparado os critérios de informação entre os modelos exponencial, Weibull, log-normal e log-logístico, a fim de obter o melhor modelo.

O Critério de Informação de Akaike (AIC) é definido como

$$AIC = -2\log(L(\theta)) + 2[(p+1)+1],$$

em que $L(\theta)$ é a função de máxima verossimilhança do modelo e p é o número de variáveis explicativas consideradas no modelo.

O Critério de Informação Bayesiano (BIC) é definido como

$$BIC = -2\log(L(\theta)) + [(p+1)+1]log(n).$$

em que n é o número de observações.

Distribuição Log-normal

Se T tem distribuição Log-normal com parâmetros μ e σ, então a variável $Y = \log(T)$ tem distribuição normal com parâmetros μ e σ.

A densidade de probabilidade da distribuição log-normal é dada por:

$$f(t) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}t\sigma} \exp\left[-\frac{1}{2}\left(\frac{\log(t)-\mu}{\sigma}\right)^2\right]$$

em que $t > 0$, μ é a média do logaritmo do tempo e σ é o desvio padrão.

A função de sobrevivência da distribuição log-normal é dada por:

$$S(t) = \Phi\left(\frac{-\log(t)+\mu}{\sigma}\right)$$

em que $\Phi(.)$ é a função de distribuição acumulada de uma distribuição normal padrão.

Análise de resíduos

Na regressão linear é comum utilizar a análise de resíduos para verificar a adequação dos modelo ajustado. Em análise de Sobrevivência, devido as censuras e os resíduos não precisarem seguir uma distribuição Normal, outros métodos são utilizados para verificar a qualidade do ajuste.

Resíduo Cox-Snell

Segundo Cox-Snell (1968), Collet (1994), Klein e Moeschberger (1997) e Colosimo e Giolo (2006), os resíduos de Cox-Snell são definidos por:

$$e_i = \hat{H}_0(t_i)\, exp\{ \sum_{k=1}^{p} x_{ip}\hat{\beta}_k \}.$$

Se o modelo estiver bem ajustado, os resíduos e_i's podem ser vistos como uma amostra cesurada de uma distribuição exponencial. Ao fazermos o gráfico da $\hat{H}_i$ versus e_i esperamos aproximadamente uma reta.

Uma das desvantagens do uso de Cox-Snell é que quando o gráfico $\hat{H}_i$ versus e_i não apresenta uma forma linear, não conseguimos identificar o tipo de falha identificado pelo modelo. Para amostras menores verifica-se uma incerteza maior na cauda direita da distribuição.

Resíduo Martingal

O resíduo Martingale é resultado de uma modificação feita no resíduo Cox-Snell e é usado para identificar discrepâncias entre um modelo ajustado e o conjunto de dados. Assim, quando os dados apresentarem censura à direita e as covariáveis não forem dependentes do tempo, O resíduo Martingale é definido por:

$$M_i = \delta_i - \hat{H}_0(t_i)\, exp\{ \sum_{k=1}^{p} x_{ip}\hat{\beta}_k \} = \delta_i - e_i.$$

Segundo Klein e Moeschberger (1997) para verificar a forma funcional da covariável, deve-se ajustar o modelo de interesse e calcular o resíduo Martingal. Então faz-se o gráfico dos e_i's versus a variável em estudo. Se o gráfico apresenta uma forma linear, não é necessário transformar as covariável. Caso contrário, é indicado categorizar a covariável.

Resíduo Deviance

Esse resíduo é uma tentativa de tornar o resíduo Martingal mais simétrico em torno do zero. Qualquer padrão ou tendência incomum no gráfico de resíduos deviance indica que o modelo ajustado pode ser inadequado. Para os modelos de regressão paramétricos, os resíduos deviance são definidos por:

$$\hat{r}_{Di} = sinal(\hat{r}_{Mi})[\, -2(\hat{r}_{Mi}) + \delta_i log(\delta_i - \hat{r}_{Mi})\,]^{1/2}.$$

Análise de Sobrevivência para os Dados

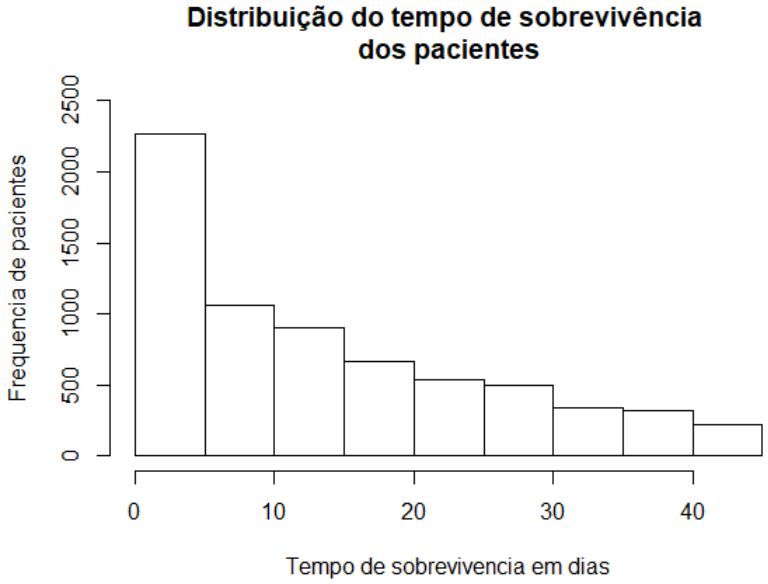
Os dados provêm de uma coorte de 6.805 pacientes que foram submetidos a hemodiálise em 67 unidades de atendimento no Rio de Janeiro, no período de janeiro de 1998 a outubro de 2001. Os dados foram originados pelo sistema Apac (Autorização de Procedimentos de Alta Complexidade – DATASUS). Uma discussão detalhada do tema e da modelagem pode ser encontrada em Carvalho e cols. (2003). Cada paciente possui um registo que apresenta dados para cada variável.

A variável resposta é denotada por **tempo**, a censura é **status** e existem algumas variáveis que influênciam a variável resposta. Elas são: *grande*, *idade* e *causa*. Sendo que a variável causa foi subdividida em variáveis dummy. Essas variáveis são formadas por doenças que são algumas causas da insuficiência renal: diabetes (cdiab), causas renais (crim), congenita (congenita), hipertensão (hip). A lista a seguir contém a descrição das variáveis.

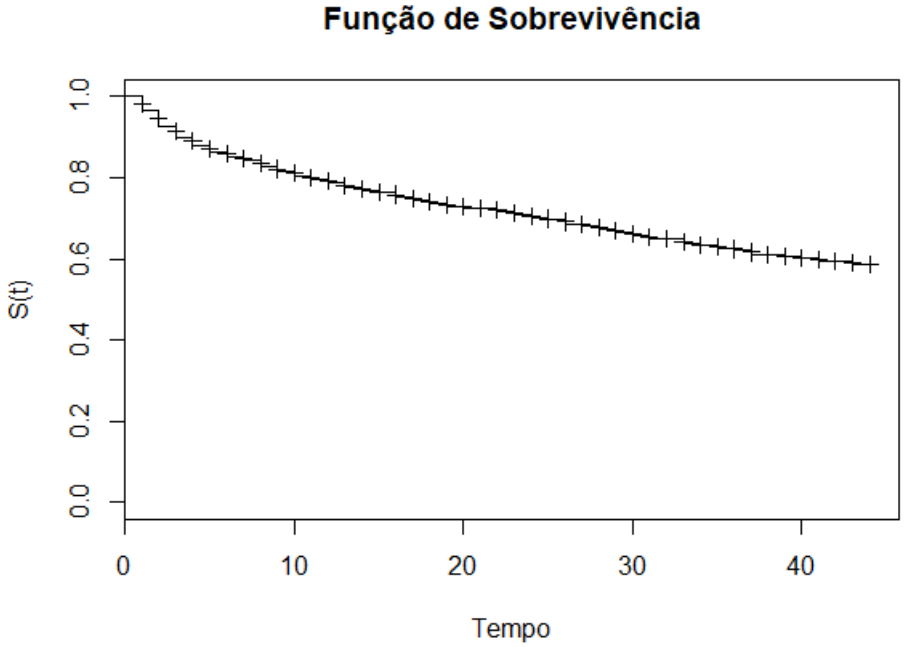
- **unidade** = número do centro de diálise
- **idade** = idade ao iniciar a diálise (0 a 97 anos)
- **inicio** = data do início da primeira diálise
- **fim** = data da interrupção do acompanhamento
- **status** = (0 = censura, 1 = óbito)
- **tempo** = tempo de sobrevivência (meses) (fim - inicio)
- **grande** = número de salas de diálise na unidade de tratamento: 0 = uma ou duas salas; e 1 = três salas ou mais
- **cdiab** = (1 = diabetes como causa da insuficiência renal e 0 = não)
- **crim** = (1 = causas renais e 0 = não)
- **congenita** = (1 = causas congênitas e 0 = não)
- **hip** = (1 = causas hipertensão e 0 = não)
- **out** = (1 = outras causas e 0 = não)

Para realizar nossa análise queremos saber que tipo de distribuição de probabilidade poderia representar a variável aleatória tempo de sobrevivência T. Para isso devemos observar algumas características. Se a variável tempo é contínua e não negativa, devemos verificar se a distribuição normal é adequada para modelar T ou ainda se apresenta forte assimetria. Para descobrirmos a distribuição, primeiramente, fizemos uma análise

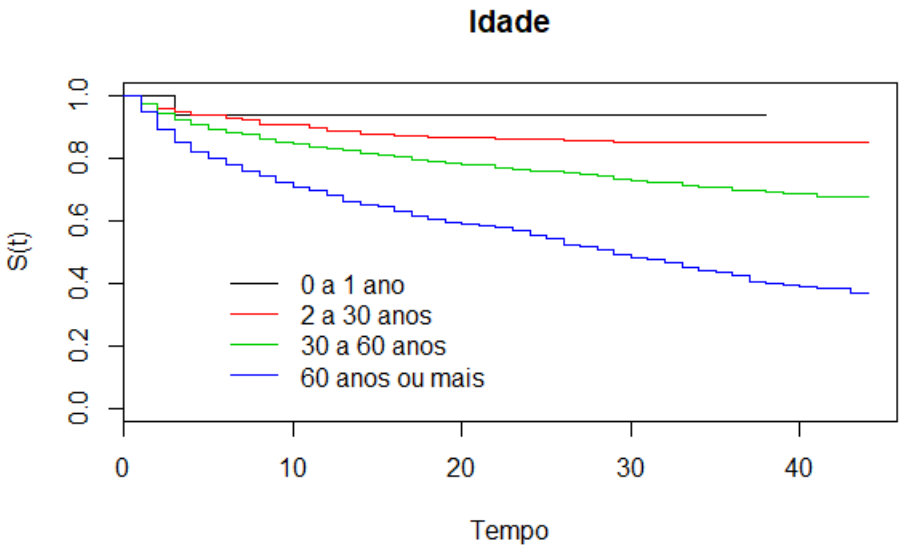
exploratória dos dados. Criamos um histograma para ver a disposição dos dados da variável tempo.

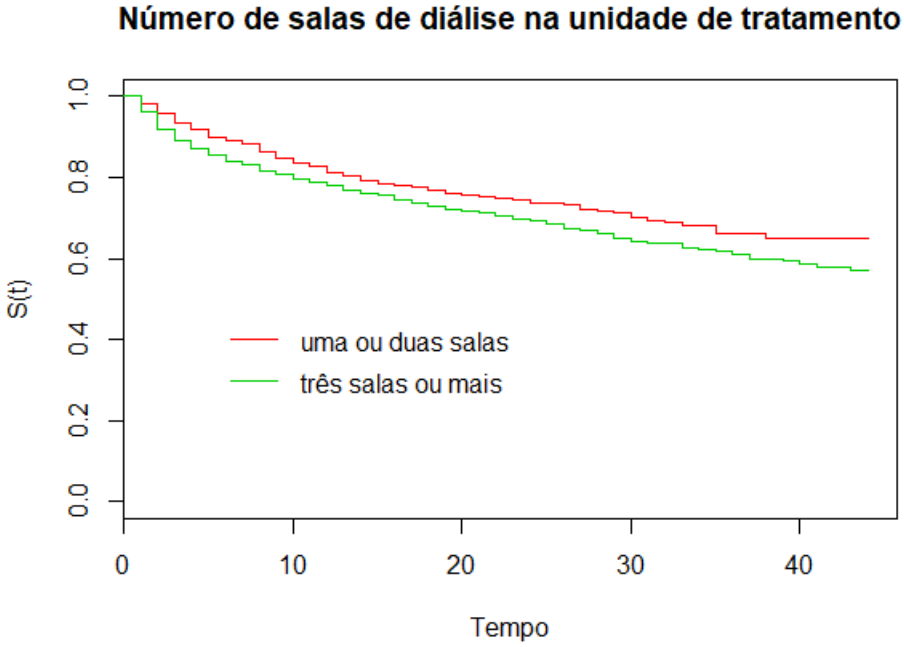


Dado que os dados apresentaram censura, utilizamos o estimador de Kaplan-Meier para estimar a função de sobrevivência. Selecionamos apenas a variável **tempo** e a variável censura representada por **status**.



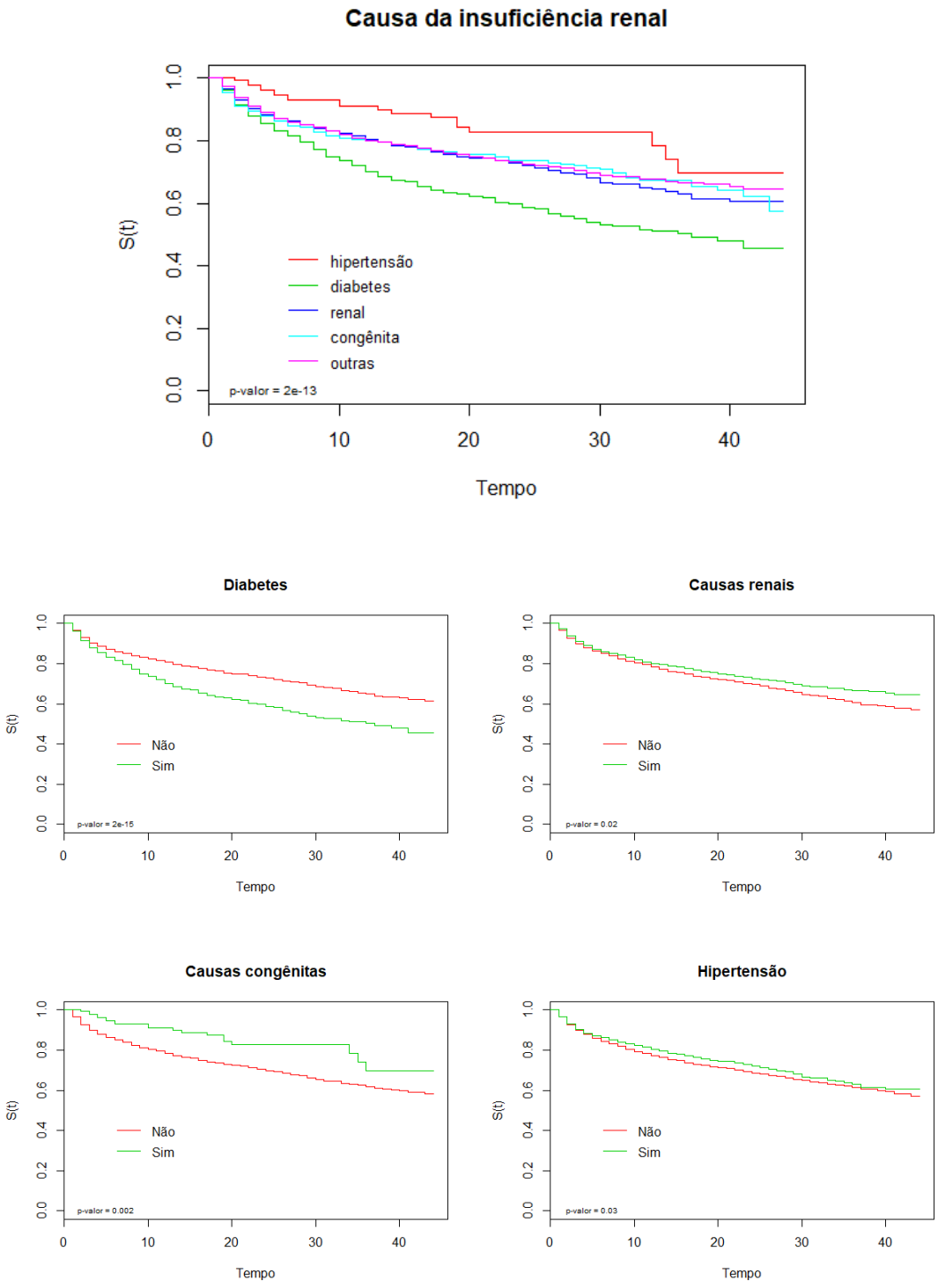
Analisando as funções de sobrevivência para as covariváveis *grande*, *idade* e *causa*, individualmente, podemos que notar que todas elas se mostraram significativas. A variável idade por ser um variável contínua, foi dividida em categorias conforme a legenda.





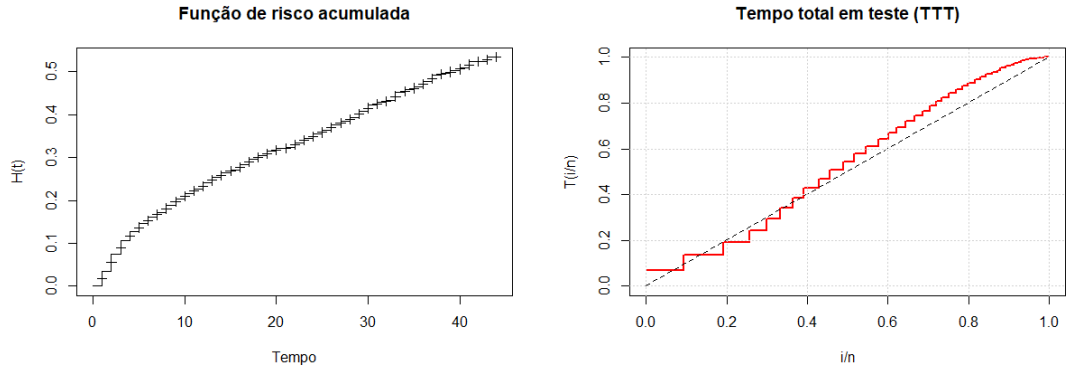
Para a variável *causa*, obter um p-valor significativo nos mostra que há diferença entre os diferentes tipos de causa no efeito do tempo de sobrevivência dos pacientes com insuficiência renal. Entretanto, só obter esse p-valor não nos revela qual causa tem maior efeito, nem especifica entre quais grupos há diferença. Dessa forma, decidimos analisar os tipos de causa de ineficiência renal uma a uma.

Observando o efeito individual de cada causa de insuficiência renal, todas se mostraram significativas. Apesar disso, num modelo de regressão, essas covariáveis podem se mostrarem não significativas.



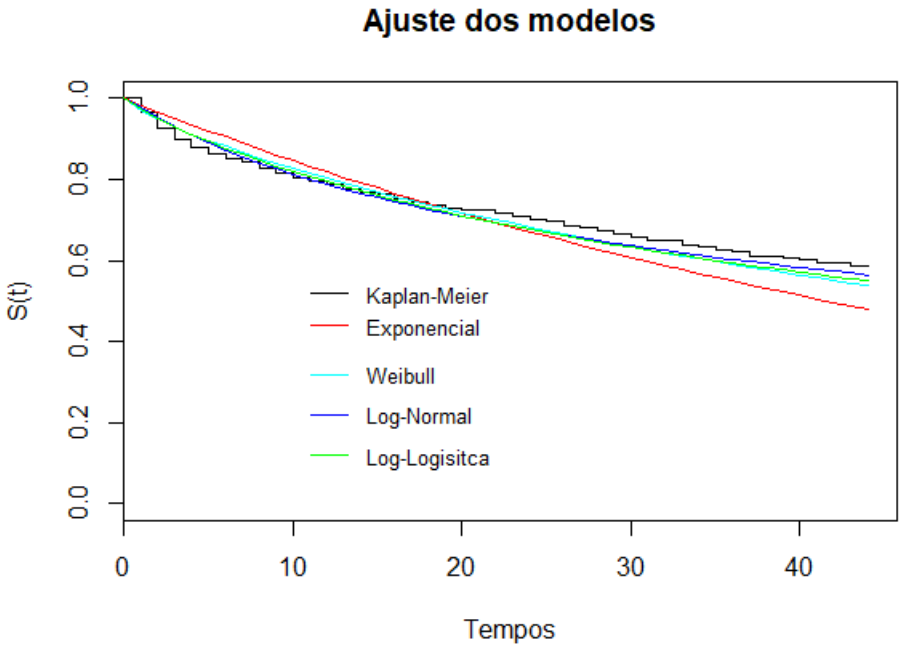
Logo após, construímos os gráficos do risco acumulado e do TTT, assim podemos ter uma ideia da distribuição mais adequada para os dados.

Como o número de censuras é grande, o gráfico da função de risco acumulada, $\hat{H}(t)$, seria uma escolha mais adequada do que o gráfico TTT. Análisamos os dois gráficos, no TTT observamos que a curva pode supor uma forma de banheira ou uma constante. Já o gráfico da risco acumulada aparenta mais um formato de curva côncava. Porém, apenas analisando



esses gráficos não obtivemos respostas muito conclusivas. Então implementamos o critério de informação de Akakike, Akaike corrigido e Bayesiano, para certificar a distribuição.

Modelamos algumas distribuições estatísticas para os tempos de sobrevivência dos dados. Graficamente, as distribuições Weibull, Log-normal e Log-logística apresentaram uma boa modelagem, mas para selecionar o modelo que apresenta o melhor ajuste, iremos considerar os valores de AIC, AICc e BIC para cada distribuição.



Distribuição	AIC	AICc	BIC
Exponencial	16339.96	16339.97	16344.67
Weibull	16212.48	16212.48	16221.89
Log-normal	16017.70	16017.72	16027.12
Log-logístico	16156.33	16156.34	16165.74

Concluída a análise exploratória dos dados, verificamos que a distribuição log-normal apresentou um melhor ajuste. Agora, passamos para etapa de selecionar as variáveis que serão incluídas na modelagem estatística.

Para verificarmos o ajuste utilizamos a função *survreg*. Primeiramente testamos as variáveis uma de cada vez. Logo após testamos o modelo completo e os modelos com a retirada de cada uma das covariáveis. Por fim testamos um modelo com interação. Verificamos que o melhor modelo encontrado foi com as variáveis **grande+cdiab+congênita+grande*congênita**. Sendo que, nesse modelo a interação de grande e congênita deu $p-valor = 0.024$, portanto, foi considerada no modelo.

- Modelo final: A variável resposta **tempo** trouxe como explicação para o modelo mais ajustado as variáveis: grande, diabetes, congenita e a interação entre grande e congenita.

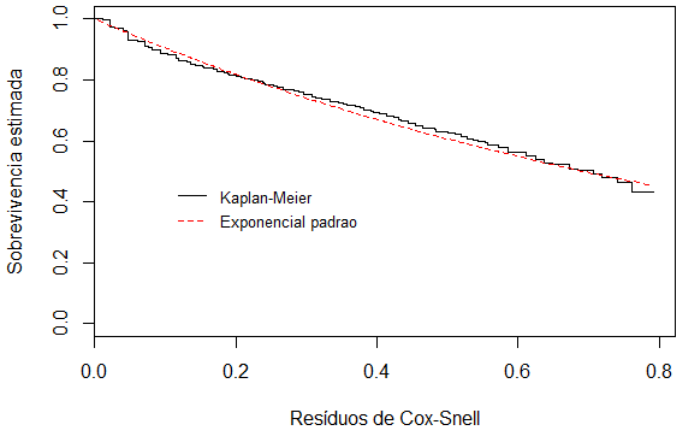
Intercepto	β_1	β_2	β_3	β_4	σ
4.5384	-0.4417	-0.5486	-0.1636	1.3842	0.7011

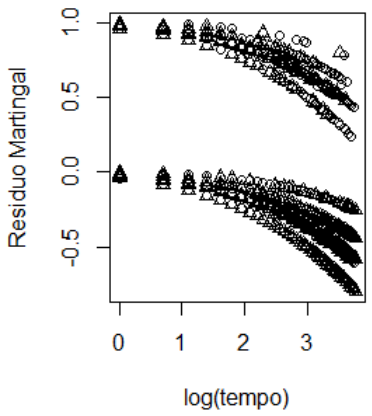
O modelo final, então, nos mostra que as diabetes e causas congênitas tem grande efeito sobre o tempo de sobrevivência de um paciente. O número de salas de diálise na unidade de tratamento também se mostrou significativa pro modelo. Acreditamos que tenha a ver com a estrutura do local, em que unidades com grande número de salas possa ter uma qualidade de atendimento pior. Por fim, a interação entre *grande* e *congênita*, apesar de ser significativa, não temos uma interpretação real. Todavia, essa interação possivelmente seria interpretada por algum profissional da saúde.

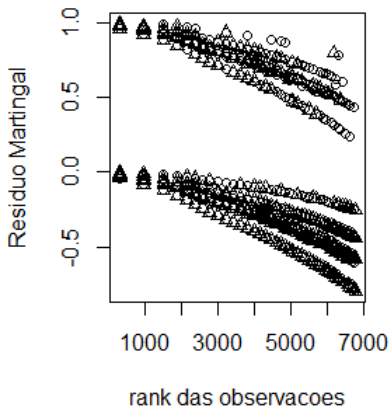
Para validação dos resultados obtidos na realização de inferências para o modelo é necessário verificar se ele é adequado. Para isso as técnicas de análise de resíduos e sensibilidade são algumas das formas para avaliar a adequabilidade dos modelos considerados.

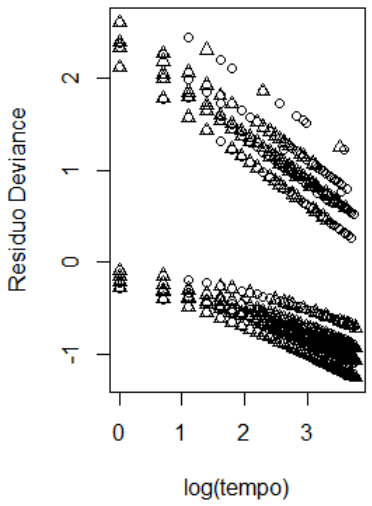
Nos resíduos de Cox-Snell vemos que a função de sobrevivência estimada se ajusta bem para uma distribuição exponencial padrão, o que nos indica uma boa validação para o modelo obtido.

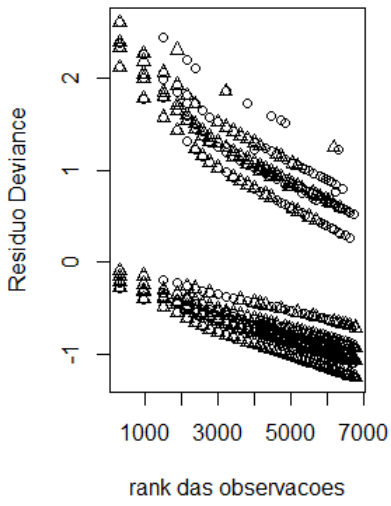
Os resíduos Martingal e Deviance, apresentaram um comportamento um pouco irregular, todavia isso se deve ao fato das variáveis do modelo serem todas categóricas.


Sobrevivencia estimada
0.0
0.2
0.4
0.6
0.8
1.0
Kaplan-Meier
Exponencial padrao
0.0
0.2
0.4
0.6
0.8
Resíduos de Cox-Snell


Residuo Martingal
-0.5
0.0
0.5
1.0
0
1
2
3
log(tempo)


Residuo Martingal
-0.5
0.0
0.5
1.0
1000
3000
5000
7000
rank das observacoes


Residuo Deviance
-1
0
1
2
0
1
2
3
log(tempo)


Residuo Deviance
-1
0
1
2
1000
3000
5000
7000
rank das observacoes

Anexo

```
dados <- read.csv('C:/Users/tonny/Downloads/dialise.csv')
head(dados)
require(survival)
require(AdequacyModel)

attach(dados)

dados$hip=as.numeric(dados$causa=="hip")
dados$out=as.numeric(dados$causa=="out")

table(idade,tempo)

summary(dados$idade)
dados$faixas<-cut(dados$idade,breaks = c(0,2,30,60,98),include.lowest = T,labels
= c("0 - 1","1 - 30","30 - 60","60 - 98"))

attach(dados)
head(dados)
hist(tempo,xlab="Tempo de sobrevivencia em dias", ylab="Frequencia de pacientes",
     main="Distribuição do tempo de sobrevivência
dos pacientes",ylim=c(0,2500))

KM<-survfit(Surv(tempo,status)~1,conf.int=T)
summary(KM)
plot(KM,conf.int=F, xlab="Tempo", ylab="S(t)", mark.time = T,col=c(1,2,2)
     ,main="Função de Sobrevivência")

plot(KM,conf.int=F, fun="cumhaz", mark.time = T,xlab="Tempo", ylab="H(t)"
     ,main="Função de risco acumulada",col=c(1,2,2))

summary(KM)

###S(t) por grupo###

KMfaixa<-survfit(Surv(tempo,status)~faixas, conf.int=F)
plot(KMfaixa,conf.int=F, xlab="Tempo", ylab="S(t)", col=c(1,2,3,4),mark.time = F
     ,main="Idade")
legend(5,0.5,lty=c(1,1,1,1),col=c(1,2,3,4),c("0 a 1 ano","2 a 30 anos",
```

```
"30 a 60 anos","60 anos ou mais"), bty="n",cex=1)
text(5,0,"p-valor = 2e-16",cex=0.6)

survdiff(Surv(tempo,status)~faixas, rho=0)

KMg<-survfit(Surv(tempo,status)~grande, conf.int=F)
plot(KMg,conf.int=F, xlab="Tempo", ylab="S(t)", col=c(2:3),mark.time = F
     ,main="Número de salas de diálise na unidade de tratamento")
legend(5,0.5,lty=c(1,1,1),col=c(2,3),c("uma ou duas salas","três salas ou mais")
, bty="n",cex=1)
text(5,0,"p-valor = 0.0002",cex=0.6)

survdiff(Surv(tempo,status)~grande, rho=0)

KMcausa<-survfit(Surv(tempo,status)~causa, conf.int=F)
plot(KMcausa,conf.int=F, xlab="Tempo", ylab="S(t)", col=c(2:6),mark.time = F
     ,main="Causa da insuficiência renal")
legend(5,0.5,lty=c(1,1,1,1,1),col=c(2:6),c("hipertensão","diabetes","renal",
                                           "congênita","outras"), bty="n",cex=0.8)
text(5,0,"p-valor = 2e-13",cex=0.6)

survdiff(Surv(tempo,status)~causa, rho=1)

KMcdiab<-survfit(Surv(tempo,status)~cdiab, conf.int=F)
plot(KMcdiab,conf.int=F, xlab="Tempo", ylab="S(t)", col=c(2:3),mark.time = F
     ,main="Diabetes")
legend(5,0.5,lty=c(1,1,1),col=c(2,3),c("Não","Sim"), bty="n",cex=1)
text(5,0,"p-valor = 2e-15",cex=0.6)

survdiff(Surv(tempo,status)~cdiab, rho=0)

KMcrim<-survfit(Surv(tempo,status)~crim, conf.int=F)
plot(KMcrim,conf.int=F, xlab="Tempo", ylab="S(t)", col=c(2:3),mark.time = F
     ,main="Causas renais")
legend(5,0.5,lty=c(1,1,1),col=c(2,3),c("Não","Sim"), bty="n",cex=1)
text(5,0,"p-valor = 0.02",cex=0.6)

survdiff(Surv(tempo,status)~crim, rho=0)
```

```
KMcongenita<-survfit(Surv(tempo,status)~congenita, conf.int=F)
plot(KMcongenita,conf.int=F, xlab="Tempo", ylab="S(t)", col=c(2:3),mark.time = F
     ,main="Causas congênitas")
legend(5,0.5,lty=c(1,1,1),col=c(2,3),c("Não","Sim"), bty="n",cex=1)
text(5,0,"p-valor = 0.002",cex=0.6)

survdiff(Surv(tempo,status)~congenita, rho=0)

KMhip<-survfit(Surv(tempo,status)~hip, conf.int=F)
plot(KMhip,conf.int=F, xlab="Tempo", ylab="S(t)", col=c(2:3),mark.time = F
     ,main="Hipertensão")
legend(5,0.5,lty=c(1,1,1),col=c(2,3),c("Não","Sim"), bty="n",cex=1)
text(5,0,"p-valor = 0.03",cex=0.6)

survdiff(Surv(tempo,status)~hip, rho=0)

KMout<-survfit(Surv(tempo,status)~out, conf.int=F)
plot(KMout,conf.int=F, xlab="Tempo", ylab="S(t)", col=c(2:3),mark.time = F
     ,main="Outras causas")
legend(5,0.5,lty=c(1,1,1),col=c(2,3),c("Não","Sim"), bty="n",cex=1)
text(5,0,"p-valor = 0.3",cex=0.6)

survdiff(Surv(tempo,status)~out, rho=1)

TTT(tempo, col="red", lwd=2, grid=TRUE, lty=2)
title("Tempo total em teste (TTT)")

dad2<-data.frame(tempo,cdiab)
tempo1<-dad2[dad2$cdiab == "0",]
TTT(tempo1$tempo, col="red", lwd=2.5, grid=TRUE, lty=2)
tempo2<-dad2[dad2$cdiab == "1",]
TTT(tempo2$tempo, col="red", lwd=2.5, grid=TRUE, lty=2)

### Definição do modelo ###
### Dist. Exponencial ###
n=length(tempo)
mexp<-survreg(Surv(tempo,status)~1, dist="exponential")
mexp
summary(mexp)
```

```
alpha<-exp(mexp$coefficients[1])
alpha

mwe<-survreg(Surv(tempo,status)~1, dist="weibull")
mwe
summary(mwe)

alphaw<-exp(mwe$coefficients[1])
alphaw

gamaw<-1/mwe$scale
gamaw

pws<-2
AICws<-(-2*mwe$loglik[1])+(2*pws)

AICcws<-AICws + ((2*pws*(pws+1))/(n-pws-1))

BICws<-(-2*mwe$loglik[1]) + pws*log(n)

medidasw<-cbind(AICws,AICcws,BICws)
medidasw

#### Dist. Log-normal ##

mlognorm<-survreg(Surv(tempo,status)~1, dist='lognorm')
mlognorm

mi<-mlognorm$coefficients[1]
sigma<-mlognorm$scale

pexp <- 1
AICexp <- (-2*mexp$loglik[1])+(2*pexp)
AICcexp<-AICexp + ((2*pexp*(pexp+1))/(n-pexp-1))

BIClns<-(-2*mexp$loglik[1]) + pexp*log(n)

medidasexp <- cbind(AICexp,AICcexp,BIClns)
```

```
medidasexp

plns<-2
AICln s<-(-2*mlognorm$loglik[1])+(2*plns)

AICclns<-AICln s + ((2*pws*(pws+1))/(n-plns-1))

BICln s<-(-2*mlognorm$loglik[1]) + plns*log(n)

medidasln<-cbind(AICln s,AICclns,BICln s)
medidasln

#### Dist. Log-logistica ##

mloglogi<-survreg(Surv(tempo,status)~1, dist='loglogistic')
summary(mloglogi)

alphall<-exp(mloglogi$coefficients[1])
alphall

gamall<- 1/mloglogi$scale
gamall

plls<-2
AICll s<-(-2*mloglogi$loglik[1])+(2*plls)

AICclls<-AICll s + ((2*pws*(pws+1))/(n-plls-1))

BICll s<-(-2*mloglogi$loglik[1]) + plls*log(n)

medidasll<-cbind(AICll s,AICclls,BICll s)
medidasll

dframe <- data.frame(t(medidasexp),t(medidasw),t(medidasll),t(medidasln))
fix(dframe)
View(dframe)

km<-survfit(Surv(tempo,status)~1)
time<-km$time
```

```
skm<-km$surv

sexp <- exp(-time/alpha)

swe<-exp(-(time/alphaw)^gamaw)

slognorm<-pnorm((-log(time)+mi)/sigma)

sloglogi<-1/(1+(time/alphall)^gamall)

plot(km,conf.int=F, xlab="Tempos", ylab="S(t)",mark.time = F)
lines(c(0,time),c(1,sexp),lty=1,col=2)
legend(10,0.6,lty=c(1,1),col=c(1,2),c("Kaplan-Meier","Exponencial"),bty="n",cex=0.8)

lines(c(0,time),c(1,swe),lty=1,col=5)
legend(10,0.4,lty=c(1,1),col=c(5),c("Weibull"),bty="n",cex=0.8)
lines(c(0,time),c(1,slognorm),lty=1,col=4)
legend(10,0.3,lty=1,col=4,c("Log-Normal"),bty="n",cex=0.8)
lines(c(0,time),c(1,sloglogi),lty=1,col="green")
legend(10,0.2,lty=1,col="green",c("Log-Logisitca"),bty="n",cex=0.8)
title("Ajuste dos modelos")

#########################################################################

### Seleção de covariaveis ###
y<-log(tempo)
mod<-survreg(Surv(y,status)~1, dist='gaussian')
summary(mod)

mod1<-survreg(Surv(y,status)~hip, dist='gaussian')
summary(mod1)

mod2<-survreg(Surv(y,status)~grande, dist='gaussian')
summary(mod2)

mod3<-survreg(Surv(y,status)~cdiab, dist='gaussian')
summary(mod3)
```

```
mod4<-survreg(Surv(y,status)~crim, dist='gaussian')
summary(mod4)

mod5<-survreg(Surv(y,status)~congenita, dist='gaussian')
summary(mod5)

mod6<-survreg(Surv(y,status)~out, dist='gaussian')
summary(mod6)

mod12345<-survreg(Surv(y,status)~grande+cdiab+crim+congenita+hip, dist='gaussian')
summary(mod12345)

mod1234<-survreg(Surv(y,status)~grande+cdiab+congenita+crim, dist='gaussian')
summary(mod1234)

mod123<-survreg(Surv(y,status)~grande+cdiab+congenita, dist='gaussian')
summary(mod123)

mod12<-survreg(Surv(y,status)~grande+cdiab, dist='gaussian')
summary(mod12)

mod13<-survreg(Surv(y,status)~grande+congenita, dist='gaussian')
summary(mod13)


mod23<-survreg(Surv(y,status)~cdiab+congenita, dist='gaussian')
summary(mod23)

#Modelo Final #

mod123_1<-survreg(Surv(y,status)~grande+cdiab+congenita+grande*congenita,
 dist='gaussian')
summary(mod123_1)

##############################################3

model.matrix(mod123_1)
x2 <- model.matrix(mod123_1)[,2]
x3 <- model.matrix(mod123_1)[,3]
```

```
x4 <- model.matrix(mod123_1)[,4]
x5 <- model.matrix(mod123_1)[,5]

mod123_1$coeff[1]

mi <- mod123_1$coeff[1] + mod123_1$coeff[2]*x2 + mod123_1$coeff[3]*x3 +
  mod123_1$coeff[4]*x4 + mod123_1$coeff[5]*x5

mip <- mod123_1$linear.predictors

### Resíduos ###
Smod <- 1-pnorm((y-mip)/mod123_1$scale)
ei <- (-log(Smod))

KMew <- survfit(Surv(ei,status)~1,conf.int=F)
te <- KMew$time
ste <- KMew$surv
sexp <- exp(-te)
summary(KMew)

plot(ste,sexp, xlab="S(ei):Kaplan-Meier", ylab="S(ei):Exponencial padrao")
abline(0,1,col="red",lty=1,lwd=2)
title()
plot(KMew,conf.int=F, xlab="Resíduos de Cox-Snell", ylab="Sobrevivencia estimada")
lines(te,sexp,lty=2, col=2)
legend(0.1,0.5,lty=c(1,2), col=c(1,2),c("Kaplan-Meier", "Exponencial padrao")
, cex=0.8, bty="n")

#Martingal#
martingal <- status-ei

par(mfrow=c(1,2))
plot(y,martingal,xlab="log(tempo)", ylab="Residuo Martingal",pch=censura+1)
plot(rank(y),martingal,xlab="rank das observacoes", ylab="Residuo Martingal"
,pch=censura+1)
title("Martingal")
par(mfrow=c(2,2))
plot(cdiab,martingal,xlab="Diabetes", ylab="Residuo Martingal",pch=censura+1)
plot(congenita,martingal,xlab="Congênitas", ylab="Residuo Martingal",pch=censura+1)
```

```
plot(crim,martingal,xlab="Renais", ylab="Residuo Martingal",pch=censura+1)
plot(factor(hip),martingal,xlab="Hipertensão", ylab="Residuo Martingal"
,pch=censura+1)

#Deviance#
par(mfrow=c(1,2))
devw<-(martingal/abs(martingal))*(-2*(martingal+status*log(status-martingal)))
^(1/2)
plot(y,devw,xlab="log(tempo)", ylab="Residuo Deviance",pch=censura+1)
plot(rank(y),devw,xlab="rank das observacoes", ylab="Residuo Deviance"
,pch=censura+1)
par(mfrow=c(1,1))
```