Document title

Subtitle

Davi Wentrick Feijó -200016806, Micael Papa - 000000000

November 29, 2023

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean ut elit odio. Donec fermentum tellus neque, vitae fringilla orci pretium vitae. Fusce maximus finibus facilisis. Donec ut ullamcorper turpis. Donec ut porta ipsum. Nullam cursus mauris a sapien ornare pulvinar. Aenean malesuada molestie erat quis mattis. Praesent scelerisque posuere faucibus. Praesent nunc nulla, ullamcorper ut ullamcorper sed, molestie ut est. Donec consequat libero nisi, non semper velit vulputate et. Quisque eleifend tincidunt ligula, bibendum finibus massa cursus eget. Curabitur aliquet vehicula quam non pulvinar. Aliquam facilisis tortor nec purus finibus, sit amet elementum eros sodales. Ut porta porttitor vestibulum.

1 Introdução

A pesquisa e o estudo da insuficiência cardíaca representam um campo crucial na área da saúde, especialmente considerando o cenário contemporâneo. A insuficiência cardíaca é uma condição crônica debilitante que afeta milhões de pessoas globalmente, resultando em um ônus significativo para os sistemas de saúde e afetando diretamente a qualidade de vida dos pacientes. Neste contexto, a compreensão aprofundada dos mecanismos subjacentes, fatores de risco, estratégias de prevenção e tratamento é fundamental.

No mundo atual, onde as doenças cardiovasculares continuam a ser uma das principais causas de morbidade e mortalidade, a insuficiência cardíaca emerge como um desafio complexo e premente. A interseção entre fatores de risco modificáveis, como dieta, estilo de vida, poluição ambiental e condições socioeconômicas, tem um impacto direto na incidência e na progressão dessa condição cardíaca.

Além disso, a crescente longevidade da população e a prevalência de comorbidades relacionadas, como diabetes e hipertensão, têm contribuído para um aumento substancial na incidência de insuficiência cardíaca. Essa realidade destaca a importância crítica de investigar não apenas os aspectos biomédicos, mas também os contextos sociais, comportamentais e ambientais que desempenham um papel na manifestação e gestão dessa condição.

Os avanços na pesquisa, diagnóstico e terapia oferecem uma perspectiva promissora, mas ainda há lacunas significativas a serem preenchidas. A exploração contínua dos mecanismos moleculares, novas terapias farmacológicas, intervenções não farmacológicas e abordagens inovadoras de gerenciamento são áreas cruciais que exigem uma atenção contínua.

Portanto, compreender a insuficiência cardíaca não apenas como uma entidade clínica isolada, mas como um desafio multifacetado que requer abordagens interdisciplinares e holísticas, torna-se essencial. Esta compreensão abrangente é crucial para orientar políticas de saúde pública, estratégias de prevenção e intervenções clínicas mais eficazes, visando não apenas tratar, mas também mitigar os fatores de risco associados a essa condição.

Em resumo, o estudo da insuficiência cardíaca é um imperativo no panorama atual da saúde, exigindo uma abordagem abrangente e colaborativa para mitigar seu impacto, melhorar a qualidade de vida dos pacientes e aliviar a carga que essa condição exerce sobre os sistemas de saúde em todo o mundo. O presente trabalho busca estudar, dadas as devidas proporções, as causas do acréscimo e da recorrência de pacientes com insuficiência cardíaca por intermédio de uma modelagem estatística imbuída de metodologias de análises de sobrevivência em conjunto com modelos lineares generalizados.

2 Metodologia

2.1 Sobre o dataset

Doenças cardiovasculares (DCVs) são a principal causa de morte globalmente, tirando uma estimativa de 17,9 milhões de vidas a cada ano, o que representa 31% de todas as mortes no mundo.

A insuficiência cardíaca é um evento comum causado por DCVs, e este conjunto de dados contém 12 características que podem ser usadas para prever a mortalidade por insuficiência cardíaca. Explicitando as características:

Variáveis booleanas:

- Death event : Se o paciente faleceu durante o período de acompanhamento.
- Smoking : Se o paciente é fumante.
- Sexo
- High blood pressure : Se o paciente tem hipertensão.
- Diabetes : Se o paciente tem diabetes.
- Anemia : Se o paciente tem anemia.

Variáveis Numéricas:

- Idade
- Creatinine phosphokinase: Nível da enzima CPK no sangue (mcg/L)
- ejection fraction : Percentual de sangue deixando o coração a cada contração (porcentagem)
- platelets : Plaquetas no sangue (quiloplaquetas/mL).
- serum creatinine : Nível de creatinina sérica no sangue (mg/dL).
- serum sodium : Nível de sódio sérico no sangue (mEq/L)
- time

A maioria das doenças cardiovasculares pode ser prevenida ao abordar fatores de risco comportamentais, como o uso de tabaco, dieta não saudável e obesidade, inatividade física e uso prejudicial de álcool, por meio de estratégias abrangentes para toda a população.

Pessoas com doenças cardiovasculares ou que estão em alto risco cardiovascular (devido à presença de um ou mais fatores de risco, como hipertensão, diabetes, hiperlipidemia ou doença já estabelecida) precisam de detecção precoce e manejo, onde um modelo de aprendizado de máquina pode ser de grande ajuda.

Dentre suas variáveis temos: - Idade - Anemia : Diminuição de glóbulos vermelhos ou hemoglobina (variável booleana) - : Nível da enzima CPK no sangue (mcg/L)

2.2 Função de sobrevivência - Log normal

A função de sobrevivência de uma distribuição log-normal descreve a probabilidade de uma variável aleatória contínua exceder um determinado valor ao longo do tempo. Na distribuição log-normal, os valores são logaritmicamente distribuídos, o que significa que o logaritmo dos dados segue uma distribuição normal.

Essa função é usada para modelar dados onde os valores têm uma distribuição assimétrica positiva e é útil em muitos contextos, como na análise de tempo de vida de produtos, estudos epidemiológicos ou financeiros. A função de sobrevivência da distribuição log-normal permite calcular a

probabilidade de um evento ocorrer além de um determinado ponto no tempo, levando em consideração a natureza dos dados logarítmicos.

$$S(t) = 1 - \Phi\left(\frac{\ln(t) - \mu}{\sigma}\right)$$

Onde:

1. S(t) é a função de sobrevivência em um tempo t ,
2. Φ é a função de distribuição acumulada da distribuição normal padrão,
3. μ é a média da distribuição log-normal,
4. σ é o desvio padrão da distribuição log-normal,
5. $\ln(t)$ é o logaritmo natural de t, o tempo.

2.3 Kaplan-Meier

O método de Kaplan-Meier é uma técnica estatística usada para estimar a função de sobrevivência a partir de dados de tempo até um evento ocorrer. É frequentemente aplicado em estudos de sobrevivência ou análise de tempo até um evento (como tempo até a morte, falha de equipamentos, etc.). Funciona calculando as estimativas de probabilidade de sobrevivência em intervalos de tempo, ajustando os cálculos à medida que os eventos ocorrem ou os participantes são censurados. Essas estimativas são representadas graficamente na forma de uma curva de sobrevivência, que mostra a probabilidade de um indivíduo sobreviver além de um determinado ponto no tempo. Além disso, o método de Kaplan-Meier permite a comparação de diferentes grupos de indivíduos para avaliar se há diferenças significativas na função de sobrevivência entre eles. Isso pode ser feito usando testes estatísticos, como o teste log-rank, para determinar se as curvas de sobrevivência são estatisticamente diferentes entre os grupos.

A função Kaplan-Meier pode ser representada em LaTeX da seguinte maneira:

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

Nesta fórmula:

1. - $\hat{S}(t)$ é a estimativa da função de sobrevivência em um tempo t .
2. - t_i representa os tempos de eventos.
3. - d_i é o número de eventos no tempo t_i .
4. n_i é o número de indivíduos em risco no tempo t_i .

2.4 Função Hazard

A função hazard, na teoria da sobrevivência e análise de sobrevivência, descreve a taxa instantânea na qual um evento (como morte, falha de equipamento, etc.) ocorre em um determinado momento, dado que o indivíduo tenha sobrevivido até aquele ponto no tempo. É uma medida da probabilidade condicional de um evento ocorrer em um pequeno intervalo de tempo, dado que o indivíduo tenha sobrevivido até esse momento.

Matematicamente, a função hazard é definida como a razão entre a densidade de probabilidade de um evento ocorrer em um determinado ponto no tempo e a probabilidade de sobrevivência até esse ponto. Em um contexto contínuo, a função hazard é representada por $\lambda(t)$ tal que :

Claro, a função hazard em um contexto contínuo é frequentemente representada da seguinte maneira em LaTeX:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}$$

Nesta fórmula:

1. - $\lambda\,(t)$ é a função hazard no tempo t
2. - $P(t \leq T < t + \Delta t \mid T \geq t)$ é a probabilidade condicional de um evento ocorrer no intervalo $[t, t + \Delta t]$ dado que o evento não ocorreu até o tempo t
3. - Δt representa um intervalo de tempo infinitesimalmente pequeno.

Assim, conseguimos expressar a taxa de risco de um evento ocorrer em um tempo específico, dado que o indivíduo sobreviveu até aquele ponto no tempo.

A interpretação da função hazard é crucial. Se a função hazard é constante ao longo do tempo, isso indica que a taxa de risco do evento é constante, o que é característico de muitos processos naturais. Por outro lado, se a função hazard aumenta ou diminui ao longo do tempo, isso indica mudanças na taxa de risco ao longo do tempo.

Uma função hazard crescente sugere que o risco de um evento aumenta com o tempo, enquanto uma função hazard decrescente indica que o risco diminui ao longo do tempo. Por exemplo, em estudos médicos, a função hazard pode mostrar como o risco de certas condições de saúde, como doenças cardiovasculares ou câncer, pode variar ao longo da vida de um paciente.

2.5 Análise dos resíduos - Cox Snell

Esses resíduos são uma maneira de avaliar a adequação do modelo ajustado aos dados e examinar como bem o modelo de riscos proporcionais está descrevendo a relação entre as variáveis explicativas e a taxa de risco.

Quando ajustamos um modelo de riscos proporcionais de Cox para dados de sobrevivência, estamos modelando como as variáveis independentes influenciam a taxa de risco (hazard) de um evento ocorrer ao longo do tempo. Os resíduos de Cox-Snell são calculados a partir da distribuição acumulada dos tempos observados versus a distribuição acumulada dos tempos esperados, conforme previstos pelo modelo.

Para calcular os resíduos de Cox-Snell, os tempos de sobrevivência observados são transformados usando as estimativas de probabilidade de sobrevivência derivadas do modelo ajustado. Em seguida, esses tempos transformados são comparados com uma distribuição teórica (normalmente uma distribuição exponencial se o modelo está bem especificado) para verificar se o modelo ajustado está adequado aos dados.

Se os resíduos de Cox-Snell se comportarem de maneira semelhante à distribuição teórica esperada (por exemplo, se seguirem uma distribuição exponencial), isso sugere que o modelo de riscos proporcionais está ajustando bem os dados observados. Por outro lado, desvios significativos dessa distribuição teórica podem indicar problemas na especificação do modelo ou falta de ajuste aos dados.

São dados por :

$$RS_i = -\ln(1 - \hat{S}(t_i))$$

Onde:

1. RS_i é o resíduo de Cox-Snell para o evento i.
2. $\hat{S}t_i$ é a estimativa da função de sobrevivência no tempo t_i.

Explicitamos assim que eles são uma medida da discrepância entre a probabilidade prevista de sobrevivência e a não ocorrência do evento até o tempo t_i, transformada para facilitar a avaliação do ajuste do modelo de riscos proporcionais de Cox aos dados de sobrevivência.

2.6 Seleção das variáveis

2.7 Escolha do modelo

2.7.1 TRV

A estatística de razão de verossimilhança (likelihood ratio test) é uma ferramenta fundamental na comparação de modelos em análise de sobrevivência, especialmente quando se trabalha com modelos de riscos proporcionais de Cox.

Ela compara a adequação de dois modelos distintos, geralmente um modelo completo (mais complexo) e um modelo reduzido (menos complexo). A diferença na verossimilhança entre esses dois modelos é usada para avaliar se o modelo mais complexo oferece um ajuste significativamente melhor em comparação com o modelo mais simples.

A ideia central é comparar as verossimilhanças dos dois modelos (o modelo completo e o modelo reduzido) para determinar se a inclusão de variáveis adicionais ou complexidade no modelo completo melhora significativamente a capacidade do modelo de explicar os dados observados.

A estatística de razão de verossimilhança é calculada como o logaritmo natural da razão entre as verossimilhanças dos dois modelos. Em um contexto de riscos proporcionais de Cox, essa estatística segue aproximadamente uma distribuição qui-quadrado, assumindo que o modelo mais simples é verdadeiro (ou seja, não há diferenças reais entre os modelos).

Se a estatística de razão de verossimilhança for grande o suficiente, ou seja, se a diferença entre os modelos for significativa, isso indica que o modelo mais complexo se ajusta significativamente melhor aos dados do que o modelo mais simples. Portanto, pode-se rejeitar a hipótese nula de que o modelo mais simples é suficiente para descrever os dados.

Em resumo, a estatística de razão de verossimilhança é uma ferramenta estatística poderosa para comparar a adequação de modelos distintos na análise de sobrevivência, permitindo determinar se a inclusão de variáveis ou complexidade adicional resulta em uma melhoria significativa na capacidade do modelo de explicar os dados observados.

$$\text{LR} = -2 \times (\ln(\mathcal{L}_{\text{reduzido}}) - \ln(\mathcal{L}_{\text{completo}}))$$

Onde :

1. LR é a estatística de razão de verossimilhança.
2. $\ln(\mathcal{L}_{\text{reduzido}})$ é o logaritmo da verossimilhança do modelo reduzido.
3. $\ln(\mathcal{L}_{\text{completo}})$ é o logaritmo da verossimilhança do modelo completo.

Representando assim a diferença entre os logaritmos das verossimilhanças dos modelos completo e reduzido, multiplicada por -2 para ajustar a distribuição da estatística de razão de verossimilhança para uma distribuição qui-quadrado, que é usada para testar a significância estatística da diferença entre os modelos.

2.7.2 BIC

O BIC é derivado da teoria da informação e é utilizado para comparar diferentes modelos com base na verossimilhança dos dados e no número de parâmetros do modelo. A ideia central é penalizar modelos mais complexos, aqueles com mais parâmetros, com o intuito de evitar o overfitting, ou seja, evitar que o modelo se ajuste excessivamente aos dados de treinamento e perca capacidade de generalização para novos dados.

A fórmula do BIC é dada por:

$$BIC = -2 \times \ln(L) + k \times \ln(n)$$

Onde:

1. $\ln(L)$ é o logaritmo da verossimilhança do modelo, ou seja, o valor máximo da função de verossimilhança atingido pelo modelo.
2. k é o número de parâmetros no modelo.
3. n é o número de observações nos dados.

O BIC penaliza modelos mais complexos (com um número maior de parâmetros) adicionando um termo proporcional a $k \times \ln(n)$ ao valor $-2 \times \ln(L)$. Isso significa que, à medida que o número de parâmetros aumenta, o BIC aumenta, mas a penalização é maior para conjuntos de dados menores, refletida pelo termo $\ln(n)$.

Ao comparar modelos, o BIC indica que o modelo com o valor mais baixo é preferível, pois alcança um bom ajuste aos dados, mas também é mais parcimonioso, evitando o sobreajuste. Portanto, o BIC é útil para a seleção de modelos, ajudando a encontrar um equilíbrio entre a capacidade de ajuste e a complexidade do modelo.

2.7.3 AIC

O AIC é baseado na ideia de encontrar um equilíbrio entre a capacidade de ajuste do modelo aos dados e a complexidade do modelo, penalizando modelos mais complexos. Ele leva em consideração tanto a habilidade do modelo em ajustar os dados quanto o número de parâmetros utilizados, buscando encontrar o modelo que melhor se ajuste aos dados sem ser excessivamente complexo. Sendo dado por:

$$AIC = -2 \times \ln(L) + 2k$$

Onde:

1. $\ln(L)$ é o logaritmo da verossimilhança do modelo.
2. k é o número de parâmetros no modelo.

O AIC penaliza modelos mais complexos adicionando $2k$ ao valor $-2 \times \ln(L)$, onde k representa o número de parâmetros no modelo. Portanto, à medida que o número de parâmetros aumenta, o AIC aumenta, mas ele também recompensa modelos com uma verossimilhança maior, refletida pelo termo $-2 \times \ln(L)$.

Ao comparar modelos, o AIC indica que o modelo com o valor mais baixo é preferível, pois alcança um bom ajuste aos dados, mas também é mais parcimonioso, evitando o sobreajuste.

2.7.4 AICc

O AIC corrigido (ou AICc) é uma versão modificada do Critério de Akaike (AIC), especialmente útil em situações em que o tamanho da amostra é pequeno em relação ao número de parâmetros do modelo. Ele ajusta o AIC para levar em consideração a amostra limitada, oferecendo uma penalização mais forte para modelos mais complexos em comparação com o AIC padrão.

O AICc adiciona um fator de correção à penalidade do AIC, levando em conta o tamanho da amostra (n) e o número de parâmetros (k) no modelo. A fórmula do AICc é:

$$AICc = AIC + \frac{2k(k+1)}{n-k-1}$$

Aqui, além da penalidade padrão do AIC ($2k$), adiciona-se o termo $\frac{2k(k+1)}{n-k-1}$ como correção, onde n é o número de observações no conjunto de dados.

O AICc é particularmente valioso em conjuntos de dados pequenos, onde o AIC padrão pode superestimar a complexidade do modelo devido à amostra limitada. Ele oferece uma penalização adicional para modelos mais complexos, ajudando na seleção de modelos quando o tamanho da amostra é pequeno em relação ao número de parâmetros do modelo.

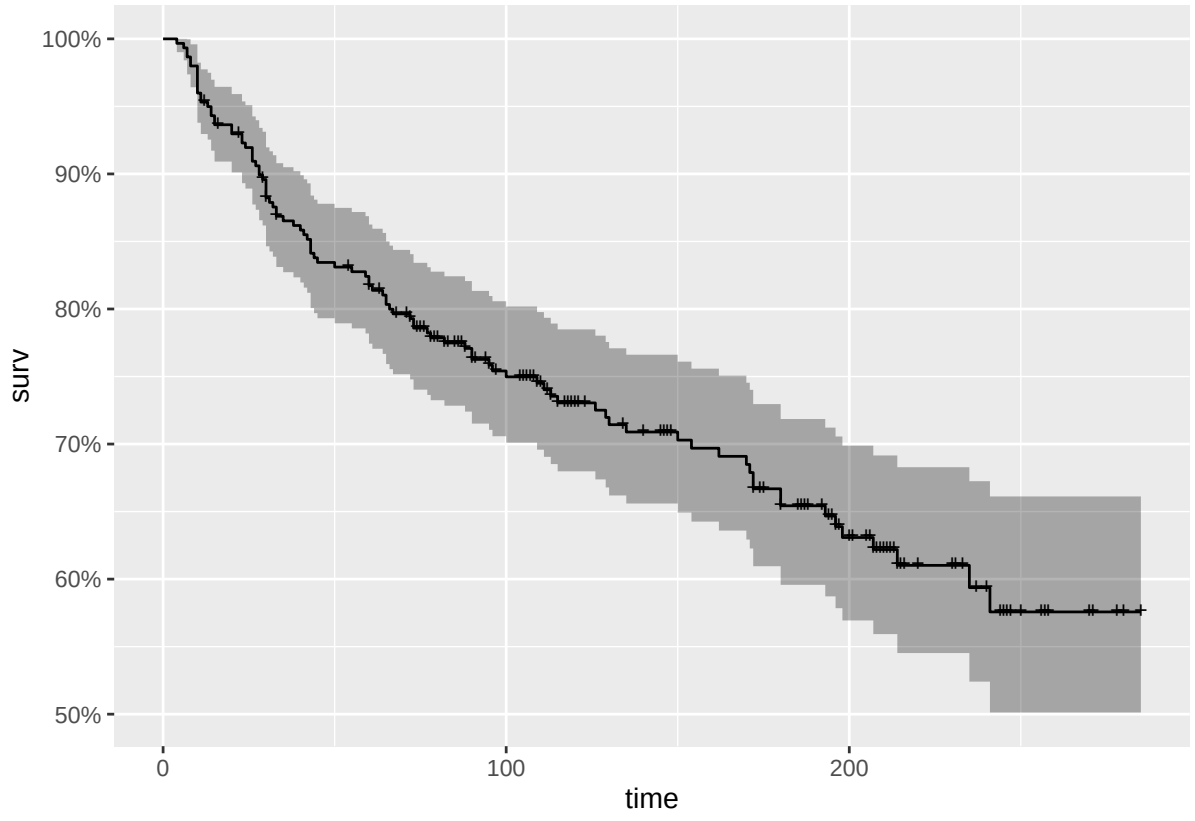
3 Resultados

3.1 Analise exploratoria

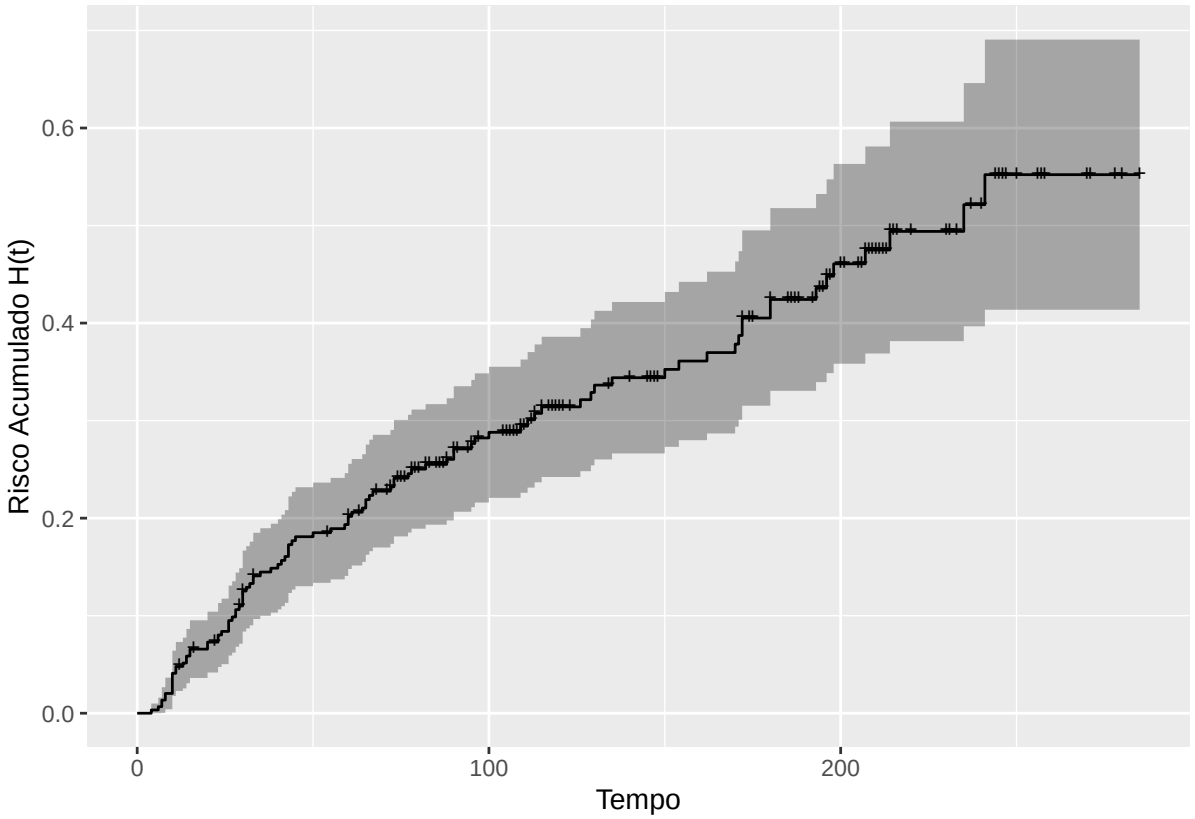
Antes de ajustar o modelo,é necessario estudarmos o comportamento dos dados antes para pode identificar qual distribuicao sera mais adequada e ja realizar uma seleção das variaveis categoricas que sao significativas para o modelo final.

```
## # A tibble: 6 x 13
##     age anaemia creatinine_phosphokinase diabetes ejection_fraction
##   <dbl>   <dbl>                    <dbl>    <dbl>              <dbl>
## 1    75       0                      582        0                 20
## 2    55       0                     7861        0                 38
## 3    65       0                      146        0                 20
## 4    50       1                      111        0                 20
## 5    65       1                      160        1                 20
## 6    90       1                       47        0                 40
## # i 8 more variables: high_blood_pressure <dbl>, platelets <dbl>,
## #   serum_creatinine <dbl>, serum_sodium <dbl>, sex <dbl>, smoking <dbl>,
## #   censura <dbl>, tempo <dbl>
```

3.1.1 Modelo de sobrevivencia nao parametrico de Kaplan-Meier

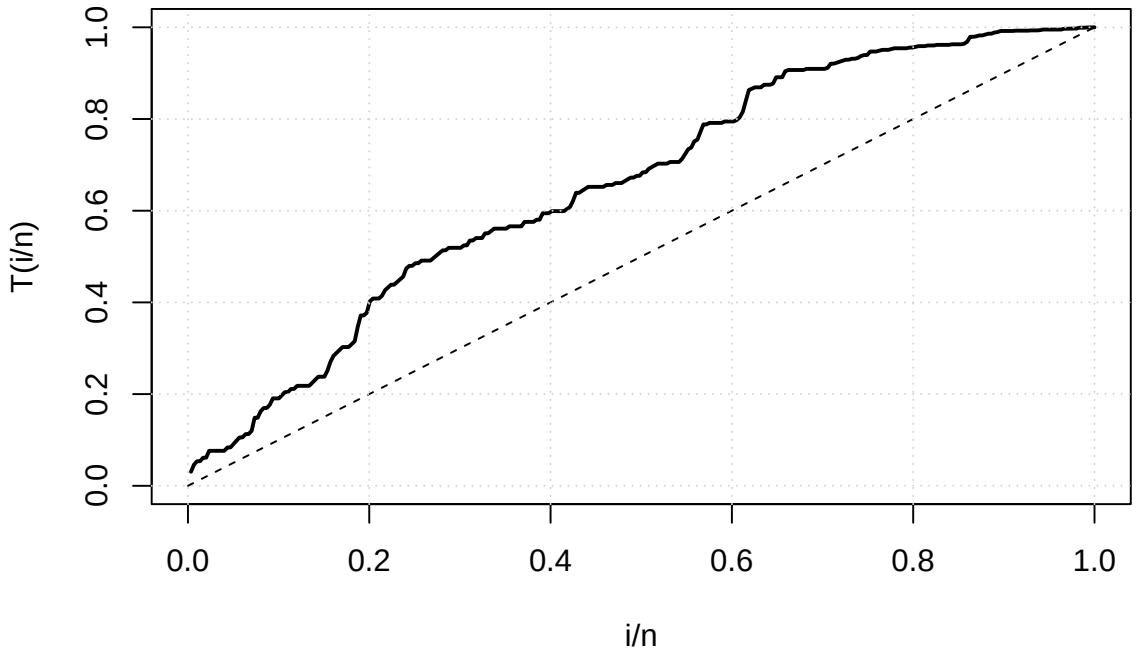


3.1.2 A funcao de risco acumulado



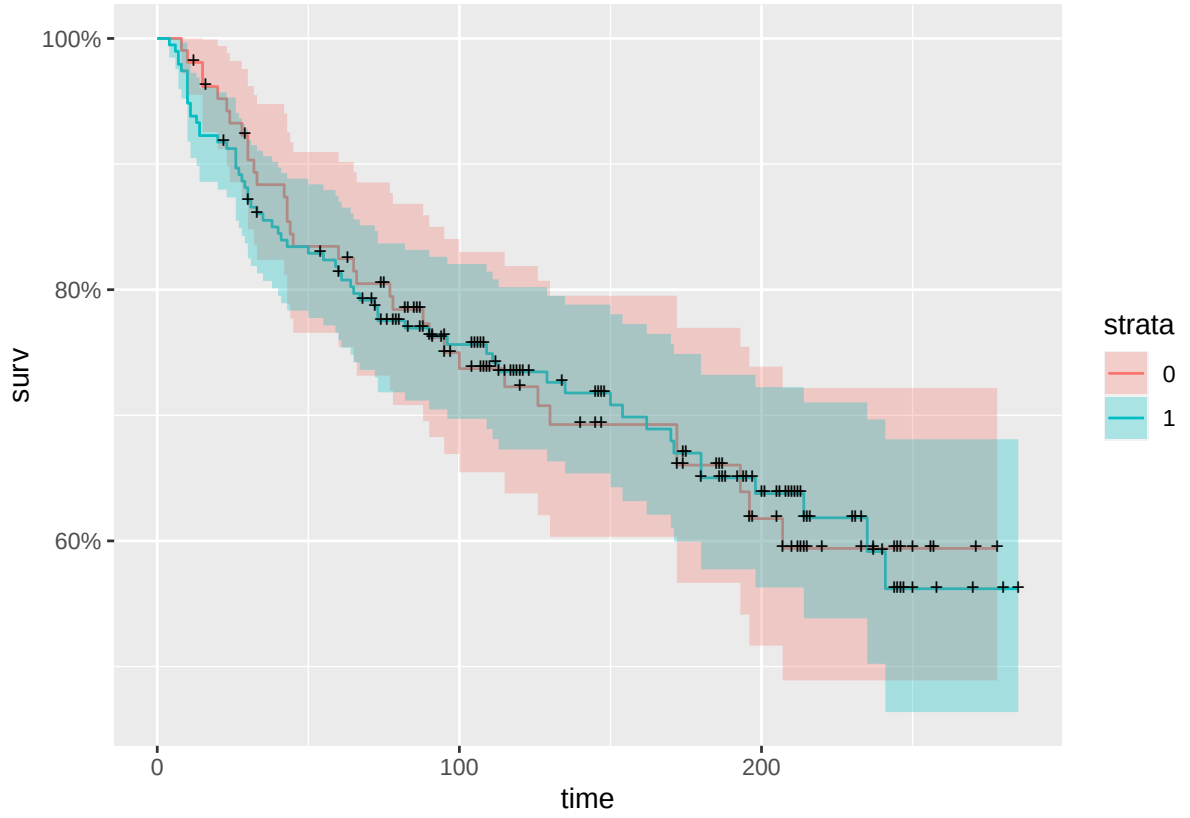
3.1.3 Curva TTT

Grafico do Tempo Total sobre Teste



3.1.4 Analise das Variaveis Categoricas

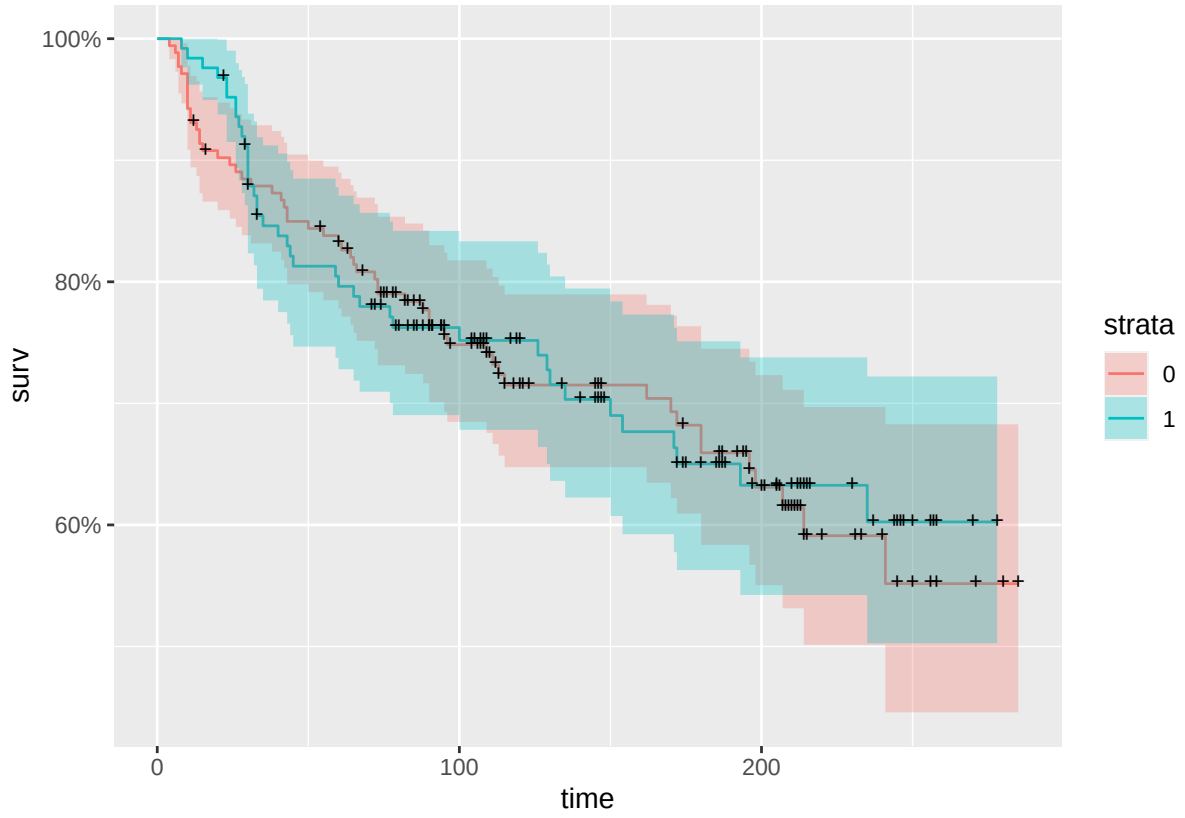
3.1.4.1 Variavel Sex Vamos comparar as curvas de sobrevivencias divididas por Sexo, com o objetivo de ver se essa variavel influencia na curva de sobrevivencia. Em seguida iremos fazer um teste para verificar a diferenca entre as curvas.




```
## Call:
## survdiff(formula = Surv(tempo, censura) ~ sex, data = dados,
##     rho = 1)
##
##         N Observed Expected (O-E)^2/E (O-E)^2/V
## sex=0 105     27.9     28.5   0.01467    0.0271
## sex=1 194     52.0     51.4   0.00814    0.0271
##
##  Chisq= 0  on 1 degrees of freedom, p= 0.9
```

Podemos notar que tanto pelo grafico quanto pelo teste, com p-valor = 0.9, que a variavel Sexo nao parece influenciar nas curvas de sobrevivencia

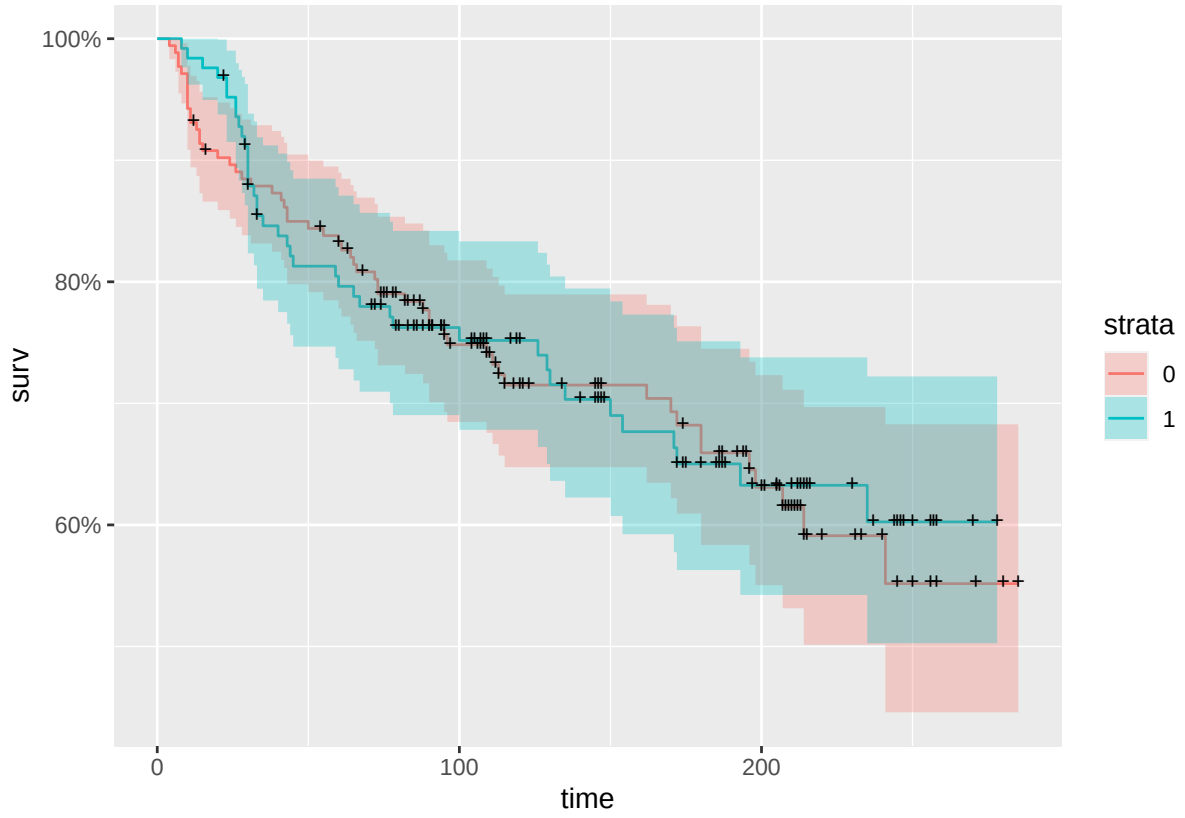
3.1.4.2 Variavel Diabetes




```
## Call:
## survdiff(formula = Surv(tempo, censura) ~ diabetes, data = dados,
##     rho = 1)
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## diabetes=0 174      46.7     45.9    0.0125    0.0349
## diabetes=1 125      33.2     34.0    0.0168    0.0349
##
##  Chisq= 0  on 1 degrees of freedom, p= 0.9
```

Pelo p-valor de 0.9 pode assumir que nao existe diferenca entre as curvas

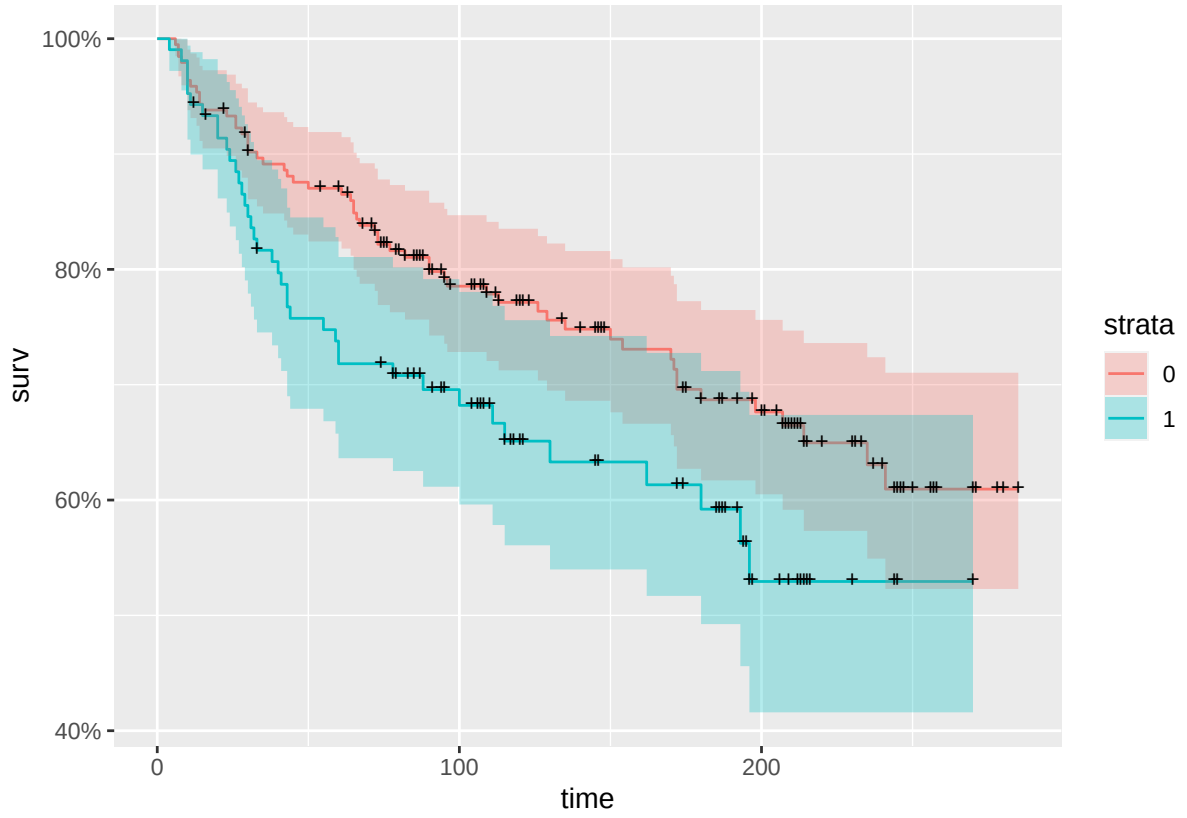
3.1.4.3 Variavel Anaemia




```
## Call:
## survdiff(formula = Surv(tempo, censura) ~ diabetes, data = dados,
##     rho = 1)
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## diabetes=0 174      46.7     45.9    0.0125    0.0349
## diabetes=1 125      33.2     34.0    0.0168    0.0349
##
##  Chisq= 0  on 1 degrees of freedom, p= 0.9
```

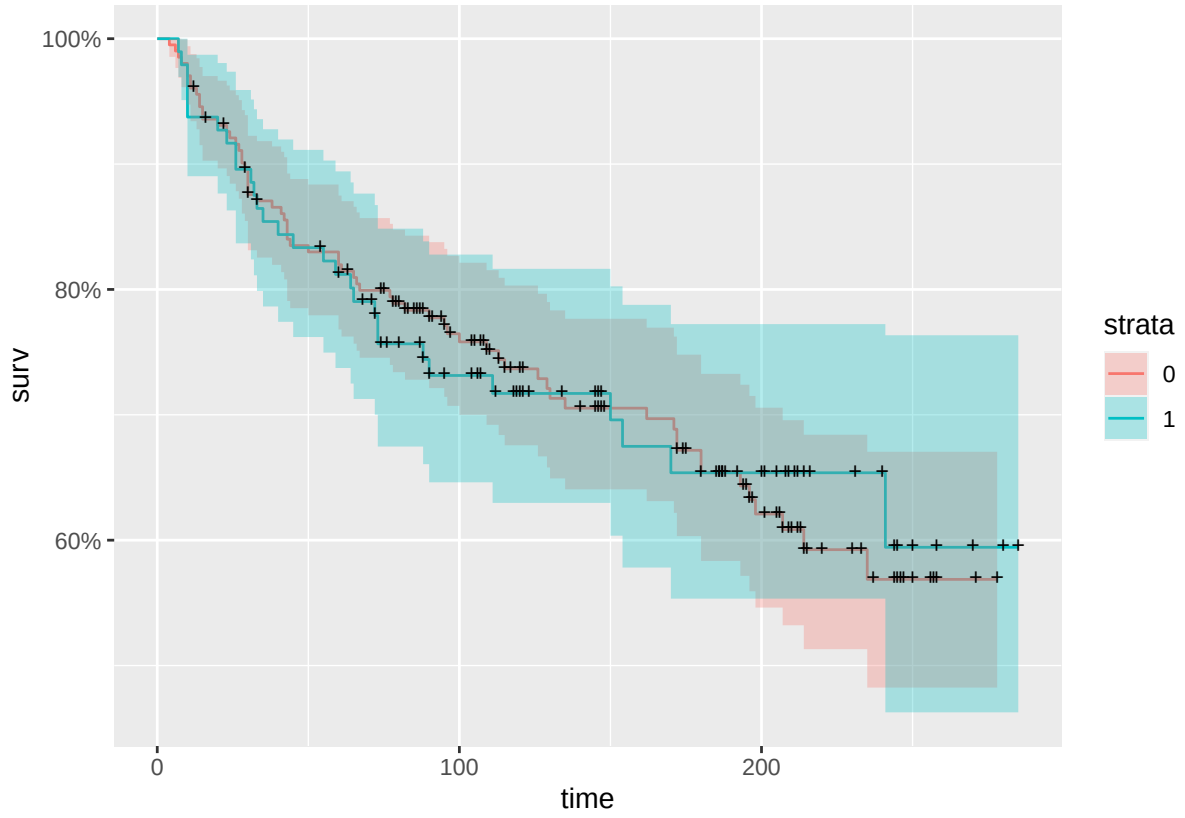
Pelo p-valor de 0.9 podemos assumir que nao ha diferenca entre as categorias

3.1.4.4 Variavel High Blood Pressure




```
## Call:
## survdiff(formula = Surv(tempo, censura) ~ high_blood_pressure,
##     data = dados, rho = 1)
##
##                           N Observed Expected (O-E)^2/E (O-E)^2/V
## high_blood_pressure=0 194     46.6     54.8      1.25      4.71
## high_blood_pressure=1 105     33.3     25.1      2.72      4.71
##
##  Chisq= 4.7  on 1 degrees of freedom, p= 0.03
```

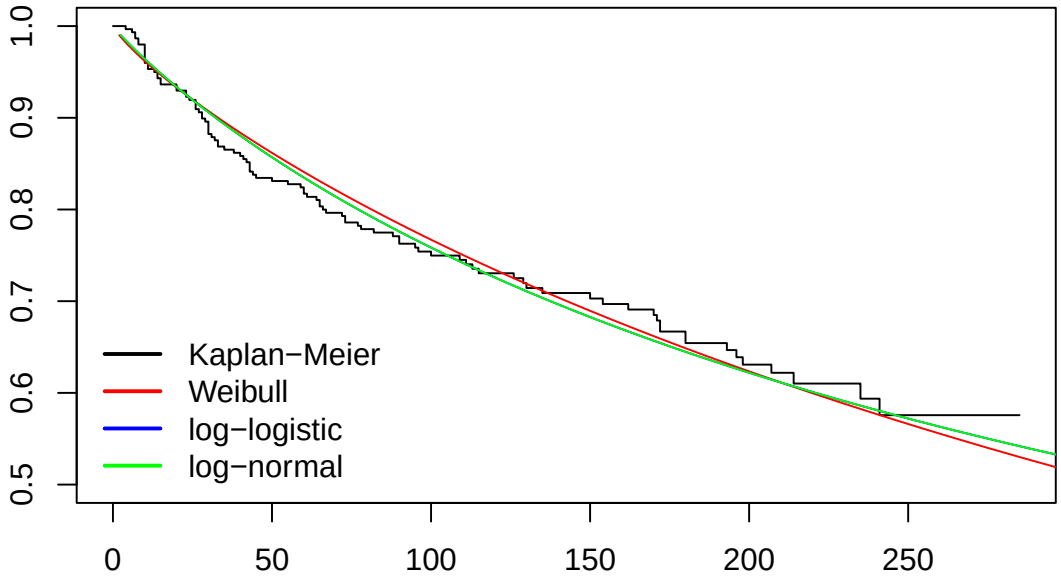
3.1.4.5 Variavel Smoking




```
## Call:
## survdiff(formula = Surv(tempo, censura) ~ smoking, data = dados,
##     rho = 1)
##
##              N Observed Expected (O-E)^2/E (O-E)^2/V
## smoking=0 203     54.5     54.7  0.000902   0.00339
## smoking=1  96     25.5     25.2  0.001955   0.00339
##
##  Chisq= 0  on 1 degrees of freedom, p= 1
```

3.2 Seleçao da distribuicao

Vamos ajustar algumas distribuicoes sobre o grafico de Kaplan-Meier para selecionar aquela que se adapta melhor a curva. Alem disso estaremos verificando os valores AIC, AIC corrigido e BIC para deicidir a distribuicao a ser utilizada.



3.2.0.1 Distribuicao Weibull

```
## Weibull ~( 0.8333038 , 491.7358 )

##          AICws    AICcws     BICws
## [1,] 1344.876 1344.916 1352.277

##
## Call:
## survreg(formula = s ~ 1, data = dados, dist = "weibull")
##                Value Std. Error     z      p
## (Intercept) 6.1979     0.1638 37.83 <2e-16
## Log(scale)  0.1824     0.0923  1.98  0.048
##
## Scale= 1.2
##
## Weibull distribution
## Loglik(model)= -670.4   Loglik(intercept only)= -670.4
## Number of Newton-Raphson Iterations: 5
## n= 299
```

3.2.0.2 Distribuicao Log-Normal

```
## Log-Normal ~( 5.914729 , 1.916652 )

##        AICln  AICcln   BICln
## [1,] 1336.546 1336.587 1343.947

##
## Call:
## survreg(formula = s ~ 1, data = dados, dist = "loglogistic")
##              Value Std. Error     z      p
## (Intercept) 5.8326     0.1613 36.17 <2e-16
## Log(scale)  0.0703     0.0897  0.78   0.43
##
## Scale= 1.07
##
## Log logistic distribution
## Loglik(model)= -669.2   Loglik(intercept only)= -669.2
## Number of Newton-Raphson Iterations: 4
## n= 299
```

3.2.0.3 Distribuicao Log-Logistica

```
## Log-Logistica ~( 0.9320725 , 341.255 )

##        AICll  AICcll   BICll
## [1,] 1342.334 1342.375 1349.735

##
## Call:
## survreg(formula = s ~ 1, data = dados, dist = "loglogistic")
##              Value Std. Error     z      p
## (Intercept) 5.8326     0.1613 36.17 <2e-16
## Log(scale)  0.0703     0.0897  0.78   0.43
##
## Scale= 1.07
##
## Log logistic distribution
## Loglik(model)= -669.2   Loglik(intercept only)= -669.2
## Number of Newton-Raphson Iterations: 4
## n= 299
```

3.2.0.4 Comparando as 3 distribuicoes

```
##         AICws   AICcws    BICws
## [1,] 1344.876 1344.916 1352.277

##        AICln  AICcln   BICln
## [1,] 1336.546 1336.587 1343.947
```

```
##         AICls  AICcls   BICls
## [1,] 1342.334 1342.375 1349.735
```

3.3 Selecao de variaveis

3.3.1 Manual

Nessa etapa vamos ajustar um modelo utilizando todas as covariaveis disponiveis e em seguida remover as que nao sao significativas para chegar num modelo reduzido e realizar o TRV para decidir qual ficamos, em seguida vamos realizar outras formas de selecao de variaveis e comparar para chegar numa final.

```
##
## Call:
## survreg(formula = s ~ age + anaemia + creatinine_phosphokinase +
##     diabetes + ejection_fraction + high_blood_pressure + platelets +
##     serum_creatinine + serum_sodium + sex + smoking, data = dados,
##     dist = "lognorm")
##                              Value Std. Error     z       p
## (Intercept)              -2.98e-01   3.68e+00 -0.08 0.93544
## age                      -4.82e-02   1.07e-02 -4.51 6.5e-06
## anaemia                  -5.24e-01   2.53e-01 -2.07 0.03852
## creatinine_phosphokinase -2.55e-04   1.17e-04 -2.18 0.02905
## diabetes                 -8.61e-02   2.55e-01 -0.34 0.73535
## ejection_fraction         4.43e-02   1.15e-02  3.84 0.00012
## high_blood_pressure      -5.04e-01   2.56e-01 -1.97 0.04888
## platelets                 7.17e-07   1.32e-06  0.54 0.58751
## serum_creatinine         -3.59e-01   1.04e-01 -3.44 0.00058
## serum_sodium              6.08e-02   2.69e-02  2.26 0.02386
## sex                       1.76e-01   2.96e-01  0.60 0.55158
## smoking                  -8.65e-02   2.92e-01 -0.30 0.76688
## Log(scale)                4.90e-01   8.01e-02  6.12 9.5e-10
##
## Scale= 1.63
##
## Log Normal distribution
## Loglik(model)= -630.7   Loglik(intercept only)= -666.3
##  Chisq= 71.17 on 11 degrees of freedom, p= 7.3e-11
## Number of Newton-Raphson Iterations: 4
## n= 299
```

```
##
## Call:
## survreg(formula = s ~ age + anaemia + creatinine_phosphokinase +
##     ejection_fraction + high_blood_pressure + serum_creatinine +
##     serum_sodium, data = dados, dist = "lognorm")
##                              Value Std. Error     z       p
## (Intercept)              -0.336261   3.634879 -0.09 0.92629
## age                      -0.046719   0.010410 -4.49 7.2e-06
## anaemia                  -0.526733   0.251203 -2.10 0.03601
## creatinine_phosphokinase -0.000244   0.000115 -2.11 0.03448
## ejection_fraction         0.043152   0.011233  3.84 0.00012
## high_blood_pressure      -0.510715   0.251881 -2.03 0.04260
```

```
## serum_creatinine         -0.357048   0.103693 -3.44 0.00057
## serum_sodium              0.062447   0.026754  2.33 0.01959
## Log(scale)                0.489603   0.080044  6.12 9.6e-10
##
## Scale= 1.63
##
## Log Normal distribution
## Loglik(model)= -631.1   Loglik(intercept only)= -666.3
##  Chisq= 70.43 on 7 degrees of freedom, p= 1.2e-12
## Number of Newton-Raphson Iterations: 4
## n= 299

## [1] -0.7344368

## [1] 4

## [1] 1
```

Nesse caso ficamos com o modelo reduzido

3.3.2 Random Forest

Nesse etapa vamos usar um algoritmo de random forest para selecionar as variaveis importantes do nosso modelo

```
## minimal depth variable selection ...
##
##
## ----------------------------------------------------------
## family             : surv
## var. selection     : Minimal Depth
## conservativeness   : medium
## x-weighting used?  : TRUE
## dimension          : 11
## sample size        : 299
## ntree              : 500
## nsplit             : 10
## mtry               : 4
## nodesize           : 20
## refitted forest    : FALSE
## model size         : 4
## depth threshold    : 4.4558
## PE (true OOB)      : 28.0334
##
##
## Top variables:
##                   depth  vimp
## ejection_fraction 1.820 0.212
## serum_creatinine  2.052 0.218
```

```
## age               2.420 0.200
## serum_sodium      4.084 0.062
## ---------------------------------------------------------
```

```
##
## Call:
## survreg(formula = s ~ ejection_fraction + serum_creatinine +
##     age + serum_sodium, data = dados, dist = "lognorm")
##                      Value Std. Error     z       p
## (Intercept)         0.4126     3.7574  0.11 0.91256
## ejection_fraction   0.0443     0.0117  3.79 0.00015
## serum_creatinine   -0.3811     0.1078 -3.54 0.00041
## age                -0.0480     0.0106 -4.50 6.7e-06
## serum_sodium        0.0536     0.0275  1.95 0.05112
## Log(scale)          0.5286     0.0803  6.58 4.6e-11
##
## Scale= 1.7
##
## Log Normal distribution
## Loglik(model)= -636.5   Loglik(intercept only)= -666.3
##  Chisq= 59.52 on 4 degrees of freedom, p= 3.7e-12
## Number of Newton-Raphson Iterations: 4
## n= 299
```

3.3.3 Stepwise

```
## [1] "high_blood_pressure" "age"
```

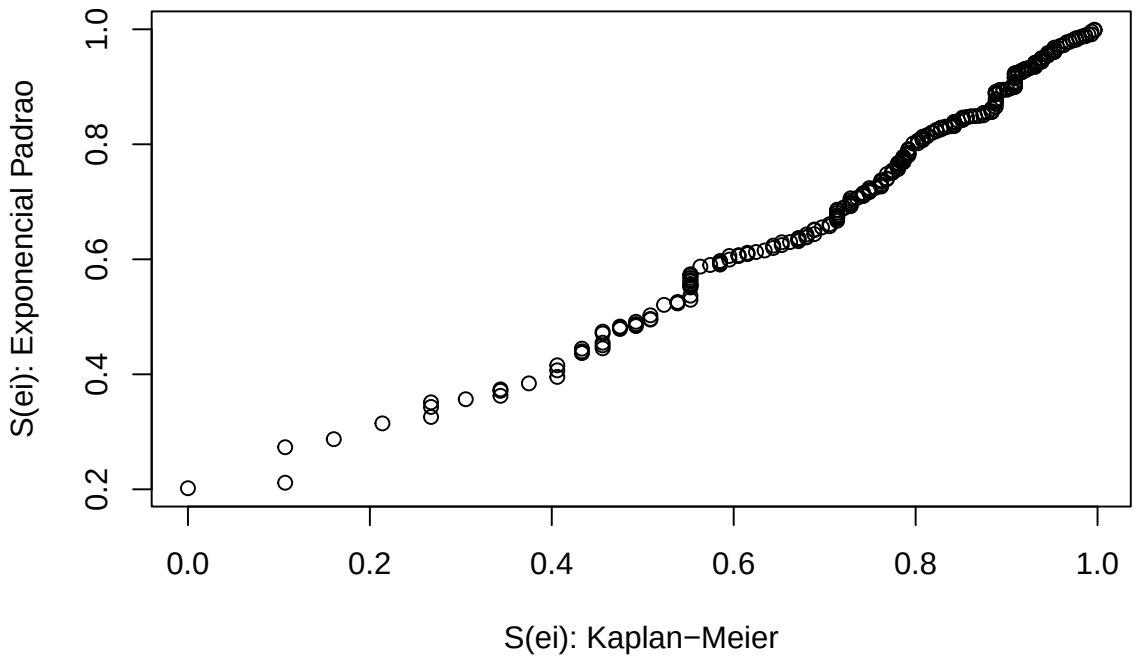
```
## [1] 6 1
```

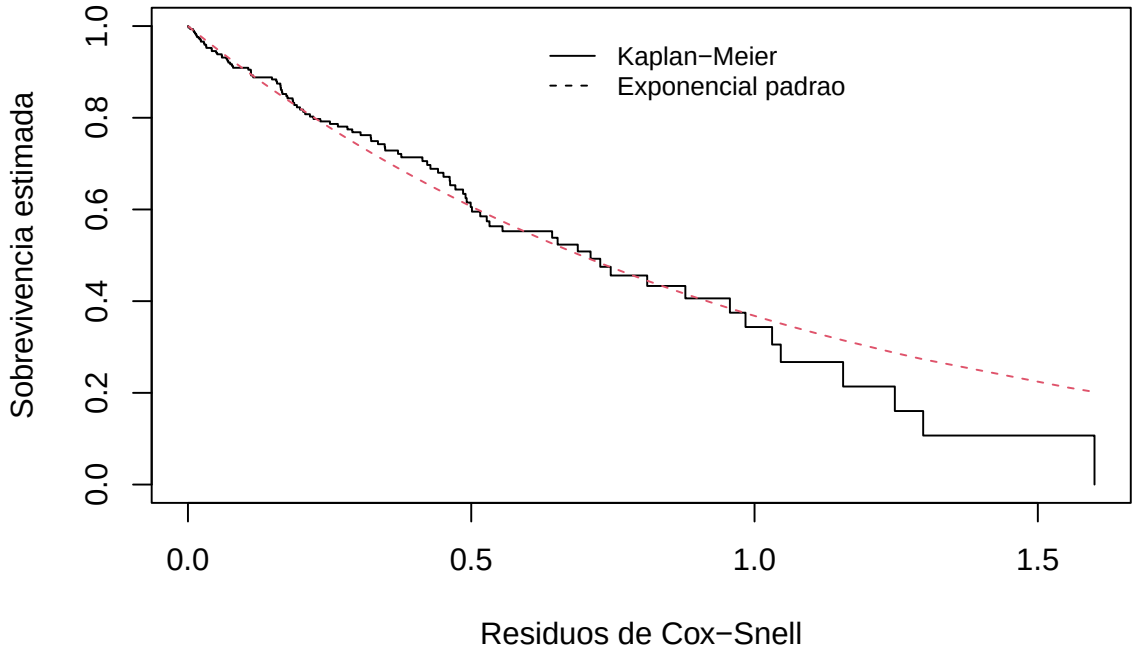
```
##
## Call:
## survreg(formula = s ~ age + high_blood_pressure, data = dados,
##     dist = "lognorm")
##                       Value Std. Error     z       p
## (Intercept)          8.9675     0.7277 12.32 < 2e-16
## age                 -0.0477     0.0108 -4.41 1.0e-05
## high_blood_pressure -0.4844     0.2656 -1.82   0.068
## Log(scale)           0.5814     0.0812  7.16 8.1e-13
##
## Scale= 1.79
##
## Log Normal distribution
## Loglik(model)= -653.8   Loglik(intercept only)= -666.3
##  Chisq= 24.87 on 2 degrees of freedom, p= 4e-06
## Number of Newton-Raphson Iterations: 4
## n= 299
```

3.4 Analise de reisudos

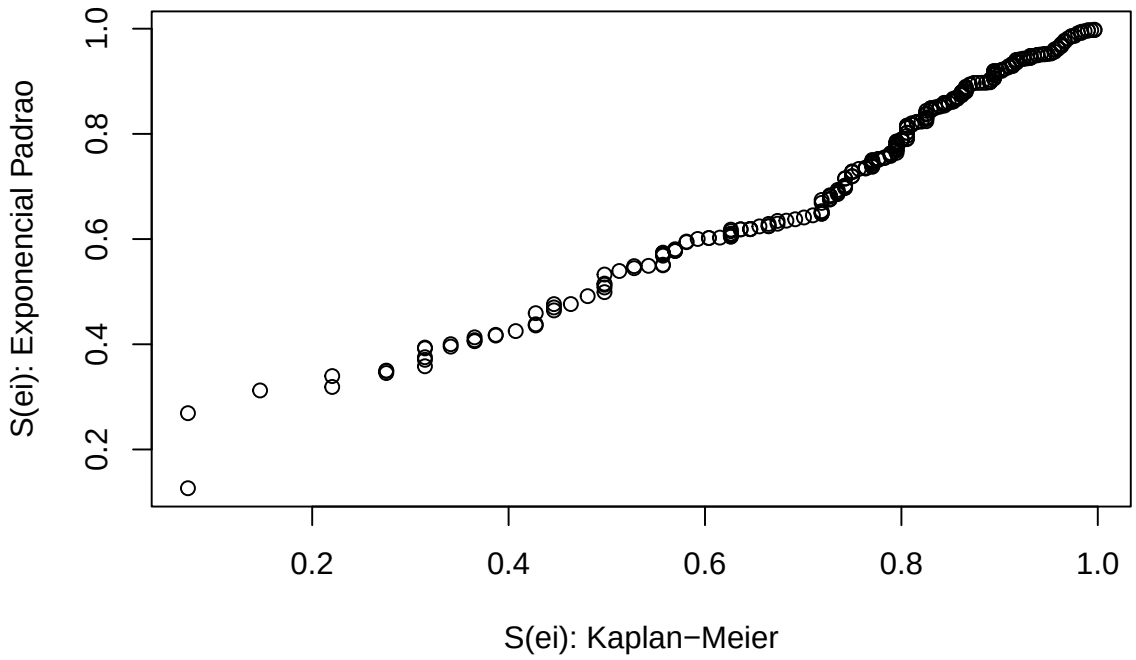
Temos 3 modelos finais e para isso vamos utilizar a analise de reisudos de cada um para ver qual se adequa melhor aos dados

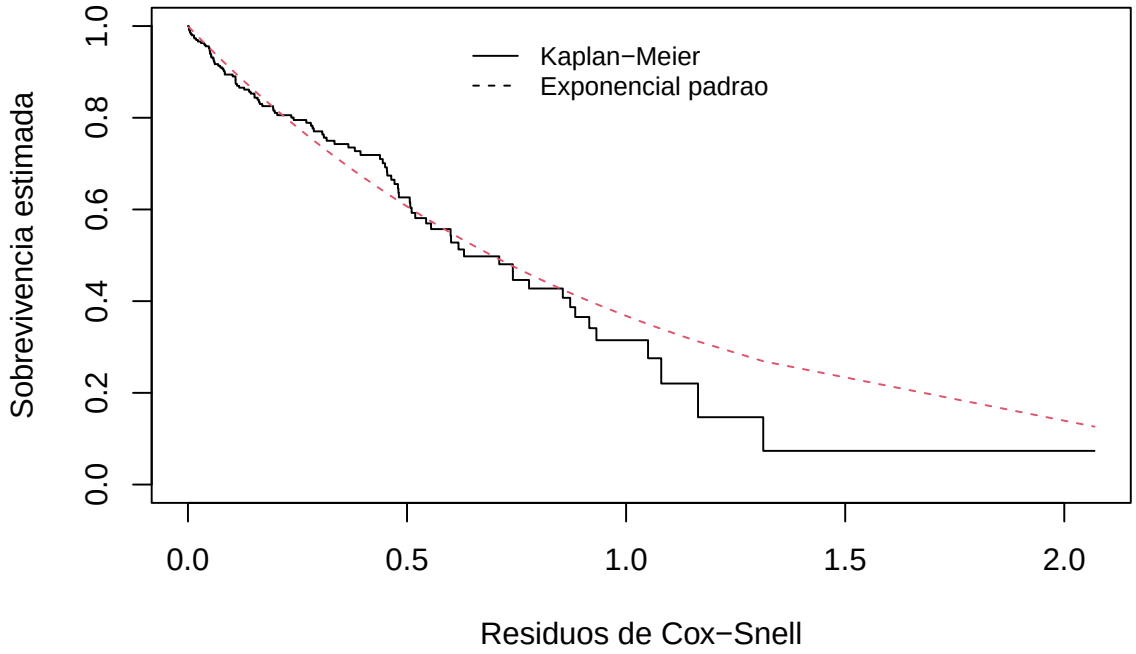
3.4.1 Modelo Manual



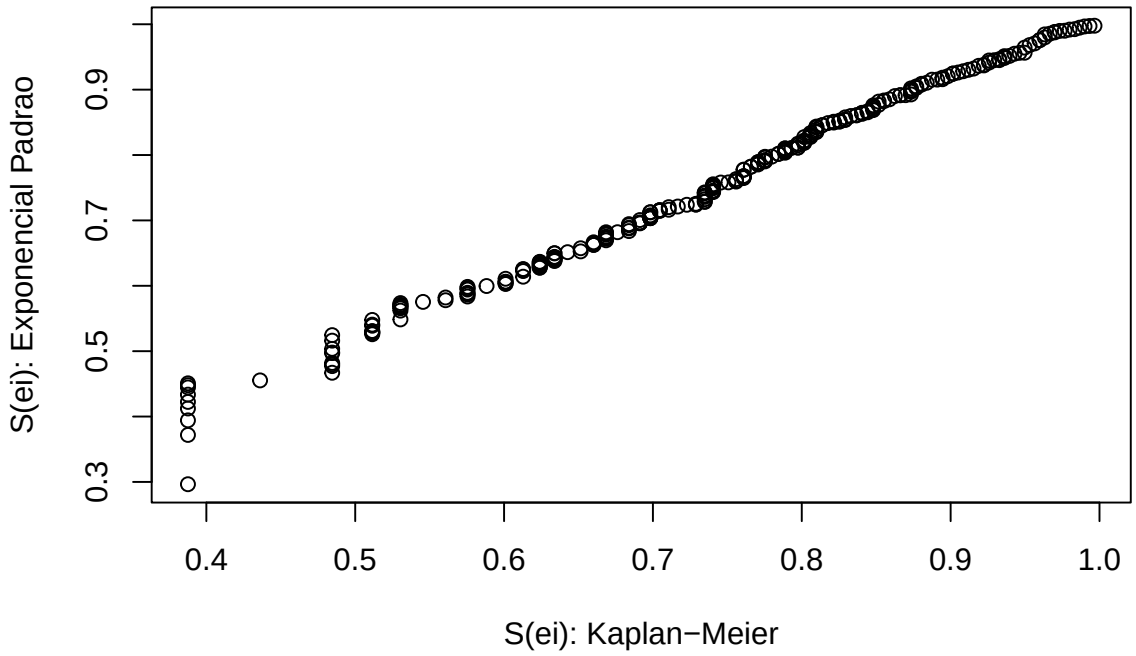


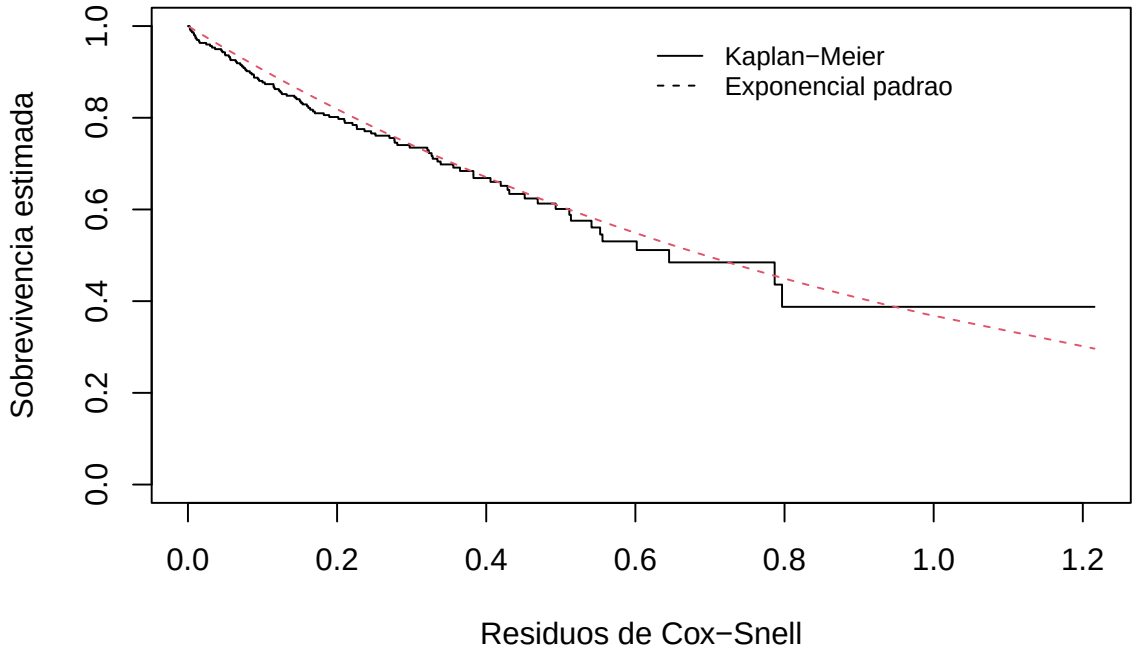
3.4.2 Modelo Random Forest





3.4.3 Modelo Stepwise





3.5 Comparando os modelos finais

```
## Valores modelo manual: 1278.113 1278.609 1307.716
```

```
## Valores modelo random forest: 1283.031 1283.236 1301.533
```

```
## Valores modelo stepwise: 1336.546 1336.587 1343.947
```

4 Conclusão