Aplicação de modelos sobrevivência mediante dados de insuficiência cardíaca

Davi Wentrick Feijó -200016806, Micael Egídio Papa da Silva - 211029236

December 18, 2023

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

2.6.1 Stepwise

O método stepwise é uma técnica usada na seleção de variáveis em modelos estatísticos, especialmente em modelos de regressão. Ele envolve a inclusão e exclusão iterativa de variáveis explicativas com base em critérios específicos, como a melhoria do ajuste do modelo ou a minimização do erro.

Existem duas formas principais de stepwise: forward stepwise e backward stepwise.

1. Forward Stepwise Selection: \
    - Começa sem variáveis no modelo. \
    - Iterativamente, adiciona variáveis uma por uma, escolhendo aquela que mais melhora o modelo com base em algum critério (como R^2, AIC, BIC, etc.). \
    - Continua adicionando variáveis até que a adição de outras não melhore significativamente o modelo. \
2. Backward Stepwise Selection:\
    - Começa com todas as variáveis no modelo. \
    - Iterativamente, remove variáveis uma por uma, escolhendo a que menos prejudica o modelo com base em algum critério. \
    - Continua removendo variáveis até que a exclusão de outras não melhore significativamente o modelo. \

Ambos os métodos podem ser combinados em um processo “stepwise” que realiza adições e remoções de variáveis no modelo até que não haja mais melhoria substancial nos critérios de avaliação.

Esse processo Stepwise é uma abordagem que combina os métodos forward e backward em um único procedimento. Ele realiza tanto adições quanto remoções de variáveis em cada iteração, considerando diferentes conjuntos de variáveis candidatas.

O processo do stepwise conjunto pode ser realizado da seguinte maneira: \

1. Inicialização: Começa com um conjunto vazio de variáveis no modelo (forward) e todas as variáveis no modelo (backward). \
2. Passos Iterativos:\
    - Forward Step: Adiciona uma variável ao modelo se melhorar o critério de seleção (como R^2, AIC, BIC, etc.). \
    - Backward Step: Remove uma variável do modelo se sua exclusão melhorar o critério de seleção.\
3. Critério de Parada: O algoritmo continua alternando entre os passos forward e backward até que não seja possível adicionar mais variáveis que melhorem o critério de seleção (forward) ou remover mais variáveis que melhorem o critério (backward). Pode ser determinado por

critérios como estabilidade do modelo, não melhoria substancial do critério de avaliação, ou outros critérios pré-definidos. \

Essa abordagem combina as vantagens do forward e backward stepwise, permitindo a exploração de diferentes conjuntos de variáveis e a busca por um modelo que se ajuste bem aos dados. No entanto, também enfrenta críticas semelhantes às dos métodos individuais, como o potencial de overfitting e a sensibilidade aos critérios de seleção. \

É importante ter em mente que, embora o método stepwise conjunto seja uma tentativa de abordar as limitações dos métodos forward e backward, a seleção de variáveis em modelos estatísticos geralmente requer uma compreensão cuidadosa do contexto do problema e a consideração de múltiplos métodos para validar os resultados. \

Onde, a abordagem stepwise não possui uma fórmula única, mas podemos exemplificar como a adição e remoção de variáveis podem ser representadas em um contexto de regressão linear múltipla.

Suponha um modelo de regressão linear múltipla onde estamos selecionando variáveis com base em algum critério de ajuste, como o AIC (Critério de Informação de Akaike):

A fórmula para o AIC é: \

$$AIC = 2k - 2\ln(\hat{L})$$

Onde: \ - k é o número de parâmetros no modelo. \ - $\hat{L}$ é a função de verossimilhança máxima do modelo. \

No processo stepwise, durante a adição de variáveis, você pode comparar dois modelos:

1. Modelo com uma variável a mais: \
   - $AIC_1 = 2k_1 - 2\ln(\hat{L}_1)$ \
2. Modelo atual: \
   - $AIC_0 = 2k_0 - 2\ln(\hat{L}_0)$ \

Se $AIC_1 < AIC_0$, a variável adicional é considerada benéfica e é adicionada ao modelo.

Durante a remoção de variáveis, você faz o processo inverso:

1. Modelo sem uma variável: \
   - $AIC_2 = 2k_2 - 2\ln(\hat{L}_2)$ \
2. Modelo atual: \
   - $AIC_0 = 2k_0 - 2\ln(\hat{L}_0)$ \

Se $AIC_2 < AIC_0$, a variável é removida do modelo.

Esses são os princípios básicos do processo stepwise em termos de comparação de critérios para adição e remoção de variáveis. No entanto, a implementação exata pode variar dependendo do critério de seleção, do tipo de modelo e de outros fatores específicos ao contexto do problema.

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.

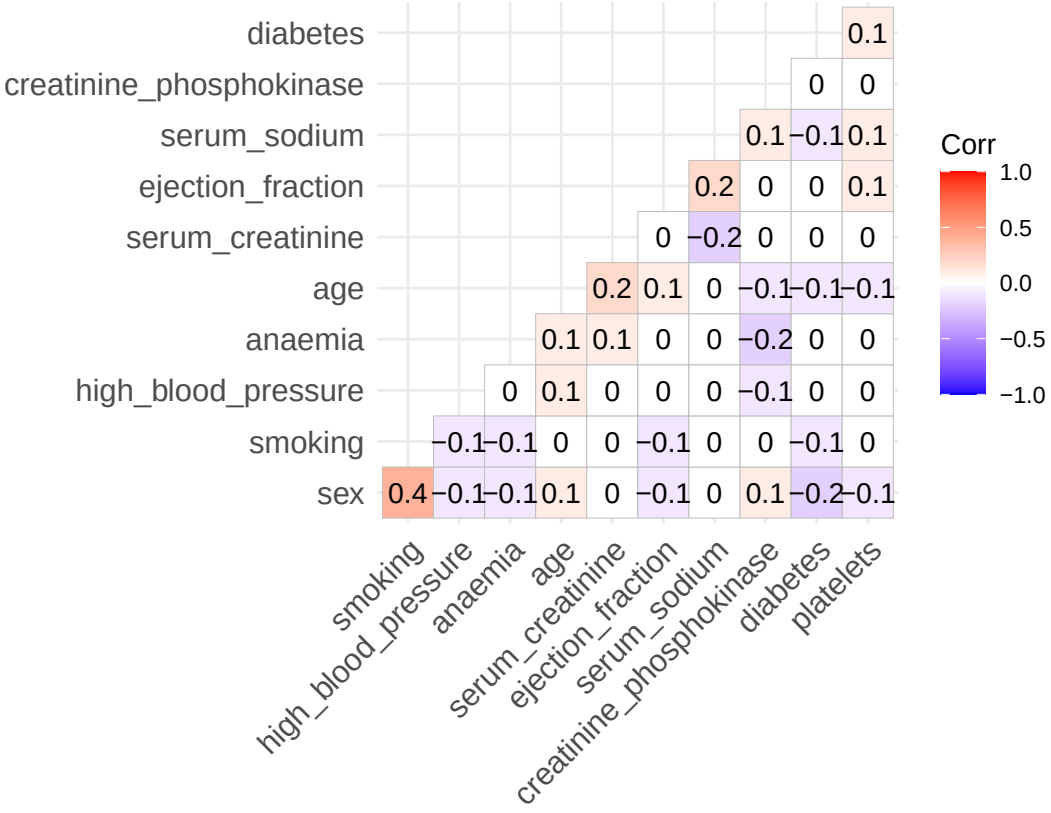
3 Resultados

3.1 Analise Exploratória

Nessa etapa, vamos estudar a distribuição das variáveis numéricas e categóricas do banco, com o objetivo de verificar sua distribuição e correlação.

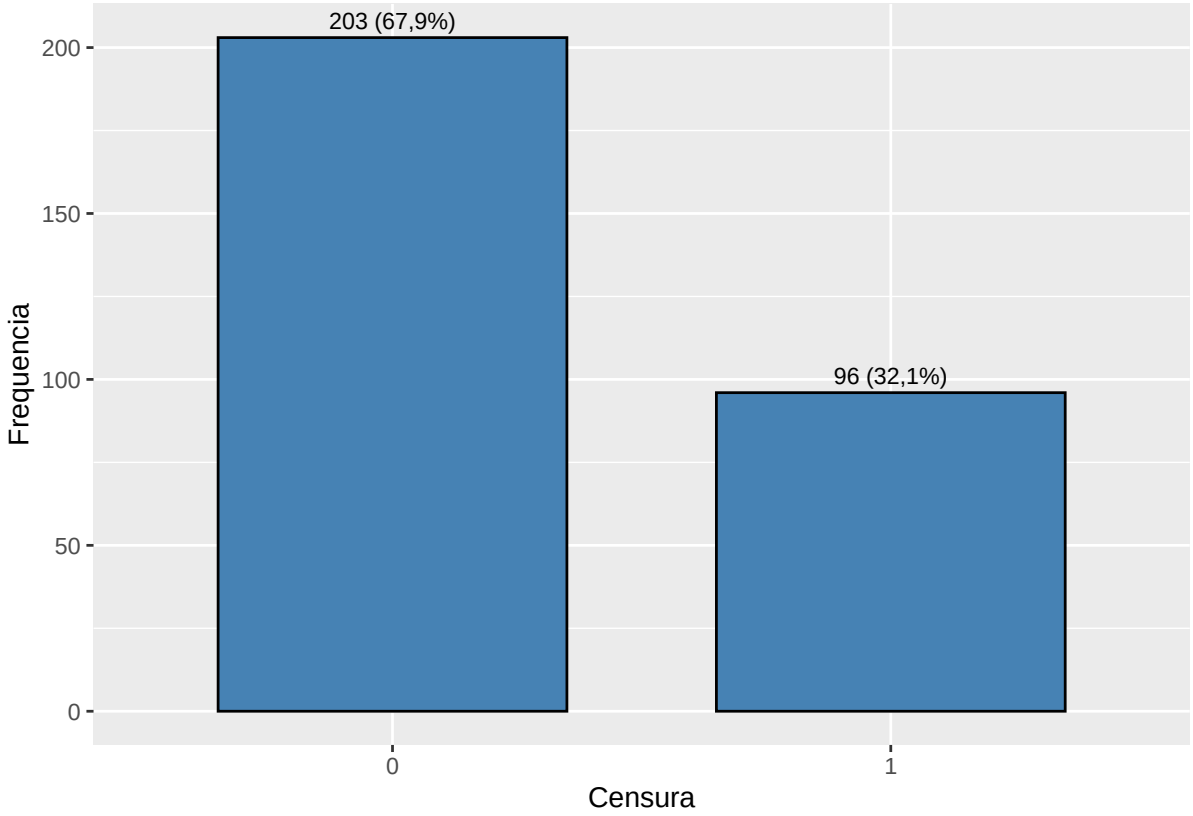
```
## # A tibble: 6 x 13
##     age anaemia creatinine_phosphokinase diabetes ejection_fraction
##   <dbl>   <dbl>                    <dbl>    <dbl>              <dbl>
## 1    75       0                      582        0                 20
## 2    55       0                     7861        0                 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 Correlação



Nenhuma das variáveis exibe uma correlação muito forte, sendo a correlação entre as variáveis `sex` e `smoking` a que mais se diferencia das demais.

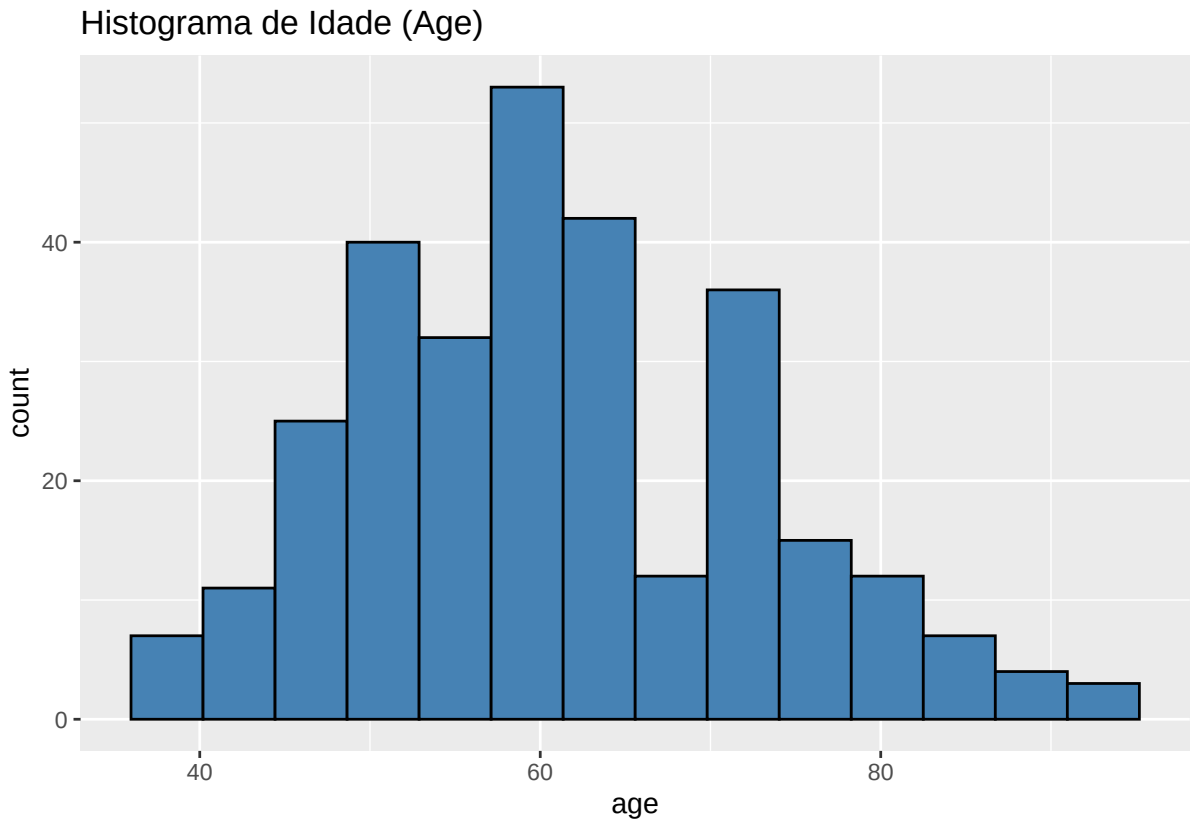
3.1.2 Analise de Censura

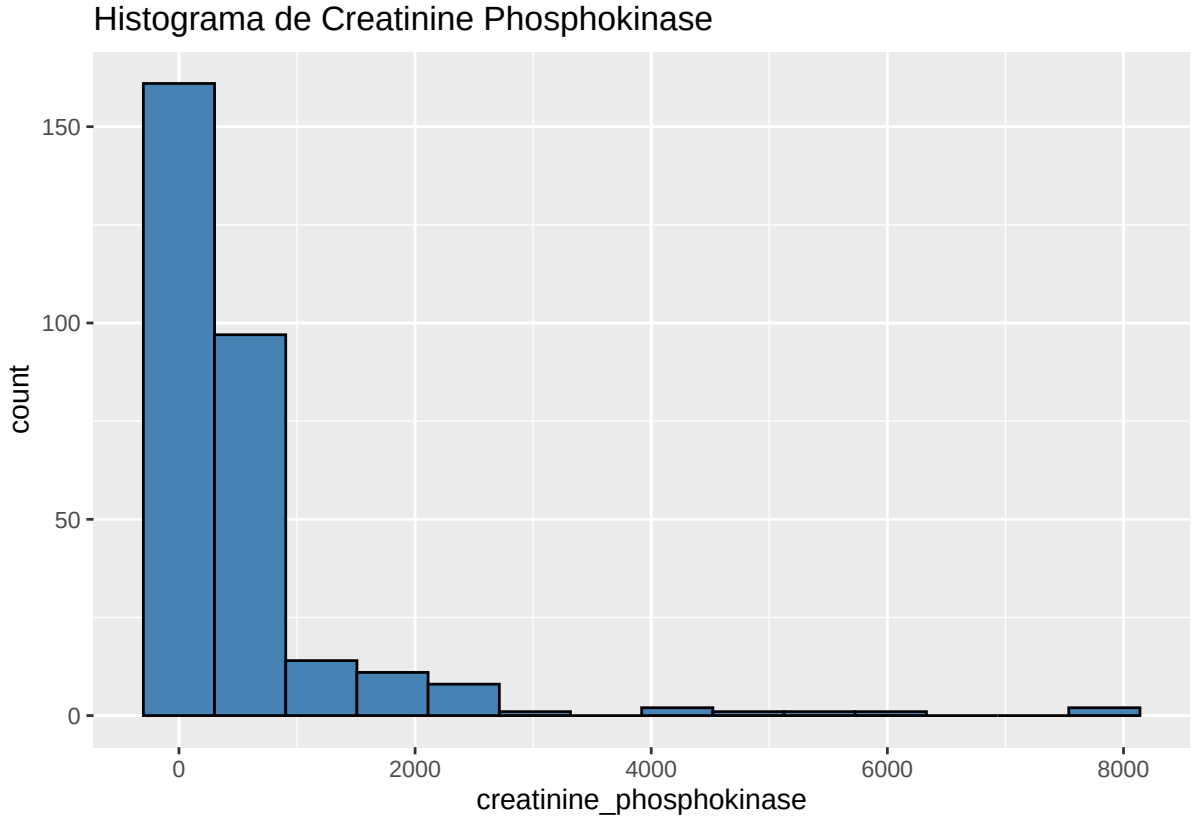


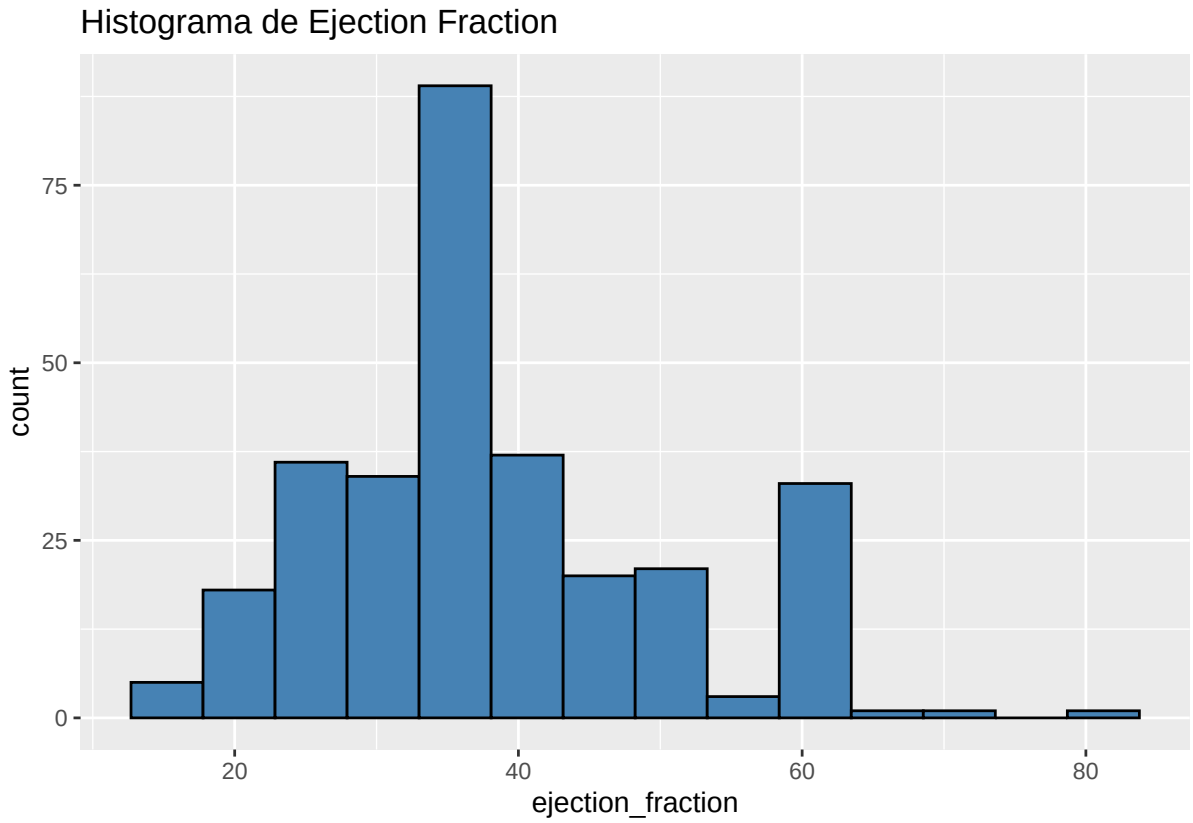
Aqui podemos notar uma grande quantidade de censura no banco, aproximadamente 68%.

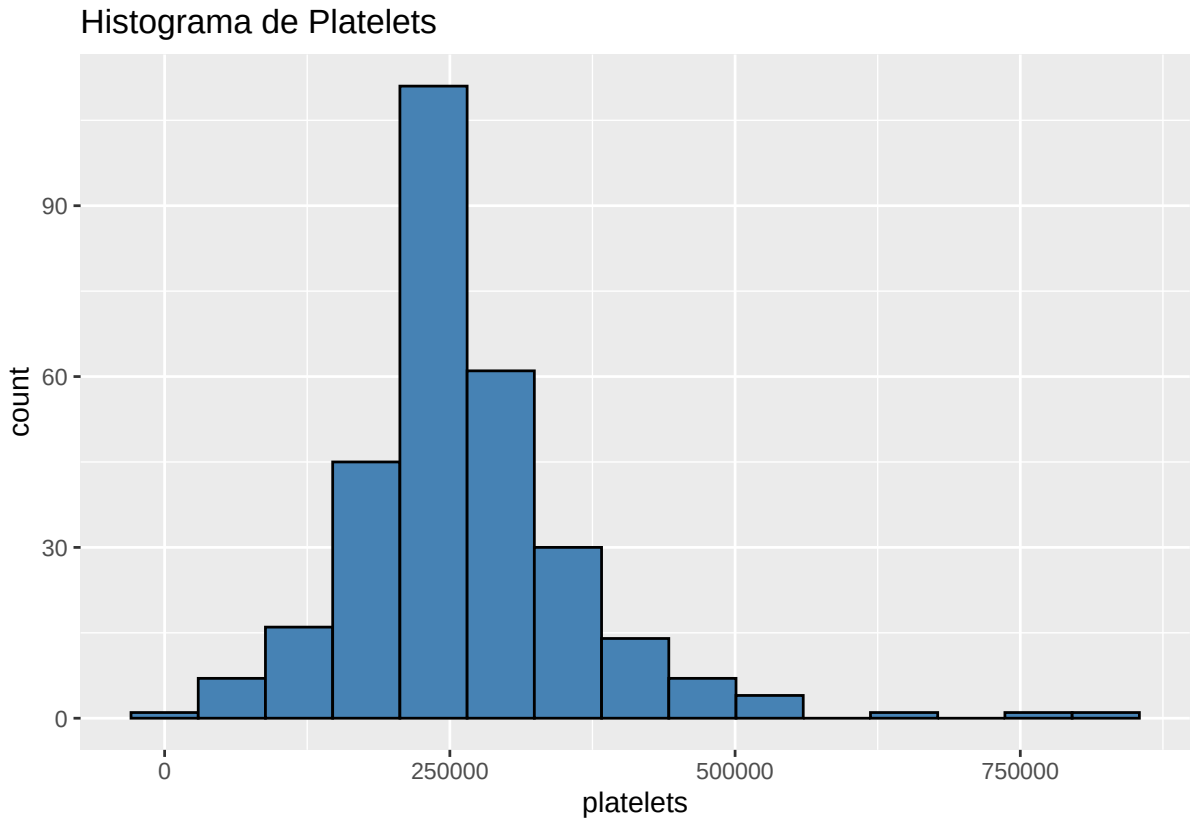
3.1.3 Análise das variáveis numéricas.

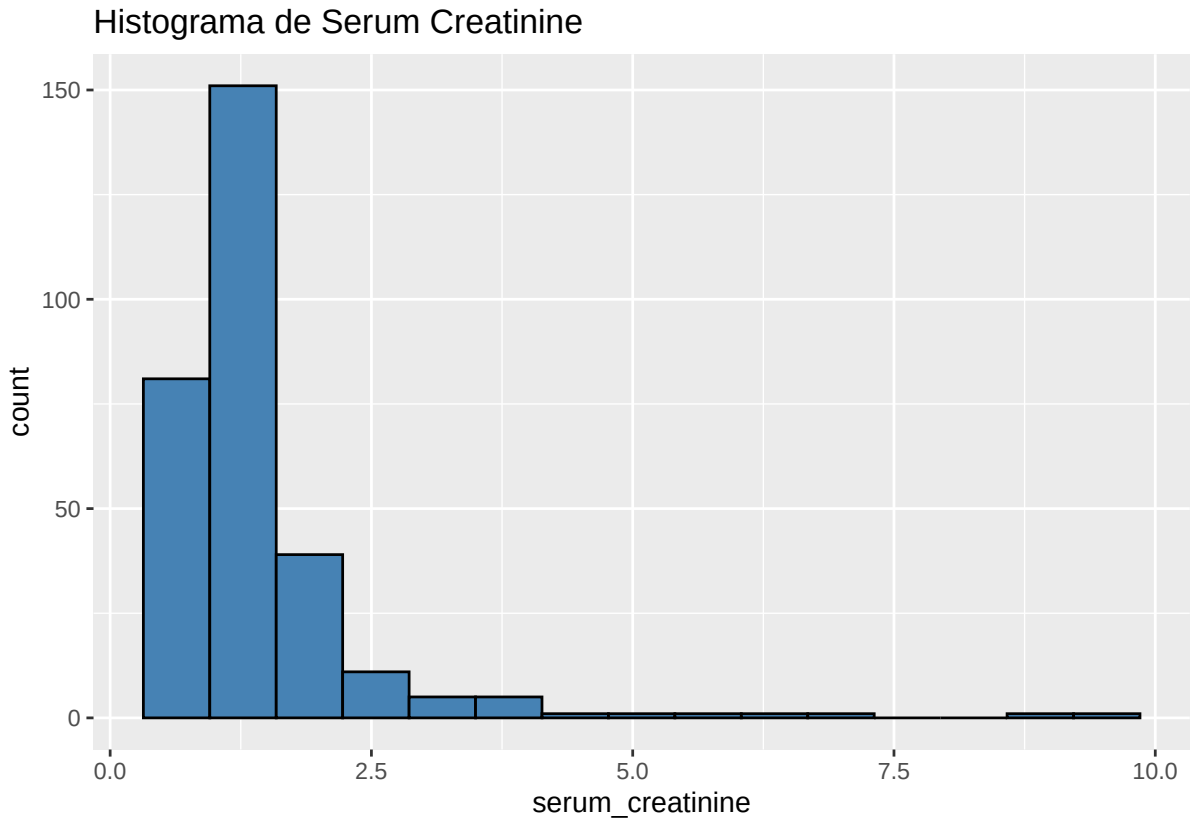
Vamos observar como é a distribuicao das variaveis numericas do banco

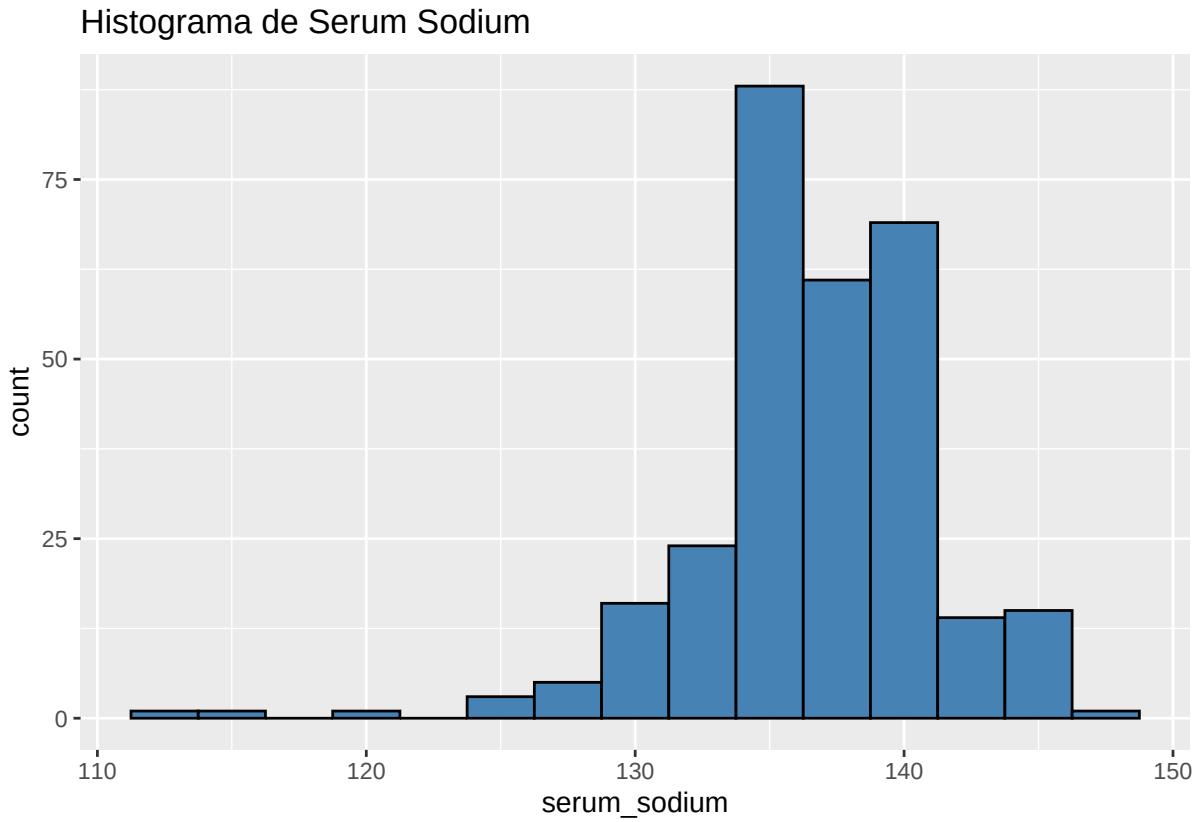




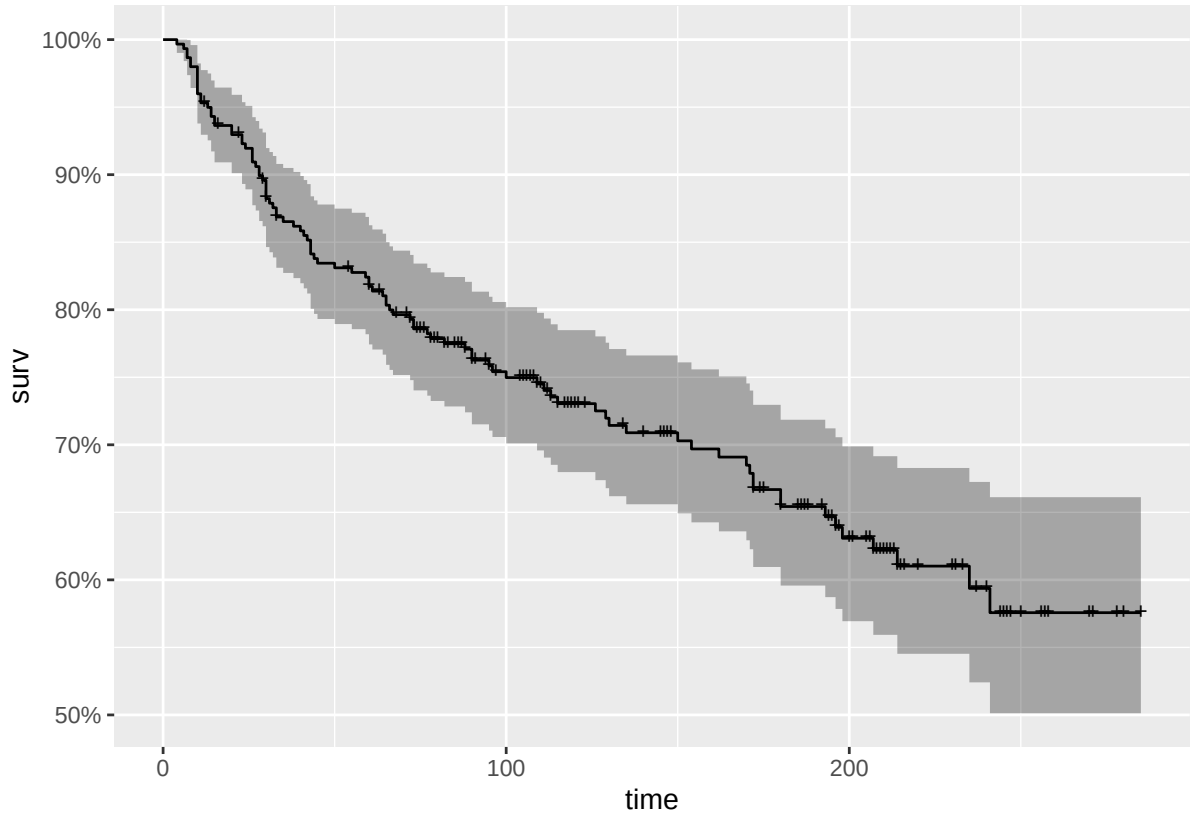

Histograma de Ejection Fraction
count
75
50
25
0
20
40
60
80
ejection_fraction


Histograma de Platelets
count
90
60
30
0
0
250000
500000
750000
platelets


Histograma de Serum Creatinine
count
150
100
50
0
0.0
2.5
5.0
7.5
10.0
serum_creatinine

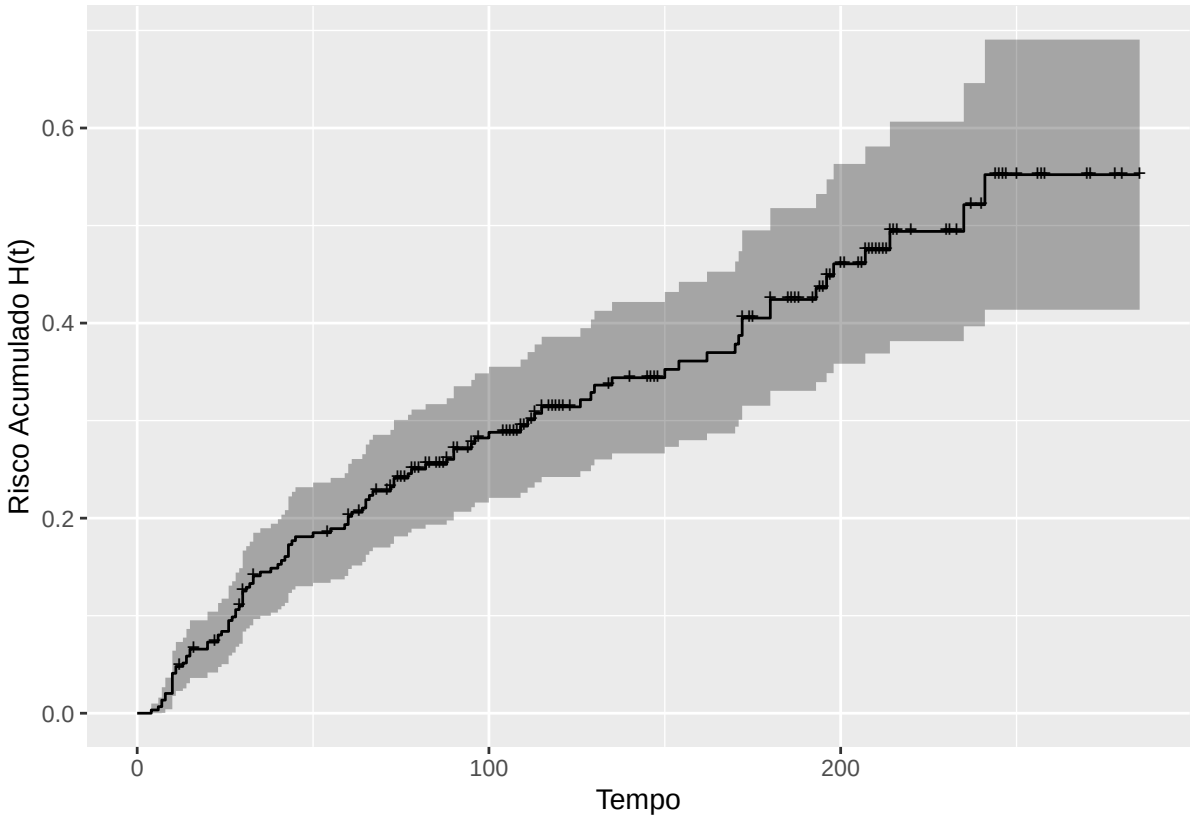

Histograma de Serum Sodium
count
75
50
25
0
110
120
130
140
150
serum_sodium

3.1.4 Modelo de sobrevivência não paramétrico de Kaplan-Meier.



Podemos notar que nosso estimador nao chega a 0, indicando a existencia de uma fracao de cura ou seja algumas observacoes nunca vao registrar o evento de interesse. No caso vamos estar comparando um modelo sem fracao de cura com um que leva isso em consideracao

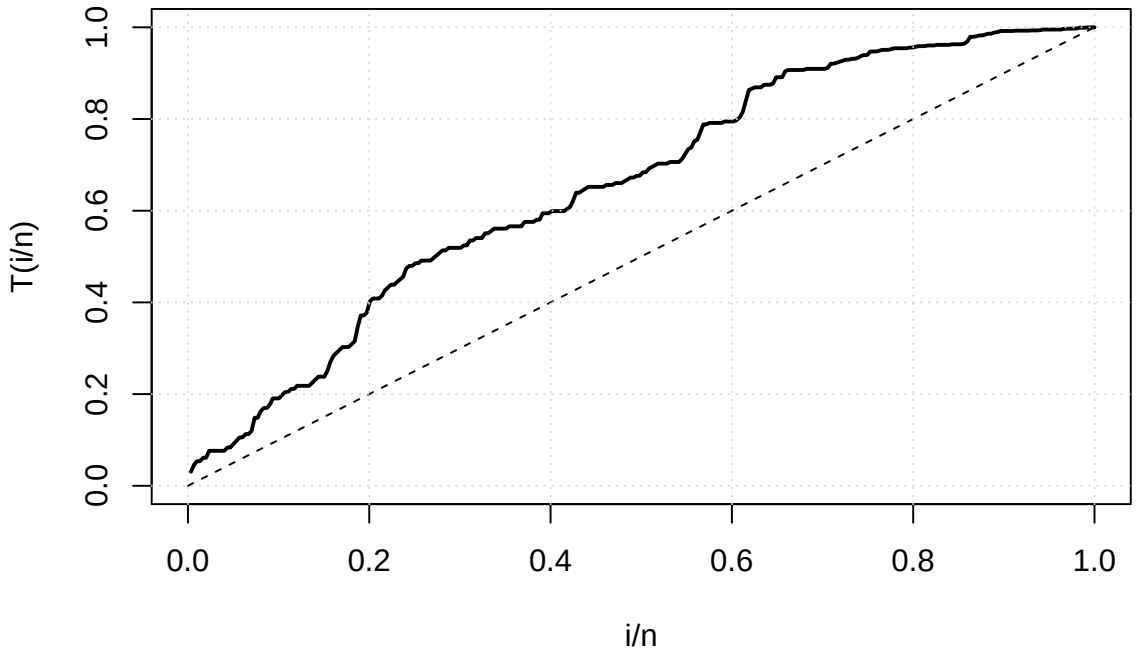
3.1.5 A função de risco acumulado



Pelo gráfico podemos notar o risco crescente do nosso estudo.

3.1.6 Curva TTT (Tempo Total sobre Teste)

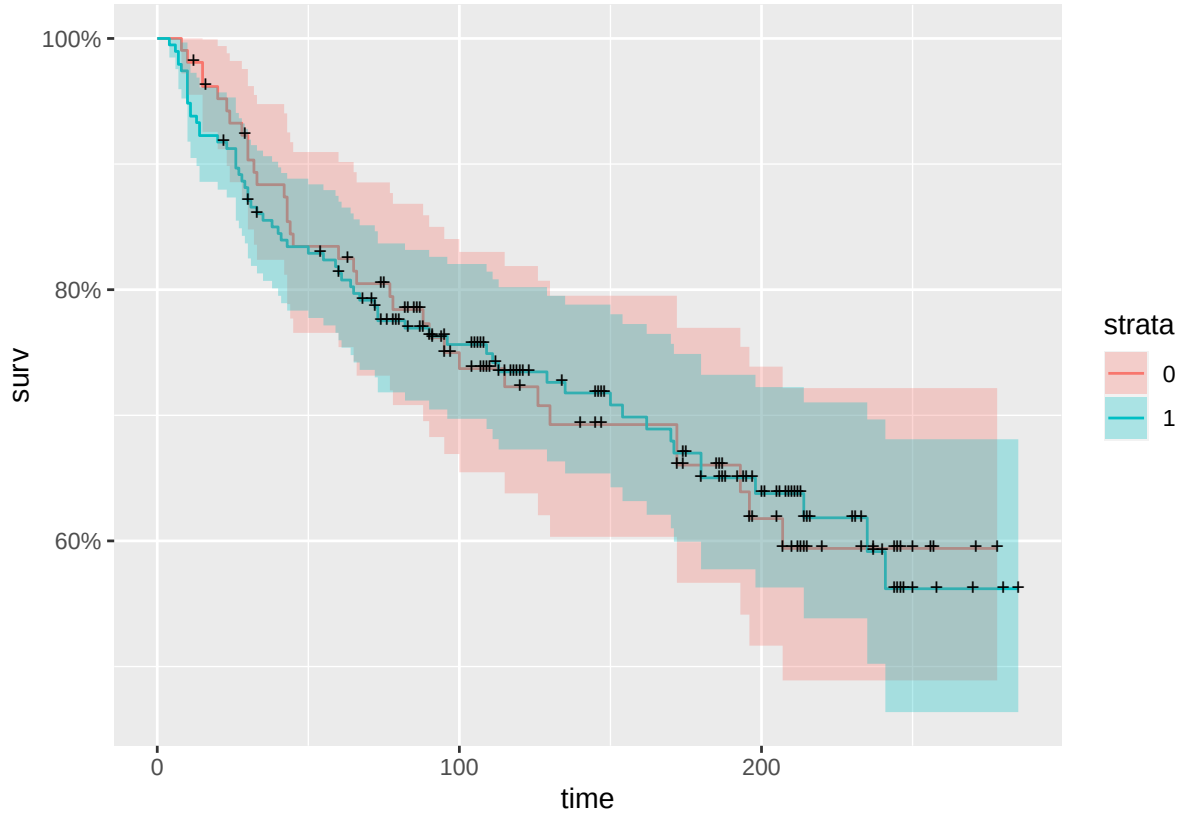
Gráfico do Tempo Total sobre Teste



Pelo formato da curva do gráfico, podemos notar que a função taxa de falha é monotonicamente crescente.

3.1.7 Analise das Variaveis Categoricas

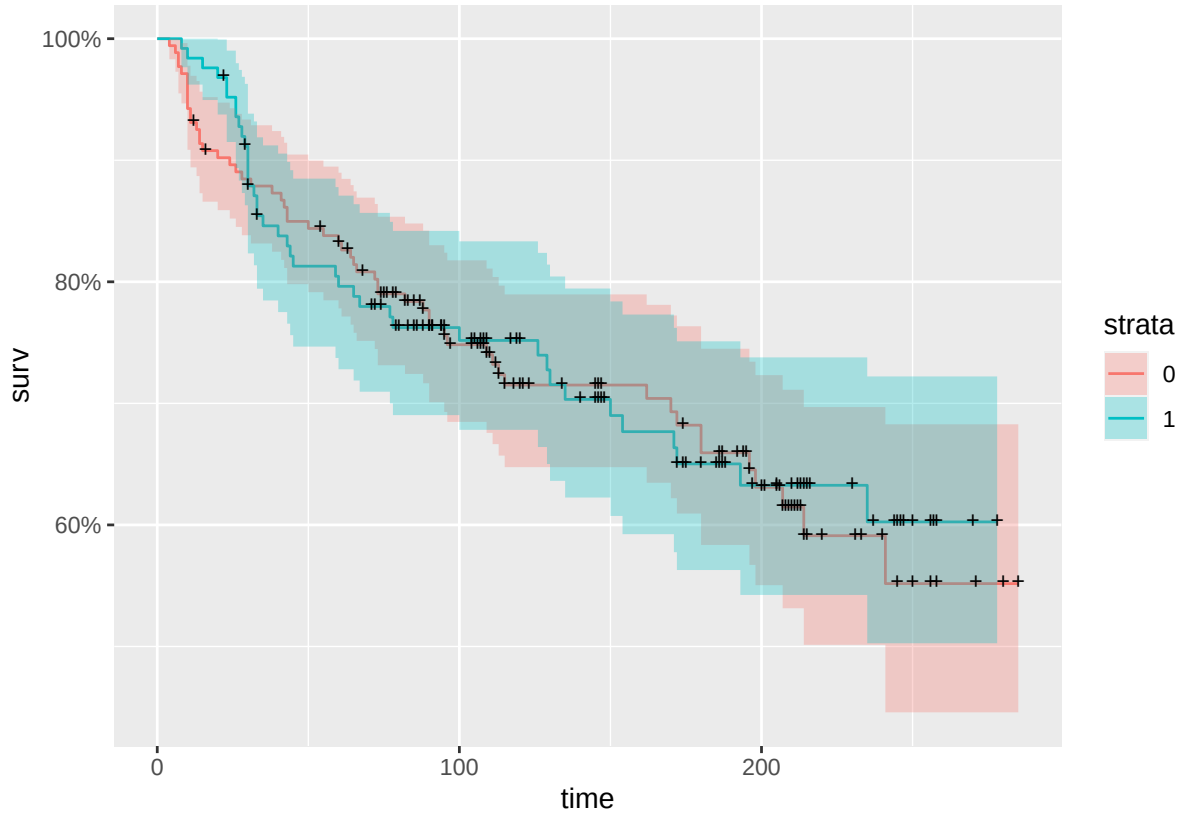
3.1.7.1 Variavel Sex Vamos comparar as curvas de sobrevivência divididas por sexo, com o objetivo de verificar se essa variável influencia na curva de sobrevivência. Em seguida, iremos realizar um teste para verificar a diferença entre as curvas.




```
## 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, tanto pelo gráfico quanto pelo teste com p-valor igual a 0.9, que a variável Sexo não parece influenciar nas curvas de sobrevivência.

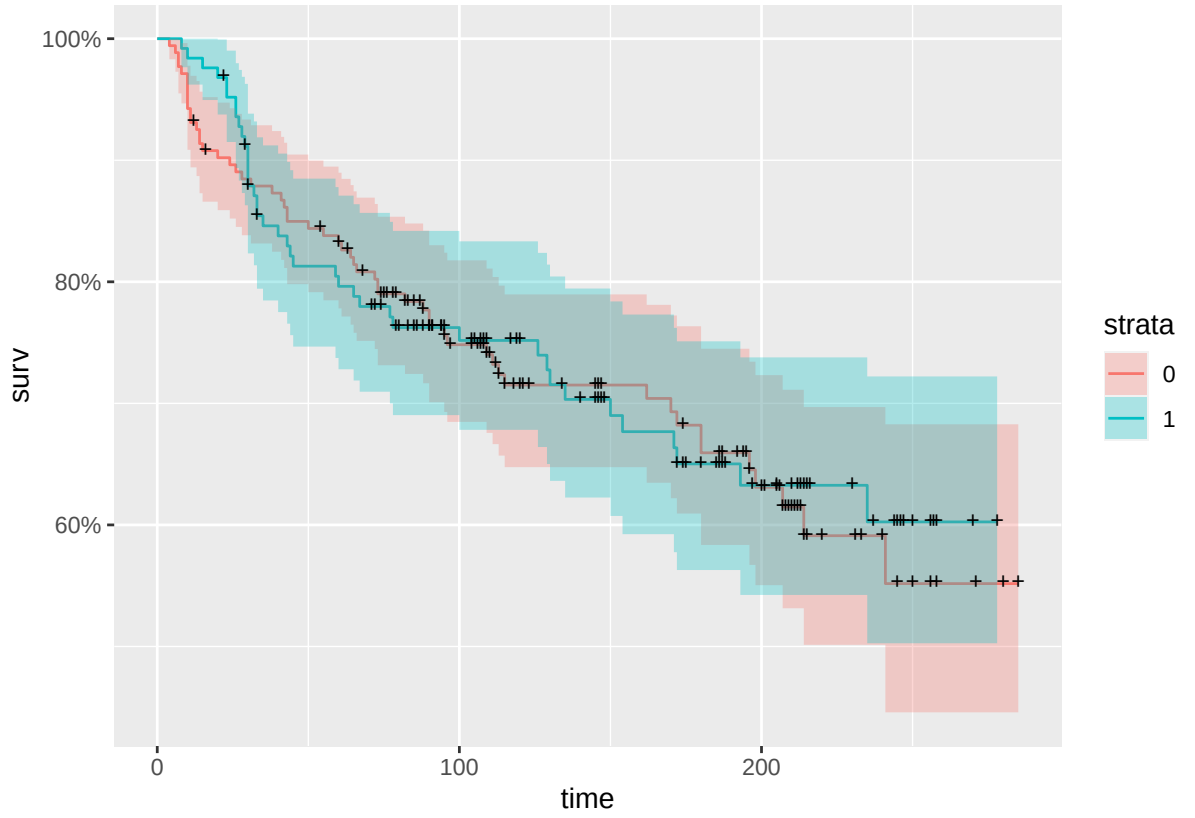
3.1.7.2 Variavel Diabetes




```
## 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
```

Com um p-valor de 0.9, podemos concluir que não há evidência estatística suficiente para afirmar a existência de diferença significativa entre as curvas.

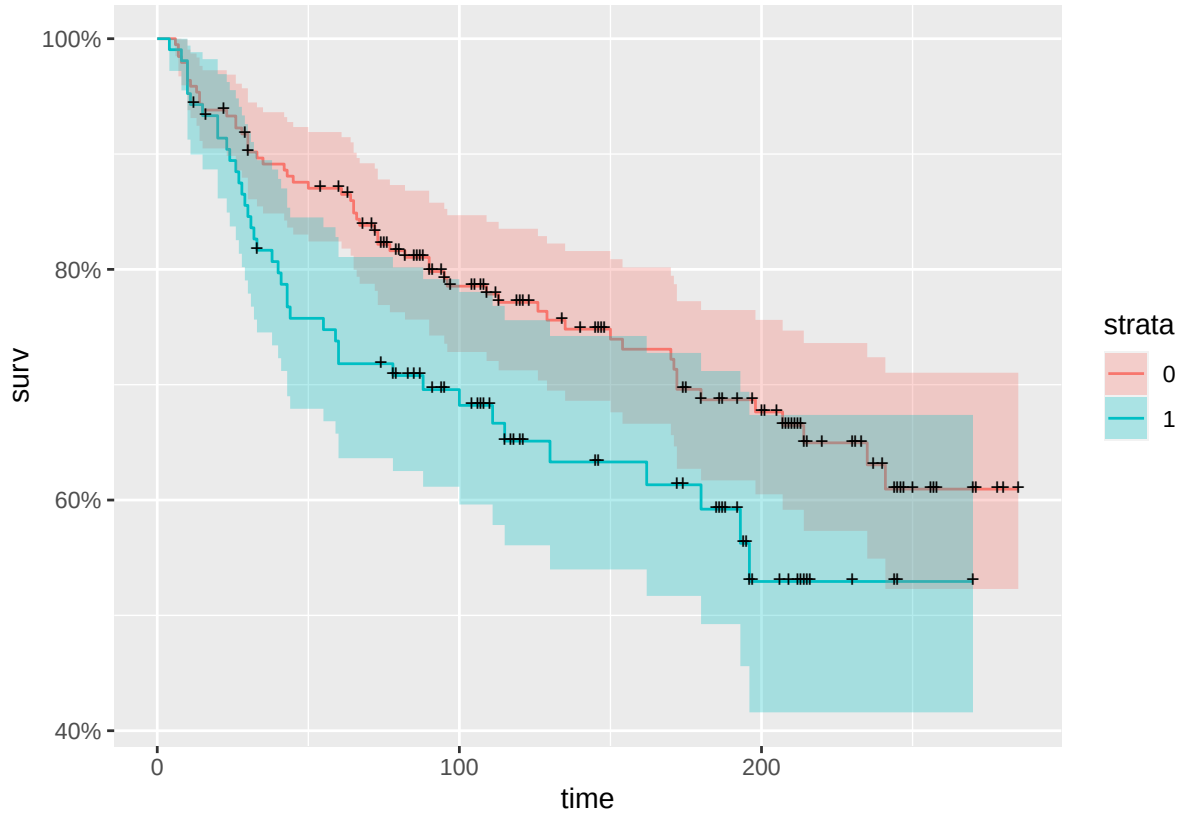
3.1.7.3 Variavel Anaemia




```
## 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
```

Com um p-valor de 0.9, podemos inferir que não há diferença estatisticamente significativa entre as categorias.

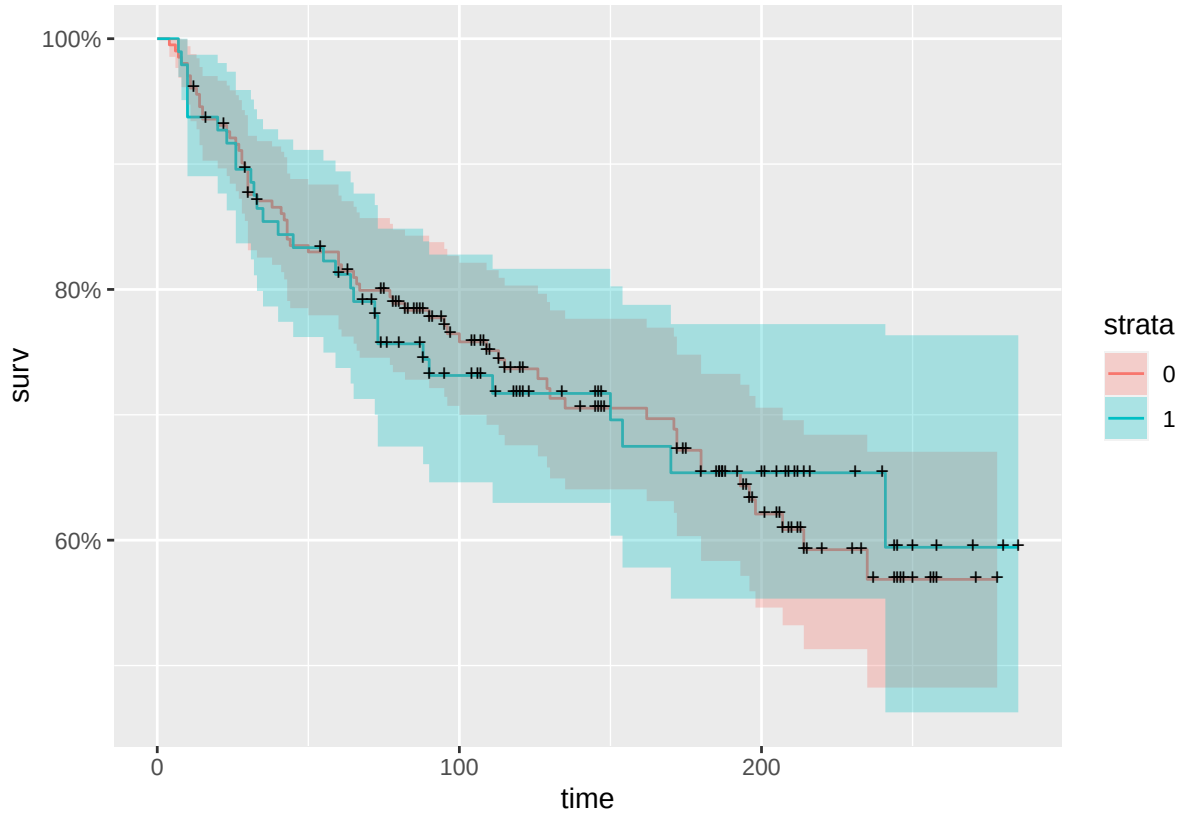
3.1.7.4 Variavel High Blood Pressure




```
## 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
```

Aqui vemos que a variavel influencia na sobrevivnecia tendo em vista as curvas mais separadas e sem cruzzamentos, ou seja essa é uma variavel que podemos estar analisando no modelo final.

3.1.7.5 Variavel Smoking

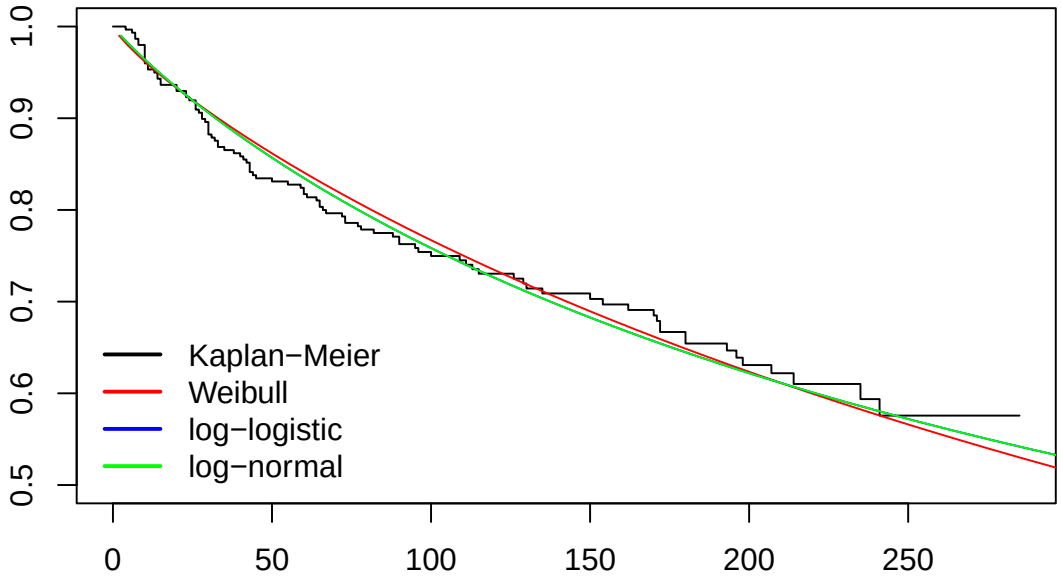



```
## Call:
## survdiff(formula = Surv(tempo, censura) ~ smoking, data = dados,
##     rho = 1)
##
##              N Observed Expected (O-E)^2/E (O-E)^2/V
## smoking=0 203     54.5     54.7  0.000902   0.00339
## smoking=1  96     25.5     25.2  0.001955   0.00339
##
##  Chisq= 0  on 1 degrees of freedom,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

Podemos notar pela curva que nao existe diferenca significativa.

3.2 Seleção da distribuição.

Vamos ajustar diferentes distribuições sobre o gráfico de Kaplan-Meier para selecionar aquela que melhor se adapta à curva. Além disso, estaremos verificando os valores de AIC, AIC corrigido e BIC para decidir a distribuição 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

##         AIClns  AICclns   BIClns
## [1,] 1336.546 1336.587 1343.947

##         AIClls  AICclls   BIClls
## [1,] 1342.334 1342.375 1349.735
```

Podemos observar pelos valores de AIC, AICc e BIC que a distribuição mais recomendada é a log-normal, no entanto, todas estão bastante próximas, indicando que também poderiam ser utilizadas com resultados semelhantes.

3.3 Selecao de variaveis

3.3.1 Manual

Vamos manualmente selecionar as variáveis que devem permanecer no modelo final. Isso será feito para posterior comparação com outro modelo selecionado por uma função já implementada no R. Ao final, pretendemos comparar os dois modelos.

3.3.1.1 Etapa 1: Ajustar os modelos com somente uma covariavel para verificar se sao significativas individualmente.

```
##
## Call:
## survreg(formula = s ~ age, data = dados, dist = "lognorm")
##               Value Std. Error     z       p
## (Intercept)  8.9137     0.7319 12.18 < 2e-16
## age         -0.0494     0.0109 -4.52 6.1e-06
## Log(scale)   0.5949     0.0813  7.32 2.5e-13
##
## Scale= 1.81
##
## Log Normal distribution
## Loglik(model)= -655.5   Loglik(intercept only)= -666.3
##  Chisq= 21.58 on 1 degrees of freedom, p= 3.4e-06
## Number of Newton-Raphson Iterations: 4
## n= 299
```

```
##
## Call:
## survreg(formula = s ~ anaemia, data = dados, dist = "lognorm")
##               Value Std. Error    z       p
## (Intercept)  6.1137     0.2218 27.6 < 2e-16
## anaemia     -0.4851     0.2701 -1.8   0.073
## Log(scale)   0.6384     0.0818  7.8 6.1e-15
##
## Scale= 1.89
##
## Log Normal distribution
## Loglik(model)= -664.7   Loglik(intercept only)= -666.3
##  Chisq= 3.21 on 1 degrees of freedom, p= 0.073
## Number of Newton-Raphson Iterations: 3
## n= 299
```

```
##
## Call:
## survreg(formula = s ~ creatinine_phosphokinase, data = dados,
##     dist = "lognorm")
##                               Value Std. Error     z       p
## (Intercept)                5.976386   0.199985 29.88 < 2e-16
## creatinine_phosphokinase -0.000103   0.000128 -0.81    0.42
```

```
## Log(scale)                0.649713   0.081890  7.93 2.1e-15
##
## Scale= 1.91
##
## Log Normal distribution
## Loglik(model)= -665.9   Loglik(intercept only)= -666.3
##  Chisq= 0.65 on 1 degrees of freedom, p= 0.42
## Number of Newton-Raphson Iterations: 3
## n= 299

##
## Call:
## survreg(formula = s ~ diabetes, data = dados, dist = "lognorm")
##              Value Std. Error     z       p
## (Intercept) 5.8724     0.2142 27.42 < 2e-16
## diabetes    0.0985     0.2751  0.36    0.72
## Log(scale)  0.6498     0.0819  7.93 2.1e-15
##
## Scale= 1.92
##
## Log Normal distribution
## Loglik(model)= -666.2   Loglik(intercept only)= -666.3
##  Chisq= 0.13 on 1 degrees of freedom, p= 0.72
## Number of Newton-Raphson Iterations: 3
## n= 299

##
## Call:
## survreg(formula = s ~ ejection_fraction, data = dados, dist = "lognorm")
##                     Value Std. Error    z       p
## (Intercept)       4.3241     0.4451 9.71 < 2e-16
## ejection_fraction 0.0430     0.0121 3.54   4e-04
## Log(scale)        0.6354     0.0816 7.78 7.1e-15
##
## Scale= 1.89
##
## Log Normal distribution
## Loglik(model)= -659.3   Loglik(intercept only)= -666.3
##  Chisq= 13.86 on 1 degrees of freedom, p= 2e-04
## Number of Newton-Raphson Iterations: 3
## n= 299

##
## Call:
## survreg(formula = s ~ high_blood_pressure, data = dados, dist = "lognorm")
##                        Value Std. Error     z       p
## (Intercept)           6.1051     0.2113 28.89 < 2e-16
## high_blood_pressure -0.5845     0.2767 -2.11   0.035
## Log(scale)           0.6339     0.0818  7.75 9.3e-15
##
```

```
## Scale= 1.88
##
## Log Normal distribution
## Loglik(model)= -664.1   Loglik(intercept only)= -666.3
##  Chisq= 4.43 on 1 degrees of freedom, p= 0.035
## Number of Newton-Raphson Iterations: 3
## n= 299

##
## Call:
## survreg(formula = s ~ platelets, data = dados, dist = "lognorm")
##                Value Std. Error     z       p
## (Intercept) 5.64e+00   4.02e-01 14.04 < 2e-16
## platelets   1.03e-06   1.40e-06  0.74    0.46
## Log(scale)  6.50e-01   8.19e-02  7.93 2.2e-15
##
## Scale= 1.91
##
## Log Normal distribution
## Loglik(model)= -666   Loglik(intercept only)= -666.3
##  Chisq= 0.55 on 1 degrees of freedom, p= 0.46
## Number of Newton-Raphson Iterations: 3
## n= 299

##
## Call:
## survreg(formula = s ~ serum_creatinine, data = dados, dist = "lognorm")
##                     Value Std. Error     z       p
## (Intercept)        6.5854     0.2663 24.73 < 2e-16
## serum_creatinine -0.4786     0.1140 -4.20 2.7e-05
## Log(scale)         0.6128     0.0814  7.53 5.1e-14
##
## Scale= 1.85
##
## Log Normal distribution
## Loglik(model)= -657.1   Loglik(intercept only)= -666.3
##  Chisq= 18.33 on 1 degrees of freedom, p= 1.9e-05
## Number of Newton-Raphson Iterations: 3
## n= 299

##
## Call:
## survreg(formula = s ~ serum_sodium, data = dados, dist = "lognorm")
##                Value Std. Error     z       p
## (Intercept)  -7.3607     3.9452 -1.87 0.06208
## serum_sodium  0.0972     0.0291  3.34 0.00085
## Log(scale)    0.6237     0.0815  7.65   2e-14
##
## Scale= 1.87
##
```

```
## Log Normal distribution
## Loglik(model)= -660.6   Loglik(intercept only)= -666.3
##  Chisq= 11.44 on 1 degrees of freedom, p= 0.00072
## Number of Newton-Raphson Iterations: 3
## n= 299

##
## Call:
## survreg(formula = s ~ sex, data = dados, dist = "lognorm")
##               Value Std. Error     z       p
## (Intercept)  5.9742     0.2589 23.07 < 2e-16
## sex         -0.0934     0.2846 -0.33    0.74
## Log(scale)   0.6497     0.0819  7.93 2.2e-15
##
## Scale= 1.92
##
## Log Normal distribution
## Loglik(model)= -666.2   Loglik(intercept only)= -666.3
##  Chisq= 0.11 on 1 degrees of freedom, p= 0.74
## Number of Newton-Raphson Iterations: 3
## n= 299

##
## Call:
## survreg(formula = s ~ smoking, data = dados, dist = "lognorm")
##               Value Std. Error     z      p
## (Intercept)  5.9207     0.2032 29.14 <2e-16
## smoking     -0.0190     0.2908 -0.07   0.95
## Log(scale)   0.6505     0.0819  7.94  2e-15
##
## Scale= 1.92
##
## Log Normal distribution
## Loglik(model)= -666.3   Loglik(intercept only)= -666.3
##  Chisq= 0 on 1 degrees of freedom, p= 0.95
## Number of Newton-Raphson Iterations: 3
## n= 299
```

Observando os p-valores de cada modelo ajustado com uma variavel sao significantes a nivel de 5%: `serum_sodium`,`serum_creatinine`,`high_blood_pressure`,`ejection_fraction` e `age`.

3.3.1.2 Etapa 2: Ajustar um modelo com todas essas variaveis significativas e comparar com modelos reduzidos removendo uma dessas variaveis.

```
## Likelihood ratio test
##
## Model 1: s ~ age + ejection_fraction + high_blood_pressure + serum_creatinine +
##     serum_sodium
## Model 2: s ~ age + ejection_fraction + high_blood_pressure + serum_creatinine
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   7 -634.69
## 2   6 -636.74 -1 4.0984    0.04292 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Likelihood ratio test
##
## Model 1: s ~ age + ejection_fraction + high_blood_pressure + serum_creatinine +
##     serum_sodium
## Model 2: s ~ age + ejection_fraction + high_blood_pressure + serum_sodium
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   7 -634.69
## 2   6 -641.07 -1 12.763  0.0003536 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Likelihood ratio test
##
## Model 1: s ~ age + ejection_fraction + high_blood_pressure + serum_creatinine +
##     serum_sodium
## Model 2: s ~ age + ejection_fraction + serum_creatinine + serum_sodium
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   7 -634.69
## 2   6 -636.52 -1 3.6491     0.0561 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Likelihood ratio test
##
## Model 1: s ~ age + ejection_fraction + high_blood_pressure + serum_creatinine +
##     serum_sodium
## Model 2: s ~ age + high_blood_pressure + serum_creatinine + serum_sodium
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   7 -634.69
## 2   6 -642.66 -1 15.928  6.578e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Likelihood ratio test
##
## Model 1: s ~ age + ejection_fraction + high_blood_pressure + serum_creatinine +
```

```
##     serum_sodium
## Model 2: s ~ ejection_fraction + high_blood_pressure + serum_creatinine +
##     serum_sodium
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   7 -634.69
## 2   6 -644.84 -1 20.302  6.615e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A única variável que não apresentou significância foi `high_blood_pressure`. Optamos por selecionar o modelo sem ela. Em seguida, deveríamos ajustar um modelo com as variáveis restantes e compará-lo com o modelo que excluiu a variável, para verificar se a exclusão conjunta tem algum impacto. No entanto, como apenas uma variável foi removida, não é necessário realizar esse procedimento, pois repetiríamos o teste já conduzido.

3.3.1.3 Etapa 3: Ajustar o modelo sem a variável excluída e compará-lo com o modelo que continha as variáveis removidas na etapa 1.

O objetivo desta etapa é verificar se alguma das variáveis excluídas na etapa 1 (`anaemia`,`creatinine_phosphokinase`,`diabetes`,`platelets`,`sex`,`smoking`) é significativa em conjunto com as outras selecionadas na etapa 2.

```
## Likelihood ratio test
##
## Model 1: s ~ age + ejection_fraction + serum_creatinine + serum_sodium
## Model 2: s ~ age + ejection_fraction + serum_creatinine + serum_sodium +
##     anaemia
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   6 -636.52
## 2   7 -635.06  1 2.9079    0.08815 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Likelihood ratio test
##
## Model 1: s ~ age + ejection_fraction + serum_creatinine + serum_sodium
## Model 2: s ~ age + ejection_fraction + serum_creatinine + serum_sodium +
##     creatinine_phosphokinase
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   6 -636.52
## 2   7 -635.28  1 2.4673     0.1162

## Likelihood ratio test
##
## Model 1: s ~ age + ejection_fraction + serum_creatinine + serum_sodium
## Model 2: s ~ age + ejection_fraction + serum_creatinine + serum_sodium +
##     diabetes
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   6 -636.52
## 2   7 -636.42  1 0.1933     0.6602

## Likelihood ratio test
##
## Model 1: s ~ age + ejection_fraction + serum_creatinine + serum_sodium
## Model 2: s ~ age + ejection_fraction + serum_creatinine + serum_sodium +
##     platelets
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   6 -636.52
## 2   7 -636.50  1 0.0295     0.8635

## Likelihood ratio test
##
## Model 1: s ~ age + ejection_fraction + serum_creatinine + serum_sodium
## Model 2: s ~ age + ejection_fraction + serum_creatinine + serum_sodium +
##     sex
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   6 -636.52
## 2   7 -636.20  1 0.6386     0.4242
```

```
## Likelihood ratio test
##
## Model 1: s ~ age + ejection_fraction + serum_creatinine + serum_sodium
## Model 2: s ~ age + ejection_fraction + serum_creatinine + serum_sodium +
##     smoking
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   6 -636.52
## 2   7 -636.47  1 0.0981     0.7542
```

Nenhuma das variáveis que foram inicialmente retiradas permanece no modelo com um nível de significância de 5%.

3.3.1.4 Etapa 4: Vamos verificar se é possível simplificar o modelo obtido na etapa 3 removendo alguma variável.

```
## Likelihood ratio test
##
## Model 1: s ~ age + ejection_fraction + serum_creatinine + serum_sodium
## Model 2: s ~ age + ejection_fraction + serum_creatinine
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   6 -636.52
## 2   5 -638.41 -1 3.7975    0.05133 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Likelihood ratio test
##
## Model 1: s ~ age + ejection_fraction + serum_creatinine + serum_sodium
## Model 2: s ~ age + ejection_fraction + serum_sodium
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   6 -636.52
## 2   5 -642.97 -1 12.899  0.0003287 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Likelihood ratio test
##
## Model 1: s ~ age + ejection_fraction + serum_creatinine + serum_sodium
## Model 2: s ~ age + serum_creatinine + serum_sodium
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   6 -636.52
## 2   5 -644.56 -1 16.095  6.024e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Likelihood ratio test
##
## Model 1: s ~ age + ejection_fraction + serum_creatinine + serum_sodium
## Model 2: s ~ ejection_fraction + serum_creatinine + serum_sodium
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   6 -636.52
## 2   5 -647.21 -1 21.398  3.733e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Com base nos resultados a um nível de significância de 5%, optamos por remover a variável `serum_sodium`. Com essa decisão, chegamos ao nosso modelo final para os efeitos principais.

3.3.1.5 Etapa 5: Vamos verificar a interação entre as variáveis restantes no modelo final obtido na etapa anterior.

```
## Likelihood ratio test
##
## Model 1: s ~ age + ejection_fraction + serum_creatinine
## Model 2: s ~ age + ejection_fraction + serum_creatinine + (age * ejection_fraction)
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   5 -638.41
## 2   6 -637.07  1 2.6874     0.1011
```

```
## Likelihood ratio test
##
## Model 1: s ~ age + ejection_fraction + serum_creatinine
## Model 2: s ~ age + ejection_fraction + serum_creatinine + (age * serum_creatinine)
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   5 -638.41
## 2   6 -638.40  1 0.0339      0.854
```

```
## Likelihood ratio test
##
## Model 1: s ~ age + ejection_fraction + serum_creatinine
## Model 2: s ~ age + ejection_fraction + serum_creatinine + (ejection_fraction *
##     serum_creatinine)
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   5 -638.41
## 2   6 -638.33  1 0.1595     0.6896
```

Nenhum dos termos de interação apresentou significância estatística para o modelo.

3.3.1.6 Etapa 6: Ajuste do modelo final

```
##
## Call:
## survreg(formula = s ~ age + ejection_fraction + serum_creatinine,
##     data = dados, dist = "lognorm")
##                     Value Std. Error     z       p
## (Intercept)        7.6559     0.7546 10.15 < 2e-16
## age               -0.0485     0.0108 -4.51 6.5e-06
## ejection_fraction  0.0490     0.0117  4.20 2.7e-05
## serum_creatinine  -0.4230     0.1077 -3.93 8.6e-05
## Log(scale)         0.5404     0.0804  6.72 1.8e-11
##
## Scale= 1.72
##
## Log Normal distribution
## Loglik(model)= -638.4   Loglik(intercept only)= -666.3
##  Chisq= 55.72 on 3 degrees of freedom, p= 4.8e-12
## Number of Newton-Raphson Iterations: 4
## n= 299
```

3.3.2 Stepwise

Nesta etapa, vamos utilizar uma função já implementada no R chamada `stepwiseCox()`, que serve para realizar a seleção de variáveis para modelos de sobrevivência.

```
##        Table 1. Summary of Parameters
##
##            Paramters               Value
## ----------------------------------------------
## Response Variable              s
## Included Variable              NULL
## Selection Method               bidirection
## Select Criterion               SL
## Entry Significance Level(sle)  0.05
## Stay Significance Level(sls)   0.05
## Method                         efron
## Multicollinearity Terms        NULL
##
##
##                                                                Table 2. Variables Type
##
##    class                                                                                variable
## ------------------------------------------------------------------------------------------------------------------------------------------------------
---------------------
## nmatrix.2  s
## numeric    age anaemia creatinine_phosphokinase diabetes ejection_fraction high_blood_pressure platelets serum_creatinine serum_sodium sex smoking
##
##
##                          Table 3. Process of Selection
##
##  Step     EnteredEffect     RemovedEffect  DF  NumberIn           SL
## ---------------------------------------------------------------------------------
## 1     age                                  1   1         1.23463804460881e-06
## 2     ejection_fraction                    1   2         6.41738636716686e-07
## 3     serum_creatinine                     1   3         1.97577379367093e-05
## 4     high_blood_pressure                  1   4         0.0285254624508617
##
##
##                      Table 4. Selected Varaibles
##
##  variables1     variables2         variables3         variables4
## ------------------------------------------------------------------------
## age         ejection_fraction  serum_creatinine  high_blood_pressure
##
##
##                                       Table 5. Coefficients of the Selected Variables
##
##       Variable               coef              exp(coef)           se(coef)                z                Pr(>|z|)
## ------------------------------------------------------------------------------------------------------------------------------
## age                  0.0441830655874345   1.04517367270842   0.00902794850348484  4.89403163635455   9.87909473896575e-07
## ejection_fraction    -0.0495889458782338  0.951620511649886  0.00996856973904786  -4.97452966436991  6.54062483178095e-07
## serum_creatinine     0.347022481997909    1.41484853363465   0.0667050008460413   5.20234581510397   1.96788607403579e-07
## high_blood_pressure  0.471204762458174    1.60192296777894   0.211409849883658    2.22886853529997   0.0258226532042563
##



##
## Call:
## survreg(formula = s ~ age + ejection_fraction + serum_creatinine +
##     high_blood_pressure, data = dados, dist = "lognorm")
##                       Value Std. Error     z       p
## (Intercept)          7.7264     0.7482 10.33 < 2e-16
## age                 -0.0469     0.0106 -4.40 1.1e-05
## ejection_fraction    0.0482     0.0115  4.21 2.6e-05
## serum_creatinine    -0.4167     0.1062 -3.92 8.7e-05
## high_blood_pressure -0.4768     0.2589 -1.84   0.065
## Log(scale)           0.5261     0.0804  6.55 5.9e-11
##
## Scale= 1.69
##
## Log Normal distribution
## Loglik(model)= -636.7   Loglik(intercept only)= -666.3
##  Chisq= 59.07 on 4 degrees of freedom, p= 4.6e-12
## Number of Newton-Raphson Iterations: 4
## n= 299
```

3.3.3 Modelo Manual com fracao de cura

Para esse modelo vamos considerar as variaveis selecionadas no modelo manual e adicionar a fracao de cura

```
## Paramaetro de Escala do Modelo (Sigma): 0.999643

## Taxa de cura do Modelo (Phi): 0.6789292

## Coeficientes do modelo:

## Intercepto 4.717952

## age: -0.01099357

## ejection_fraction -0.005385327

## serum_creatinine -0.0009745832

##        AICln  AICcln   BICln
## [1,] 1384.954 1384.995 1392.355
```

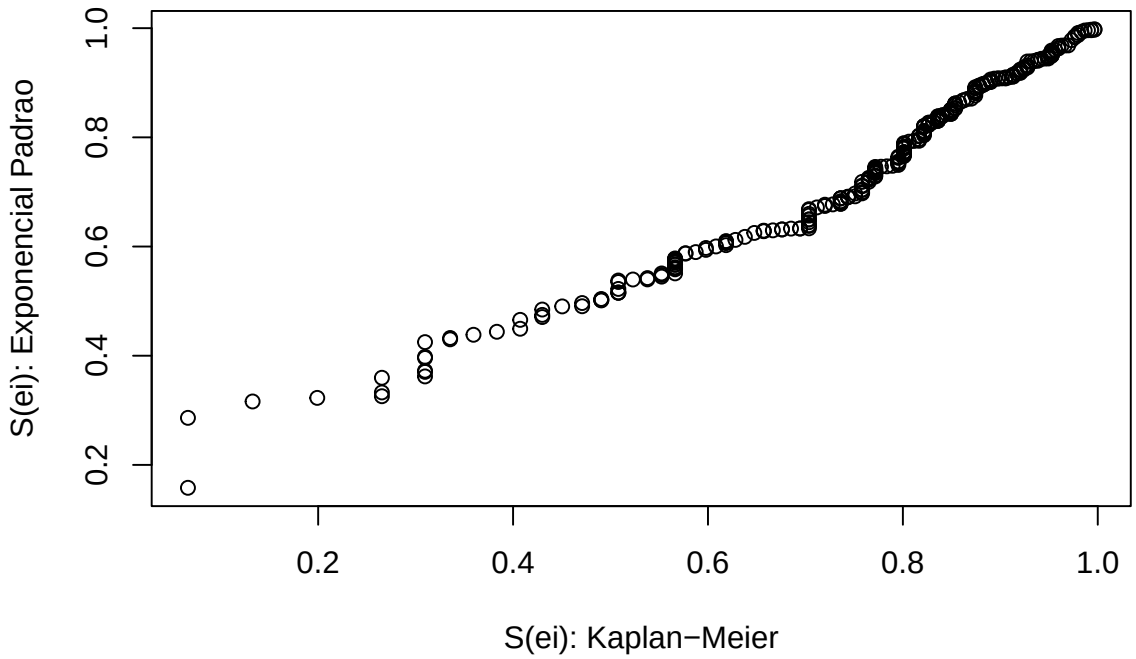
3.3.4 Modelo Stepwise com fracao de cura

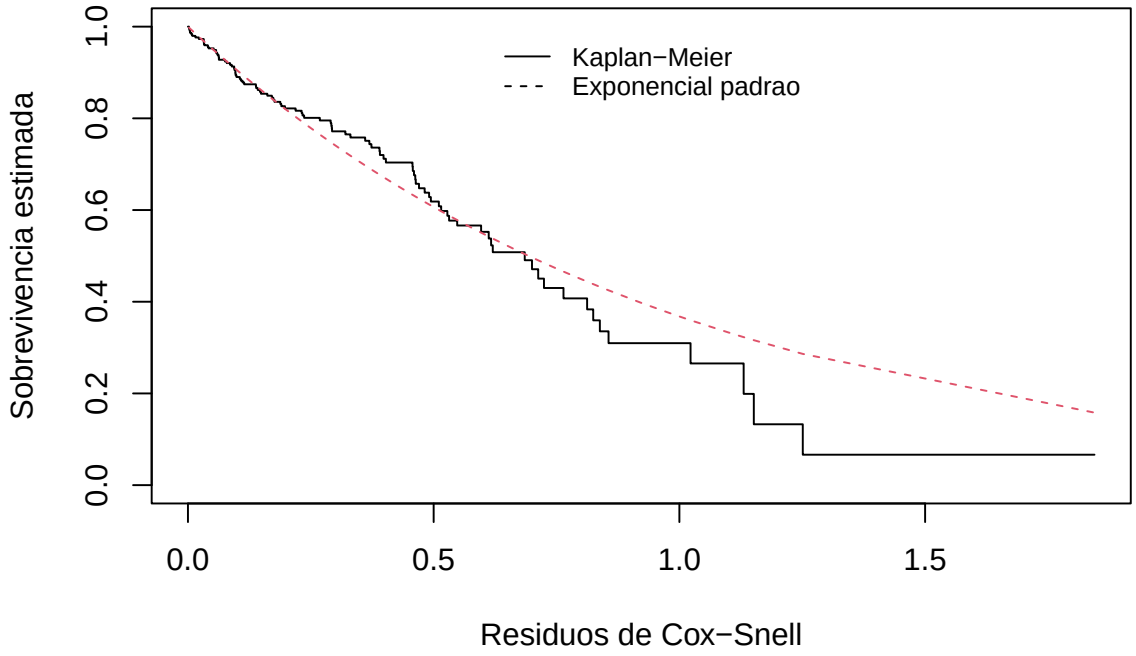
Para esse modelo vamos considerar as variaveis selecionadas no modelo Stepwise e adicionar a fracao de cura

```
## Paramaetro de Escala do Modelo (Sigma): 0.989679

## Taxa de cura do Modelo (Phi): 0.6789277

## Coeficientes do modelo:

## Intercepto 4.800037

## age: -0.01032949

## ejection_fraction -0.006028014

## high_blood_pressure -0.2882529

## serum_creatinine 0.006226618

##        AICln  AICcln   BICln
## [1,] 1383.028 1383.069 1390.429
```

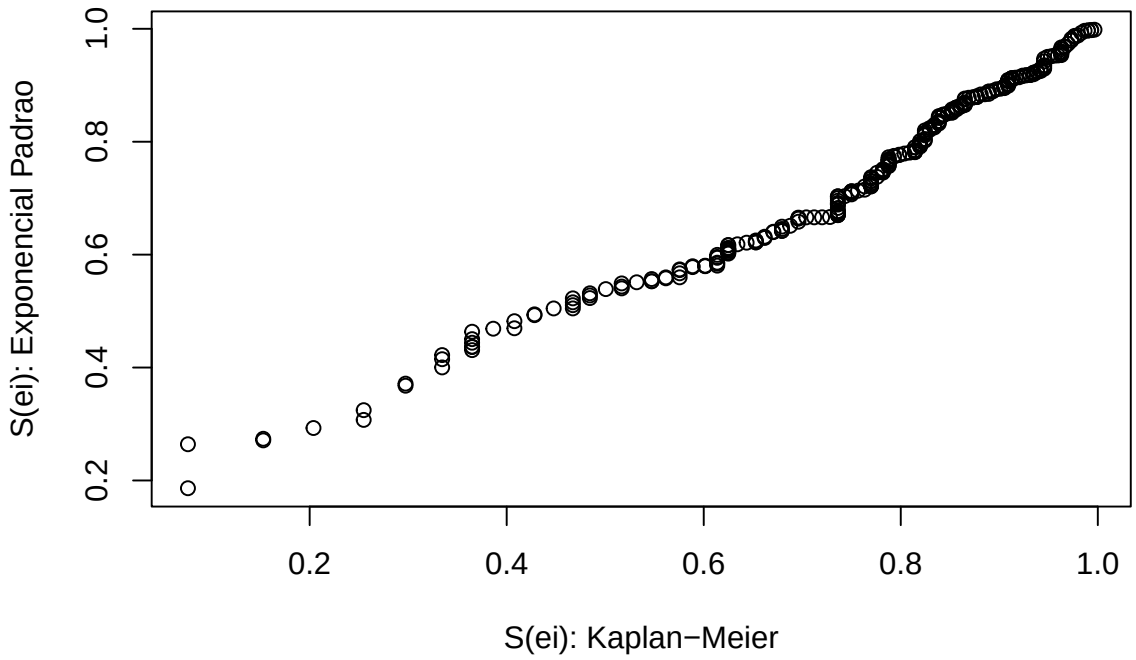
3.4 Analise de residuos

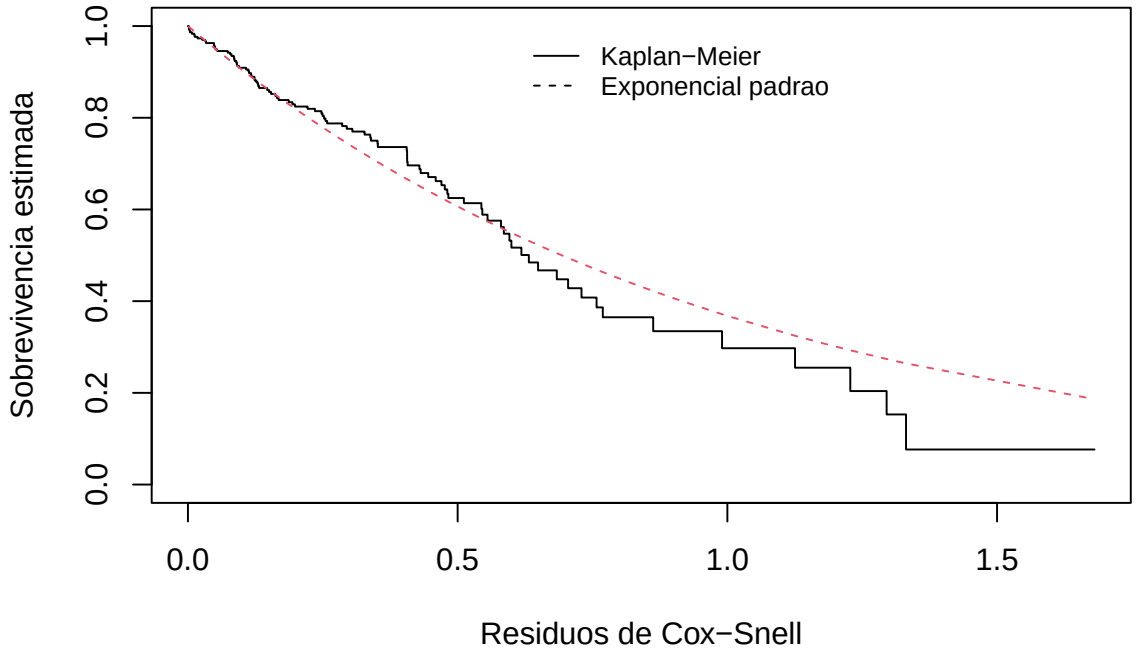
3.4.1 Modelo Manual



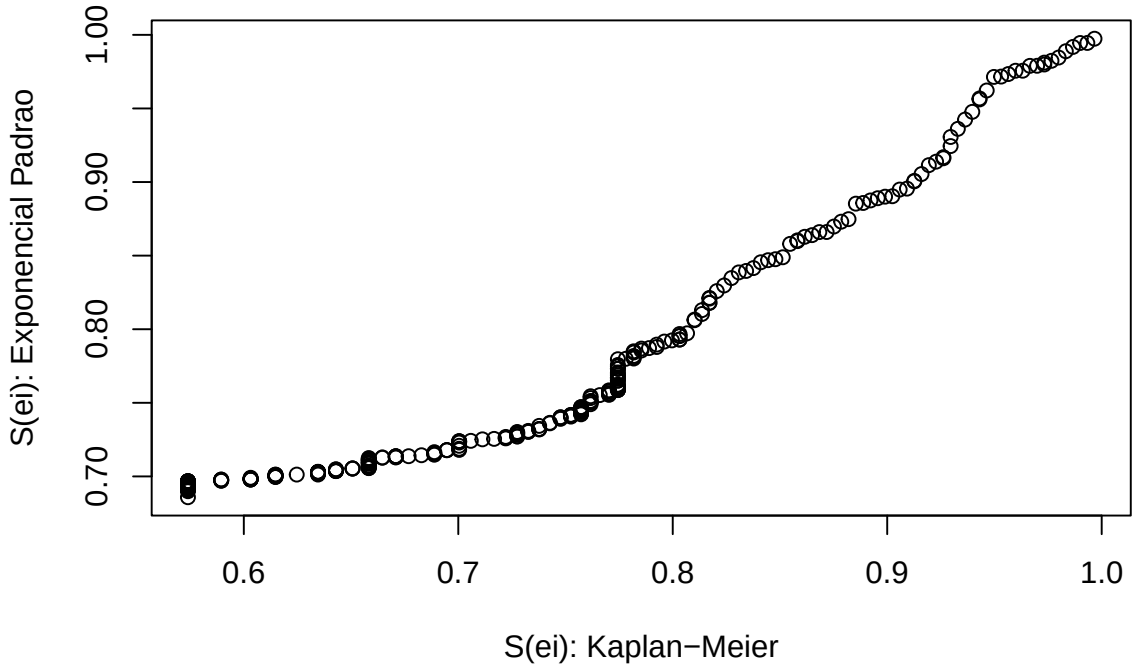


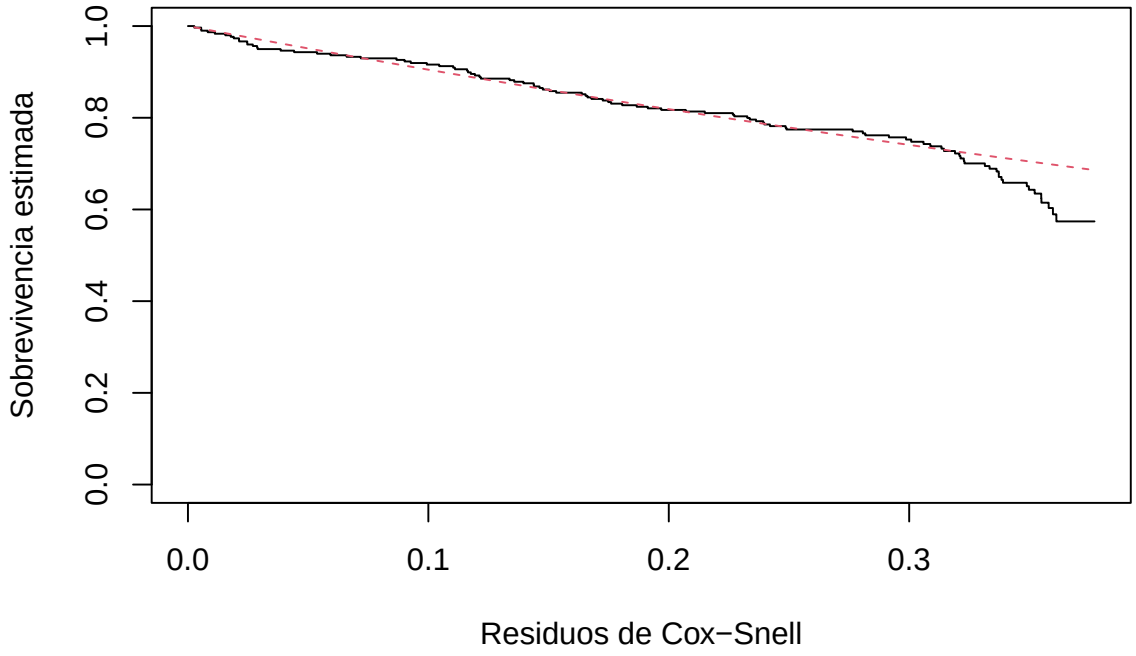
3.4.2 Modelo Stepwise



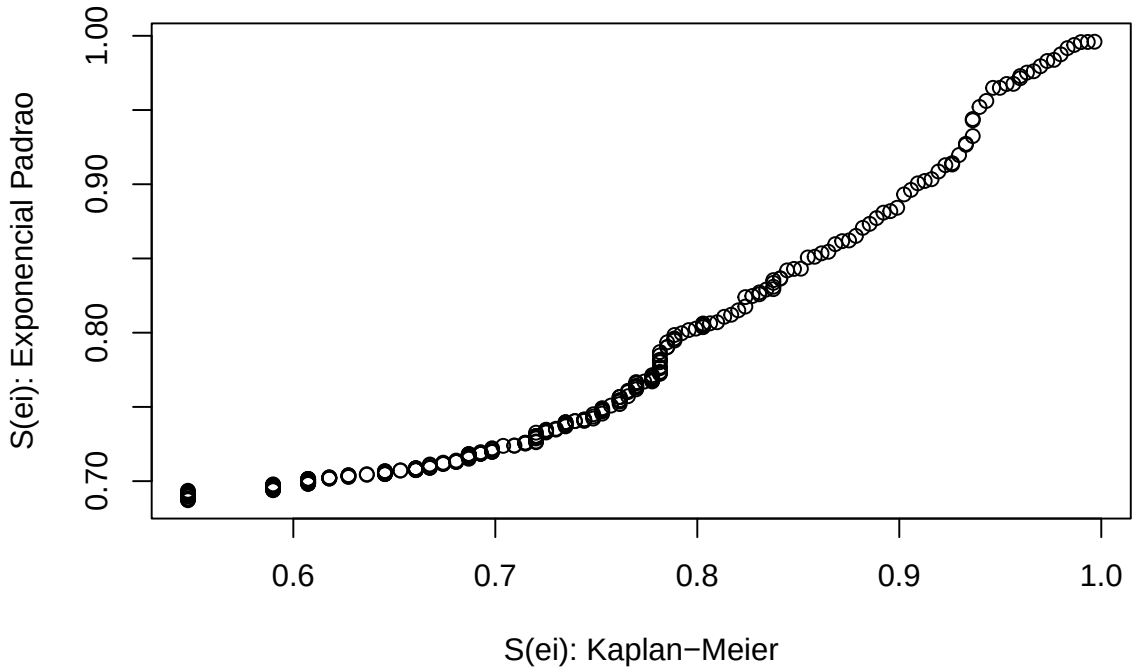


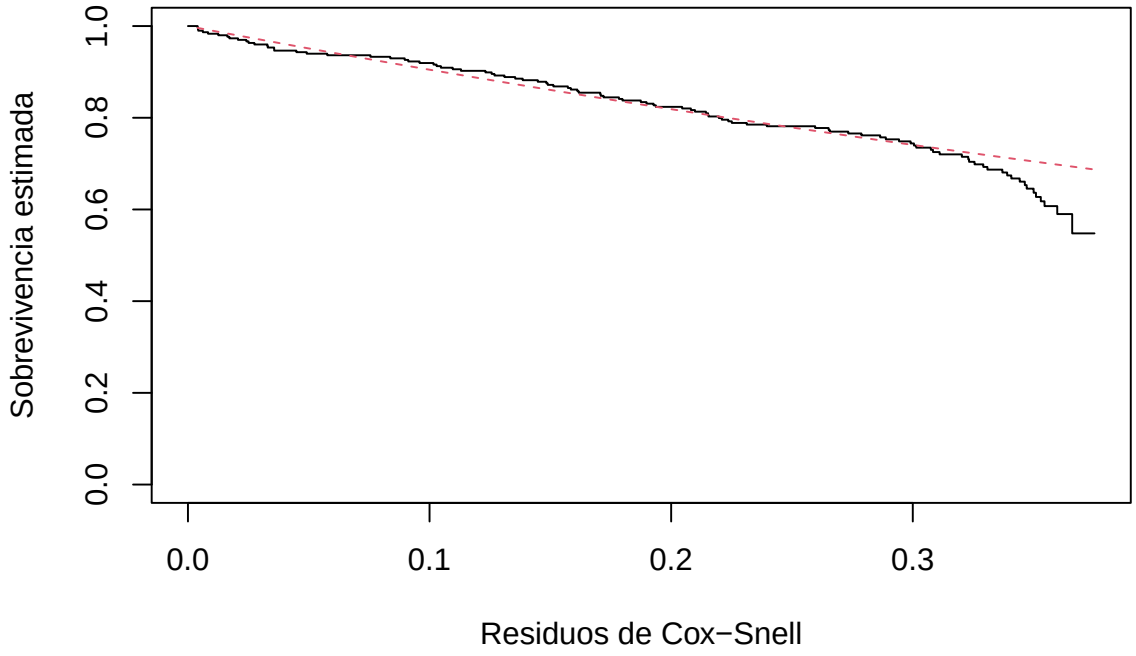
3.4.3 Modelo Manual com fracao de cura





3.4.4 Modelo Stepwise com fracao de cura





3.5 Comparando os modelos finais

```
## Valores modelo manual: 1286.829 1286.965 1305.331

## Valores modelo stepwise: 1285.48 1285.685 1307.683

## Valores modelo manual com fracao de cura: 1384.954 1384.995 1392.355

## Valores modelo stepwise com fracao de cura: 1383.028 1383.069 1390.429
```

A avaliação do gráfico de resíduos de Cox-Snell sugere que, em geral, os modelos estão bem ajustados. No entanto, ao observarmos o gráfico de Kaplan-Meier, fica evidente que seria mais apropriado utilizar uma fração de cura, já que a estimativa não atinge o valor 0. Notamos que os modelos ajustados com a fração de cura conseguem seguir de perto a curva esperada dos resíduos, divergindo apenas no final. Isso nos indica que o mais adequado seria utilizar os modelos com fração de cura. No entanto, temos dois modelos, sendo a única diferença a variável `high_blood_pressure`. Conforme observado pelo Kaplan-Meier, essa variável apresenta diferenças nas curvas de sobrevivência. Portanto, o modelo final escolhido é o modelo stepwise com fração de cura por ter essa variavel.

4 Conclusão

Ao interpretar o modelo final, observamos que os coeficientes das variáveis `high_blood_pressure`, `ejection_fraction` e `age` contribuem para a redução da chance de sobrevivência do indivíduo, ou seja, aumentam a probabilidade de ocorrência de uma falha cardíaca. Por outro lado, o coeficiente da variável `serum_creatinine` auxilia na diminuição da probabilidade de falha, aumentando assim a chance de sobrevivência.

5 Referencia Bibliografica

COLOSIMO, E. A.; GIOLO, S. R. Análise de Sobrevivência Aplicada. Edgard Blucher, São Paulo, 2006.

CORDEIRO, G. M.; ORTEGA, E. M. M.; SILVA, G. O. Modelos de Regressão Estendidos em Análise de Sobrevivência. XII Escola de Modelos de Regressão, 2011.

KALBFLEISH, J. D.; PRENTICE, R. L. The Statistical Analysis of Failure Time Data. John Wiley & Sons, New York, 2002.

LAI, C. D.; XIE, M.; MURTHY, D. N. P. Modified Weibull model. IEEE Transactions on Reliability, v. 52, p. 33-37, 2003.

LAWLESS, J. F. Statistical Methods and Models for Lifetime Data. John Wiley & Sons, New York, 1982.