Especialização em *Data Science* e Estatística Aplicada

Módulo IV - Análise de Sobrevivência

Prof. Dr. Eder Angelo Milani

Goiânia, 2025

IME
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
FEN
FACULDADE DE ENFERMAGEM
UFG
UNIVERSIDADE FEDERAL DE GOIÁS

Conteúdo Programático

- Conceitos básicos. *(Aula 1)*
- Técnicas não-paramétricas. *(Aula 1)*
- Modelos probabilísticos em análise de sobrevivência. *(Aula 2)*
- Modelos de regressão paramétrico. *(Aula 2)*
- Modelo semiparamétrico de riscos proporcionais de Cox.
- Métodos para verificação do modelo ajustado.
- Modelo de Cox estratificado.

Conteúdo - Aula 3

1. Modelo semiparamétrico de riscos proporcionais de Cox
   - Introdução
   - Modelo de Cox
   - Interpretação dos coeficientes

2. Adequação do modelo de Cox
   - Avaliação da proporcionalidade
   - Avaliação da qualidade geral do modelo ajustado
   - Resíduos martingale
   - Probabilidade de concordância
   - Teste de hipóteses

Introdução

O modelo de regressão de Cox (Cox, 1972) abriu uma nova fase na modelagem de dados clínicos.

Stigler (1994) apresenta que entre 1987 e 1989 o artigo Cox (1972) foi o segundo artigo mais citado na literatura estatística, somente ultrapassado pelo artigo de Kaplan-Meier (1958). Em números, média de 600 citações por ano, o que representa aproximadamente 25% das citações anuais ao *Journal of the Royal Statistical Society B*, a revista que publicou o artigo.

Introdução

Atualmente, o artigo tem mais de 63 mil citações (informação obtida do site do Google Acadêmico em 08/03/2025).

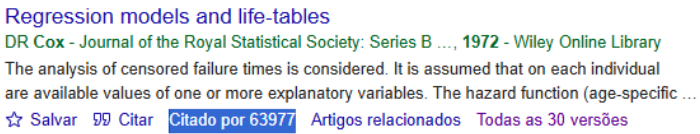

Figura 1: Imagem do Google Acadêmico retirada em 08/03/2025

Modelo de Cox

O modelo de regressão de Cox permite a análise de dados provenientes de estudos de tempo de vida em que a resposta é o tempo até a ocorrência de um evento de interesse, ajustando covariáveis.

No caso especial em que a única covariável é um indicador de grupos, o modelo de Cox assume a sua forma mais simples. Este caso é apresentado a seguir, para introduzir a forma do modelo de Cox.

Suponha um estudo controlado que consiste na **comparação dos tempos de falha de dois grupos em que os pacientes são selecionados aleatoriamente para receber o tratamento padrão (grupo 0) ou o novo tratamento (grupo 1)**.

Modelo de Cox

Representando a função de taxa de falha do primeiro grupo por $\lambda_0(t)$ e a do segundo grupo por $\lambda_1(t)$ e, assumindo proporcionalidade entre estas funções, tem-se que

$$\frac{\lambda_1(t)}{\lambda_0(t)} = K,$$

em que K é a razão das taxas de falha, constante para todo tempo t de acompanhamento do estudo. Se x é a variável indicadora de grupo, em que

$$\begin{cases} 0, & \text{se grupo 0;} \\ 1, & \text{se grupo 1,} \end{cases}$$

e $K = \exp(x\beta)$, então

$$\lambda(t|x) = \lambda_0(t)\exp(x\beta),$$

Modelo de Cox

Ou seja,

$$\lambda(t|x) = \begin{cases} \lambda_1(t) = \lambda_0(t)\exp(x\beta), & \text{se } x = 1; \\ \lambda_0(t), & \text{se } x = 0, \end{cases}$$

A expressão

$$\lambda(t|x) = \lambda_0(t)\exp(x\beta),$$

defini o modelo de Cox para uma única covariável.

Modelo de Cox

De forma genérica, considere p covariáveis, de modo que $\boldsymbol{x}$ seja um vetor com os componentes $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)'$. **A expressão geral do modelo de regressão de Cox considera**

$$\lambda(t|\boldsymbol{x}) = \lambda_0(t) g(\boldsymbol{x}'\boldsymbol{\beta}),$$

em que $g(\boldsymbol{x}'\boldsymbol{\beta})$ é uma função não-negativa que deve ser especificada, tal que $g(0) = 1$.

Este modelo é composto pelo produto de dois componentes, um não-paramétrico e outro paramétrico. O componente não-paramétrico, $\lambda_0(t)$, não é especificado e é uma função não-negativa do tempo. Ele é usualmente denominado função de taxa de falha de base, pois $\lambda(t|\boldsymbol{x}) = \lambda_0(t)$, quando $\boldsymbol{x} = 0$.

Modelo de Cox

O componente paramétrico é frequentemente utilizado na seguinte forma multiplicativa

$$g(\boldsymbol{x}'\boldsymbol{\beta}) = \exp(\boldsymbol{x}'\boldsymbol{\beta}) = \exp(\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p),$$

em que $\boldsymbol{\beta}$ é o vetor de parâmetros associados às covariáveis. Esta forma garante que $\lambda(t|\boldsymbol{x})$ seja sempre não-negativa.

Observe que a constante β_0, presente nos modelos paramétricos, não aparece no componente mostrado acima, isto ocorrer devido à presença do componente não-paramétrico no modelo que absorve este termo constante.

Modelo de Cox

Este modelo é também denominado modelo de taxas de falha proporcionais (do inglês, *hazard ratio***)**, pois a razão das taxas de dois indivíduos diferentes é constante no tempo. Isto é, a razão das funções de taxa de falha para os indivíduos i e j é dada por

$$\frac{\lambda(t|\boldsymbol{x}_i)}{\lambda(t|\boldsymbol{x}_j)} = \frac{\lambda_0(t)\exp(\boldsymbol{x}_i{}'\boldsymbol{\beta})}{\lambda_0(t)\exp(\boldsymbol{x}_j{}'\boldsymbol{\beta})} = \exp(\boldsymbol{x}_i{}'\boldsymbol{\beta} - \boldsymbol{x}_j{}'\boldsymbol{\beta}) = \exp[(\boldsymbol{x}_i{}' - \boldsymbol{x}_j{}')\boldsymbol{\beta}]$$

e não depender do tempo.

Por exemplo, se um indivíduo no início do estudo tem uma taxa de falha igual a duas vezes a de um segundo indivíduo, então, esta razão de taxas de falha será a mesma para todo o período de acompanhamento.

Modelo de Cox

A suposição básica para o uso do modelo de regressão de Cox é, portanto, que as taxas de falha sejam proporcionais ou, de forma equivalente, que as taxas de falha acumulada sejam também proporcionais.

Modelo de Cox

A Figura 2 apresenta uma situação que o uso desse modelo é inadequado. Esta Figura mostra as curvas das taxas de falha acumulada para dois grupos na escala logarítmica.

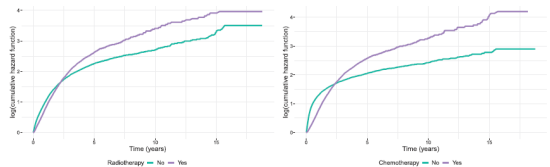

Figura 2: Gráfico do logaritmo do risco acumulado versus o tempo de acompanhamento, para as variáveis radioterapia (esquerda) e quimioterapia (direita)

Fonte: Gazon et *al.*, 2021

Modelo de Cox

Note que o grupo "*No*" tem uma taxa de mortalidade acumulada mais alta no início do acompanhamento. Esta taxa fica, contudo, menor do que a taxa acumulada do grupo "*Yes*" no restante do tempo.

Neste caso, as taxas de falha não são proporcionais e, portanto, violam a suposição básica do modelo. As curvas seriam proporcionais se elas mantivessem uma diferença constante ao longo do período de acompanhamento em uma escala logarítmica.

Modelo de Cox

O modelo de regressão de Cox é utilizado extensivamente em estudos médicos. A principal razão desta popularidade é a presença do componente não-paramétrico, que torna o modelo bastante flexível.

A partir da expressão do modelo de Cox,

$$\lambda(t|\boldsymbol{x}) = \lambda_0(t)\exp(\boldsymbol{x}'\boldsymbol{\beta}),$$

pode-se observar que o **efeito das covariáveis é de acelerar ou desacelerar a função de taxa de falha. Se $x'\beta > 0$ tem-se uma aceleração, enquanto que $x'\beta < 0$, tem-se então uma desaceleração da taxa de falha basal.**

Interpretação dos coeficientes

A propriedade de taxas de falha proporcional do modelo é geralmente utilizada para interpretar os coeficientes estimados.

Tomando-se a razão das taxas de falha de dois indivíduos, i e j, que têm os mesmos valores para as covariáveis com exceção da l-ésima, tem-se

$$\frac{\lambda(t|\boldsymbol{x}_i)}{\lambda(t|\boldsymbol{x}_j)} = \exp((x_{il} - x_{jl})\beta_l),$$

que **pode ser interpretado como a razão de taxas de falha instantânea no tempo t. Entretanto, como esta razão é constante para todo o acompanhamento, pode-se suprimir a palavra instantânea da interpretação**.

Interpretação dos coeficientes

Por exemplo, suponha que x_l seja uma covariável dicotômica indicando pacientes hipertensos, ou seja, se $x_l = 1$ indica paciente hipertenso e $x_l = 0$ indica paciente com pressão normal. Então, a taxa de morte entre os hipertensos é $\exp(\beta_l)$ vezes a de pacientes com pressão normal, mantida fixa as outras covariáveis.

Curiosidade: Estimativa pontual para $\exp(\beta_l)$ pode ser obtida utilizando-se a propriedade de invariância do estimador de máxima verossimilhança. Para obtenção da estimativa intervalar, é necessário obter uma estimativa do erro-padrão de $\exp(\beta_l)$. Isto pode ser feito utilizando o método delta. **Geralmente, os softwares utilizados para a modelagem já trazem essas informações, como é o caso do software R**.

Interpretação dos coeficientes

Importante!
Se o valor 1 pertence ao intervalo estimado para $\exp(\beta_l)$, indica não haver evidências de que as taxas de falha dos pacientes hipertensos e com pressão normal apresentam diferenças significativas.

Interpretação dos coeficientes

Considere que a covariável grupo com três níveis (0 se controle, 1 se grupo 1 e 2 se grupo 2) seja representada por duas covariáveis indicadoras com o grupo controle como referência.

O termo referente a esta covariável no modelo é $x_1\beta_1 + x_2\beta_2$, em que x_1 é o indicador do grupo 1 e x_2 é o indicador do grupo 2.

As estimativas pontuais de máxima verossimilhança e por intervalo são: $\exp(\hat{\beta}_1) = 2,0\ (1,5; 2,5)$ e $\exp(\hat{\beta}_2) = 1,2\ (0,7; 1,8)$.

Interpretação dos coeficientes

Neste caso, existe diferença significativa entre os grupos controle e 1, mas não existe entre os grupos controle e 2. A interpretação é a seguinte: a taxa de morte para os pacientes do grupo 1 é duas vezes a dos pacientes do grupo controle com um intervalo de 95% de confiança de (1,5; 2,5).

Uma interpretação similar é obtida para covariáveis contínuas. Por exemplo, se o efeito de idade for significativo e $\exp(\hat{\beta}) = 1,05$ para este termo, tem-se que, ao aumentarmos em 1 ano a idade, a taxa de morte fica aumentada em 5%.

Aplicação

Aplicação no *software* **R.**

Adequação do modelo de Cox

O modelo de regressão de Cox é bastante flexível devido à presença do componente não-paramétrico. Mesmo assim, ele não se ajusta a qualquer situação clínica e, como qualquer outro modelo estatístico, requer o uso de técnicas para avaliar a sua adequação.

Como mencionado anteriormente, **o modelo de Cox tem uma suposição básica que é a de taxas de falha proporcionais. A violação desta suposição pode acarretar séries vícios na estimação dos coeficientes do modelo**. Sendo assim, vamos iniciar a verificação da adequação do ajuste do modelo de Cox avaliando a suposição de riscos proporcionais.

Avaliação da proporcionalidade

Para avaliar a suposição de taxas de falha proporcionais no modelo de Cox, algumas técnicas gráficas e testes estatísticos encontram-se propostos na literatura.

A seguir vamos apresentar uma técnica gráfica e o teste de Schoenfeld.

Método gráfico descritivo

Este método consiste em construir um gráfico da seguinte forma

- Dividir os dados em m estratos, usualmente de acordo com alguma covariável, por exemplo, dividir os dados em dois estratos de acordo com a covariável sexo;
- Estimar $\hat{\Lambda}_{0j}(t)$ para cada estrato;
- Construir o gráfico logaritmo de $\hat{\Lambda}_{0j}(t)$ *versus* t, ou $\log(t)$.

Se a suposição de proporcionalidade for válida o gráfico deve apresentar diferenças aproximadamente constantes ao longo do tempo.

Curvas não paralelas significam desvios da suposição de taxas de falha proporcionais.

Método gráfico descritivo

É razoável construir este gráfico para cada covariável incluída no estudo.

Se a covariável for de natureza contínua, uma sugestão é agrupá-la em um pequeno número de categorias.

Uma vantagem dessa técnica gráfica é a de indicar a covariável que estaria gerando a violação da suposição, caso isto ocorra.

Uma desvantagem é que a **conclusão sobre a proporcionalidade das taxas de falha é subjetiva**, pois depende da interpretação dos gráficos.

Método gráfico descritivo

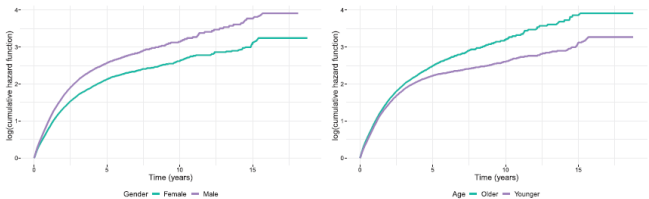

Figura 3: Gráfico do logaritmo do risco acumulado versus o tempo de acompanhamento, para as variáveis sexo (esquerda) e idade (direita)

Fonte: Gazon et *al.*, 2021

Método gráfico descritivo

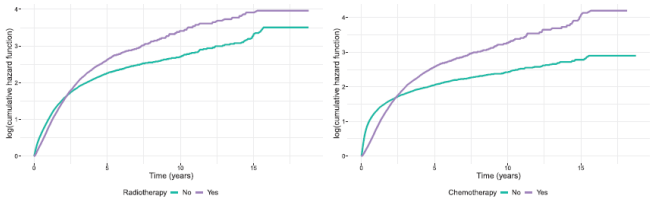

Figura 4: Gráfico do logaritmo do risco acumulado versus o tempo de acompanhamento, para as variáveis radioterapia (esquerda) e quimioterapia (direita)

Fonte: Gazon et *al.*, 2021

Resíduos de Schoenfeld

Como responder à seguinte pergunta: **o efeito - o risco relativo ou** *hazard ratio* **- de uma covariável é sempre o mesmo durante todo o tempo de observação? Caso não seja, o efeito da covariável é tempo-dependente**.

Para investigar a proporcionalidade de cada covariável k utiliza-se os resíduos de Schoenfeld definidos para cada indivíduo i:

$$r_{ik} = \delta_i(x_{ik} - a_{ik}),$$

em que δ_i é o indicador de ocorrência de evento no indivíduo i, e por isso quando ocorre censura o resíduo é nulo. Defini-se a_{ik} como uma média ponderada dos valores das covariáveis dos indivíduos em risco no tempo t:

$$a_{ik} = \frac{\sum_{j \in R(t_i)} x_{jk} \exp(\boldsymbol{x_j}\hat{\boldsymbol{\beta}})}{\sum_{j \in R(t_i)} \exp(\boldsymbol{x_j}\hat{\boldsymbol{\beta}})}.$$

Resíduos de Schoenfeld

Observe que $R(t_i)$ é o conjunto de indivíduos em risco no tempo t_i e x_{jk} representa o valor da covariável k do indivíduo j pertencente ao grupo de risco.

Pode-se dizer que esse resíduo é a diferença entre o valor da covariável x_{ik} e essa média ponderada dos valores das covariáveis dos indivíduos em risco naquele instante.

Cabe ressaltar que haverá tantos vetores de resíduos quanto covariáveis ajustadas no modelo, e que esses são definidos somente nos tempos j em que ocorreu um evento.

Resíduos de Schoenfeld

Demonstra-se que o valor esperado desse resíduo padronizado é aproximadamente igual à parte de β que varia no tempo. Assim, o gráfico dos resíduos padronizados de Schoenfeld contra os tempos de sobrevivência permite verificar se estes estão distribuídos igualmente ao longo do tempo. Em outras palavras, se a suposição de riscos proporcionais for satisfeita não deverá existir tendência sistemática no gráfico.

A interpretação pode ser facilitada adicionando-se ao gráfico uma linha que permite melhor visualização da tendência, tal como o *spline* - função não paramétrica de suavização. Para essa curva pode-se estimar o intervalo de confiança, de forma a verificar se as oscilações estão ou não significativamente afastadas do valor do coeficiente em cada período de tempo.

Resíduos de Schoenfeld

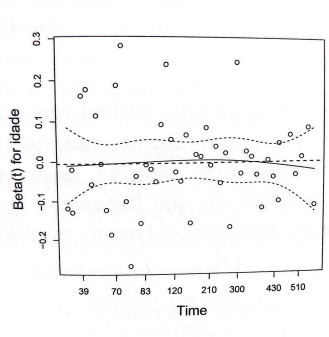

Figura 5: Exemplo do gráfico dos resíduos de Schoenfeld

Fonte: Carvalho et *al.*, 2011

Teste de hipóteses

Além da análise gráfica, pode-se testar a presença de correlação linear entre o tempo de sobrevivência e o resíduo. Sob a hipótese nula , de correlação igual a zero, tem-se que a distribuição do teste é uma qui-quadrado com um grau de liberdade.

Se a hipótese nula não é rejeitada, a premissa de proporcionalidade dos riscos não é rejeitada.

Importante!!!

O que fazer quanto os riscos não são proporcionais?

Antes de mais nada, deve-se verificar se a não proporcionalidade encontrada é realmente importante, pois a variação em $\hat{\beta}_k(t)$ pode ser pequena em relação ao parâmetro estimado $\hat{\beta}_k$. A segunda questão é verificar se a não proporcionalidade é real, pois alguns poucos pontos *outliers* podem estar influenciando a significância do teste.

Caso a não proporcionalidade seja importante, uma estratégia bastante utilizada é estratificar o modelo pela respectiva covariável, veremos como fazer isso na próxima aula. Uma outra alternativa é utilizar outro tipo de modelo, existem modelos próprios para analisar riscos não proporcionais, como os apresentados na Aula 2.

Avaliação da qualidade geral do modelo ajustado

Assim como nos modelos paramétricos, os resíduos de Cox-Snell (1968) também são utilizados com o propósito de avaliar a qualidade geral de ajuste do modelo de Cox.

Para este modelo, os resíduos de Cox-Snell são definidos por

$$\hat{e}_i = \hat{\Lambda}_0(t_i) \exp\left(\sum_{k=1}^{p} x_{ik}\hat{\beta}_k\right),$$

em que $i = 1, \ldots, n$.

Se o modelo estiver bem ajustado, os $\hat{e}_i$ devem ser olhados como uma amostra censurada de uma distribuição exponencial padrão.

Avaliação da qualidade geral do modelo ajustado

A análise gráfica desses resíduos não fornece, contudo, informações sobre o tipo de problema que estaria ocorrendo caso o ajuste não se apresentar satisfatório.

Gráficos envolvendo esses resíduos não são recomendados para avaliação da suposição de taxas de falha proporcionais.

Os mesmos comentários feitos anteriormente sobre os resíduos de Cox-Snell quanto aos cuidados e desvantagens de sua utilização, são também válidos para o modelo de Cox.

Resíduos martingale

Os resíduos martingale, M_i, são baseados no processo de contagem individual e definido por:

$$M_i = N_i - E_i,$$

tal que N_i é igual ao número de eventos observados no intervalo $[0, \infty)$ e E_i é o número de eventos esperados sob o modelo ajustado no intervalo $[0, \infty)$. **Estima-se o resíduo martigale como:**

$$\hat{M}_i = \delta_i - \hat{\Lambda}_0(t_i) \exp(\boldsymbol{x'}\hat{\boldsymbol{\beta}}) = \delta_i - rc_i,$$

em que rc_i é o resíduo de Cox-Snell.

Resíduos martingale

São propriedades dos resíduos martigale, algumas das quais semelhantes aos resíduos dos modelos de regressão linear:

(i) o valor esperado de M_i é 0, quando avaliado no valor verdadeiro (e desconhecido) do vetor de parâmetros $\boldsymbol{\beta}$;

(ii) os resíduos M_i não são simetricamente distribuídos em torno de 0, variando de $(-\infty, 1]$ e quando o tempo de sobrevivência é censurado o resíduo é negativo;

(iii) o somatório dos resíduos observados baseados no valor estimado de $\boldsymbol{\beta}$ é igual a 0; e

(iv) os resíduos M_i calculados usando o verdadeiro vetor de parâmetros $\boldsymbol{\beta}$ são não correlacionados, mas as estimativas $\hat{M}_i$ são negativamente correlacionadas, ainda que fracamente.

Resíduos martingale

Apesar da semelhança com a decomposição dos dados utilizada em modelos de regressão linear (resíduo = valor observado - valor esperado), os resíduos martingale possuem propriedades diferentes dos resíduos dos modelos lineares, são elas:

(i) **a soma de quadrados dos resíduos não auxilia na avaliação do ajuste global do modelo**, ou seja, o melhor modelo de Cox não necessariamente apresenta a menor soma dos quadrados dos resíduos martingale;

(ii) em modelos lineares assume-se que a distribuição dos resíduos deve ser aproximadamente normal, o que também não acontece. E nem os resíduos seguem uma distribuição exponencial, como poderia se supor devido à forma do modelo de Cox.

Resíduos martingale

Resíduos martingale

Os resíduos martingale são úteis na avaliação da qualidade de ajuste do modelo em duas situações importantes, observadas graficamente:

- M_i contra o índice do indivíduo - permite revelar indivíduos mal ajustados pelo modelo (pontos aberrantes ou *outliers*);
- M_i do modelo nulo (sem covariáveis) contra covariável com a superposição de uma curva de suavização - sugere a forma funcional de uma covariável contínua.

Resíduos martingale

O primeiro tipo de uso do resíduo martingale serve para apontar possíveis valores aberrantes no estudo, ou seja, indivíduos que demoram muito tempo para sofrer o evento ou que sofrem o evento muito rapidamente, dadas as covariáveis. Valores de M_i positivos indicam que o número de eventos observados é maior que o estimado pelo modelo e vice-versa.

O resíduo martingale do modelo nulo permite explorar a forma funcional de uma covariável contínua, isto é, se é linear ou se alguma transformação é necessária, por exemplo, fazendo uso da raiz quadrada ou do logatimo da covariável, ou ainda se uma forma não paramétrica deve ser usada.

Resíduos martingale

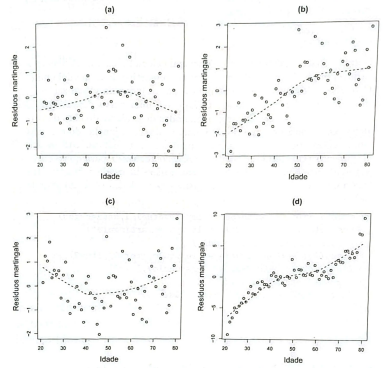

Figura 6: Resíduo do modelo nulo contra covariável para investigação da forma funcional

Fonte: Carvalho et *al.*, 2011

Resíduos martingale

Da Figura 6, tem-se que

- da figura (a), um exemplo de caso em que **não há associação entre o tempo de sobrevivência e a idade**;
- da figura (b), **existe uma relação e ela é linear**, isto é, a covariável não precisa de nenhuma transformação;
- da figura (c), **a relação não é linear, parece ser quadrática, e a raiz quadrada da variável poderia linearizar a relação**;
- da figura (d), **a relação não parece linear**, e neste caso não existe uma simples transformação de linearização, sugere-se neste caso o **uso de função não paramétrica de suavização**.

Probabilidade de concordância

Uma medida global de ajuste útil quando o objetivo do estudo é obter um modelo preditivo, é a estimativa da probabilidade de concordância. Mais especificamente, essa medida é usada para avaliar o poder discriminatório e a acurácia preditiva do modelo de Cox.

Concordância significa que, ao selecionar aleatoriamente duas observações, **a que possui menor tempo de sobrevivência é também aquela que possui o maior risco estimado (preditivo) pelo modelo de Cox**.

Probabilidade de concordância

Como regra geral, pode-se considerar que se a probabilidade de concordância estimada pelo modelo de Cox:

- estiver entre 0,3 e 0,4 tem-se que o modelo possui um baixo poder preditivo (ou descriminatório);
- for a 0,5 significa que a concordância pode ser por acaso;
- estiver entre 0,6 e 0,7, tem-se um resultado comum para na análise de dados de sobrevivência;
- estiver entre 0,7 e 0,8, tem-se um resultado discriminatório muito bom;
- estiver entre 0,8 e 0,9, tem-se um resultado excelente.

Teste da razão de verossimilhanças

O teste da razão de verossimilhanças compara modelos aninhados avaliando se a inclusão de uma ou mais covariáveis no modelo aumenta de modo significativo a verossimilhança de um modelo em relação ao modelo mais parcimonioso. A estatística do teste é definida pela razão das verossimilhanças ou pela diferença entre os logaritmos da função de verossimilhança da seguinte forma:

$$RV = 2 \times (l_{\text{maior}} - l_{\text{menor}}),$$

que, sob a **hipótese nula (H_0) de que não há diferença entre os modelos**, segue uma distribuição χ^2 com graus de liberdade iguais à diferença no número de covariáveis dos modelos em questão.

Outros testes de hipóteses também são utilizados na prática, tais como de Wald e o Score.

Aplicação

Aplicação no *software* **R.**

Especialização em *Data Science* e Estatística Aplicada

Módulo IV - Análise de Sobrevivência

Prof. Dr. Eder Angelo Milani

edermilani@ufg.br

IME INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

FEN FACULDADE DE ENFERMAGEM

UFG UNIVERSIDADE FEDERAL DE GOIÁS