Modelo de regressão logística e Modelos de risco proporcionais (Cox)

Modelos prognósticos

Modelos prognósticos

Prognóstico significa prever, predizer ou estimar a probabilidade ou risco de **condições futuras**.

Na área da saúde, prognóstico comumente relaciona-se à probabilidade ou risco de um **indivíduo** desenvolver um particular estado de saúde (um desfecho) ao longo de um período de tempo específico, baseado na presença ou ausência de um perfil clínico.

Regressão linear, **logística** e **de risco** (análise de sobrevida) são métodos comuns utilizados na pesquisa clínica para relatar covariáveis e desfechos.

Modelos prognósticos

A **regressão linear** é o método padrão para **desfechos contínuos**.

A **regressão logística** é adequada para **desfechos binários**

Quando desfechos binários são medidos **prospectivamente**, eles também são associados com um tempo para o evento. Neste caso, a **regressão logística** ou a **análise de sobrevida** podem ser adequadas.

Modelos de regressão linear

Na regressão linear, ajustamos um modelo do tipo

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_i X_i + \varepsilon$$

- **Pressuposto importante:** a variável Y era de natureza **contínua** e seguia uma distribuição normal.

O modelo se preocupava em estimar (ou predizer) o valor médio de Y dado um certo conjunto de valores das variáveis explicativas.

Modelos de regressão linear

Em um modelo de regressão linear, as variáveis explicativas podem ser **contínuas** ou **dicotômicas**.

O **coeficiente** para uma covariável contínua depende, em grande parte, das unidades, e a **magnitude** só pode ser interpretada em um **contexto clínico**.

Se um pesquisador quer saber a **mudança esperada na resposta** para um aumento de 10 unidades (ou qualquer outro aumento) em uma covariável contínua, ele pode simplesmente multiplicar o coeficiente (e os limites de confiança) por 10.

Para covariáveis **dicotômicas**, o coeficiente é interpretado como a **diferença na resposta** que seria vista, em média, entre os dois níveis da covariável.

Exemplo

Descrição da base de dados

O NHANES (National Health and Nutrition Examination Survey) é um estudo populacional dos EUA que contém dados de saúde.

Vamos estimar um modelo de regressão linear múltiplo usando os dados do NHANES, incluindo covariáveis clínicas e demográficas.

Descrição da base de dados

Usaremos a pressão arterial sistólica (PAS) como variável dependente e as seguintes variáveis independentes:

- IMC (quanto maior o IMC, maior pode ser a pressão arterial?)
- Idade (o envelhecimento está associado ao aumento da PAS?)
- Sexo (homens têm maior PAS que mulheres?)
- Horas de sono à noite (SleepHrsNight) (dormir mais pode diminuir a PAS?)
- Nível de colesterol total (TotChol) (colesterol elevado está associado a maior PAS?)
- Tabagismo (SmokeNow) (fumantes têm PAS mais alta?)

Preparação

Carregando os pacotes R

```
pacman::p_load(
NHANES,
tidyverse,
car,
lmtest,
mfx
  )
```

Carregando os dados

```
# Selecionar variáveis relevantes e remover NAs
dados <- NHANES %>%
  filter(!is.na(BPSysAve) & !is.na(BMI) & !is.na(Age) & !is
           !is.na(TotChol) & !is.na(SleepHrsNight) & !is.na
  dplyr::select(BPSysAve, BMI, Age, Gender, TotChol,SleepHr
```

Manipulação dos dados

```
# Renomear variáveis para facilitar
colnames(dados) <- c("PAS", "IMC", "Idade", "Sexo", "Colest

# Transformar Sexo e Fumante em fatores
dados$Sexo <- as.factor(dados$Sexo)
dados$Fumante <- as.factor(dados$Fumante)

# Estrutura dos dados

dados %>%
  str()
```

```
tibble [2,940 x 7] (S3: tbl_df/tbl/data.frame)
 $ PAS              : int [1:2940] 113 113 113 112 111 127 1
 $ IMC              : num [1:2940] 32.2 32.2 32.2 30.6 23.7
 $ Idade            : int [1:2940] 34 34 34 49 66 58 33 60 5
 $ Sexo             : Factor w/ 2 levels "female","male": 2
 $ Colesterol       : num [1:2940] 3.49 3.49 3.49 6.7 4.99 4
 $ HorasDeSonoNoite: int [1:2940] 4 4 4 8 7 5 6 6 8 4
```

Ajuste do modelo

```r
# Ajustar modelo de regressão linear múltiplo
modelo_multiplo <- lm(PAS ~ IMC + Idade + Sexo + Colesterol

# Resumo do modelo
summary(modelo_multiplo)
```

```

Call:
lm(formula = PAS ~ IMC + Idade + Sexo + Colesterol + HorasI
    Fumante, data = dados)

Residuals:
    Min       1Q   Median       3Q      Max 
-56.440  -9.003  -1.070   8.170  83.997 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      88.82356    2.63293  33.736  < 2e-16 ***
IMC               0.17376    0.04634   3.750 0.000181 ***
```

Interpretação do modelo

Variável	Coeficiente ( )	p-valor	
Intercepto	88,82	$< 0{,}001$	V
IMC	0,17	$< 0{,}001$	A cada aumento de 1 uni
Idade	0,42	$< 0{,}001$	
Sexo (Masculino)	4,91	$< 0{,}001$	O sexo masculin
Colesterol	1,47	$< 0{,}001$	Para c
Horas de sono	-0,42	0,047	Para
Fumante (Sim)	0,64	0,312	Não significativo (p $>$

Variável dependente (desfecho) binária

E se a variável dependente y for **binária**?

- Doença (presente $= 1/$ausente $= 0$)
- Morto $= 1/$Vivo $= 0$

Aqui, $Y = 1$ corresponde ao sucesso, ou seja, a ocorrência do evento e $Y = 0$ corresponde ao fracasso, ou seja, à não ocorrência do evento.

Temos então que a média de Y é igual a p, sendo p a proporção de vezes que Y assume o valor 1. Assim,

$$p = P(Y = 1) = P(\text{sucesso})$$

Variável dependente (desfecho) binária

A regressão logística é um modelo estatístico que permite estimar a probabilidade p da ocorrência de um determinado desfecho categórico (Y) em função de um ou mais preditores (X), que podem ser contínuos ou categóricos.

Vamos a um exemplo...

Variável dependente (desfecho) binária

Considere a população de bebês com baixo peso ao nascer (definido como < 1750g) que estão confinados em uma unidade de tratamento intensivo neonatal, entubados durante as primeiras 12 semanas de vida e sobreviventes por, no mínimo, 28 dias.

Na amostra de 223 bebês extraída da população original, 76 foram diagnosticados com *displasia broncopulmonar* (BPD). Os restantes 147 não tinham a doença.

Seja Y uma variável aleatória dicotômica, de forma que

$$Y = \begin{cases} 0 & \text{Ausência de BPD} \\ 1 & \text{Presença de BPD} \end{cases}$$

Variável dependente (desfecho) binária

A probabilidade estimada de que um bebê dessa população desenvolva BPD é a **proporção amostral** de bebês com BPD, ou seja,

$$\hat{p} = \frac{76}{223} = 0,341$$

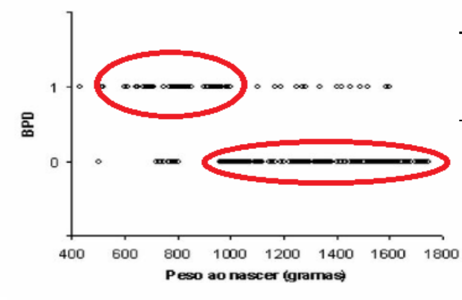
Podemos suspeitar que alguns fatores, maternos ou neonatais, devem afetar a probabilidade de um bebê em particular desenvolver BPD.

O conhecimento da presença ou ausência desses fatores pode:

- aumentar a precisão de nossa estimativa de p,
- desenvolverintervenções para reduzir essa probabilidade.

Variável dependente (desfecho) binária

- Um fator interessante poderia ser o peso de nascimento de um bebê, que chamaremos de X.
- Se a variável Y fosse **contínua**, poderíamos começar a análise contruindo um diagrama de dispersão entre as variáveis X e Y.



Peso(g)	n	BPD	p
0-950	68	49	0,721
951-1350	80	18	0,225
1351-1750	75	9	0,120
	223	76	0,341

Variável dependente (desfecho) binária

- Parece que a probabilidade de desenvolver BPD aumenta à medida que o peso do bebê ao nascer diminui e vice-versa.
- Como parece haver uma relação entre estas duas variáveis, podemos usar o peso ao nascer de uma criança para nos ajudar a prever a probabilidade de que ela desenvolva BPD.

Função logística

A primeira estratégia poderia ser ajustar um modelo da forma

$$p = \beta_0 + \beta_1 x$$

onde x representa o peso ao nascer.

Sob inspeção, este modelo não é viável, uma vez que p é uma probabilidade, podendo assumir, portanto, valores entre 0 e 1.

O termo $\beta_0 + \beta_1 x$, ao contrário, pode assumir valores fora desse intervalo.

Função logística

Uma alternativa seria ajustar o modelo

$$p = e^{\beta_0 + \beta_1 x}$$

Essa expressão garante que a estimativa de p é sempre positiva.

No entanto, este modelo também é inadequado, uma vez que pode produzir valores maiores que 1.

Função logística

Podemos então, ajustar um modelo da forma

$$p = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

Esta expressão, conhecida como **função logística**, não admite valores negativos nem maiores que 1.

Função logística

Lembre-se de que, se um evento ocorre com probabilidade p, a **chance** a seu favor é de $\dfrac{p}{1-p}$ para 1.

Assim, se um **sucesso** ocorre com probabilidade

$$p = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}},$$

Função logística

a **chance** em favor de sucesso é

$$\frac{p}{1-p} = \frac{\dfrac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}}{\dfrac{1}{1+e^{\beta_0+\beta_1 x}}} = e^{\beta_0+\beta_1 x}$$

Tomando o logaritmo natural de cada lado dessa equação,

$$\underbrace{\ln\left(\frac{p}{1-p}\right)}_{\text{logit}} = \ln(e^{\beta_0+\beta_1 x}) = \beta_0 + \beta_1 x$$

Modelo de regressão logística

Modelar uma probabilidade p com uma função logística é **equivalente** a ajustar um modelo de regressão linear onde a variável dependente contínua y foi substituída pelo logaritmo neperiano da **chance** de ocorrência de um evento **dicotômico**.

Em vez de assumir que a relação entre p e X é **linear**, assume-se que a relação entre $\ln\left(\frac{p}{1-p}\right)$ e X é **linear**.

Essa técnica é conhecida como **regressão logística**.

Modelo de regressão logística

Os parâmetros do modelo (β's) são estimados usando o **método de máxima verossimilhança**, que busca maximizar a probabilidade de observar os dados dados os parâmetros.

Este processo envolve iterativamente ajustar os coeficientes para melhor se alinhar com os dados observados.

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{\beta}_0 + \hat{\beta}_1 x$$

Modelo de regressão logística

Voltando ao exemplo, para a amostra de 223 bebês com baixo peso ao nascer, a equação da regressão logística estimada é

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = 4,0343 - 0,0042x$$

- **Interpretação:** O coeficiente do peso indica que, para cada aumento de 1 grama no peso ao nascer, o logaritmo da chance de que um bebê desenvolva BPD diminui de 0,0042, em média.

Modelo de regressão logística

Qual a probabilidade de que um bebê, retirado desta população pesando 750g ao nascer, irá desenvolver BPD?

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = 4,0343 - 0,0042x$$

Trocando-se x por 750, temos

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = 4,0343 - 0,0042 \times 750 = 0,8843$$

Modelo de regressão logística

Aplicando a função exponencial em ambos os lados, temos

$$\begin{aligned}
\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) &= 0,8843 \\
\frac{\hat{p}}{1-\hat{p}} &= e^{0,8843} = 2,4113
\end{aligned}$$

Modelo de regressão logística

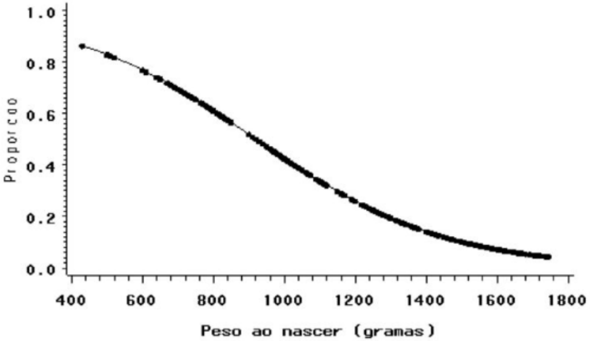
Isolando $\hat{p}$:

$$\begin{aligned}
\frac{\hat{p}}{1-\hat{p}} &= 2,4113 \\
\hat{p} &= 2,4113 - 2,4113\hat{p} \\
\hat{p} + 2,4113\hat{p} &= 2,4113 \\
(1+2,4113)\hat{p} &= 2,4113 \\
\hat{p} &= \frac{2,4113}{1+2,4113} = 0,708
\end{aligned}$$

Modelo de regressão logística

A probabilidade estimada de que uma criança, que pesa 750g ao nascer, desenvolva BPD é de 0,708.

Se calculássemos a probabilidade estimada $\hat{p}$ para cada valor observado dos pesos ao nascer e plotássemos $\hat{p} \times$ o peso, teríamos



Exemplo

Descrição da base de dados

O NHANES (National Health and Nutrition Examination Survey) é um estudo populacional dos EUA que contém dados de saúde.

Vamos estimar um modelo de regressão logística usando os dados do NHANES, incluindo covariáveis clínicas e demográficas.

Descrição da base de dados

Usaremos *Diabetes* como variável dependente e as seguintes variáveis independentes:

- PAS (a pressão arterial sistólica está associada à diabetes?)
- IMC (IMC está associado à diabetes?)
- Idade (o envelhecimento está associado à diabetes?)
- Sexo (homens têm maior risco de ter diabetes que mulheres?)
- Horas de sono à noite (SleepHrsNight) (dormir mais influencia no fato de ter diabetes?)
- Nível de colesterol total (TotChol) (colesterol elevado está associado à diabetes?)
- Tabagismo (SmokeNow) (fumantes têm maior risco de ter diabetes?)

Preparação

Carregando os pacotes R

```
pacman::p_load(
NHANES,
tidyverse,
car,
lmtest
  )
```

Carregando os dados

```
# Selecionar variáveis relevantes e remover NAs
dados <- NHANES %>%
  filter(!is.na(BPSysAve) & !is.na(BMI) & !is.na(Age) & !is
           !is.na(TotChol) & !is.na(SleepHrsNight) & !is.na
  dplyr::select(BPSysAve, BMI, Age, Gender, TotChol,SleepHr
```

Manipulação dos dados

```
# Renomear variáveis para facilitar
colnames(dados) <- c("PAS", "IMC", "Idade", "Sexo", "Colest

# Transformar Sexo e Fumante em fatores
dados$Sexo <- as.factor(dados$Sexo)
dados$Fumante <- as.factor(dados$Fumante)
dados$Diabetes <- as.factor(dados$Diabetes)

# Estrutura dos dados

dados %>%
  str()
```

```
tibble [2,938 x 8] (S3: tbl_df/tbl/data.frame)
 $ PAS          : int [1:2938] 113 113 113 112 111 127 1
 $ IMC          : num [1:2938] 32.2 32.2 32.2 30.6 23.7
 $ Idade        : int [1:2938] 34 34 34 49 66 58 33 60 5
 $ Sexo         : Factor w/ 2 levels "female","male": 2
 $ Colesterol   : num [1:2938] 3.49 3.49 3.49 6.7 4.99 4
```

Ajuste do modelo

```r
# Ajustar modelo de regressão logistica
modelo_logistico <- glm(Diabetes ~ ., data = dados, family

# Resumo do modelo
summary(modelo_logistico)
```

```
Call:
glm(formula = Diabetes ~ ., family = binomial(link = "logit
    data = dados)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -7.746757   0.721204 -10.741  < 2e-16 ***
PAS               0.003845   0.003409   1.128  0.25925
IMC               0.104682   0.009341  11.206  < 2e-16 ***
Idade             0.051445   0.004731  10.874  < 2e-16 ***
Sexomale          0.317999   0.131212   2.424  0.01537 *
Colesterol        0.198636   0.061478   3.231  0.00123 **
```

Avaliando *odds ratio*

```
logitor(Diabetes ~., data=dados)
```

```
Call:
logitor(formula = Diabetes ~ ., data = dados)

Odds Ratio:
                 OddsRatio Std. Err.       z     P>|z|
PAS              1.0038527 0.0034216  1.1282  0.259253
IMC              1.1103572 0.0103724 11.2061 < 2.2e-16 ***
Idade            1.0527909 0.0049805 10.8744 < 2.2e-16 ***
Sexomale         1.3743748 0.1803346  2.4235  0.015370 *
Colesterol       0.8198486 0.0504025 -3.2310  0.001234 **
HorasDeSonoNoite 1.0234136 0.0460907  0.5139  0.607328
FumanteYes       0.8321009 0.1211769 -1.2621  0.206900
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Interpretação do modelo

Variável	OddsRatio	p-valor	
PAS	1,004	0,259	Não significativo (p-v
IMC	1,110	$< 0{,}001$	A cada aumento de 1 unid
Idade	1,05	$< 0{,}001$	A cada ano a mais d
Sexo (Masculino)	1,37	$< 0{,}015$	A char
Colesterol	0,82	0,001	Para cada aumento de 1 mg/
Horas de sono	1,02	0,607	Não significativo ($p > 0.05$),
Fumante (Sim)	0,83	0,207	Não significativo ($p > 0.0$

Modelo de riscos proporcionais (Cox)

Análise de sobrevivência

Quando desfechos são associados com **tempo até o evento**, não estamos limitados a estudar um **ponto específico no tempo**, como é feito em estudos transversais.

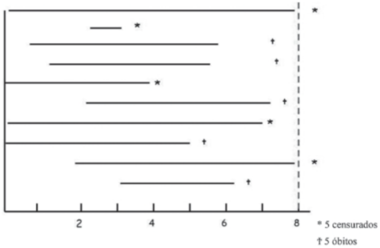
Em vez disso, podemos estar interessados em saber se a probabilidade do evento tende a ser maior (ou menor) ao **longo de todo período de acompanhamento**.

A **análise de sobrevivência** é utilizada para responder essa questão mais ampla.

Em estudos de análise de sobrevivência, o problema-chave é o dado considerado **censurado**, ou seja, quando o evento (desfecho) **não ocorreu** quer seja porque o estudo terminou antes da ocorrência do desfecho ou porque se perdeu o acompanhamento do caso.

Análise de sobrevivência

Dados censurados



Dados censurados à direita: maioria

Análise de sobrevivência

Indicada para estudos longitudinais (coorte, ensaios clínicos), ou seja, o mesmo grupo de indivíduos é acompanhado durante um **intervalo de tempo** preestabelecido pelo pesquisador.

A **grande vantagem** nesse tipo de análise é que se permite utilizar informações de **todos** os participantes até o momento da ocorrência do evento ou quando são censurados.

Regressão de riscos

O modelo de sobrevida mais comum é o **modelo de riscos proporcionais de Cox**. Este modelo permite que as covariáveis, numéricas ou categóricas, variem com o tempo.

Assume-se, nesse modelo, que os tempos t_i, $i = 1, \cdots, n$, são independentes e que a taxa de falha (risco) tem a seguinte forma:

$$\lambda(t) = \lambda_0(t) \exp(\beta_1 X_1 + \cdots + \beta_p X_p)$$

O componente não-paramétrico, $\lambda_0(t)$, não é especificado e é uma função não-negativa do tempo. Ele é usualmente chamado de função de base ou risco basal.

Regressão de riscos

O componente paramétrico $\exp(\mathbf{x}^{\mathbf{t}}\ )$ é o nosso interesse, em especial no vetor de parâmetros .

O modelo é conhecido por ter taxas de falhas proporcionais. Este fato é conveniente na sua interpretação, ou seja, a razão das taxas de falha de dois indivíduos diferentes i e j é dada por

$$\frac{\lambda_i(t)}{\lambda_j(t)} = \frac{\exp(\mathbf{x}_{\mathbf{i}}^{\mathbf{t}}\ )}{\exp(\mathbf{x}_{\mathbf{j}}^{\mathbf{t}}\ )} = \exp(\mathbf{x}_{\mathbf{i}}^{\mathbf{t}}\ - \mathbf{x}_{\mathbf{j}}^{\mathbf{t}}\ ),$$

que não depende do tempo.

Assim, se um indivíduo no início do estudo tem um risco de morte igual a duas vezes o risco de um segundo indivíduo, entao esta razão de riscos será a mesma para todo o período de acompanhamento.

Hazard Ratio (HR)

O Hazard Ratio (HR) é uma medida estatística amplamente utilizada em estudos de sobrevivência e análises de risco, especialmente em pesquisas médicas e epidemiológicas.

Em um modelo de Cox, a razão de risco (HR) é usada para medir o efeito de uma variável preditora no risco do evento.

É dada por

$$HR = \exp(\beta_i), i = 1, \cdots, p,$$

sendo β_i os coeficientes estimados do modelo de regressão de cox.

Hazard Ratio (HR)

A HR representa o risco relativo do evento ocorrer para uma dada mudança de unidade na variável preditora, com uma HR maior que 1 indicando um risco aumentado e uma HR menor que 1 indicando um risco diminuído.

Exemplo

Descrição da base de dados

O conjunto de dados contém casos de um estudo que foi conduzido entre 1958 e 1970 no Hospital Billings da Universidade de Chicago sobre a sobrevivência de pacientes que foram submetidos a cirurgia de mama. Os dados incluem:

- Idade do paciente no momento da operação (numérico)
- Ano da operação do paciente (ano - 1900, numérico)
- Número de nódulos axilares positivos detectados (numérico)
- Status de sobrevivência (dicotômica)
  - $1 =$ o paciente sobreviveu 5 anos ou mais
  - $2 =$ o paciente morreu em 5 anos

Preparação

Carregando os pacotes R

```
pacman::p_load(
  rio,
  tidyverse,
  survival,
  survminer
)
```

Carregando os dados

```r
dados = rio::import("https://raw.githubusercontent.com/tiag

dados %>%
  str()
```

```
'data.frame':   306 obs. of  4 variables:
 $ idade  : int  30 30 30 31 31 33 33 34 34 34 ...
 $ anoOper: int  64 62 65 59 65 58 60 59 66 58 ...
 $ numNod : int  1 3 0 2 4 10 0 0 9 30 ...
 $ status : int  1 1 1 1 1 1 1 2 2 1 ...
```

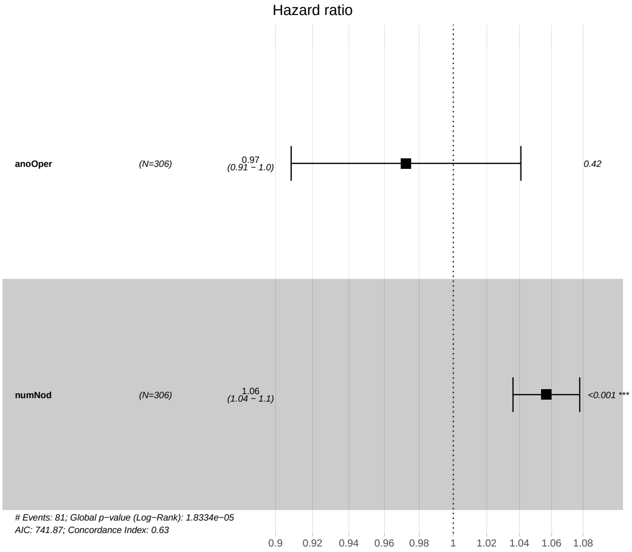
Ajuste do modelo

```r
# Ajustando o modelo de Cox
modelo_cox <- coxph(Surv(idade, status=='2') ~ ., data = da

# Resumo do modelo
summary(modelo_cox)
```

```
Call:
coxph(formula = Surv(idade, status == "2") ~ ., data = dad

  n= 306, number of events= 81

            coef exp(coef) se(coef)      z Pr(>|z|)
anoOper -0.02801   0.97238  0.03474 -0.806     0.42
numNod   0.05517   1.05672  0.01009  5.469 4.52e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '

        exp(coef) exp(-coef) lower .95 upper .95
anoOper    0.9724     1.0284    0.9084     1.041
```

Hazard Ratio (HR)

```
ggforest(modelo_cox, data = dados)
```



Interpretação

O ano da operação do paciente não tem relação significativa com o desfecho do paciente ($p - valor = 0,42 > 0,05$) ao nível de 5% de significância.

Já o número de nódulos axilares positivos detectados têm efeito significativo na sobrevida dos pacientes. Cada aumento de um nódulo detectado foi associado um aumento médio de 5,67% no risco de mortalidade do paciente em 5 anos, após considerar outras covariáveis.

O IC de 95% para esse aumento ficou entre 3,6% e 7,8% ($p - valor << 0,001$)