Diego da Costa Medeiros

UMA ANÁLISE DE SOBREVIDA E DE TENDÊNCIA DA INCIDÊNCIA DO CÂNCER DE COLO DO ÚTERO, EM PACIENTES ATENDIDAS PELA LIGA NORTERIOGRANDENSE CONTRA O CÂNCER, NO PERÍODO DE 2007 A 2013

Natal - RN

13 de junho de 2019

Diego da Costa Medeiros

UMA ANÁLISE DE SOBREVIDA E DE TENDÊNCIA DA INCIDÊNCIA DO CÂNCER DE COLO DO ÚTERO, EM PACIENTES ATENDIDAS PELA LIGA NORTERIOGRANDENSE CONTRA O CÂNCER, NO PERÍODO DE 2007 A 2013

Monografia de Graduação apresentada ao Departamento de Estatística do Centro de Ciências Exatas e da Terra da Universidade Federal do Rio Grande do Norte como requisito parcial para a obtenção do grau de Bacharel em Estatística.

Universidade Federal do Rio Grande do Norte

Centro de Ciências Exatas e da Terra

Departamento de Estatística

Orientador: Prof. Dr. Paulo Roberto Medeiros de Azevedo

Natal - RN

13 de junho de 2019

Diego da Costa Medeiros

UMA ANÁLISE DE SOBREVIDA E DE TENDÊNCIA DA INCIDÊNCIA DO CÂNCER DE COLO DO ÚTERO, EM PACIENTES ATENDIDAS PELA LIGA NORTERIOGRANDENSE CONTRA O CÂNCER, NO PERÍODO DE 2007 A 2013

Monografia de Graduação apresentada ao Departamento de Estatística do Centro de Ciências Exatas e da Terra da Universidade Federal do Rio Grande do Norte como requisito parcial para a obtenção do grau de Bacharel em Estatística.

Aprovado em de de .

Prof. Dr. Paulo Roberto Medeiros de Azevedo
Orientador

Profa. Dra. Jeanete Alves Moreira
Examinador

Prof. Dr. José Veríssimo Fernandes
Examinador

Natal - RN
13 de junho de 2019

Universidade Federal do Rio Grande do Norte - UFRN
Sistema de Bibliotecas - SISBI
Catalogação de Publicação na Fonte. UFRN - Biblioteca Setorial Prof. Ronaldo Xavier de Arruda - CCET

Medeiros, Diego da Costa.
Uma análise de sobrevida e de tendência da incidência do câncer de colo do útero, em pacientes atendidas pela Liga Norteriograndense Contra o Câncer, no período de 2007 a 2013 / Diego da Costa Medeiros. - 2019.
34f.: il.

Monografia (Bacharelado em Estatística) - Universidade Federal do Rio Grande do Norte, Centro de Ciências Exatas e da Terra, Departamento de Estatística. Natal, 2019.
Orientadora: Jeanete Alves Moreira.
Coorientador: Paulo Roberto Medeiros de Azevedo.

1. Estatística - Monografia. 2. Regressão linear - Monografia. 3. Coeficiente de incidência - Monografia. 4. Estimador de Kaplan-Meyer - Monografia. 5. Modelo de Cox - Monografia. I. Moreira, Jeanete Alves. II. Azevedo, Paulo Roberto Medeiros de. III. Título.

RN/UF/CCET CDU 519.2

Elaborado por Joseneide Ferreira Dantas - CRB-15/324

À minha família, que sempre acreditou em mim e me apoiou.

Agradecimentos

Agradeço a meus pais, Fábio Medeiros e Ilucilane Pereira, a minha irmã Beatriz e a minha namorada que também se chama Beatriz, por todo o suporte que me deram para que eu conseguisse chegar ao final do curso feliz.

Aos amigos que fiz no decorrer do curso: Antony Iury, Kaio Breno, Igor Antônio e Ronne Von, sem vocês a graduação teria sido muito mais difícil.

Agradeço a Paulo Roberto por toda ajuda e dedicação na construção desse trabalho. E a todos os meus professores do departamento de estatística, que com excelência ensinaram-me muita coisa.

"Sem dados você é apenas mais uma pessoa com uma opinião."
– William Edwards Deming

Resumo

Introdução: O câncer do colo do útero é uma doença que acomete mulheres no mundo inteiro, com taxas de incidência e mortalidade variando de acordo com o grau de desenvolvimento da região. **Objetivo:** Descrever as taxas de incidência, bruta e padronizada pela idade, analisar suas tendências, bem como estimar curvas de sobrevida e ajustar o modelo de Cox para avaliar a razão de risco (RR) de covariáveis, dos casos de câncer de colo de útero atendidos pela Liga Norteriograndense Contra o Câncer, no período de 2007 a 2013. **Métodos:** Os cálculos dos coeficientes de incidência padronizados foram realizados pelo método direto, utilizando como padrão a população mundial de Segi de 1960. Na análise de tendência das séries dos coeficientes de incidência ajustou-se o modelo de regressão linear simples. As curvas de sobrevida foram estimadas através do estimador de Kaplan-Meyer e o teste *logrank* foi utilizado na avaliação de possíveis diferenças entre essas curvas. O modelo de Cox é ajustado para estimar a razão de risco nos casos univariado e múltiplo. **Resultados:** O coeficiente de incidência padronizado pela idade obtido foi de 13,26 por 100 mil mulheres. A série dos coeficientes de incidência não padronizados, para mulheres com idade abaixo de 40 anos, foi classificada como tendo tendência de crescimento, enquanto que, para as mulheres com 60 anos ou mais, essa série foi classificada como decrescente. As demais séries do estudo foram classificadas como estacionárias. A probabilidade acumulada de sobrevida após cinco anos, estimada para todas as mulheres do estudo foi de 76,6%, com o tempo mediano estimado de sobrevida de 17,1 meses. **Conclusão:** Neste estudo o câncer de colo do útero apresentou coeficiente de incidência com tendência crescente na faixa etária de até 39 anos, estável naquelas com idade entre 40 e 59 anos e decrescente após os 60 anos, no período estudado. As covariáveis faixa etária e escolaridade foram identificadas como fatores de prognóstico independentes.

Palavras-chave: Regressão linear; Coeficiente de incidência; Estimador de Kaplan-Meyer; Modelo de Cox.

Abstract

Introduction: Cervical cancer is a disease that affects women worldwide, with incidence rates and mortality varying according to the degree of development of the region. **Objective:** To describe gross and age-standardized incidence rates, to analyze their trends, as well as to estimate survival curves and to adjust the Cox model to evaluate the risk ratio (RR) of covariates, cases of cervical cancer attended by the Liga Norteriograndense Contra o Câncer, from 2007 to 2013. **Methods:** The calculation of the standardized incidence coefficients was performed using the direct method, using the world Segi population of 1960 as standard. In the trend analysis of the series of incidence coefficients, the simple linear regression model was adjusted. Survival curves were estimated using the Kaplan-Meyer estimator and the *logrank* test was used to evaluate possible differences between these curves. The Cox model is adjusted to estimate the risk ratio in univariate and multiple cases. **Results:** The standardized incidence coefficient was 13.26 per 100.000 women. The series of non-standardized incidence coefficients for women under the age of 40 years was classified as having a growth trend, while for women aged 60 years or over, this series was classified as decreasing. The other series of the study were classified as stationary. The cumulative five-year survival probability, estimated for all women in the study was 76.6%, with an estimated median survival time of 17.1 months. **Conclusion:** In this study, cervical cancer presented an incidence coefficient with an increasing tendency in the age group up to 39 years old, stable in those aged between 40 and 59 years and decreasing after 60 years, during the period studied. Age and school covariables were identified as independent prognostic factors.

Keywords: Linear regression; Incidence coefficient; Kaplan-Meyer estimator; Cox model.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

O câncer de colo uterino (CCU) é um tumor que tem origem a partir de células que revestem o epitélio da cérvice uterina e afeta mulheres de todo o mundo (TORRE et al., 2015). Uma estimativa global realizada por Ordikhani et al. (2016) em 2012, mostra que naquele ano foram notificados 528.000 novos casos da doença, com registro de 266.000 mortes, tornando-se o quarto mais comum tipo de câncer que afeta as mulheres no mundo, estando depois apenas dos cânceres de mama, colo-retal e de pulmão. Nos países menos desenvolvidos é o segundo tipo de câncer mais comum, e a terceira causa de morte por câncer em mulheres, com taxas de incidência e de mortalidade maiores principalmente na África subsaariana, Melanésia, América Latina e Caribe (ORDIKHANI et al., 2016).

No Brasil, segundo estimativa do Instituto Nacional do Câncer (INCA), excluindo-se os tumores de pele não melanoma, o CCU destaca-se como o terceiro câncer mais frequente entre as mulheres, depois dos cânceres de mama e colo-retal. A incidência é heterogênea, variando nas diferentes regiões do país, sendo mais incidente na região Norte, com taxa bruta de 23,97 por 100 mil mulheres, seguido pelas regiões Centro-oeste, com 20,72, Nordeste com 19,49, Sul com 15,17 e Sudeste com 11,30. Para o estado do Rio Grande do Norte a taxa bruta estimada foi de 17,25, situando-se um pouco abaixo da média estimada para a região Nordeste (INSTITUTO NACIONAL DE CÂNCER JOSÉ ALENCAR GOMES DA SILVA, 2016).

As taxas de sobrevida do câncer do colo do útero após cinco anos do diagnóstico variam, estando fortememente relacionadas com o grau de desenvolvimento do país, observando-se taxas mais elevadas nas regiões mais desenvolvidas. Entre os países desenvolvidos, Estados Unidos da América (EUA), Alemanha e Espanha têm uma taxa de sobrevida após cinco anos em torno de 60%. Na Inglaterra e no País de Gales a taxa de sobrevida acima de cinco anos é de 67,4%. Em países em desenvolvimento como China e Tailândia foram relatadas taxas de sobrevida após cinco anos superiores a 50%. Por outro lado, países como Gâmbia e Uganda têm taxa de sobrevida após cinco anos de menos de 25% (MUHAMAD et al., 2015). No Brasil, para o período de 2005 a 2009, a sobrevida após cinco anos ficou em torno de 61% (INSTITUTO NACIONAL DE CÂNCER JOSÉ ALENCAR GOMES DA SILVA, 2016).

Os dados deste trabalho foram obtidos na Liga Norte Riograndense Contra o Câncer, no período de 2007 a 2013. O hospital está situada na cidade de Natal, e os dados são referentes aos casos de câncer de colo do útero de todo o estado do Rio Grande do Norte.

1.1 Objetivos

O presente estudo tem como objetivos calcular os coeficientes de incidência, padronizados e não padronizados, analisar as tendências desses coeficientes, ao longo do tempo, estimar a probabilidade acumulada de sobrevida após 5 anos e ajustar o modelo de regressão de Cox, para os casos de câncer de colo do útero em pacientes atendidas no Hospital Dr. Luiz Antonio, Natal/RN, entre 2007 e 2013.

1.2 Justificativa

O conhecimento da incidência e da sobrevida de pacientes com câncer de colo do útero certamente favorecerá o planejamento de ações preventivas visando o diagnóstico precoce, avaliação e acompanhamento de atividades e de tratamento visando melhorar a sobrevida de pacientes dessa doença. Ou seja, este trabalho ajudará, na medida em que produzirá indicadores sobre incidência e prognóstico dos referidos casos de câncer atendidos no hospital Dr. Luiz Antonio, Natal/RN.

2 Metodologia

Neste capítulo são apresentados os conceitos básicos sobre a análise de regressão linear simples e sobre coeficientes de incidência, bruto e padronizado. E por último é mostrado conceitos de análise de sobrevivência.

2.1 Análise de Regressão

Análise de Regressão é um método estatístico que utiliza a relação entre duas ou mais variáveis, de modo que a variável resposta (variável dependente) possa ser prevista a partir de uma ou mais variáveis explicativas (variável independente ou explicativa).

Uma relação funcional entre duas variáveis é dada por $Y = f(X)$, onde X é a variável independente , Y a variável dependente e f indica a relação entre elas. Esse trata-se de um modelo matemático, ou seja, todos os pontos estão sobre a curva. Em uma relação estatística, diferente da relação funcional, os pontos não estão perfeitamente localizados na curva de relação.

O modelo de regressão é uma relação estatística que é caracterizada por duas propriedades, quais sejam: na população de onde se retiram os dados, tem-se uma distribuição de probabilidade de Y, para cada nível de X; e as médias dessas distribuições variam de forma sistemática com X.

2.1.1 Regressão Linear Simples

O modelo de regressão linear simples acontece quando se tem uma única variável independente. Esse modelo é dado da senguinte forma:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1,2 \ldots ,n, \tag{2.1}$$

sendo que:

- Y_i: é a variável resposta;
- X_i: é uma constante previamente conhecida;
- β_0: é um parâmetro que representa o intercepto da reta $\beta_0 + \beta_1 X_i$ com o eixo vertical;
- β_1: é um parâmetro que representa a inclinação da reta $\beta_0 + \beta_1 X_i$;
- ε_i: é um erro aleatório, com $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$, $\forall i = 1 \ldots n$, e $Cov(\varepsilon_i,\varepsilon_j) = 0$, $\forall i \neq j$.

O valor esperado de Y é chamado de função de regressão, de forma que, utilizando as hipótese sobre os erros, temos:

$$E(Y_i) = \beta_0 + \beta_1 X_i, \tag{2.2}$$

com $Var(Y_i) = \sigma^2$, $\forall i$ e $Cov(Y_i,Y_j) = 0$, $\forall i \neq j$. O parâmetro β_1 representa a mudança de $E(Y)$, para cada unidade acrescida em X. Quando o escopo do modelo cobre $X = 0$, β_0 representa a média da distribuição de probabilidade de Y em $X = 0$. Quando o escopo do modelo não cobre $X = 0$, β_0 não tem nenhum significado particular no modelo de regressão.

2.2 Estimadores dos Parâmetros

Os parâmetros do modelo são estimados a partir de uma amostra aleatória (X_i,Y_i), com $i = 1,2,\ldots,n$. Uma maneira de estimar os parâmetros β_0 e β_1 é através do método de mínimos quadrados.

2.2.1 Método de Mínimos Quadrados

O método de mínimos quadrados considera, para cada observação (X_i,Y_i), o desvio de Y_i do seu valor esperado:

$$Y_i - (\beta_0 + \beta_1 X_i). \tag{2.3}$$

Em particular, o método considera a soma dos n desvios ao quadrado, aqui denotado por Q:

$$Q = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2. \tag{2.4}$$

O objetivo principal da técnica é encontrar estimadores para os parâmetros, tal que minimizem Q, para as observações da amostra $(X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n)$.

Derivando-se Q em relação à β_0 e β_1, obtém-se:

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i); \tag{2.5}$$

$$\frac{\partial Q}{\partial \beta_1} = -2 \sum_{i=1}^{n} X_i (Y_i - \beta_0 - \beta_1 X_i). \tag{2.6}$$

Igualando as equações 2.5 e 2.6 a zero e usando b_0 e b_1 como os respectivos valores de β_0 e β_1 que minimizam Q, temos:

$$-2 \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i) = 0; \tag{2.7}$$

$$-2\sum_{i=1}^{n} X_i(Y_i - b_0 - b_1X_i) = 0. \tag{2.8}$$

Resolvendo o sistema acima, obtém-se o seguinte resultado:

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}; \tag{2.9}$$

$$b_0 = \bar{Y} - b_1\bar{X}. \tag{2.10}$$

Portanto, $\hat{Y} = b_0 + b_1X$ é o estimador de mínimos quadrados da função de regressão $E(Y) = \beta_0 + \beta_1X$, sendo b_0 e b_1 os estimadores de β_0 e β_1, respectivamente.

2.3 Teste Sobre os Parâmetros

2.3.1 Teste sobre β_1

Para o modelo de regressão linear simples, $Y_i = \beta_0 + \beta_1X_i + \varepsilon_i$, com erros normalmente distribuídos, b_1 tem distribuição normal com média e variância dadas por:

$$E(b_1) = \beta_1 \quad \text{e} \quad Var(b_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

Para estimar a variância de b_1 é utilizado a soma dos quadrados residuais (ou soma dos quadrados dos erros), que é definida da seguinte forma:

$$SQE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)^2. \tag{2.11}$$

Dividindo-se SQE por $n-2$, obtém-se o quadrado médio do resíduo (ou quadrado médio do erro):

$$QME = \frac{SQE}{n-2}. \tag{2.12}$$

Pode-se mostrar que QME é um estimador não viciado de σ^2, de forma que um estimador não viciado da variância de b_1 é dado por:

$$S^2(b_1) = \frac{QME}{\sum_{i=1}^{n}(X_i - \bar{X})^2}. \tag{2.13}$$

Sobre a distribuição amostral de $(b_1 - \beta_1)/S(b_1)$, temos:

$$\frac{(b_1 - \beta_1)}{S(b_1)} = \frac{\frac{b_1 - \beta_1}{\sqrt{Var(b_1)}}}{\frac{S(b_1)}{\sqrt{Var(b_1)}}},$$

de forma que, sendo os erros são normalmente distribuíduos, $(b_1 - \beta_1)/\sqrt{Var(b_1)}$ tem distribuição normal padrão e:

$$\frac{S^2(b_1)}{Var(b_1)} = \frac{\frac{QME}{\sum_{i=1}^{n}(X_i-\bar{X})^2}}{\frac{\sigma^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}} = \frac{QME}{\sigma^2} = \frac{SQE}{(n-2)\sigma^2},$$

em que pode-se mostrar que SQE/σ^2 tem ditribuição Qui-quadrado com $n-2$ graus de liberdade e é independente de b_0 e b_1. Portanto:

$$\frac{S^2(b_1)}{Var(b1)} = \frac{\chi^2_{(n-2)}}{n-2}. \tag{2.14}$$

Assim:

$$\frac{(b_1 - \beta_1)}{S(b_1)} = \frac{N(0,1)}{\sqrt{\frac{\chi^2_{(n-2)}}{n-2}}}, \tag{2.15}$$

e como o numerador é independente do denominador, temos que esta distribuição é uma t de Student com $n-2$ graus de liberdade.

Para verificar se não existe relação linear entre a X e Y, testamos:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

de forma que não rejeitar H_0 significa a não existência de relação linear entre X e Y. Para isso, temos que a estatística do teste é:

$$t = \frac{b_1}{S(b_1)}, \tag{2.16}$$

cuja distribuição é uma t de Student com $n-2$ graus de liberdade, supondo H_0 verdadeira.

2.3.2 Teste sobre β_0

Para o modelo de regressão linear simples, $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, com erros normalmente distribuídos, $b_0 = \bar{Y} - b_1\bar{X}$ tem distribuição normal com média e variância dadas por:

$$E(b_0) = \beta_0 \quad \text{e} \quad Var(b_0) = \sigma^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right].$$

Analogamente ao que foi visto para b_1, temos que a distribuição de $(b_0 - \beta_0)/S(b_0)$ é uma t de Student com n-2 graus de liberdade sendo:

$$S^2(b_0) = QME\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right] \tag{2.17}$$

um estimador não viciado de $Var(b_0)$.

Da mesma forma, temos então que $t = b_0/S(b_0)$ é a estatística para testar:

$$H_0 : \beta_0 = 0$$
$$H_1 : \beta_0 \neq 0$$

cuja distribuição também é uma t de Student com $n - 2$ graus de liberdade, quando H_0 é verdadeira.

2.4 Análise Residual

O resíduo é definido para uma dada observação como sendo a diferença entre o valor observado e o valor estimado, aqui o i-ésimo resíduo é dado por $e_i = Y_i - \hat{Y}_i$, ou seja, para o modelo de regressão 2.1, o resíduo e_i torna-se $e_i = Y_i - (b_0 - b_1 X_i)$.

Os resíduos são importantes para verificar se o modelo ajustado é apropiado para os dados. Essa análise é feita a partir da verificação das suposições do modelo. Normalmente usam-se os gráficos de resíduos para detectar, de maneira informal, problemas no modelo de regressão linear, tais como:

- A função de regressão não é linear.
- Os erros não têm variância constante.
- Presença de *outliers*.

A função de regressão linear não é adequada para os dados quando os resíduos não se distribuem aleatoriamente em torno do zero, indicando uma não-linearidade na relação entre X e Y. Outra importante característica sobre os resíduos que esse gráfico pode mostrar é se variância dos erros é constante, ou seja, se a variância dos erros não aumenta (ou diminui) quando os valores de X crescem.

A identificação de *outliers* também é importante, pois a presença desses pode causar prejuízos para o ajuste de uma reta de regressão, dado que a reta é puxada desproporcionalmente para esses pontos, quando o método de estimação dos parâmetros é o de mínimos quadrados. Porém, um valor como esse poderá apenas ser retirado, se for constatado que sua prensença é resultado de um erro grosseiro na fase da amostragem. Uma maneira de identificar a presença de *outliers* nas análises é considerando o gráfico de resíduos padronizados versus a variável independente, identificando-se esses pontos quando sua distância para o zero é superior a três.

Temos que $d_i = e_i/\sqrt{QME}$ é o i-ésimo resíduo padronizado pelo desvio padrão, em que QME é uma variância amostral dos resíduos, ou seja:

$$\frac{\sum_{i=1}^{n}(e_i - \bar{e})^2}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{SQE}{n-2} = QME. \tag{2.18}$$

2.5 Coeficientes de Incidência

O cálculo da taxa de incidência de câncer é de extrema importância para estudar o risco de câncer em pessoas na área de registro em comparação com outros lugares, ou para comparar diferentes subgrupos na mesma área de estudo.

O número de casos de uma doença não é suficiente para termos ideia do risco dessa doença em um local, sem levar em consideração o tamanho da população desse local. Para representar o risco de ocorrência de uma determinada doença, é utilizado muitas vezes a taxa de incidência bruta (C). Convencionalmente, as taxas de incidência de câncer são expressas como casos por 100000 pessoas, pois isso evita o uso de números bastante pequenos.

2.5.1 Taxa de Incidência Bruta

Suponha que existam A faixas etárias para as quais o número de casos possam ser avaliados. Vamos denotar por r_i o número de casos de câncer que ocorreram na i-ésima faixa etária. Se todos os casos são de idade conhecida, então o número total de casos (R) pode ser escrito como:

$$R = \sum_{i=1}^{A} r_i = r_1 + r_2 + r_3 + \cdots + r_A. \tag{2.19}$$

Da mesma forma, denotando por n_i o total populacional na mesma i-ésima faixa etária durante um mesmo período de tempo em que os casos foram contados. O total populacional de todas as faixas de idade, no período de tempo considerado, é:

$$N = \sum_{i=1}^{A} n_i = n_1 + n_2 + n_3 + \cdots + n_A. \tag{2.20}$$

A taxa bruta para todas as idades, no período de tempo considerado, por 100000 habitantes, é facilmente calculada dividindo-se o número total de casos (R) pelo total populacional (N), multiplicando-se o resultado por 100000.

$$\text{Taxa bruta} = C = \frac{R}{N} \times 100000. \tag{2.21}$$

A taxa específica para a faixa etária i, a qual denota-se por a_i, também é facilmente calculada, como uma taxa por 100000, dividindo o número de casos nessa faixa etária (r_i) pelo correspondente total populacional da faixa etária i (n_i) e multiplicando o resultado por 100000:

$$a_i = \frac{r_i}{n_i} \times 100000. \tag{2.22}$$

2.5.2 Taxa de Incidência Padronizada Pela Idade

Um dos problemas mais frequentes em estudos epidemiológicos sobre câncer envolve a comparação das taxas de incidência de um câncer em particular entre duas populações diferentes, ou para a mesma população ao longo do tempo. A comparação da taxa bruta pode dar um ideia falsa devido as diferenças entre as distribuições de idade nas duas populações. Por exemplo, se uma população é, em média, mais jovem do que outra, então mais casos de câncer tendem a aparecer na população mais velha.

Assim, ao comparar os níveis incidência de câncer entre duas áreas, ou ao investigar o padrão de câncer ao longo do tempo para a mesma área, é importante levar em consideração a diferença na estrutura etária. Isso é realizado pela padronização da incidência pela idade.

Na padronização da taxa de incidência pela idade, é tomado uma população como referência e essa taxa padronizada pela idade é a taxa teórica que teria ocorrido se as taxas específicas para cada faixa etária observada fossem aplicadas nessa população de referência.

Em cada faixa etária da população padrão são definidos os pesos a serem usados no processo de padronização. Muitos conjuntos possíveis de pesos, w_i, podem ser usados. A padronização mais utilizada é a da População Padrão Mundial (ver Tabela 2.1), modificada por Doll e Hill (1966), daquela proposta por Segi et al. (1960). Seu uso difundido facilita muito a comparação dos níveis de incidência de câncer entre as áreas.

Tabela 2.1 – População Padrão Mundial

Índice da faixa etária (i)	Faixa etária	População (w_i)
1	$0-4$	12000
2	$5-9$	10000
3	$10-14$	9000
4	$15-19$	9000
5	$20-24$	8000
6	$25-29$	8000
7	$30-34$	6000
8	$35-39$	6000
9	$40-44$	6000
10	$45-49$	6000
11	$50-54$	5000
12	$55-59$	4000
13	$60-64$	4000
14	$65-69$	3000
15	$70-74$	2000
16	$75-79$	1000
17	$80-84$	500
18	85+	500
		100000

É denotato por w_i a população presente em uma faixa etária da população padrão, no qual, assim como mostrado anteriormente, $i = 1,2,\ldots,A$, com a_i representando novamente a taxa específica na faixa etária i. A taxa padronizada (age-standardized rate (ASR)) é calculada da seguinte forma:

$$\text{ASR} = \frac{\sum_{i=1}^{A} a_i w_i}{\sum_{i=1}^{A} w_i}. \tag{2.23}$$

2.6 Análise de Sobrevivência

A análise de sobrevivência é uma das áreas da estatística que mais cresceu nos últimos anos, principalmente na área médica. Na análise de sobrevivência a variável resposta é o tempo até a ocorrenência de um evento de interesse, sendo esse denominado de tempo de falha. Esse pode ser o tempo até a morte do paciente, tempo até a cura ou tempo até a recidiva de uma doença, por exemplo. Em estudos de câncer é usual esse tempo ser da data do diagnóstico até a cura, ou até a morte do paciente. Neste trabalho o tempo de falha é do diagnóstico até a morte.

Uma das principais características de dados de sobrevivência é a presença de censura, que é a observação parcial da resposta. A cencura ocorre quando o acompanhamento do paciente é interrompido por algum motivo, como por exemplo por mudança de cidade do paciente, término do estudo, entre outras razões. As censuras são divididas em três tipos:

- **Censura à direita:** ocorre quando o tempo entre o início e o evento de interesse é maior do que o tempo registrado;
- **Censura à esquerda:** ocorre quando o evento de interesse já aconteceu quando o indivíduo foi observado;
- **Censura intervalar:** ocorre quando o tempo exato de ocorrência do evento é desconhecido e o que se sabe é que ele ocorreu num determinado intervalo.

Na análise de sobrevivência cada indivíduo i é representado pelo par (t_i,δ_i), com $i = 1,\ldots,n$, sendo t_i o tempo até o evento de interesse ou censura e δ_i é a variável indicadora de falha/censura, em que $\delta_i = 1$, se t_i é o tempo observado do evento, ou $\delta_i = 0$, se t_i é um tempo de censura.

O principal componente na análise de sobrevivência é a função de sobrevivência, que é definida como a probabilidade de uma observação não falhar até o tempo t e é escrita como:

$$S(t) = P(T \geq t),$$

então o procedimento inicial é encontrar uma estimativa para a função de sobrevivência, para então estimar as estatísticas de interesse. O estimador não paramétrico Kaplan-Meier

é bastante utilizado na literatura e a partir dele quantidades como mediana e média podem ser obtidos.

2.6.1 O estimador de Kaplan-Meier

O estimador não paramétrico de Kaplan-Meier, proposto por Kaplan e Meier (1958) para estimar a função de sobrevivência, é uma adaptação da sobrevivência empírica, que na ausência de censura, é definida como:

$$\hat{S}(t) = \frac{\text{número de observações que não falharam até o tempo } t}{\text{número total de indivíduos no estudo}}, \tag{2.24}$$

em que $\hat{S}(t)$ é uma função escada com degraus nos tempos observados de falha de tamanho $1/n$, no qual n é o tamanho da amostra. Quando existem empates em um determinado tempo t, o tamanho fica multiplicado pelo número de empates.

O estimador de Kaplan Meier, na sua construção, considera tantos intervalos de tempo quantos forem o número de falhas distintas.

Para qualquer t, $S(t)$ pode ser escrito em termos de probabilidades condicionais. Suponhamos n pacientes no estudo e $k(\leq n)$ falhas distintas nos tempos $t_1 < t_2 < \cdots < t_k$. Considerando $S(t)$ uma função discreta com probabilidade maior que zero somente nos tempos de falha t_j, $j = 1, \ldots, k$, tem-se:

$$S(t_j) = (1 - q1)(1 - q2) \cdots (1 - q_j), \tag{2.25}$$

em que q_j é a probabilidade de um indivíduo falhar no intervalo $[t_{j-1}, t_j)$ dado que ele não falhou até t_{j-1} e considerando $t_0 = 0$. Então, pode-se escrever q_j da seuinte forma:

$$q_j = P(T \in [t_{j-1}, t_j) \,|T \geq t_{j-1}). \tag{2.26}$$

Portanto, o estimador de Kaplan-Meier é estimado a partir da estimação de q_j, que adaptado da expressão 2.24, obtem-se:

$$\hat{q_j} = \frac{d_j}{n_j}, \tag{2.27}$$

em que d_j é o número de falhas em t_j e n_j o número de indivíduos sob risco em t_j, ou seja, o número de indivíduos que não falharam e não foram censurados até o instante t_{j-1}, para $j = 1, \ldots, k$.

Considerando $t_1 < t_2 < \cdots < t_k$, d_j e n_j o estimador de Kaplan-Meier é definido como:

$$\hat{S}(t) = \prod_{j:t_j<t} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{j:t_j<t} \left(1 - \frac{d_j}{n_j} \right). \tag{2.28}$$

As principais propriedades do estimador de Kaplan-Meier são basicamente as seguintes:

1. é não-viciado para amostras grandes;

2. converge assintoticamente para um processo gaussiano;

3. é estimador de máxima verossimilhança de $S(t)$.

2.6.2 Teste de logrank

A estatística mais comumente usada como generalizações para dados censurados em análise de sobrevivência é o teste de *logrank* (MANTEL, 1966), que se trata de um teste não paramétrico que é utilizado para avaliar se existe diferença significativa entre duas ou mais curvas de sobrevivências.

A estatística deste teste é obtida atrazés da diferença entre o número de falhas em cada grupo e uma quantidade que pode ser pensada como o número esperado de falhas sob a hipótese nula.

Considerando o teste para duas funções de sobrevivência $S_1(t)$ e $S_2(t)$. Sejam $t_1 < t_2 < \cdots < t_k$ os tempos de falha distintos da amostra formada pela combinação das duas amostras individuais. Supondo que no tempo t_j aconteçam d_j falhas e que n_j indivíduos estejam sob risco em um tempo imediatamente anterior a t_j na amostra combinada e, respectivamente, d_{ij} e n_{ij} na amostra i, com $i = 1{,}2$ e $j = 1, \ldots, k$. Em cada tempo de falha t_j, os dados podem ser dispostos em forma de uma tabela de contigência 2×2 com d_{ij} falhas e $n_{ij} - d_{ij}$ sobreviventes na coluna i, como é mostrado na Tabela 2.2.

Tabela 2.2 – Tabela de contigência gerada no tempo t_j.

	Grupos		
	1	2	
Falha	d_{1j}	d_{2j}	d_j
Não Falha	$n_{1j} - d_{1j}$	$n_{2j} - d_{2j}$	$n_j - d_j$
	n_{1j}	n_{2j}	n_j

Fixando as marginais de coluna e de linha, ou seja, condicinalmente à experiência de falha e cencura até o tempo t_j e ao número de falhas no tempo t_j, a distribuição de d_{2j} é dada por uma hipergeométrica:

$$\frac{\binom{n_{1j}}{d_{1j}}\binom{n_{2j}}{d_{2j}}}{\binom{n_j}{d_j}}.$$

A média de d_{2j} é $w_{2j} = n_{2j} d_j n_j^{-1}$ com variância obtida a partir da distribuição hipergeométrica, igual a $(V_j)_2 = n_{2j}(n_j - n_{2j}) d_j (n_j - d_j) n_j^{-2} (n_j - 1)^{-1}$.

Então, a estatística $d_{2j} - w_{2j}$ tem média zero e variância $(V_j)_2$. Se as k tabelas de contigência forem independentes, um teste aproximado para avaliar a igualdade de das

duas funções de sobrevivência pode ser baseado na estatística:

$$T = \frac{\left[\sum_{j=1}^{k}(d_{2j} - w_{2j})\right]^2}{\sum_{j=1}^{k}(V_j)_2}, \tag{2.29}$$

que, sob a hipótese nula $H_0 : S_1(t) = S_2(t)$, para todo t no período de acompanhamento, tem uma distribuição qui-quadrado com 1 grau de liberdade, para amostras grandes.

Na generalização do teste *logrank* para a igualdade de $r > 2$ funções de sobrevivência, considerando que o i pode variar entre 1 e r, os dados podem ser arranjados em forma de uma tabela de contigência $2 \times r$, com $n_{ij} - d_{ij}$ sobreviventes na coluna i.

Fixando novamente as marginais de coluna e de linha, a distribuição conjunta de $d_{1j}, \ldots, d_{rj}$ é dada por uma hipergeométrica multivariada:

$$\frac{\prod_{j=1}^{r}\binom{n_{ij}}{d_{ij}}}{\binom{n_j}{d_j}}.$$

A média de d_{ij} é $w_{ij} = n_{ij}d_j n_j^{-1}$, e a variância de d_{ij} e a covariância entre d_{ij} e d_{lj} são respectivamente,

$$(V_j)_{ii} = n_{ij}(n_j - n_{ij})d_j(n_j - d_j)n_j^{-2}(n_j - 1)^{-1}$$

e

$$(V_j)_{il} = -n_{ij}n_{lj}d_j(n_j - d_j)n_j^{-2}(n_j - 1)^{-1}.$$

Então a estatística $\mathbf{v}'_j = (d_{2j} - w_{2j}, \ldots, d_{rj} - w_{rj})$ tem como média um vetor de zeros e matriz de variância-covariância $\mathbf{V}_j$ de dimensão $(r-1) \times (r-1)$. Pode-se então formar a estatística v, somando sobre todos os tempos distintos de falhas:

$$\mathbf{v} = \sum_{j}^{k} \mathbf{v}_j,$$

com $\mathbf{v}$ um vetor de dimensão $(r-1) \times 1$, cujos elementos são as diferenças entre os totais observados e esperados da falha.

Considerando novamente que as k tabelas de contigência são independentes, assim a matriz variância-covariância da estatística $\mathbf{v}$ será $\mathbf{V} = \mathbf{V}_1 + \cdots + \mathbf{V}_k$. Um teste aproximado para a igualdade das r funções de sobrevivência é baseado na estatística:

$$T = \mathbf{v}'\mathbf{V}^{-1}\mathbf{v}, \tag{2.30}$$

que sob H_0 (igualdade das curvas), tem uma distribuição qui-quadrado com $r-1$ graus de liberdade, para amostras grandes.

2.6.3 Função de Taxa de Falha ou de Risco

A probabilidade da ocorrência de falha num intervalo de tempo $[t_1,t_2)$ pode ser calculado da seguinte forma:

$$S(t_1) - S(t_2).$$

A taxa de falha no intervalo $[t_1,t_2)$ é definida pela probabilidade acima, dado que a falha não ocorreu até o instante anterior a t_1, dividida pelo comprimento do intervalo. Assim, tem-se:

$$\frac{S(t_1) - S(t_2)}{(t_2 - t_1)S(t_1)}. \tag{2.31}$$

De modo geral, considerando o intervalo $[t,t + \Delta t)$, a expressão 2.31 é dada por:

$$\frac{S(t_1) - S(t + \Delta t)}{\Delta t \quad S(t_1)}.$$

Considerando Δt bem pequeno, a expressão acima representa a taxa de falha instantânea no tempo t condicional à sobrevivência até o tempo t. A função de taxa de falha de T é então definida como:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}. \tag{2.32}$$

2.6.4 Modelo de Regressão de Cox

O modelo de Cox (COX, 1972) é um dos mais utilizados em estudos clínicos devido a sua versatilidade. O modelo de Cox é utilizado para analisar dados de estudos de tempo de vida em que a resposta é o tempo até a ocorrência de um evento de interesse, sendo dependente de covariáveis.

Supondo um vetor com p covariáveis $\mathbf{X} = (X_1,X_2,\ldots,X_p)'$, a expressão geral do modelo de regressão de Cox é dada da seguinte forma:

$$\lambda(t) = \lambda_0(t)g(\mathbf{X}'\boldsymbol{\beta}), \tag{2.33}$$

em que $\boldsymbol{\beta} = (\beta_1,\beta_2,\ldots,\beta_p)'$ é o vetor dos parâmetros associados às covariáveis, a função g é um componente paramétrico e $\lambda_0(t)$ um componente não-paramétrico. A função g é não-negativa e deve ser especificada, tal que $g(0) = 1$, enquanto que $\lambda_0(t)$ é uma função não negativa do tempo, como visto em 2.32, sendo conhecida como função base, pois $\lambda(t) = \lambda_0(t)$ quando $\mathbf{X} = \mathbf{0}$. O componente paramétrico mais utilisado é o multiplicativo:

$$g(\mathbf{X}'\boldsymbol{\beta}) = \exp\{\mathbf{X}'\boldsymbol{\beta}\} = \exp\{\beta_1 X_1 + \cdots + \beta_p X_p\}$$

garantindo que $\lambda(t)$ seja sempre não-negativa.

Este modelo também é conhecido como modelo de risco proporcionais, pois a razão das taxas de falha para dois indivíduos diferentes é constante no decorrer do tempo. A razão entre as funções de taxa de falha dos indivíduos i e j é dada por:

$$\frac{\lambda_i(t)}{\lambda_j(t)} = \frac{\lambda_0(t)\exp\{\mathbf{X}_i'\boldsymbol{\beta}\}}{\lambda_0(t)\exp\{\mathbf{X}_j'\boldsymbol{\beta}\}} = \exp\{\mathbf{X}_i'\boldsymbol{\beta} - \mathbf{X}_i'\boldsymbol{\beta}\},$$

observando-se aasim que essa razão não depende do tempo. Porntato, a suposição básica para utilização do modelo de regressão de Cox é que as taxas de falha sejam proporcionais.

2.6.5 Ajustando o Modelo de Cox

O modelo de Cox é caracterizado pelos β's, que medem o impacto da covariável sobre a função de taxa de falha. Para estimação desses parâmetros o método de máxima verossimilhança (COX; HINKLEY,) é inapropriado devido ao componente não-paramétrico $\lambda_0(t)$ na função de verossimilhança.

Sabe-se que a função de verossimilhança é dada da seguinte maneira:

$$\boldsymbol{L}(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\lambda(t_i|\mathbf{X}_i)\right]^{\delta_i} S(t_i|\mathbf{X}_i) \tag{2.34}$$

No modelo de Cox verifica-se que:

$$S(t_i|\mathbf{X}_i) = \exp\left\{-\int_0^{t_i} \lambda_0(u)\exp\{\mathbf{X}_i'\boldsymbol{\beta}\}du\right\} = \left[S_0(t_i)\right]^{\exp\left\{\mathbf{X}_i'\boldsymbol{\beta}\right\}},$$

assim aplicando-se este resultado em 2.34, temos:

$$\boldsymbol{L}(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\lambda_0(t_i)\exp\left\{\mathbf{X}_i'\boldsymbol{\beta}\right\}\right]^{\delta_i} \left[S_0(t_i)\right]^{\exp\left\{\mathbf{X}_i'\boldsymbol{\beta}\right\}},$$

que é função do componente não paramétrico $\lambda_0(t)$. Para resolver o problema desta função de pertubação da verossimilhança, Cox propôs no seu artigo (COX, 1975), um método denominado de máxima verossimilhança parcial, que consiste em condicionar a construção da função de verossimilhança ao conhecimento da história passada de falhas e censuras, de forma a eliminar $\lambda_0(t)$.

2.7 Fonte dos dados e Métodos Utilizados

Os dados foram obtidos através do Registro de Câncer do hospital Dr. Luiz Antônio (Liga Norte Riograndense Contra o Câncer). Neste estudo foram analisadas a incidência e a sobrevida global total, até cinco anos de acompanhamento, dos casos registrados de 2007 a 2013.

O coeficiente de incidência padronizado foi obtido através do método direto, utilizando como padrão a população mundial de Segi et al. (1960). Na análise de tendência das séries dos coeficientes de incidência ajustou-se o modelo de regressão linear simples. Neste modelo, a variável resposta é o coeficiente de incidência e a variável explicativa é o tempo (anos). A avaliação da existência de tendência na série baseou-se no teste cujas hipóteses nula e alternativa são, respectivamente, de que o coeficiente da variável explicativa é zero e diferente de zero. Ou seja, a série é considerada estável quando a hipótese nula não é rejeitada (p-valor do teste é maior que 0,05). Se a hipótese nula é rejeitada, a série é classificada como tendo tendência crescente ou decrescente, conforme seja o sinal positivo ou negativo, respectivamente, da estimativa obtida para o coeficiente da variável explicativa. Em cada ajuste realizado foi feita análise de resíduos, sendo avaliadas as hipóteses de normalidade, de variância constante e de não correlação entre os erros.

Na análise de sobrevida tomou-se como evento de interesse o óbito da paciente e foram considerados dados censurados as mulheres que não foram a óbito até o término do estudo ou que não possuíam acompanhamento atualizado. Foi considerada como data limite do acompanhamento 31/12/2013.

Foi obtida a curva da função de probabilidade acumulada de sobrevida para cada uma das categorias das variáveis: faixa etária, escolaridade, raça/cor, classificação histológica e procedência. As diferenças entre as curvas de sobrevida foram analisadas pelo teste *logrank*, considerando-se como tendo diferenças estatisticamente significativas os casos em que o p-valor é menor que 0,05. O modelo de riscos proporcionais de Cox foi utilizado para estimativas das razões de riscos.

Na construção do arquivo e tabulação dos dados foi utilizado o *software* de planilhas *Microsoft Excel (2016)*. Para as análises de tendências e os cálculos das probabilidades acumuladas de sobrevida e suas apresentações gráficas, para a realização do teste log-rank e o ajuste do modelo de riscos proporcionais de Cox foi utilizado o *software* estatístico *R Core Team (2018)*.

3 Resultados

Entre 2007 e 2013 foram registrados 1.649 novos casos de câncer do colo do útero atendidos pela Liga Norte Riograndense Contra o Câncer. A taxa de incidência padronizada pela idade obtida foi de 13,26 por 100 mil mulheres. As taxas de incidência não padronizadas para as faixas etárias consideradas variaram de 6,76 a 35,96 por 100 mil mulheres. Estes resultados estão resumidos na Tabela 3.1 a seguir.

Tabela 3.1 – Distribuição do número, da porcentagem e das taxas de incidência dos casos de câncer de colo do útero atendidos pela Liga Norte Riograndense Contra o Câncer, de 2007 a 2013.

Faixa etária	**Casos novos registrados**		**Taxa de incidência por 100.000 mulheres**
	Nº	**%**	
≤ 39	534	32,38	6,76
$40 - 59$	660	40,02	27,26
≥ 60	455	27,59	35,96
Total	1649	100,00	13,26*

* Taxa padronizada.

A série das taxas de incidência dos casos com idade de 60 anos ou mais está com a tendência classificada como decrescente, ao longo do tempo, visto que o p-valor do respectivo teste sobre o coeficiente da variável explicativa (tempo) é menor que 0,05 e a estimativa desse coeficiente é negativa. Por outro lado, a série das taxas de incidência dos casos com 39 anos ou menos está com tendência de crescimento, pois o p-valor do respectivo teste sobre o coeficiente da variável tempo é menor que 0,05 e sua estimativa é positiva. As demais séries são classificadas como estáveis. Estes resultados estão na Tabela 3.2 a seguir.

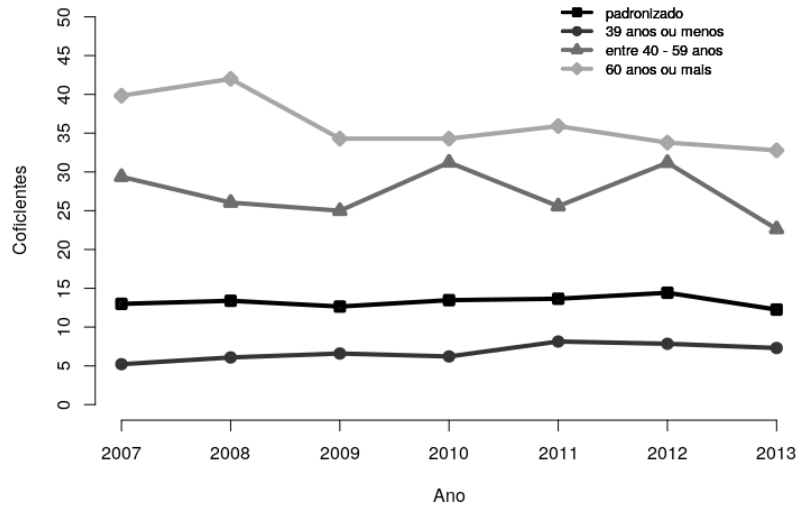
Tabela 3.2 – Análise de tendência dos casos de câncer de colo do útero atendidos pela Liga Norte Riograndense Contra o Câncer, de 2007 à 2013.

Faixa etária	**Modelo estimado**	***p*-valor**	**Tendência**
≤ 39	$5{,}1501 + 0{,}4047t$	0,0188	crescente
$40 - 59$	$28{,}6260 - 0{,}3330t$	0,6413	estável
≥ 60	$41{,}2649 - 1.2641t$	0,0297	decrescente
Total	$13{,}1387 - 0.0309t$	0,8410	estável*

* Tendência da série das taxas de incidência padronizadas.

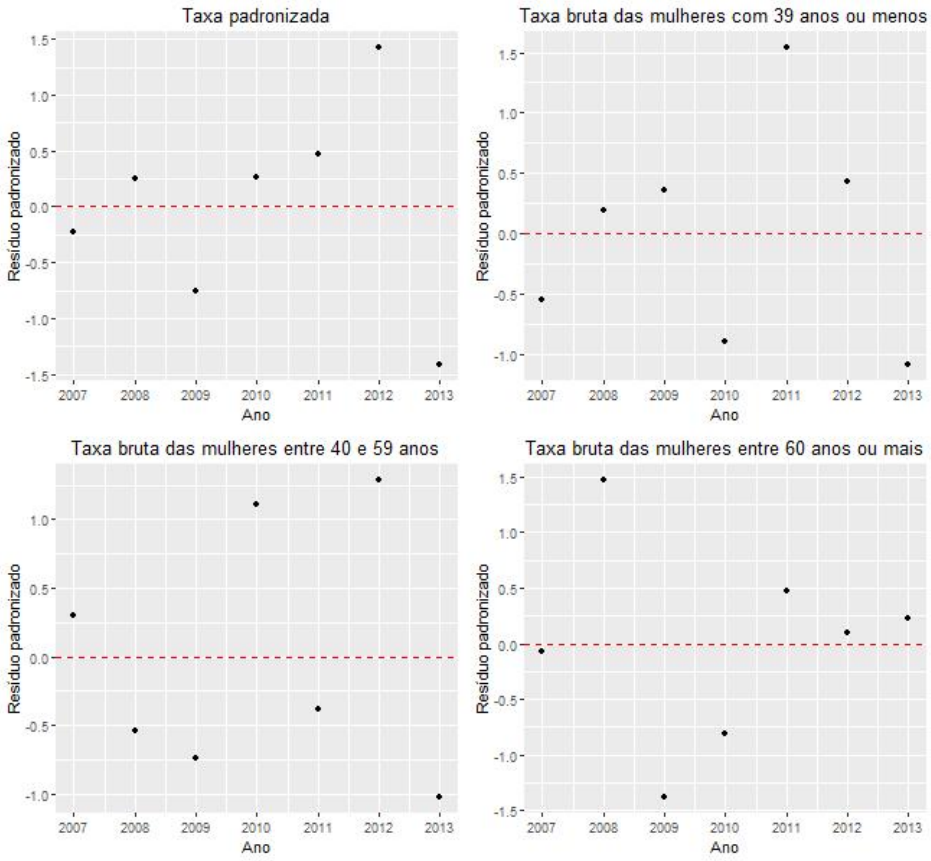
A Figura 3.1 ilustra estes resultados da Tabela 3.2.

Figura 3.1 – Gráfico das séries das incidências, padronizadas e não padronizadas, 2007-2013.



Na Figura 3.2 estão ilustrados os resíduos dos ajustes das séries das incidências padronizadas e brutas (não padronizadas).

Figura 3.2 – Gráficos dos resíduos dos ajustes das regressões das taxas padronizadas e brutas.



Através dos gráficos acima, verifica-se que não há indícios de não adequação dos ajustes realizados dos modelos de regressão.

Na Tabela 3.3 a seguir tem-se a distribuição dos casos segundo a faixa etária, escolaridade, raça/cor, classificação histológica e procedência.

Tabela 3.3 – Distribuição do número e porcentagem dos casos de câncer de colo do útero, segundo a faixa etária, escolaridade, raça/cor, histologia e procedência, das pacientes atendidas pela Liga Norte Riograndense Contra o Câncer, de 2007 à 2013.

Variável	**Categoria**	**Nº**	**%**
Faixa etária	≤ 39	534	32,38
	$40 - 59$	660	40,02
	≥ 60	455	27,60
Escolaridade	Sem escolaridade	286	17,34
	Até o Fundamental	605	36,69
	Ensino Médio	184	11,16
	Ensino Superior	54	3,28
	Sem informação	520	31,53
Raça/Cor	Branca	393	23,83
	Não branca	1205	73,07
	Sem informação	51	3,10
Classificação histológica	Carcinoma Epidermoide de Células Escamosas	1486	90,12
	Adenocarcinoma	163	9,88
Procedência	Natal ou municípios limítrofes	526	31,90
	Interior do Estado	687	41,66
	Sem informação	436	26,44

A Tabela 3.4 apresenta o tempo mediano e as probabilidades acumuladas de sobrevida segundo faixa etária, escolaridade, raça/cor, classificação histológica e procedência. O tempo mediano de sobrevida obtido para todos os casos foi de 17,13 meses, com probabilidade acumulada após 5 anos de 76,6%. As probabilidades acumuladas de sobrevida não tiveram diferença estatisticamente significativa nas comparações entre as categorias das variáveis raça/cor e procedência, com $p = 0{,}480$ e $p = 0{,}254$, respectivamente. Por outro lado, foram obtidas diferenças estatisticamente significativas nas comparações das curvas de sobrevida das faixas etárias, da escolaridade e da classificação histológica, com $p = 0{,}000$, $p = 0{,}021$ e $p = 0{,}011$, respectivamente. Assim, com relação as faixas etárias, a menor das probabilidades acumuladas de sobrevida é das pacientes acima de 60 anos; com respeito a escolaridade, as pacientes com ensino superior apresentam as menores probabilidades acumuladas de sobrevida e, com relação a classificação histológica, o adenocarcinoma tem a curva de sobrevida inferior a do carcinoma epidermoide de células escamosas.

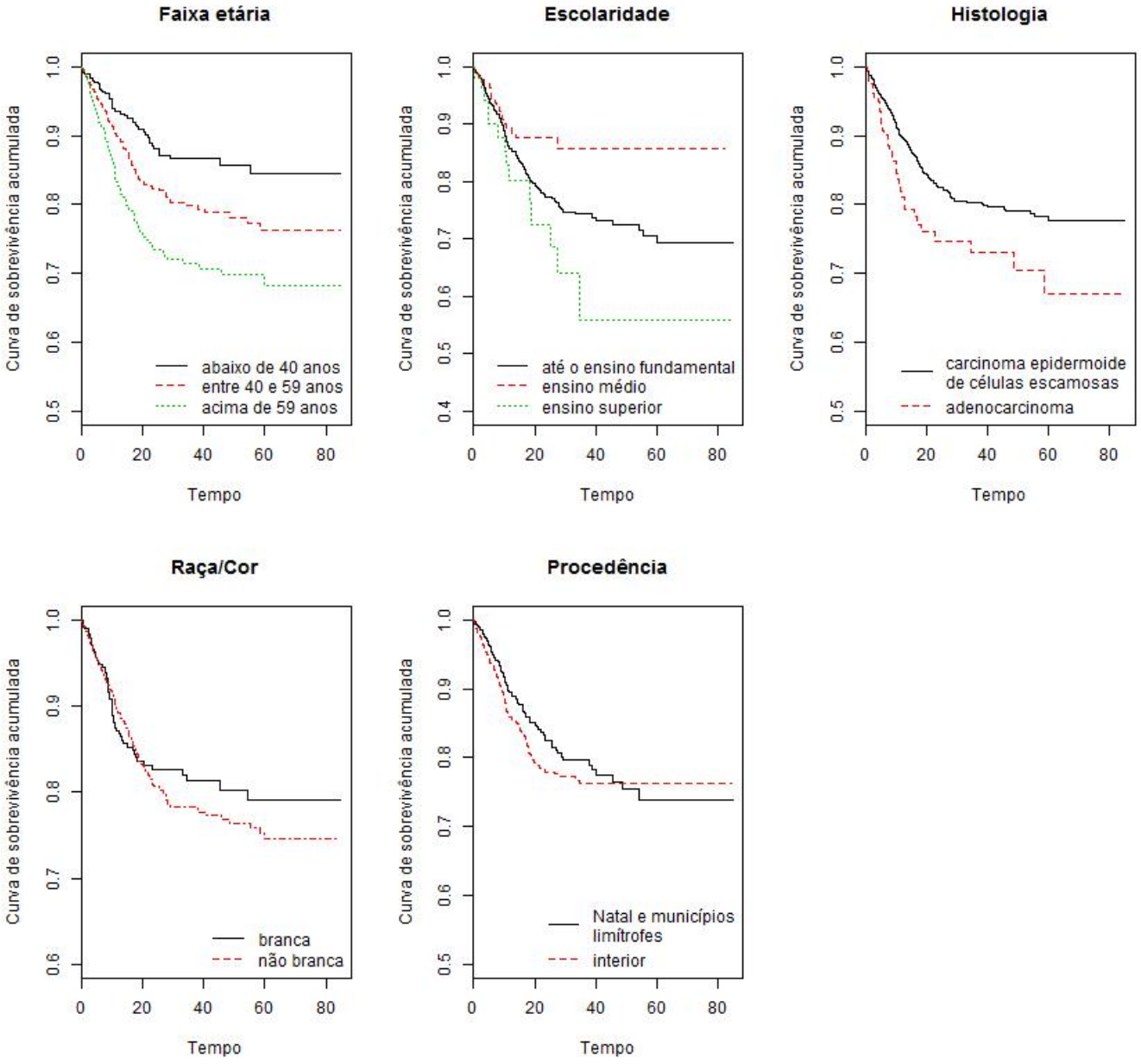
Tabela 3.4 – Tempo mediano e probabilidades acumuladas de sobrevida para o câncer de colo do útero, segundo variáveis do estudo, de 2007 a 2013.

Variável	**Tempo mediano de sobrevida**	**Probabilidade acumulada de sobrevida**				***p****
		12 meses	**24 meses**	**36 meses**	**60 meses**	
Faixa etária						0,000
≤ 39	17,71	0,932	0,876	0,857	0,845	
$40 - 59$	17,95	0,893	0,821	0,793	0,763	
≥ 60	15,90	0,825	0,731	0,707	0,682	
Escolaridade						0,021
Até o fundamental	15,20	0,855	0,770	0,738	0,715	
Médio	16,23	0,885	0,858	0,858	0,858	
Superior	13,85	0,764	0,685	0,560	0,560	
Raça/Cor						0,480
Branca	17,60	0,868	0,820	0,803	0,791	
Não branca	16,73	0,890	0,806	0,779	0,746	
Classificação histológica						0,011
Carcinoma Epidermoide de Células Escamosas	17,60	0,896	0,823	0,801	0,777	
Adenocarcinoma	14,70	0,801	0,731	0,705	0,669	
Procedência						0,254
Natal ou municípios limítrofes	13,20	0,892	0,821	0,790	0,741	
Interior do estado	12,57	0,857	0,776	0,762	0,762	
Total	17,13	0,888	0,816	0,794	0,766	0,000

* p-valor do teste *logrank* para igualdade entre estratos.

Através da Figura 3.3 a seguir é possível observar o comportamento das curvas de sobrevidas estimadas para cada covariável. Observa-se que as curvas para Raça/Cor e Procedência se entrelaçam ao longo do tempo, diferentemente das curvas das covariáveis Faixa etária, Escolaridade e Histologia, estas últimas indicando proporcionalidade entre suas funções de risco, ou seja, para estas últimas covariáveis há evidências de paralelismo entre suas curvas de sobrevida.

Figura 3.3 – Gráficos das curvas de sobrevivência das covariáveis consideradas no estudo.



Através do modelo de regressão de Cox foram feitas estimativas a fim de verificar os riscos associados as variáveis faixa etária, escolaridade e classificação histológica, de acordo com os resultados da Tabela 3.5 a seguir.

Tabela 3.5 – Fatores prognósticos definidos pelas regressões de Cox, univariada e múltipla.

Variável	**Regressão univariada**			**Regressão múltipla**		
	Razão de Risco (RR)	**IC$_{95\%}$(RR)**	***p***	**Razão de Risco (RR)**	**IC$_{95\%}$(RR)**	***p***
Faixa etária						
≤ 39	1,000	–	–	1,000	–	–
$40 - 59$	1,369	[0,934 – 2,009]	0,108	1,286	[0,873 – 1,894]	0,203
≥ 60	1,935	[1,321 – 2,835]	0,001	1,797	[1,213 – 2,661]	0,003
Escolaridade						
Até o fundamental	1,717	[1,067 – 2,762]	0,026	1,495	[0,918 – 2,432]	0,106
Ensino médio	1,000	–	–	1,000	–	–
Superior	2,511	[1,259 – 5,008]	0,009	2,352	[1,177 – 4,698]	0,015
Classificação histológica						
Carcinoma epidermóide	1,000	–	–	1,000	–	–
Adenocarcinoma	1,480	[0,997 – 2,197]	0,052	1,425	[0,957 – 2,121]	0,081

Para estas análises não foram considerados os casos em que não havia a informação sobre a escolaridade da paciente, o que corresponde a 520 casos, conforme os resultados da Tabela 3.3. Na regressão univariada foram obtidas como significativas as variáveis faixa etária e escolaridade. A variável classificação histológica não é significativa ao nível de 5% (p-valor = 0,052), porém, seria significativa caso fosse considerado um nível de significância de 6%, por exemplo. A faixa etária aparece como de valor prognóstico (ou variável preditora independente de sobrevida), por ser significativa nas regressões univariada e múltipla. Ou seja, através da análise múltipla verifica-se, com relação a idade, que as pacientes com 60 anos ou mais têm 1,797 vezes mais risco de morrer do que aquelas com 39 anos ou menos. Observa-se que a variável escolaridade também é preditora independente de sobrevida. Isto é, através da análise múltipla verifica-se que as pacientes com ensino superior completo têm 2,352 vezes mais risco de morte do que aquelas com ensino médio. Por outro lado, a variável classificação histológica não é significativa ao nível de 5%, porém, caso fosse considerado um nível de significância de 10%, esta também seria considerada como variável de valor prognóstico.

4 Considerações Finais

De acordo com o modelo de Cox, ao nível de significância de 5%, as variáveis faixa etária e escolaridade foram consideradas como variáveis preditoras independentes de sobrevida. As pacientes com 60 anos ou mais tem cerca de 2 vezes mais risco de morrer do que as pacientes com 39 anos ou menos. Quanto a variável escolaridade, estima-se que as mulheres com ensino superior tem mais de duas vezes risco de morte do que as mulheres com ensino médio.

A série da taxa de incidência para mulheres com 39 anos ou menos apresentou tendência crescente, enquanto a taxa para as mulheres com 60 anos ou mais mostrou tendência decrescente. A série geral das incidências padronizadas e a série das incidências brutas para a faixa etária entre 40 e 59 anos apresentaram estabilidade.

Nossos resultados sugerem a necessidade de implementação de ações preventivas tais como melhorias no programa de rastreio para o diagnóstico e tratamento precoce das lesões precursoras do câncer do colo do útero, visando reduzir a sua incidência. Além disso, são necessárias políticas públicas que melhore o acesso das mulheres ao tratamento adequado e o acompanhamento após o tratamento. Tais medidas são necessárias para aumentar a sobrevida e tornar decrescentes as séries que apresentaram estabilidade ou tendência crescente das taxas de incidência.

Referências

COX, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 34, n. 2, p. 187–202, 1972.

COX, D. R. Partial likelihood. *Biometrika*, Oxford University Press, v. 62, n. 2, p. 269–276, 1975.

COX, D. R.; HINKLEY, D. V. Theoretical statistics. 1974. *New York: Chapman&Hall/CRC*.

DOLL, R.; HILL, A. B. Mortality of british doctors in relation to smoking: observations on coronary thrombosis. *National cancer institute monograph*, v. 19, p. 205–268, 1966.

INSTITUTO NACIONAL DE CÂNCER JOSÉ ALENCAR GOMES DA SILVA. *Estimativa 2016: Incidência de câncer no Brasil*. 2016. Disponível em: <http://www.inca.gov.br/wcm/dncc/2015/index.asp>. Acesso em: 27 nov. 2018.

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, Taylor & Francis, v. 53, n. 282, p. 457–481, 1958.

MANTEL, N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep*, v. 50, p. 163–170, 1966.

Microsoft Excel. [S.l.], 2016.

MUHAMAD, N. A. et al. Survival rates of cervical cancer patients in malaysia. *Asian Pacific journal of cancer prevention: APJCP*, v. 16, n. 7, p. 3067–72, 2015.

ORDIKHANI, F. et al. Drug delivery approaches for the treatment of cervical cancer. *Pharmaceutics*, Multidisciplinary Digital Publishing Institute, v. 8, n. 3, p. 23, 2016.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2018. Disponível em: <https://www.R-project.org/>.

SEGI, M. et al. The age-adjusted death rates for malignant neoplasms in some selected sites in 23 countries in 1954-1955 and their geographical correlation. *The Tohoku journal of experimental medicine*, Tohoku University Medical Press, v. 72, n. 1, p. 91–103, 1960.

TORRE, L. A. et al. Global cancer statistics, 2012. *CA: a cancer journal for clinicians*, Wiley Online Library, v. 65, n. 2, p. 87–108, 2015.