Universidade de Brasília - UnB
Instituto de Ciências - IE
Departamento de Estatística - EST

Aplicação de Modelos de Regressão Logística em um Estudo de Neurocirurgia

TAIAN CRISTAL FERREIRA SALLES

Orientador: Professor Eduardo Freitas

Brasília
2018

Sumário

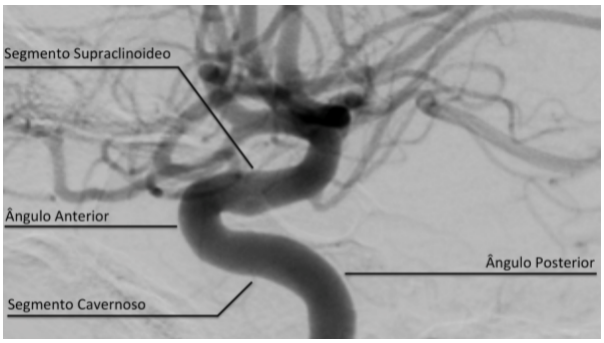
1 Introdução

Nos dias atuais com o avanço e disponibilidade da tecnologia, a medicina utiliza em suas pesquisas conhecimentos estatísticos que facilitam na hora de justificar e prevenir problemas biológicos. Este estudo vai tratar de um problema encontrado por pesquisadores da área médica que obtêm seus dados de forma binária: estimação de parâmetros e obtenção de estimativas adequadas e confiáveis.

Será analisado o problema de um Neurocirurgião que deseja saber se existe ou não uma relação entre o ângulo do sifão anterior da artéria cerebral com o rompimento do aneurisma cerebral (AC).

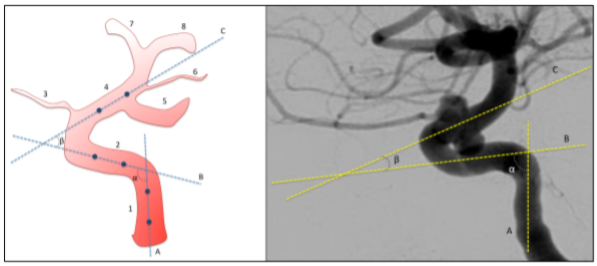
Aneurismas cerebrais são lesões adquiridas, caraterizadas como dilatações da parede das artérias intracranianas. Trata-se de uma patologia geralmente assintomática que ataca de 5 a 10 por cento da população, sendo mais comum em mulheres, o que pode ser justificado pela decorrente alteração hormonal, com pico de incidência entre 50 e 60 anos de idade.

O sifão carotídeo merece especial atenção já que constitui a porta de entrada do fluxo sanguíneo da circulação cerebral anterior e é característico por ser a porção mais sinuosa da carótida interna tornando-se assim o albergueiro de cerca de 80 por cento dos aneurismas cerebrais. A principal função fisiológica dessas sucessivas curvaturas é a atenuação da força vetorial do fluxo sanguíneo.



Admite-se que a perda de energia cinética do fluxo sanguíneo linear, ao colidir com a parede endotelial das curvaturas, forçando a mudança de direção do fluxo sanguíneo e transformando o fluxo normalmente linear em helicoidal, esteja relacionada à transformação endotelial.

O ângulo do sifão anterior da artéria cerebral é medido da seguinte forma:



Será utilizado um modelo de regressão logística para investigar uma possível relação entre o ângulo do sifão anterior da artéria cerebral e a ruptura do aneurisma cerebral.

2 Objetivos

2.1 Objetivo Geral

Aplicar regressão logística em um banco de dados na área de neurocirurgia.

2.2 Objetivos Específicos

Estudar o modelo de Regressão Logística Múltipla com resposta dicotômica.

Verificar a associação do tamanho do ângulo do sifão anterior da artéria cerebral com o rompimento do aneurisma cerebral.

3 Metodologia

3.1 Regressão Logística Simples

A preocupação aqui é estudar a relação entre a variável resposta e uma ou mais variáveis explicativas. Quando as variáveis respostas assumem apenas dois valores é comum utilizarmos a regressão logística. A diferença principal entre a regressão logística e a regressão linear é que na logística a variável resposta é binária, o que reflete no processo de estimação e nas suposições do modelo.

3.1.1 Modelo

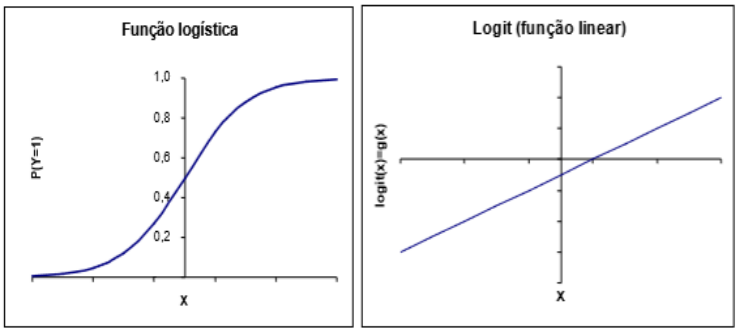
O interesse aqui é achar o valor médio da variável resposta dado o valor da variável explicativa, denotado por $E(Y/x)$. E no caso da regressão logística na qual a variável resposta é dicotômica esse valor médio é uma proporção que varia de 0 a 1, ou seja, $0 < E(Y/x) < 1$. A curva de $E(Y/x)$ tem forma de S, pois conforme a variável explicativa diminui $E(Y/x)$ gradativamente se aproxima de 0 e quando a variável explicativa aumenta $E(Y/x)$ gradativamente se aproxima de 1.

Para a modelagem desse tipo de curva escolhe-se a distribuição logística. A notação utilizada é $\pi(x) = E(Y/x)$ para representar a média condicional de Y dado x e ela é definida como:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

O logaritmo da média condicional é o tópico de interesse no estudo e ela é dada por:

$$g(x) = ln\frac{\pi(x)}{1-\pi(x)} = \beta_0 + \beta_1 x$$



A variável resposta Y pode ser escrita como $Y = \pi(x) + \varepsilon$, onde ε é o erro. Para uma variável resposta dicotômica Y, ε assume um de dois valores possíveis. Se $y = 1$, $\varepsilon = 1 - \pi(x)$ com probabilidade $\pi(x)$, e se $y = 0$, $\varepsilon = -\pi(x)$ com probabilidade igual a $1 - \pi(x)$. Disso ε tem distribuição com média 0 (zero) e variância $\pi(x)*(1-\pi(x))$. Portanto, $Y|x$ segue uma Binomial com probabilidade $\pi(x)$.

3.1.2 Estimação dos Parâmetros

Considerando uma amostra com n observações independentes do par (yi, xi), $i = 1, 2, ..., n$, onde yi é o valor da i-ésima variável resposta binária e xi o valor da i-ésima variável explicativa, é necessário estimar os valores de β_0 e β_1 para ajustar um modelo. O método da máxima verossimilhança encontra estimadores que maximizam a probabilidade de obter os dados observados da amostra.

Como, de pressuposto, as observações são independentes, a função de verossimilhança é obtida pelo produtório das contribuições de cada par (yi, xi) indicado acima. Então, tem-se que:

$$L(\beta) = \prod \pi(xi)^{yi}[1 - \pi(xi)]^{1-yi}$$

O objetivo é estimar β que maximize a equação acima. Então utilizamos o *log* da verossimilhança que é definido por:

$$l(\beta) = ln(L(\beta)) = \sum_{i=1}^{n} \{yi\, ln(\pi(x)) + (1 - yi)\, ln(1 - \pi(x))\}$$

Logo depois deriva-se $l(\beta)$ em relação a β_0 e β_1 e iguala-se o resultado a zero. Chegando as duas equações:

$$\sum_{i=1}^{n} [yi - \pi(xi)] = 0$$

e

$$\sum_{i=1}^{n} x_i\, [y_i - \pi(x_i)] = 0$$

Os valores de β são estimadores de máxima verossimilhança, $\hat{\beta}$.

3.1.3 Teste de Significância do Estimador

Depois de estimado β, é interessante saber se a variável que teve o β estimado é relevante ou não na análise, ou seja, se o modelo com a variável explicativa em questão explica a variável resposta melhor que o do modelo sem ela.

Os métodos em regressão logística seguem o mesmo princípio que em regressão linear: comparar os valores observados com os valores preditos da variável resposta. Essa comparação é baseada na função de verossimilhança e é amplamente conhecida como teste da razão de verossimilhança, que é baseada na função abaixo:

$$D = -2\,ln\left[\frac{verossimilhanca\,do\,modelo\,ajustado}{verossimilhanca\,do\,modelo\,saturado}\right]$$

$$= -2\sum_{i=1}^{n}\left[yi\,ln\left(\frac{\hat{\pi}}{yi}\right) + (1-yi)\,ln\left(\frac{1-\hat{\pi}_i}{1-yi}\right)\right]$$

Onde $\pi_i = \pi(xi)$.

Para avaliar a significância de uma variável explicativa, comparam-se os valores de D com e sem a variável em questão e verifica se é significativo. Para isso usa-se:

$$G = D(modelo\,sem\,a\,variavel) - D(modelo\,com\,a\,variavel)$$

$$= -2\,ln\left[\frac{verossimilhanca\,sem\,a\,variavel}{verossimilhanca\,com\,a\,variavel}\right]$$

Sob $H_0 : \beta_1 = 0$, G segue uma Qui-Quadrado com 1 grau de liberdade.

3.1.4 Teste de Wald

O teste de Wald é obtido pela comparação do estimador de máxima verossimilhança β com a estimação do seu erro.

$$W = \frac{\hat{\beta}_1}{\hat{SE}(\hat{\beta}_1)}$$

Onde $\hat{SE}(\hat{\beta}_1)$ é a estimativa do erro padrão do parâmetro estimado. Sob H_0: $\beta_1 = 0$, W segue uma normal padrão.

3.1.5 Intervalo de Confiança

Em determinados casos é interessante formular intervalos de confiança para $\hat{\beta}$. A base para construção deles é a mesma dos testes de significância, em particular o teste de Wald.

3.1.6 Estimação da Razão de Chances (Odds Ratio)

A interpretação dos parâmetros de um modelo de regressão logística é obtida comparando a probabilidade de sucesso com a probabilidade de fracasso, usando a função odds ratio - OR (razão de chances).

Utilizando-se um exemplo simples de medida de associação para tabelas de contingencia 2x2 temos:

Table 1: Tabela de Contingencia 2x2

	Y		
Grupo	Sucesso	Insucesso	Total
1	n_{11}	n_{12}	$n_{1.}$
2	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	N

Sendo:
π_1 - probabilidade de sucesso na linha 1 (Grupo 1)
π_2 - probabilidade de sucesso na linha 2 (Grupo 2)

A função *odds* é uma medida de associações dentro dos grupos relacionando sucessos com os fracassos e é dada por grupo:

$$odds_1 = \frac{\pi_1}{(1-\pi_1)}$$

$$odds_2 = \frac{\pi_2}{(1-\pi_2)}$$

Em qualquer das linhas a probabilidade de sucesso é uma função da *odds*.

$$odds = \frac{\pi}{1-\pi} \Leftrightarrow \pi = \frac{odds}{odds+1}$$

E quando $\pi_1 = \pi_2$ as *odds* satisfazem $odds_1 = odds_2$ e as variáveis são independentes. A odds ratio é a razão das "odds" de dois grupos (linhas):

$$\theta = \frac{odds_1}{odds_2} = \frac{\frac{\pi_1}{(1-\pi_1)}}{\frac{\pi_2}{(1-\pi_2)}}$$

3.2 Regressão Logística Múltipla

Como visto até agora, foi introduzida a regressão logística no caso univariado. Porém, a confiabilidade de uma técnica de modelagem consiste em utilizar quantas variáveis forem necessárias, inclusive variáveis em diferentes escalas de mensuração. A abordagem de estimação e modelagem seguirá o mesmo raciocínio utilizado na regressão logística simples.

3.2.1 Modelo

Considere o conjunto de p variáveis independentes descritas pelo vetor $x = (x_1, x_2, ., x_p)$ e a probabilidade condicional da variável resposta estar presente é denotada por $\pi(x) = P(Y = 1|x)$. O logito do modelo de regressão logística é dado pela equação:

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$$

tal que o modelo de regressão logística fica:

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}}$$

No caso da inclusão de variáveis de escala nominal não é correto usá-las como se fossem variáveis de escala intervalar. Os números usados para representa-las não possuem nenhuma significância numérica, eles são apenas identificadores. A maioria dos softwares estatísticos geram as variáveis identificadoras quando indicadas as variáveis com escala nominal. Em geral, se a variável de escala nominal possui k categorias, será necessário o uso de $k-1$ variáveis indicadoras para a variável em estudo.

3.2.2 Estimação dos Parâmetros

O método usado para estimação dos parâmetros será o mesmo do caso univariado, o método da máxima verossimilhança. A função de verossimilhança é a mesma da regressão logística simples com o fato de que $\pi(x)$ é definido como:

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}}$$

Como agora, o estudo é generalizado para o caso multivariado, olha-se para a abordagem do erro padrão dos estimadores com maiores detalhes.

O método de estimação das variâncias e covariâncias dos coeficientes estimados vem de uma teoria de estimação por máxima verossimilhança, segundo Rao (1973). Essa teoria mostra que os estimadores são obtidos da matriz de segundas derivadas parciais da função de *log* verossimilhança e são da forma:

$$\frac{\partial^2 L(\beta)}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij}^2 \, \pi_i \, (1 - \pi_i)$$

e

$$\frac{\partial^2 L(\beta)}{\partial \beta_j \beta_l} = -\sum_{i=1}^{n} x_{ij} \, x_{il} \, \pi i \, (1 - \pi_i)$$

para j, $l = 0, 1, 2, ..., p$ onde π_i denota $\pi(x_i)$. A matriz $(p+1)x(p+1)$ que contém o negativo dos termos das equações acima será denotada por $I(\beta)$, que é a matriz de informação observada. As variâncias e covariâncias são obtidas da inversa da matriz $I(\beta)$.

Uma formulação da matriz de informação que será útil na discussão e avaliação de modelagem é:

$$\hat{l}(\hat{\beta}) = X'VX$$

onde X é uma matriz n por $(p+1)$ contendo os dados de cada variável explicativa e V é a matriz n por n com diagonal $\hat{\pi_i} * (1 - \hat{\pi})$. Ou seja, a matriz X é:

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

e a matriz V é:

$$V = \begin{bmatrix} \hat{\pi_1} * (1 - \hat{\pi_1}) & 0 & \dots & 0 \\ 0 & \hat{\pi_2} * (1 - \hat{\pi_2}) & \dots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & \dots & \dots & \hat{\pi_n} * (1 - \hat{\pi_n}) \end{bmatrix}$$

3.2.3 Teste de Significância do Estimador

A avaliação das variáveis explicativas que vão compor o modelo segue da mesma forma do caso univariado. O teste da razão de verossimilhança, que é baseado na estatística G, é usado para avaliar os p coeficientes das variáveis explicativas. A diferença é que os valores ajustados de $\hat{\pi}$ são baseados nos $(p+1)$

parâmetros. As hipóteses nulas são de que os p coeficientes das covariáveis são nulos e G segue uma distribuição qui-quadrado com p graus de liberdade.

A mesma abordagem é feita no Teste de Wald que compara a estimativa de máxima verossimilhança do parâmetro $\hat{\beta}_1$ com a estimativa do seu erro padrão. Tendo como hipótese um coeficiente ser igual a zero e a estatística W (fórmula abaixo) segue uma normal padrão.

$$W_j = \frac{\hat{\beta}_j}{\hat{SE}(\hat{\beta}_j)}$$

O problema do caso múltiplo é que o teste de Wald fornece estimativas individuais para os coeficientes, mas a variável esta decomposta em outras indicadoras, que aqui são variáveis diferentes. Portanto o teste similar, porém multivariado, é definido por:

$$W = \hat{\beta}'[\hat{Var}(\hat{\beta})]^{-1}\hat{\beta} = \hat{\beta}'(X'VX)\hat{\beta}$$

Que segue uma Qui-Quadrado sob a hipótese nula de que cada um dos p+1 coeficientes são iguais à zero.

3.2.4 Intervalo de Confiança

Os intervalos de $100\,(1-\alpha)\%$ de confiança para β_i e β_0 são:

$$\hat{\beta}_i \pm z_{1-\frac{\alpha}{2}}\,\hat{SE}(\hat{\beta}_i)$$

e

$$\hat{\beta}_0 \pm z_{1-\frac{\alpha}{2}}\,\hat{SE}(\hat{\beta}_0)$$

para $i = 1, 2, ..., p$.

Um meio de expressar o estimador *logito* é , onde $\hat{\beta}$ é o vetor dos $p+1$ coeficientes e x' é o vetor que representa as constantes e os valores das p covariáveis do modelo, onde $x_0 = 1$. Sabendo que:

$$\hat{Var}(\hat{\beta}) = (X'VX)^{-1}$$

Segue que:

$$\hat{Var}[\hat{g}(x)] = x'\hat{Var}(\hat{\beta})x = x'(X'VX)^{-1}x$$

3.2.5 Interpretação do Modelo de Regressão Ajustado

Para estudar essa seção supõe-se que a regressão logística foi ajustada e que as variáveis do modelo são estatisticamente significantes. A interpretação envolve a associação entre a variável resposta e a variável independente, e definir adequadamente a unidade de mudança para a variável independente. Em regressão logística, o coeficiente angular indica a mudança no *logito* correspondente à mudança de uma unidade de variável independente, ou seja:

$$\beta_1 = g(x+1) - g(x).$$

3.3 Variável Independente Dicotômica

Começamos as considerações para interpretação dos coeficientes da regressão logística quando a variável independente tem escala nominal e dicotômica (duas possibilidades). Neste caso assumimos que a variável independente x é ou zero ou um e a diferença do *logito* caso $x = 0$ ou $x = 1$ é:

$$g(1) - g(0) = [\beta_0 + \beta_1] - [\beta_0] = \beta_1$$

Algebricamente essa equação é bastante simples e a diferença dos *logitos* é exatamente o β_1. A *odds* dos resultados para $x = 1$ é definida por $\pi(1)/[1-\pi(1)]$ e em paralelo a *odds* dos resultados para $x = 0$ é definida por $\pi(0)/[1 - \pi(0)]$. A razão de chances, denotada por OR (Odds Ratio), é definida pela razão entre os *odds* quando $x = 1$ e quando $x = 0$ e é dada pela equação:

$$OR = \frac{\pi(1)/[1 - \pi(1)]}{\pi(0)/[1 - \pi(0)]}$$

Substituindo as expressões pelos logitos do modelo de regressão logística obtemos:

$$OR = \frac{\left(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right)}{\left(\frac{1}{1+e^{\beta_0+\beta_1}}\right)} \Bigg/ \frac{\left(\frac{e^{\beta_0}}{1+e^{\beta_0}}\right)}{\frac{1}{1+e^{\beta_0}}}$$

$$OR = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}}$$

$$= e^{(\beta_0+\beta_1)-\beta_0}$$

$$= e^{\beta_1}$$

Então a relação entre a variável independente dicotômica (0 ou 1) e a razão de chances do coeficiente da regressão é:

$$OR = e^{\beta_1}$$

Essa simples relação entre o coeficiente e a *odds ratio* é a principal razão pela qual a regressão logística provou ser uma ferramenta de pesquisa analítica tão poderosa.

A *odds ratio* é uma medida de associação de amplo uso, especialmente em epidemiologia, pois aproxima o quanto mais provável (ou improvável) é o resultado estar presente entre aqueles com $x = 1$ do que entre aqueles com $x = 0$. Por exemplo, se y denota a presença ou ausência de câncer de pulmão e se x denota se a pessoa é fumante, então $OR^{=}2$ estima que o câncer de pulmão tem duas vezes mais chances de ocorrer entre fumantes do que entre não fumantes da população estudada.

A interpretação dada para a *odds ratio* é baseada no fato de que em muitos casos ela se aproxima do risco relativo. Isso ocorre quando as probabilidades de sucesso no grupo 1 e no grupo 2 são muito próximas de zero fazendo com que $[1 - p_2]/[1 - p_1]$ se aproxime de um.

$$OR = RR\left(\frac{1-p_2}{1-p_1}\right)$$

Então:

$$OR = RR$$

A *odds ratio* é geralmente o parâmetro de maior interesse na regressão logística devido a sua facilidade de interpretação. Junto com a estimativa pontual é interessante fazer o intervalo de confiança para acrescentar informação a ela:

$$exp\left[\hat{\beta}_1 \pm z_{1-\frac{\alpha}{2}}\, \hat{SE}(\hat{\beta}_1)\right]$$

3.4 Variável Independente Politômica

Variáveis qualitativas nominais são aquelas que mais limitam a possibilidade de utilização de técnicas estatísticas, especialmente quando o número de categorias excede dois. Em geral, se a variável qualitativa tem p níveis possíveis, então são necessárias $p - 1$ variáveis "dummy". As variáveis serão designadas como Du e os coeficientes dessas variáveis serão designados por β_u, $u = 1, ..., p-1$. E o modelo expresso em termos da função *logito* é:

$$logito(\pi) = ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \sum_{u=1}^{p-1} \beta_u D_u$$

Este método faz todas as variáveis *dummy* iguais à zero para o nível considerado de referência e então aloca uma única *dummy* igual a um para cada um dos outros níveis de X.

3.5 Variável Independente Contínua

Quando o modelo de regressão logística contém uma variável independente contínua, a interpretação do coeficiente estimado depende de como ela entrou no modelo e a unidade particular da variável. Supondo que o *logito* é linear na variável contínua X, a equação do *logito* é:

$$logito(\pi) = \beta_0 + \beta_1 X$$

O coeficiente β_1 fornece a mudança no *log* da chance para um aumento de uma unidade de X. Ou seja:

$$\beta_1 = logito(x+1) - logito(x)$$

O *log* da chance para uma mudança de c unidades em X é obtida pela diferença dos *logitos*:

$$logito(X+c) - logito(X) = c\beta_1$$

E a razão de chances é dada por:

$$exp(c\beta_1)$$

E o intervalo de confiança com $(1-\alpha)$ de confiança é dado por:

$$exp\left[c\hat{\beta}_1 \pm z_{1-\frac{\alpha}{2}} c\hat{EP}(\hat{\beta}_1)\right]$$

3.6 Estimação da Razão de Chances quando há Presença de Interação

Um método adequado que levará ao estimador correto baseado no modelo possui três passos. O primeiro é escrever o *logito* do fator de risco nos dois níveis que serão comparados; o segundo simplifica algebricamente a diferença entre esses *logitos*; e, finalmente, aplicar exponencial no valor do segundo passo.

Para facilitar aplica-se o método acima apenas com duas variáveis e sua interação. O fator de risco será representado por F e a variável por X. O *logito* para o modelo avaliado quando $F = f$ e $X = x$ é:

$$g(f,x) = \beta_0 + \beta_1 f + \beta_2\, x + \beta_3 f\, x$$

O almejado é a razão de chances comparando dois níveis de F, f_1 e f_0, onde $X = x$. Seguindo os procedimentos temos:

$$g(f_1, x) = \beta_0 + \beta_1 f_1 + \beta_2\, x + \beta_3 f_1\, x$$

e

$$g(f_0, x) = \beta_0 + \beta_1 f_0 + \beta_2\, x + \beta_3 f_0\, x$$

Depois obtêm-se o *log* da razão de chances:

$$ln\left[OR(F = f_1, F = f_0, X = x)\right] = g(f_1, x) - g(f_0, x)$$

$$= (\beta_0 + \beta_1 f_1 + \beta_2\, x + \beta_3 f_1\, x) - (\beta_0 + \beta_1 f_0 + \beta_2\, x + \beta_3 f_0\, x)$$

$$= \beta_1(f_1 - f_0) + \beta_3 x(f_1 - f_0)$$

No último passo aplica-se a exponencial no valor encontrado anteriormente:

$$OR = exp[\beta_1(f_1 - f_0) + \beta_3 x(f_1 - f_0)]$$

O estimador da razão de chances é obtido substituindo os parâmetros pelos seus estimadores.

Para obter o intervalo de confiança para o estimador da razão de chances encontrado a abordagem é igual para os modelos sem interação. Usando métodos para calcular a variância de uma soma, o estimador da variância é:

$$\hat{Var}\left\{ln[\hat{OR}(F = f_1, F = f_0, X = x)]\right\}$$
$$= (f_1 - f_0)^2\, \hat{Var}(\hat{\beta}_1)[x(f_1 - f_0)]^2\, \hat{Var}(\hat{\beta}_3) + 2x(f_1 - f_0)^2\, \hat{Cov}(\hat{\beta}_1, \hat{\beta}_3)$$

Substituindo os estimadores da variância e covariância encontra-se o estimador da variância do *log* da razão de chances. O intervalo com $100x(1-\alpha)\%$ de confiança para o *log* da razão de chances é:

$$[\hat{\beta}_1(f_1 - f_0) + \hat{\beta}_3 x(f_1 - f_0)] \pm z_{1-\frac{\alpha}{2}}\, \hat{SE}\left\{ln[\hat{OR}(F = f_1, F = f_0, X = x)]\right\}$$

Para obter o intervalo para a razão de chances basta aplicar a exponencial na equação anterior.

4 Banco de Dados

O estudo foi realizado em pacientes com aneurisma cerebral que foram atendidos entre Janeiro de 2007 e Dezembro de 2016 no departamento de Neurorradiologia Intervencionista do Centro Hospitalar Universitário de Limoges – França. Todos os pacientes deveriam ter sido submetidos a angiografias digitais com substração pré e pós-operatórias com padrões previamente estabelecidos para avaliação do sifão carotídeo.

Os critérios para exclusão foram: 1. Aneurismas da circulação posterior; 2. Pacientes com angiografia digital fora dos padrões; 3. Documentos de arquivos incompletos; 4. Impossibilidade de seguimento; 5. Ausência de assinatura do termo de consentimento livre e informado; 6. Negativa em participação do estudo.

Durante esse período (2007 a 2016), 703 pacientes com aneurisma cerebral foram tratados no referido departamento de Neurorradiologia dos quais 640 pacientes preencheram os critérios de inclusão e foram selecionados para esse estudo. Foram analisados um total de 755 aneurismas intracranianos (podendo haver mais de um aneurisma em um único paciente), sendo 692 aneurismas da circulação anterior e 63 da circulação posterior (excluídos do estudo conforme os critérios de exclusão). O que nos dá uma amostra com 692 observações dentro dos padrões estabelecidos.

As informações que foram coletadas dos 692 casos de aneurismas cerebrais foram o sexo do paciente, a idade do mesmo, a localização do aneurisma (sifão proximal, sifão distal e pós sifão), se houve ou não a ruptura da artéria, o tamanho da lesão (medida em milímetros), o ângulo anterior (medida em graus) e o ângulo posterior (também medido em graus).

5 Resultados

A análise do banco de dados será feita utilizando-se o software estatístico *Statistical Analysis System – SAS*. Toda a programação estará disponível em anexo neste trabalho. Primeiramente será feita uma análise descritiva dos dados e então um modelo de regressão logística será ajustado.

As variáveis consideradas na análise serão: sexo, idade, localização do aneurisma, ruptura, tamanho da lesão, ângulo anterior e ângulo posterior. Totalizando 7 variáveis utilizadas no estudo.

5.1 Análise Descritiva

Para a análise descritiva das variáveis foi utilizado o procedimento $survey freq$ no SAS que auxilia na interpretação dos resultados. Esse procedimento estima a frequência populacional, fornece intervalos de confiança de cada variável em uma tabela que facilita a compreensão. Para as variáveis mais relevantes serão feitos alguns breves comentários a respeito dos resultados.

5.1.1 Sexo

Table 2: Frequências da Variável Sexo

Sexo	Frequência	Percentual	Intervalo de Confiança de 95% para o percentual
Masculino	235	33.95954	30.4224; 37.4967
Feminino	457	66.04046	62.5033; 69.5776
Total	692	100.0000	

Na tabela pode-se analisar que dos pacientes portadores da lesão a maioria é do sexo feminino com 66% da amostra ($sexo = 1$). Foram constatados 457 aneurismas cerebrais femininos contra 235 masculinos.

5.1.2 Idade

Table 3: Frequências da Variável Idade

Idade	Frequência	Percentual	Intervalo de Confiança de 95% para o percentual
≤ 55	372	53.7572	50.0332; 57.4812
> 55	320	46.2428	42.5188; 49.9668
Total	692	100.0000	

Na amostra observamos pacientes com idades entre 17 e 85 anos de vida, mas é só a partir dos 40 anos que é constatada uma frequência maior de pacientes. A variável foi categorizada em dois grupos (≤ 55 e > 55) pra facilitar na visualização. O número 55 foi escolhido levando-se em consideração a mediana $(53,78)$ e a média $(54,59)$.

5.1.3 Localização do Aneurisma

Table 4: Frequências da Variável Localização

Localização	Frequência	Percentual	Intervalo de Confiança de 95% para o percentual
Pós Sifão	474	68.49711	65.0275; 71.9667
Sifão Distal	173	25.00000	21.7658; 28.2342
Sifão Proximal	45	6.50289	4.6612; 8.3446
Total	692	100.0000	

É nítido que nos pacientes observados o local mais frequente aonde o aneurisma se instala é o Pós Sifão $(68,49\%$ dos casos) e o local mais raro de se observar o aneurisma é o Sifão Proximal $(6,50\%)$. Será que há um local em que a energia cinética sanguínea é maior ou existe um revestimento arterial mais frágil?

5.1.4 Ruptura

Table 5: Frequências da Variável Ruptura

Ruptura	Frequência	Percentual	Intervalo de Confiança de 95% para o percentual
Sim	225	32.51445	29.0157; 36.0132
Não	467	67.48555	63.9868; 70.9843
Total	692	100.0000	

Sobre a variável ruptura o que se pode dizer é que felizmente os resultados apontam para a prevalência na não ruptura da artéria cerebral $(67,48\%$ da amostra). Tem-se dentre os pacientes 225 casos de ruptura contra 467 casos de não ruptura.

5.1.5 Tamanho da Lesão

Table 6: Frequências da Variável Tamanho da Lesão

Tamanho da Lesão	Frequência	Percentual	Intervalo de Confiança de 95% para o percentual
$x \leq 6$	418	60.4046	56.7518; 64.0574
$x > 6$	274	39.59	35.9426; 43.2482
Total	692	100.0000	

Para falar sobre o tamanho da lesão é importante explicitar primeiramente que essa variável foi medida em *milímetros*, que o mínimo registrado foi $1mm$ e que o máximo foi $52mm$. Logo em seguida, para facilitar a visualização, a variável foi categorizada em dois grupos ($x \leq 6\, e\, > 6$) a pedido do médico pesquisador. Observou-se que aneurismas com tamanho inferior a $6mm$ são mais frequentes (60.40% dos casos).

5.1.6 Ângulo Anterior

Table 7: Frequências da Variável Ângulo Anterior

Ângulo Anterior	Frequência	Percentual	Intervalo de Confiança de 95% para o percentual
≤ 15.40	346	50.0000	46.2654; 53.7346
> 15.40	346	50.0000	46.2654; 53.7346
Total	692	100.0000	

É interessante começar essa análise dizendo que o menor ângulo observado foi de 45° negativos e o maior ângulo foi 91° positivos. A partir dai pode-se analisar que a mediana foi de aproximadamente 15°, ou seja, metade das observações estão abaixo de 15° e a outra metade acima. A média observada foi de 14°36" mostrando uma simetria entre as observações.

5.1.7 Ângulo Posterior

Table 8: Frequências da Variável Ângulo Posterior

Ângulo Posterior	Frequência	Percentual	Intervalo de Confiança de 95% para o percentual
≤ 88.15	346	50.0000	46.2654; 53.7346
> 88.15	346	50.0000	46.2654; 53.7346
Total	692	100.0000	

Para dar início a análise descritiva da variável ângulo posterior observa-se que sua mediana se aproxima de 88° e sua média é aproximadamente 83°. Dentre os 692 pacientes observou-se que o menor ângulo sendo quase 0° e o maior sendo 173°. Para facilitar a análise a variável foi categorizada em dois grupos (≤ 88.15 e > 88.15).

5.1.8 Análise Bivariada

Na tabela abaixo encontram-se as prevalências de ruptura da artéria cerebral para cada variável independente. Observa-se que para a variável sexo a prevalência de ruptura é do gênero feminino, para a variável idade pacientes com idade igual ou inferior a 55 anos de idade, já para a variável tamanho da lesão destaca-se uma prevalência de ruptura para lesões iguais e inferiores a $6mm$. Tratando-se do valor do ângulo anterior constata-se que há prevalência de ruptura quando o ângulo é maior que 15.40°. Para rodar a regressão logística no SAS utilizou-se, na maior parte das variáveis, como resposta de referência as com maior prevalência. Os casos contrários podem ser justificados pelo próprio médico pesquisador.

Table 9: Prevalência de ruptura da artéria cerebral

Variáveis	Frequência	Prevalência de ruptura(%)	IC 95%
Sexo			
Feminino	144	34.46	28.37;40.55
Masculino	81	31.50	27.24;35.77
Idade			
≤ 55	126	33.87	29.04;38.69
> 55	99	30.93	25.86;36.01
Localização			
Pós Sifão	171	36.07	31.74;40.40
Sifão Distal	53	30.63	23.74;37.52
Sifão Proximal	1	2.22	0.00;6.53
Tamanho da Lesão			
≤ 6	147	35.16	30.57;39.75
> 6	78	28.46	23.11;33.82
Ângulo Anterior			
≤ 15.40	92	26.58	21.92;31.25
> 15.40	133	38.43	33.30;43.57
Ângulo Posterior			
≤ 88.15	110	31.79	26.87;36.71
> 88.15	115	33.23	28.26;38.21

5.2 Razão de chances do Modelo de Regressão Simples

Na tabela 10 se encontram as variáveis, a razão de chances associada à ela, seus respectivos intervalos de confiança e o $p - valor$ relativo à significância da variável em explicar a variável ruptura. Para isso foi gerado um modelo diferente para cada variável – regressão logística simples.

Table 10: Prevalência de ruptura da artéria cerebral

	Razão de Chances Bruta	
Variáveis	OR (IC 95%)	p-valor
Sexo		0.4316
Feminino	1	-
Masculino	1.143(0.819;1.596)	0.4316
Idade		0.4116
≤ 55	1.143 (0.830;1.574)	0.4116
> 55	1	-
Localização		0.0035
Pós Sifão	1	-
Sifão Distal	0.783 (0.539;1.137)	0.1985
Sifão Proximal	0.040 (0.006;0.295	0.0016
Tamanho da Lesão		0.0662
≤ 6	1.363 (0.980;1.897)	0.0662
> 6	1	-
Ângulo Anterior		0.0009
≤ 15.40	1	-
> 15.40	1.724 (1.249;2.379)	0.0009
Ângulo Posterior		0.6849
≤ 88.15	1	-
> 88.15	1.068 (0.777;1.468)	0.6849

Com a tabela acima (análise de regressão simples) é possível chegar a algumas conclusões sobre o efeito que cada variável tem na variável ruptura (admitindo-se as demais variáveis constantes). Primeiramente constatamos, através do $p - valor$, que apenas as variáveis localização e ângulo anterior são significantes, ou seja, essas variáveis são as que podem explicar melhor a variável dependente (ruptura).

Falando sobre a variável ângulo anterior, podemos dizer primeiramente que através da estimativa pontual da razão de chance $(1,724)$ – os pacientes que possuem ângulos anteriores superiores a $15,40°$ tem $1,724$ mais chances de romper o aneurisma do que aqueles que possuem esse ângulo menor que $15,40°$.

A respeito da variável localização é um tanto mais complicado visualizar, mas a interpretação é basicamente que quando o AC está localizado no Sifão Distal ($RC = 0.783$) a chance de haver ruptura é $21,7\%$ menor do que se ele estivesse localizado no Pós Sifão. Além disso, a tabela mostra que para a localização de Sifão Proximal ($RC = 0,040$) as chances de ruptura são 96% menores do que no Pós Sifão.

5.3 Regressão Logística Múltipla

Uma análise de regressão logística múltipla gerada nos proporciona muitas informações. Para facilitar a compreensão, a tabela 11 disponibiliza as variáveis consideradas, suas razões de chances junto aos seus intervalos de confiança e seus respectivos p-valores.

Table 11: Razão de Chances da Regressão Múltipla

	Razão de Chances da Regressão Múltipla	
Variáveis	OR (IC 95%)	p-valor
Sexo		0.9214
Feminino	1	-
Masculino	1.018 (0.718;1.443)	0.9214
Idade		0.6123
≤ 55	1.090 (0.781;1.520)	0.6123
> 55	1	-
Localização		0.0092
Pós Sifão	1	-
Sifão Distal	0.850 (0.578;1.250)	0.109
Sifão Proximal	0.047 (0.006;0.349	0.0035
Tamanho da Lesão		0.2587
≤ 6	1.219 (0.864;1.720)	0.2587
> 6	1	-
Ângulo Anterior		0.0047
≤ 15.40	1	-
> 15.40	1.610 (1.158;2.238)	0.0047
Ângulo Posterior		0.7888
≤ 88.15	1	-
> 88.15	1.046 (0.754;1.451)	0.7888

Com isso comprovou-se o que vimos na análise de regressão logística simples, as variáveis que estatisticamente apresentam efeito sobre a variável resposta – significativas estatisticamente – são as variáveis localização e ângulo anterior, dessa vez em uma análise multivariada.

Na hora de incluir as variáveis no modelo múltiplo deve ser considerada a significância estatística aferida pela razão de chances e seus respectivos intervalos de confiança, sendo incluídas aquelas variáveis que apresentaram $p < 0.05$. Razão de chances ajustadas e os respectivos intervalos de confiança foram obtidos.

É interessante observar que quando o valor um está contido no intervalo de confiança tem-se um p-valor superior a 0.05.

Portanto, o modelo final conta com as variáveis: localização do aneurisma cerebral (variável *dummy*) e ângulo anterior. Os valores dos coeficientes estimados do modelo encontram-se acima para cada variável.

6 Coclusão

Como o objetivo deste trabalho foi analisar a relação entre o ângulo anterior da artéria cerebral e a ruptura do AC, constatou-se indícios de que quanto maior for o ângulo anterior da artéria maior será a chance de que esse aneurisma se rompa. Tentando achar alguma explicação para tal informação encontrou-se como justificativa plausível constatar que os ângulos menores fazem com que a energia cinética do sangue (velocidade para correr nas artérias) diminua quando se faz uma "curva" muito fechada (ângulos menores) diminuindo as chances de ruptura do aneurisma cerebral.

Em paralelo há uma outra variável interessante de ser observada, a localização da lesão fornece indícios de que seria mais provável um rompimento do aneurisma caso ele esteja localizado no pós sifão, dando aos neurocirurgiões informações valiosas.

7 Programação

```
/*IMPORTAR DADOS*/

proc import out=mysasdata
            datafile=" /home/taiancristal0/dados_Taian.xlsx "
            dbms=xlsx replace;
   sheet='Plan1';
   run;


   data a ;
set mysasdata;
run;

/*FORMATAR DADOS*/

proc format;

value Sexo 1 = 'Feminino'
           0 = 'Masculino';

value Local 1 = 'Pós Sifão'
                    2 = 'Sifão Distal'
3 = 'Sifão Proximal';

value Ruptura 0 = 'Não'
          1 = 'Sim';

value Anterior low-15.40 = '<= 15.40'
 15.401-high = '> 15.40';

value Idade low-55 = '<= 55'
          55.0001-high = '> 55';

value Posterior low-88.15 = '<= 88.15'
 88.16-high = '> 88.15';

value Tamanho low-6 = '<= 6'
           6.001-high = '> 6';

run;
/*ANALISE DESCRITIVA*/
```

```
/*FREQUENCIAS*/

/*SEXO*/
proc surveyfreq data=a;
tables SEXO/cl row;
format sexo sexo.;
run;

PROC SGPLOT DATA = a;
VBAR sexo;
format sexo sexo.
RUN;

/*IDADE*/
proc surveyfreq data=a;
tables idade/cl row;
format idade idade.;
run;


/*LOCAL*/
proc surveyfreq data=a;
tables local/cl row;
format local local.;
run;

PROC SGPLOT DATA = a;
VBAR local;
format local local.
RUN;

/*RUPTURA*/
proc surveyfreq data=a;
tables ruptura/cl row;
format ruptura ruptura.;
run;

PROC SGPLOT DATA = a;
VBAR ruptura;
format ruptura ruptura.
RUN;

/*TAMANHO*/
proc surveyfreq data=a;
tables tamanho/cl row;
format tamanho tamanho.;
```

```
run;

/*ANTERIOR*/
proc surveyfreq data=a;
tables anterior/cl row;
format anterior anterior.;
run;

/*POSTERIOR*/
proc surveyfreq data=a;
tables posterior/cl row;
format posterior posterior.;
run;



/* RESPOSTA * EXPLICATIVAS  -  ruptura *  resto */

proc surveyfreq data=a;
tables (sexo idade local Tamanho Anterior posterior )*ruptura/cl row;
format local local. sexo sexo. Anterior anterior. Posterior posterior.  Tamanho tamanho. Ida
run;


/*REGRESSÃO SIMPLES*/

/*RUPTURA * SEXO */
Proc logistic data = mysasdata descending;
Class Sexo (ref='Feminino') / param=ref ;
Model Ruptura = Sexo ;
Format Sexo sexo. ;
Run ;



/*RUPTURA * IDADE*/
Proc logistic data = mysasdata descending;
Class Idade (ref='> 55') / param=ref ;
Model Ruptura = Idade ;
Format Idade idade. ;
Run ;



/*RUPTURA * LOCAL*/
```

```
Proc logistic data = mysasdata descending;
Class local (ref='Pós Sifão') / param=ref ;
Model Ruptura = Local ;
Format Local local. ;
Run ;

/*RUPTURA * ANTERIOR*/
Proc logistic data = mysasdata descending ;
Class tamanho (ref='> 6') / param=ref ;
Model Ruptura = tamanho ;
Format tamanho tamanho. ;
Run ;


/*RUPTURA * ANTERIOR*/
Proc logistic data = mysasdata descending ;
Class anterior (ref='<= 15.40') / param=ref ;
Model Ruptura = Anterior ;
Format Anterior anterior. ;
Run ;


/*RUPTURA*POSTERIOR*/
Proc logistic data = mysasdata descending;
Class posterior (ref='<= 88.15') / param=ref ;
Model Ruptura = Posterior ;
Format Posterior posterior. ;
Run ;

/*REGRESSÃO LOGISTICA MULTIPLA*/

Proc logistic data=a descending;
Class Sexo (ref='Feminino')
Idade (ref='> 55')
Local (ref='Pós Sifão')
Tamanho (ref='> 6')
Anterior (ref='<= 15.40')
Posterior (ref='<= 88.15') ;

model Ruptura = Sexo Idade Local Tamanho Anterior Posterior/ LACKFIT;

format   Local local.
Sexo sexo.
Anterior anterior.
Posterior posterior.
Idade idade.
```

```
Tamanho tamanho.;
run;
```

8 Referências Bibliográficas

1. Hosmer, D. W., and Lemeshow, S. . Applied Logistic Regression, second Edition. Wiley, New York, (2000).

2. McCullagh, P., and Nelder, J. A. . Generalized Linear Models, Second Edition. Chapman Hall, London, (1989).

6. Mohamed Reda Abonazel. Generalized Random Coefficient Estimators of Panel Data Models: Asymptotic and Small Sample Properties. American Journal of Applied Mathematics and Statistics, (2016).

3. Silva Neto ÂR, Câmara RL, Valença MM. Carotid siphon geometry and variants of the circle of Willis in the origin of carotid aneurysms. Arq Neuropsiquiatr, (2012).

5. Siqueira Waihrich, Eduardo. Influência da Anatomia do Sifão Carotídeo na Apresentação e Resposta ao Tratamento de Aneurismas Cerebrais / Eduardo Siqueira Waihrich; Brasília, (2017).

4. Stokes, Maura E., Charles S. Davis, and Gary G. Koch. Categorical Data Analysis Using the SAS System, Second Edition. Cary, NC: SAS Institute Inc, (2000).