UNIVERSIDADE FEDERAL DO MARANHÃO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
CURSO DE MATEMÁTICA – LICENCIATURA

RANNUF NUNES DE ABREU JÚNIOR

ESTIMAÇÃO DOS PARÂMETROS DO MODELO DE REGRESSÃO LOGÍSTICA BINÁRIA

SÃO LUIS – MA
2023

RANNUF NUNES DE ABREU JÚNIOR

ESTIMAÇÃO DOS PARÂMETROS DO MODELO DE REGRESSÃO LOGÍSTICA BINÁRIA

Monografia apresentada à Coordenação do curso de Matemática, da Universidade Federal do Maranhão – UFMA, como requisito parcial para obtenção de grau em Licenciatura em Matemática.

SÃO LUIS – MA
2023

DEMAT / UFMA

Ficha gerada por meio do SIGAA/Biblioteca com dados fornecidos pelo(a) autor(a).
Diretoria Integrada

Nunes de Abreu Júnior, Rannuf.
Estimação dos Parâmetros do Modelo de Regresssão Logística Binária / Rannuf Nunes de Abreu Júnior. - 2023.
41 p.

Orientador(a): Josenildo de Souza Chaves.Monografia(Graduação)-Curso de Matemática, Universidade Federal do Maranhão, UFMA, 2023.

1. Critérios de estimativas. 2.Modelo de Regressão Logística. 3. Exemplos. 4. Considerações. I.de Souza Chaves, Josenildo. II. Título.

RANNUF NUNES DE ABREU JÚNIOR

ESTIMAÇÃO DOS PARÂMETROS DO MODELO DE REGRESSÃO LOGÍSTICA BINÁRIA

Monografia apresentada à Coordenação do curso de Matemática, da Universidade Federal do Maranhão – UFMA, como requisito parcial para obtenção de grau em Licenciatura em Matemática.

Trabalho **APROVADO** em: 18 / 07 / 2023

Prof. Dr. Josenildo de Souza Chaves

Orientador

DEMAT / UFMA

Prof. Me. Cléber Cavalcanti

Primeiro Examinador

DEMAT / UFMA

Prof. Dr. Valeska Martins de Souza

Segunda Examinadora

DEMAT / UFMA

Agradecimentos

Dedico esse trabalho primeiramente a Deus por ter me abençoado, ajudado e guiado até este momento final.

Agradeço aos meus pais por todo amor e suporte.

Agradeço aos meus amigos que são presentes dados pela UFMA, por toda consideração e companheirismo.

Agradecendo também ao Professor Josenildo por toda sabedoria, conhecimento e bondade em todo momento de orientação.

RESUMO

Este trabalho aborda diversos conceitos estatísticos, incluindo estimadores de máxima verossimilhança, critérios de estimativas, distribuição em grandes amostras e invariância. Os estimadores de máxima verossimilhança são usados para encontrar o valor do parâmetro que maximiza a probabilidade de observar os dados coletados. Existem vários critérios para avaliar a qualidade de uma estimativa, incluindo o critério de máxima verossimilhança e o critério de mínimos quadrados.

A distribuição em grandes amostras é usada para avaliar a distribuição de uma estimativa quando o tamanho da amostra é grande. A invariância é uma propriedade importante dos estimadores de máxima verossimilhança, que significa que se aplicarmos uma transformação monótona contínua ao parâmetro, a estimativa de máxima verossimilhança da transformação é a transformação da estimativa de máxima verossimilhança original. Esses conceitos são úteis para a análise estatística de dados e para a tomada de decisões baseadas em evidências empíricas.

O modelo de regressão logística é uma técnica estatística usada para modelar a relação entre uma variável binária (ou categórica) e um conjunto de variáveis explicativas. Ele é amplamente utilizado em áreas como medicina, biologia, marketing e finanças. A função logística é usada para modelar a probabilidade de sucesso (ou fracasso) em termos de uma combinação linear das variáveis explicativas. O modelo de regressão logística é ajustado usando o método de máxima verossimilhança, que encontra os valores dos parâmetros que maximizam a probabilidade de observar os dados coletados. Os resultados do modelo de regressão logística são interpretados em termos de odds ratios, que são razões de chances as quais medem a mudança na probabilidade de sucesso para uma unidade de mudança nas variáveis explicativas. O modelo de regressão logística é uma ferramenta poderosa para a análise de dados binários e pode ser usado para prever a probabilidade de sucesso em uma variedade de situações.

Palavras-chaves: Estimativas. Parâmetros. Regressão Logística.

ABSTRACT

This work addresses several statistical concepts, including maximum likelihood estimators, estimation criteria, large-sample distribution, and invariance. Maximum likelihood estimators are used to find the parameter value that maximizes the probability of observing the collected data. There are various criteria for evaluating the quality of an estimate, including the maximum likelihood criterion and the least squares criterion.

The large-sample distribution is used to assess the distribution of an estimate when the sample size is large. Invariance is an important property of maximum likelihood estimators, which means that if we apply a continuous monotonic transformation to the parameter, the maximum likelihood estimate of the transformed parameter is the transformation of the original maximum likelihood estimate. These concepts are useful for the statistical analysis of data and for making decisions based on empirical evidence.

The logistic regression model is a statistical technique used to model the relationship between a binary (or categorical) variable and a set of explanatory variables. It is widely used in areas such as medicine, biology, marketing, and finance. The logistic function is used to model the probability of success (or failure) in terms of a linear combination of the explanatory variables. The logistic regression model is fitted using the maximum likelihood method, which finds the parameter values that maximize the probability of observing the collected data. The results of the logistic regression model are interpreted in terms of odds ratios, which are measures of the change in the probability of success for a unit change in the explanatory variables. The logistic regression model is a powerful tool for the analysis of binary data and can be used to predict the probability of success in a variety of situations.

Keywords: Estimates. Parameters. Logistic Regression.

Sumário

Capítulo 1 - Introdução 7
1.1. Introdução 7
1.2. Organização dos Capítulos 8
Capítulo 2 – Critérios de Estimativas para o Modelo de Regressão Logística Binária . 9
2.1. Estimadores de Máxima Verossimilhança 9
2.2. Critérios de estimativas 9
2.3. O Método de Máxima Verossimilhança 11
2.4. Propriedades dos Estimadores de Máxima Verossimilhança 12
2.4.1. Invariância 12
2.5. Distribuição em grandes amostras 12
2.6. Verossimilhança para Amostras Independentes. 13
Capítulo 3 – Modelo de Regressão Logística Binária 14
3.1. Introdução 14
3.2. Estimação dos Parâmetros do Modelo de Regressão Logística 14
3.3. Testando a Significância dos Coeficientes 17
3.4. Estimação dos Intervalos de Confiança 21
3.5. Interpretando um Modelo Logístico Ajustado 23
3.6. Variáveis Independentes Dicotômicas 24
Capítulo 4 – Aplicações do Modelo de Regressão Logística com uso da Linguagem R 30
4.1. Introdução 30
5. Considerações Finais 34
Apêndice A 35
Apêndice B 36
Referências Bibliográficas 41

Capítulo 1 - Introdução

1.1. Introdução

Para iniciar, podemos fazer um breve comentário que norteia a sequência desse trabalho, tendo em vista que será importante para a construção do ideal proposto.

Normalmente, em Estatística, trabalha-se com uma amostra representativa da população. Por exemplo, se desejarmos tirar conclusões sobre os resultados das eleições gerais, é impossível perguntar a toda a população do país. Para resolver este problema, uma amostra variada e representativa é escolhida. Graças a qual uma estimativa do resultado final pode ser extraída.

A regressão logística é uma técnica estatística que tem como objetivo produzir, a partir de um conjunto de observações, um modelo que permita a predição de valores tomados por uma variável categórica, frequentemente binária, mediante um conjunto de variáveis explicativas. O livro "Applied Linear Regression" (Hosmer, et.al, 2013) explora os princípios e aplicações dessa técnica.

A regressão linear simples busca encontrar a reta que minimize a diferença entre os valores observados e os valores previstos pelo modelo. A regressão linear também permite fazer previsões usando o modelo ajustado. No entanto, é importante avaliar se as suposições do modelo são atendidas, como a linearidade.

Abordamos os principais tópicos de forma clara e apresentando exemplos práticos para ilustrar a aplicação dos conceitos. Além disso, o artigo destaca a importância desses conceitos para a análise de dados e a tomada de decisões baseadas em evidências empíricas. A justificativa para a produção desse artigo é, portanto, fornecer uma base sólida para a compreensão dos conceitos estatísticos fundamentais e sua aplicação prática em diversas áreas.

O objetivo geral deste trabalho é mostrar vários aspectos que estão inseridos no modelo de regressão logística, com as suas atribuições e propriedades, que, como foi visto, estão sendo baseados desde os métodos de estimação de parâmetros e com os

intervalos de confiança juntos com seus teoremas e definições. Destacamos os seguintes objetivos específicos:

- Revisar os critérios de estimação de parâmetros.
- Interpretar os resultados do modelo de regressão logística.
- Desenvolver aplicações da metodologia usando a linguagem R (R Core Team, 2020).

1.2. Organização dos Capítulos

Esse trabalho está organizado em cinco capítulos. Apresentamos, no Capítulo 2, definições sobre como determinar as estimativas necessárias para o modelo de Regressão Logística. Dentre os assuntos, estão os Critérios, Método de Máxima Verossimilhança e suas propriedades, Distribuição para Grandes Amostras, Casos Multiparamétrico, Invariância. No Capítulo 3, são expostos temas ligados diretamente ao funcionamento do Modelo de Regressão Logística. Foram mostrados a ideia principal do Modelo, a significância dos coeficientes, o modo de interpretar o Modelo, as variáveis e aplicação do Modelo. No Capítulo 4, apresentamos exemplos do Modelo de Regressão Logística aplicados em dois exemplos de natureza distintas. E por último, temos as considerações finais do trabalho.

Capítulo 2 – Critérios de Estimativas para o Modelo de Regressão Logística Binária

2.1. Estimadores de Máxima Verossimilhança

Para a fundamentação deste capítulo, as referências principais foram Meyer (1983) e Bolfarine & Sandoval (2001).

Suponha-se que desejamos estimar a proporção p de indivíduos numa população que possuem uma determinada característica. Em geral, não conhecemos o verdadeiro valor de p, e não faz sentido dizer que uma estimativa $\hat{p}$ seja "correta". Além disso, se tivermos duas estimativas de p, por exemplo $\hat{p}_1$ e $\hat{p}_2$, devemos achar alguma maneira de decidir qual delas é "melhor". Isto significa que deveremos estabelecer algum critério para decidir se uma estimativa deve ser preferida a outra.

2.2. Critérios de estimativas

Definição 2.1 Seja X uma variável aleatória com alguma distribuição de probabilidade que dependa de um parâmetro desconhecidoθ. Seja $X_1,\ldots,X_n$ uma amostra de X, e sejam $x_1, \ldots, x_n$, os correspondentes valores amostrais. Se $g(x_n, \ldots, x_n)$ for uma função de amostra a ser empregada para estimação de θ, nos referiremos a g como um estimador de θ. O valor que g assume, isto é, $g(x_1, \ldots, x_n)$, será referido como uma estimativa de θ e é usualmente escrito assim: $\hat{\theta} = g(x_1, \ldots, x_n)$.

Definição 2.2 Seja $\hat{\theta}$ uma estimativa do parâmetro desconhecido θ associado com a distribuição da variável aleatória X. Neste caso, θ será um estimador não-tendencioso (ou uma estimativa não-tendenciosa) de θ, se for $E(\hat{\theta}) = \theta$ para todo θ.

Definição 2.3 Seja $\hat{\theta}$ uma estimativa não-tendenciosa de θ. Diremos que $\hat{\theta}$ é uma estimativa não-tendenciosa, de variância mínima de θ, se para todas as estimativas θ^* tais que $E(\theta^*) = \theta$, tivermos $Var(\hat{\theta}) < Var(\theta^*)$ para todo θ^*. Isto é, dentre todas as estimativas não-tendenciosas de θ, $\hat{\theta}$ tem a variância menor de todas.

Outro critério para julgar estimativas é um tanto mais difícil de formular e é baseado na seguinte definição:

Definição 2.4 Seja $\hat{\theta}$ uma estimativa baseada em uma amostra $X_1, \ldots, X_n$ do parâmetro θ. Então $\hat{\theta}$ é um estimador consistente de $\hat{\theta}$ se

$$P(\left|\hat{\theta} - \theta\right| > \varepsilon) = 0; \qquad para\ todo\ \varepsilon > 0 \tag{2.1}$$

ou, equivalentemente, se

$$P(\left|\hat{\theta} - \theta\right| \leq \varepsilon) = 1; \qquad para\ todo\ \varepsilon > 0 \tag{2.2}$$

Teorema 2.1 (MEYER,1983) Seja $\hat{\theta}$ um estimador de θ baseado em uma amostra de tamanho n. Se$E(\hat{\theta}) = \theta$, e se $Var(\hat{\theta}) = 0$, então $\hat{\theta}$é um estimador consistente de θ.

Demonstração: Empregando a desigualdade de Chebyshev

$$P\left[\left|\hat{\theta} - \theta\right| \geq \varepsilon\right] \leq \frac{1}{\varepsilon^2} E\left[\hat{\theta} - \theta\right]^2.$$

Segue-se que,

$$\begin{aligned} P\left[\left|\hat{\theta} - \theta\right| \geq \varepsilon\right] \leq \frac{1}{\varepsilon^2} E\left[\hat{\theta} - \theta\right]^2 &= \frac{1}{\varepsilon^2} E\left[\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta\right]^2 \\ &= \frac{1}{\varepsilon^2} E\left\{\left[\hat{\theta} - E(\hat{\theta})\right]^2 + 2\left[\hat{\theta} - E(\hat{\theta})\right]\left[\hat{\theta} - E(\hat{\theta})\right] + \left[\hat{\theta} - E(\hat{\theta})\right]^2\right\} \\ &= \frac{1}{\varepsilon^2} \{ Var\ \hat{\theta} + \left[\hat{\theta} - E(\hat{\theta})\right]^2 \end{aligned} \tag{2.3}$$

Portanto, fazendo $n \to \infty$ e empregando as hipóteses do teorema, segue-se que

$$P(\left|\hat{\theta} - \theta\right| \geq \varepsilon) \leq 0 \tag{2.4}$$

e, por isso, igual a 0.

A seguir, apresentamos um método bastante geral que fornecerá boas estimativas em um grande número de problemas, no sentido de que elas satisfazem a um ou a mais dos critérios acima.

2.3. O Método de Máxima Verossimilhança

Definição 2.6 Sejam $X_1, \ldots, X_n$ uma amostra aleatória de tamanho n da variável aleatória X com função de densidade (ou de probabilidade) $f(x|\theta)$. Com θ pertencente ao espaço paramétrico Θ. A função de verossimilhança de θ correspondente à amostra aleatória observada é dada por

$$L(\theta, x) = \prod_{i=1}^{n} f(x_i|\theta). \tag{2.5}$$

Definição 2.7 O estimador de máxima verossimilhança (MV) de θ é o valor $\hat{\theta} \in \Theta$ que maximiza a função de verossimilhança $L(\theta;\ x)$.

O logaritmo natural da função de verossimilhança de θ é denotado por

$$l(\theta; x) = logL(\theta; x). \tag{2.6}$$

Desde que $log\ log\ x$ é uma função crescente x, o valor de $\hat{\theta}$ que maximiza a função de verossimilhança $L\ (\theta;\ x)$ também maximiza $l(\theta;\ x)$. Sob condições bastante gerais, admitindo-se que θ seja um número real e que $l(\theta;\ x)$ seja uma função derivável de θ, obteremos a estimativa de MV $\hat{\theta}$ de θ pela resolução da equação

$$l'(\theta; x) = \frac{\partial t(\theta; x)}{\partial x} = 0. \tag{2.7}$$

Em alguns exemplos, a solução da equação de verossimilhança pode ser obtida explicitamente. Em situações mais complicadas, a solução da equação (2.7) será em geral obtida por procedimentos numéricos. Para se concluir que a solução da equação é um ponto de máximo, pode ser necessário verificar se

$$l''(\theta; x) = \frac{\partial^2 L(\theta; x)}{\partial^2 x}|_{\theta=\hat{\theta}} < 0. \tag{2.8}$$

2.4. Propriedades dos Estimadores de Máxima Verossimilhança

O estimador de MV pode ser tendencioso. Muito frequentemente, tal tendenciosidade pode ser eliminada pela multiplicação por uma constante apropriada.

Sob condições bastante gerais, os estimadores de MV são consistentes. Isto é, se o tamanho da amostra sobre a qual essas estimativas foram calculadas for grande, a estimativa de MV será "próxima" do valor do parâmetro a ser estimado. As estimativas de MV possuem outra propriedade de "grandes amostras" muito importante, a qual apresentaremos a seguir.

Os estimadores de MV apresentam a seguinte propriedade de invariância muito importante: Suponha-se que $\hat{\theta}$ seja o estimador de MV de θ. Nesse caso, pode-se mostrar que o estimador de MV de $g(\theta)$, em que g uma função monótona contínua, é $g(\hat{\theta})$.

2.4.1. Invariância

Teorema 2.3(*O princípio da invariância*).Sejam $X_1, \ldots, X_n$ amostra aleatória da variável aleatória X com função de densidade (ou de probabilidade) $f(x|\theta)$. Se $\hat{\theta}$ é um estimador de máxima verossimilhança de θ, então $g(\hat{\theta})$ é um estimador de máxima verossimilhança de $g(\theta)$.

Prova. A prova deste teorema pode ser vista em (Bolfarine & Sandoval, 2000).

2.5. Distribuição em grandes amostras

No caso em que o tamanho da amostra é grande e as condições de regularidade estão satisfeitas, temos que

$$\sqrt{n}(\theta - \hat{\theta}) \approx N\left(0, \frac{1}{I_F(\theta)}\right), \tag{2.9}$$

$$\sqrt{n}\left(g(\theta) - g(\hat{\theta})\right) \approx N\left(0, \frac{g'(\theta)^2}{I_F(\theta)}\right), \tag{2.10}$$

e

$$I_F(\theta) = nE\left[\frac{\partial}{\partial\theta} ln\, ln\, f(X;\theta)\right]^2.$$

sendo que" ≈ " significa distribuição assintótica. Para amostras de tamanho grande, os estimadores de máxima verossimilhança deθ e $g(\theta)$ são aproximadamente não viciados, cujas variâncias coincidem com os correspondentes limites inferiores das variâncias dos estimadores não viciados deθ e $g(\theta)$.

2.6. Verossimilhança para Amostras Independentes.

Existem situações em que temos duas ou mais amostras independentes de distribuições que dependem de um parâmetro de interesse. No caso de duas amostras aleatórias independentes, $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_n$, podemos escrever

$$L(\theta; x, y) = L(\theta; x)L(\theta; y), \tag{2.11}$$

devido à independência entre as amostras. Portanto, a verossimilhança conjunta é igual ao produto da verossimilhança correspondente à amostra $X_1, \ldots, X_n$pela verossimilhança correspondente à amostra$Y_1, \ldots, Y_n$. De (2.26), podemos escrever

$$l(\theta; x, y) = l(\theta; x) + l(\theta; y) \tag{2.12}$$

de modo que o logaritmo da verossimilhança conjunta é igual à soma do logaritmo das verossimilhanças correspondentes a cada uma das amostras.

Capítulo 3 – Modelo de Regressão Logística Binária

3.1. Introdução

Este capítulo foi desenvolvido tendo como referência principal Lemeshow (2008). Os métodos de regressão linear são usados para modelar a relação entre uma variável resposta quantitativa e uma ou mais variáveis explicativas. A seguir, descrevemos um método semelhante que é utilizado quando a variável resposta é uma variável categórica com dois valores possíveis.

De acordo com essa ideia, é adotado "sucesso" e "falha" como resultado da variável de interesse. Ao "sucesso", é atribuído o valor 1 e, a "falha", o valor 0. Antes de iniciar um estudo minucioso do modelo de regressão logística, é importante entender que o objetivo de uma análise utilizando esse modelo é o mesmo de qualquer outro modelo de regressão, ou seja, encontrar o melhor e mais parcimonioso modelo, para descrever a relação entre uma variável resposta e um conjunto de variáveis independentes (preditoras ou explicativas). As variáveis independentes são frequentemente chamadas de covariáveis. O exemplo mais comum de modelos de regressão é o linear em que a variável resposta é assumida como contínua.

O que distingue um modelo de regressão logística de um modelo de regressão linear é que a variável resposta na regressão logística é binária ou dicotômica. Essa diferença entre a regressão logística e a linear se reflete tanto na forma do modelo quanto em suas premissas. Assim como na regressão linear múltipla, as variáveis explicativas podem ser categóricas ou quantitativas. A regressão logística é um método estatístico para descrever esse tipo de relacionamento.

3.2. Estimação dos Parâmetros do Modelo de Regressão Logística

Na regressão linear, o método mais utilizado para estimar os parâmetros é o de mínimos quadrados. Neste método, são escolhidos os valores de β_0 e $\beta_1$que minimizam a soma dos quadrados dos desvios dos valores observados da variável resposta Yem relação aos valores previstos com base no modelo. Sob as premissas usuais de regressão

linear, o método de mínimos quadrados produz estimadores com uma série de propriedades estatísticas desejáveis. Infelizmente, quando o método de mínimos quadrados é aplicado a um modelo com uma resposta dicotômica, os estimadores não têm mais essas mesmas propriedades.

Em qualquer problema de regressão, a quantidade chave é o valor médio da variável resposta, dado o valor da variável independente. Essa quantidade é chamada de média condicional e é expressa como $E(x)$, em que Y denota a variável resposta e x denota um valor específico da variável independente. A quantidade $E(Y|x)$representa o valor esperado de Y, dado o valor x. Na regressão linear, assumimos que essa média pode ser expressa como uma equação linear em x, tal que

$$E(x) = \beta_0 + \beta_1 x.$$

Esta expressão implica que é possível para $E(Y|x)$ assumir qualquer valor $-\infty$ e $+\infty$.

Seja $\pi(x) = E(Y|x)$. O modelo de regressão logística binário é dado por

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}. \qquad (3.1)$$

Uma transformação de $\pi(x)$, que é central na regressão logística, é a transformação logit,

$$g(x) = log\left(\frac{\pi(x)}{1-\pi(x)}\right) \qquad (3.2)$$

que é linear nos parâmetros e assume valores em $(-\infty, +\infty)$.

No modelo de Regressão Linear, a variável resposta pode ser expressa na forma $y = E(x) + e$. Em geral, é assumido $e \sim N(0, \sigma_e^2)$. Então, $Y \sim N(0, \sigma_Y^2)$. Esse não é caso de uma variável dicotômica. Nesta situação, a variável resposta é expressa como $y = \pi(x) + e$.

A seguir, a estimação dos parâmetros do modelo de regressão logística é desenvolvida pelo método de máxima verossimilhança.

Procedemos da seguinte maneira: codificamos Y como 0 ou 1, então a expressão $\pi(x)$ já dada prevê (para um valor arbitrário de $\beta = (\beta_0, \beta_1)$ o vetor dos parâmetros); a probabilidade condicional que Yé igual a 1 dado x. Isso é denotado como $\pi(x)$. Segue-

se que a quantidade $1 - \pi(x)$ dá a probabilidade condicional que Y é igual a zero dada x, $P(Y = 0|x)$. Assim, para aqueles pares (x_i, y_i), nos quais $y_i = 1$, a contribuição para a função de probabilidade é $\pi(x_i)$, e para aqueles pares em que $y_i = 0$, a contribuição para a função de probabilidade é $1 - \pi(x_i)$, na qual a quantidade $\pi(x_i)$ denota o valor $\pi(x)$ do computado em x_i. Uma maneira conveniente de expressar a contribuição para a função de verossimilhança para o par (x_i, y_i) é através da expressão.

$$\pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i} \tag{3.3}$$

Como as observações são independentes, a função de verossimilhança é obtida como produto dos termos dados na equação (3.3) da seguinte forma

$$L(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i} \tag{3.4}$$

A partir da equação (3.4), a função de log-verossimilhança é dada por

$$l(\beta) = log\ L(\beta) = \sum_{i=1}^{n} \left[y_i\ log\ \log \pi(x_i) + (1 - y_i)\ log\ \log\left(1 - \pi(x_i)\right)\right] \tag{3.5}$$

Para encontrar o valor $\hat{\beta}$, o valor que torna máxima $l(\beta)$, derivamos $l(\beta)$ em relação a β_0 e β_1. As equações de log-verossimilhança são dadas por

$$\sum [y_i - \pi(x_i)] = 0 \tag{3.6}$$

e

$$\sum x_i[y_i - \pi(x_i)] = 0 \tag{3.7}$$

Na regressão linear, as equações de verossimilhança, obtidas diferenciando a função de desvios quadrados em relação ao que diz respeito a β, são lineares nos parâmetros desconhecidos e, portanto, são facilmente resolvidas. Para a regressão logística, as expressões nas equações (3.4) e (3.5) são não-lineares em β_0 eβ_1, logo, requerem métodos especiais para sua solução. Esses métodos são de natureza iterativa e foram programados em software de regressão logística. No momento, não precisamos nos preocupar com esses métodos iterativos e vê-los como um detalhe computacional que é cuidado para nós.

O valor β, dado pela solução para equações (3.5) e (3.6), é chamado de estimativa de máxima verossimilhança e é denotado como $\hat{\beta}$. A quantidade $\hat{\pi}(x_i)$ é a estimativa de verossimilhança de $\pi(x_i)$. Ela fornece uma estimativa da probabilidade condicional que Y é igual a 1, dado que x igual a x_i. Como tal, representa o valor ajustado ou previsto para o modelo de regressão logística. Uma consequência interessante da equação (3.6) é que

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{\pi}(x_i) \tag{3.8}$$

3.3. Testando a Significância dos Coeficientes

Após a estimativa dos coeficientes, nossa primeira vista no modelo ajustado geralmente diz respeito a uma avaliação da significância das variáveis no modelo. Isso geralmente envolve formulação e teste de uma hipótese estatística para determinar se as variáveis independentes no modelo estão "significativamente" relacionadas ao variável desfecho. O método para a realização desse teste é bastante geral e difere de um tipo de modelo para o outro apenas nos detalhes específicos.

Dessa forma, começamos por testes para a significância dos coeficientes. Será que o modelo que inclui a variável em questão nos diz mais sobre a variável resultado (ou resposta) do que um modelo que não inclui essa variável? Essa pergunta é respondida comparando os valores observados da variável resposta com os previstos por

cada um dos dois modelos; o primeiro com e o segundo sem a variável em questão. A função matemática utilizada para comparar os valores observados e previstos depende do problema particular. Se os valores previstos com a variável no modelo são melhores ou mais precisos em algum sentido do que quando a variável não está no modelo, então, sentimos que a variável em questão é "significativa". É importante notar que não estamos considerando a questão de se os valores previstos são uma representação precisa dos valores observados em um sentido absoluto (isso é chamado de bondade do ajuste)

O método geral de avaliação da significância das variáveis é facilmente ilustrado no modelo de regressão linear, e seu uso lá motiva a abordagem utilizada para a regressão logística. Uma comparação das duas abordagens destaca as diferenças entre as modelagens contínua e os dicotômicos variáveis de resposta.

Na regressão linear, avalia-se a significância do coeficiente de inclinação, formando o que é chamado de análise da tabela de variância. Esta tabela divide os desvios totais de observações sobre sua média em duas partes: (1) a soma dos desvios quadrados das observações sobre a linha de regressão SSE (ou soma residual de quadrados) e (2) a soma dos quadrados dos valores previstos, com base no modelo de regressão, sobre a média da variável dependente SSR (ou a devida regressão sum-of-squares). Esta é apenas uma maneira conveniente de exibir a comparação dos valores observados com os valores previstos em dois modelos. Na regressão linear, a comparação dos valores observados e previstos baseia-se no quadrado da distância entre os dois. Se y_i denotar o valor observado e $\hat{y}_i$ denotar o valor previsto para o indivíduo ith sob o modelo, então a estatística utilizada para avaliar essa comparação é:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{3.9}$$

Sobre o modelo que não contém a variável independente em questão, o único parâmetro é β_0, e $\beta_0 = y$, a média da variável resposta. Neste caso, $\hat{y}_i = y$ e SSE é igual à soma total de quadrados. Quando incluímos a variável independente no modelo, qualquer diminuição da SSE deve-se ao fato de que o coeficiente de inclinação para a

variável independente não é zero. A alteração no valor da SSE deve-se à fonte de regressão da variabilidade, denotada SSR. Isto é,

$$SSR = \left[\sum_{i=1}^{n}(y_i - \underline{y_i})^2\right] - \left[\sum_{i=1}^{n}(y_i - \hat{y_i})^2\right] \tag{3.10}$$

Na regressão linear, o interesse se concentra no tamanho de SSR. Um grande valor sugere que a variável independente é importante, enquanto um pequeno valor sugere que a variável independente não é útil na previsão da resposta.

O princípio norteador com regressão logística é o mesmo: comparar valores observados da variável resposta aos valores previstos obtidos, a partir de modelos, com e sem a variável em questão. Na regressão logística, a comparação dos valores observados com os previstos baseia-se na função de probabilidade de tronco definida na equação (3.5). Para entender melhor essa comparação, é útil conceitualmente pensar em um valor observado da variável resposta como sendo também um valor previsto resultante de um modelo saturado. Um modelo saturado é aquele que contém tantos parâmetros quanto há pontos de dados, um exemplo simples de um modelo saturado é encaixar um modelo de regressão linear quando há apenas dois pontos de dados, $n = 2$.

A comparação dos valores observados com os previstos, utilizando a função de probabilidade, baseia-se na seguinte expressão:

$$D = -2\ log\ log\ \left(\frac{L(modelo\ ajustado)}{L(modelo\ saturado)}\right). \tag{3.11}$$

A quantidade dentro dos grandes parênteses na expressão (3.11) acima é chamada de razão de verossimilhança. Usando menos o dobro de seu registro é necessário para obter uma quantidade cuja distribuição é conhecida e, portanto, pode ser usada para fins de teste de hipóteses. Tal teste é chamado de teste de razão de verossimilhança. Usando equação (3.5), a equação (3.11) torna-se

$$D = -2 \sum_{i=1}^{n} [y_i \, ln \, (\frac{\widehat{\pi_i}}{y_i}) + (1 - y_i) \, ln \, (\frac{1 - \widehat{\pi_i}}{1 - y_i}) \tag{3.12}$$

em que $\widehat{\pi_i} = \widehat{\pi}(x_i)$.

A estatística D, na equação (3.12), é chamada de deviance, e, para regressão logística, desempenha o mesmo papel que a soma residual de quadrados desempenha na regressão linear. De fato, o desvio, como mostrado na equação (3.12), quando computado para regressão linear, é idêntico ao SSE.

Além disso, onde os valores do variável desfecho são 0 ou 1, a probabilidade do modelo saturado é idêntica a 1,0. Especificamente, ele segue a partir da definição de um modelo saturado que $\widehat{\pi_i} = y_i$ e a probabilidade é $l(modelo\ saturado)$:

$$l(modelo\ saturado) = \prod_{i=1}^{n} y^{y_i}{}_i (1 - y_i)^{(1-y_i)} = 1.$$

Assim, segue-se da equação (3.10) que o desvio é

$$D = -2 \, ln(prob. do\ modelo\ ajustado) \tag{3.13}$$

Alguns pacotes de software relatam o valor do desvio na equação (3.13) em vez da log-verossimilhança para o modelo montado. No contexto de testes para a significância de um modelo, queremos enfatizar que pensamos no desvio da mesma forma que vemos na soma de quadrados residual em regressão linear.

Em particular, para avaliar a significância de uma variável independente comparamos o valor de D com e sem a variável independente na equação. A mudança em D devido à inclusão da variável independente no modelo é:

$$G = D(modelo\ sem\ a\ var.) - D(modelo\ com\ a\ var.)$$

Essa estatística G desempenha o mesmo papel na regressão logística que o numerador do teste F parcial faz na regressão linear. Como a probabilidade do modelo saturado é sempre comum a ambos os valores de D sendo diferentes, G pode ser expressa como:

$$G = -2\frac{ln(prob.sem\ a\ var.)}{(prob.com\ a\ var.)} \tag{3.14}$$

Para o caso específico de uma única variável independente, é fácil mostrar que, quando a variável não está no modelo, a estimativa de máxima verossimilhança de β_0é $ln(n_1/n_0)$ onde $n_1 = \sum\ y_i\ e\ n_0 = \sum\ (1 - y_i)$ e a probabilidade prevista para todos as unidades amostrais é constante e igual a $n1/n$. Nesta configuração, o valor de G é:

$$G = -2ln\left[\frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{n}\widehat{\pi_i}^{y_i}(1-\widehat{\pi_i})^{(1-y_i)}}\right] \tag{3.15}$$

ou

$$G = \sum_{i=1}^{n}[y_i ln\,(\widehat{\pi_i}) + (1 - y_i)ln\,(1 - \widehat{\pi_i})] - [\,n_1 ln\,(n_1) + n_0 ln\,(n_0) + nln\,(n)] \tag{3.16}$$

Sob a hipótese de que β_1 é igual a zero, a estatística G segue uma distribuição qui-quadrado com 1 grau de liberdade. São necessárias suposições matemáticas adicionais; no entanto, para o caso acima, eles são bastante não-restritivos e envolvem ter um tamanho amostral suficientemente grande, n, unidades amostrais suficientes com ambos $y = 0$ e $y = 1$.

3.4. Estimação dos Intervalos de Confiança

A base para a construção dos estimadores de intervalo é a mesma teoria estatística que usamos para formular os testes para significância do modelo. Em particular, os estimadores do intervalo de confiança para a inclinação e o intercepto são, na maioria das vezes, baseados em seus respectivos testes de Wald e, às vezes, são referidos como intervalos de confiança baseados em Wald. O intervalo de confiança $100(1 - \alpha)\%$ para a inclinação é dado por

$$\hat{\beta}_1 \pm z_{1-\frac{\alpha}{2}} \widehat{SE}(\hat{\beta}_1) \tag{3.17}$$

e, para o intercepto,

$$\hat{\beta}_0 \pm z_{1-\frac{\alpha}{2}} \widehat{SE}(\hat{\beta}_0) \tag{3.18}$$

em que $z_{1-\frac{\alpha}{2}}$ é o percentil $100(1 - \frac{\alpha}{2})$ da distribuição normal padrão e $\widehat{SE}(\cdot)$ denota o erro padrão do estimador do respectivo parâmetro.

O estimador do intervalo de confiança para um coeficiente pode ser concisamente descrito como o intervalo de valores para o qual o teste de razão de verossimilhança não rejeitaria a hipótese, H_0: $\beta = \beta^*$, no nível de significância $1 - \alpha$. Os dois pontos finais βlower e βupper deste intervalo para um coeficiente são definidos da seguinte forma:

$$2\left[l(\hat{\beta}) - l_p(\beta_{upper})\right] = 2\left[l(\hat{\beta}) - l_p(\beta_{lower})\right] = X^2_{1-\alpha}(1) \ , \tag{3.19}$$

em que $l(\hat{\beta})$ é o valor da log-verossimilhança do modelo ajustado e $l(\beta)$ é o valor da log-verossimilhança perfilada. Um valor da log-verossimilhança perfilada é calculado primeiro fixando um valor para o coeficiente de interesse, por exemplo, o coeficiente angular e, em seguida, encontrar o valor do intercepto.

3.5. Interpretando um Modelo Logístico Ajustado

Anteriormente, discutimos os métodos de adequação e teste para a importância do modelo de regressão logística. Após a montagem de um modelo, a ênfase muda da computação e avaliação da significância dos coeficientes estimados para a interpretação de seus valores. Assim, iniciamos este capítulo assumindo que um modelo de regressão logística tem sido adequado, no qual as variáveis no modelo são significativas em um sentido clínico ou estatístico, e que o modelo se encaixa de acordo com alguma medida estatística de ajuste.

A interpretação de qualquer modelo ajustado exige que possamos extrair inferências práticas dos coeficientes estimados no modelo. A questão que está sendo abordada é: o que os coeficientes estimados no modelo nos dizem sobre as questões de pesquisa que motivaram o estudo? Para a maioria dos modelos estatísticos, isso envolve os coeficientes estimados para as variáveis independentes no modelo. Na maioria dos casos, o coeficiente angular é de pouco interesse. Os coeficientes estimados para as variáveis independentes representam a inclinação (ou seja, taxa de alteração) de uma função da variável dependente por unidade de mudança na variável independente. Assim, a interpretação envolve duas questões: determinar a relação funcional entre a variável dependente e a variável independente, e definir adequadamente a unidade de mudança para a variável independente.

O primeiro passo é determinar qual função da variável dependente produz uma função linear das variáveis independentes, e isso é chamado de função de ligação. No modelo de regressão logística, a função de ligação é a transformação do logit:

$$g(x) = ln\{\pi(x)/[1-\pi(x)]\} = \beta_0 + \beta_1 x \qquad (3.20)$$

Para um modelo de regressão linear, lembre-se que o coeficiente de inclinação, β_1,é igual à diferença entre o valor da variável dependente $x + 1$, e o valor da variável dependente em x para qualquer valor de x. Por exemplo, o modelo de regressão linear x é $y(x) = \beta_0 + \beta_1 x$ que o coeficiente de inclinação é $\beta_1 = y(x+1) - y(x)$. Neste caso, a interpretação do coeficiente de inclinação é que é a mudança na variável desfecho correspondente a uma mudança de uma unidade na variável independente.

No modelo de regressão logística, o coeficiente de inclinação é a alteração no logit correspondente a uma mudança de uma unidade na variável independente, ou seja, $\beta_1 = y(x+1) - y(x)$. A interpretação adequada do coeficiente em um modelo de regressão logística depende de poder colocar sentido na diferença entre dois valores da função logit. Essa diferença é discutida detalhadamente caso a caso, pois se relaciona diretamente com a definição e o significado de uma mudança de uma unidade na variável independente.

3.6. Variáveis Independentes Dicotômicas

Começamos discutindo a interpretação dos coeficientes de regressão logística na situação em que a variável independente é nominal escalada e dicotômica (ou seja, medida em dois níveis). Este caso fornece a base conceitual para todas as outras situações.

Presumimos que a variável independente, x , é codificada como 0 ou 1. A diferença no logit para $x = 1$ e $x = 0$ é

$$g(1) - g(0) - (\beta_0 + \beta_1 \times 1) - (\beta_0 + \beta_1 \times 0) - (\beta_0 - \beta_1) - (\beta_0) - (\beta_1) \quad (3.21)$$

A álgebra mostrada nesta equação é bastante simples. A justificativa para apresentá-la neste nível de detalhamento é enfatizar que são necessárias quatro etapas para obter a expressão correta do coeficiente e, portanto, a interpretação correta do coeficiente(s).

As três primeiras das quatro etapas são: (1) definir os dois valores da covariada a serem comparados ($e.g., x = 1 \; and \; x = 0$); (2) substituir esses dois valores na equação para o logit [$e.g., g(1) \; and \; g(0)$], e (3) calcular a diferença nas duas equações [$e.g., g(1) - g(0)$]. Como mostrado, para uma covariada dicotômica codificada 0 e 1, o resultado no final da etapa 3 é igual a β_1. Assim, o coeficiente de inclinação, ou diferença de logit, é a diferença entre o log das chances (odds) quando $x = 1$ e o registro das probabilidades quando $x = 0$. O problema prático é que a mudança na escala das probabilidades de registro é difícil de explicar e pode não ser especialmente significativa para um público sujeito-assunto. Para fornecer uma interpretação mais

significativa, precisamos introduzir a razão de probabilidades como medida de associação.

As chances de o resultado estar presente entre os indivíduos com $x = 1$ é $\frac{\pi(1)}{[1-\pi(1)]}$. Da mesma forma, as chances de o resultado estar presente entre os indivíduos com $x = 0$ é $\pi(0)/[1-\pi(0)]$. A razão de chances (odds), denotada OR, é a razão das probabilidades para $x = 1$ e as probabilidades para $x = 0$, e é dada pela equação

$$OR = \left\{\frac{\pi(1)}{[1-\pi(1)]}\right\} / \left\{\frac{\pi(0)}{[1-\pi(0)]}\right\}. \tag{3.22}$$

Substituindo as expressões para as probabilidades do modelo de regressão logística na equação (3.19) que obtemos

$$OR = \frac{\left(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right)}{\left(\frac{1}{1+e^{\beta_0+\beta_1}}\right)} \Bigg/ \frac{\left(\frac{e^{\beta_0}}{1+e^{\beta_0}}\right)}{\left(\frac{1}{1+e^{\beta_0}}\right)} \tag{3.22 *}$$

$$OR = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1} \tag{3.22 **}$$

Assim, para um modelo de regressão logística com uma variável independente dicotômica codificada 0 e 1, a relação entre a razão de chances e o coeficiente de regressão é

$$OR = e^{\beta_1} \tag{3.23}$$

A razão de chances é amplamente utilizada como medida de associação, pois se aproxima de quão mais provável ou improvável (em termos de probabilidades) é que o

resultado esteja presente entre os sujeitos $x = 1$ em comparação com esses sujeitos com $x = 0$.

Em certos cenários, a razão de chances pode aproximar outra medida de associação chamada de risco relativo, que é a razão das duas probabilidades de desfecho, $RR = \pi(1)/\pi(0)$. Segue-se a equação (3.18) que a razão de chances aproxima o risco relativo se $[1 - \pi(0)]/[1 - \pi(1)] \approx 1$. Isso vale quando $\pi(x)$ é pequeno para ambos $x = 0$ e$x = 1$.

Junto com a estimativa pontual de um parâmetro, é sempre uma boa ideia usar uma estimativa de intervalo de confiança para fornecer informações adicionais sobre o valor do parâmetro. No caso da razão de chances de uma tabela de 2 × 2 (correspondente a um modelo de regressão logística equipada com uma covariada dicotômica única), há uma extensa literatura focada no problema da estimativa do intervalo de confiança para a razão de chances quando o tamanho da amostra é pequeno.

Como observamos anteriormente, a razão de chances é geralmente o parâmetro de interesse derivado de uma regressão logística ajustada devido a sua facilidade de interpretação. No entanto, seu estimador $\widehat{OR}$ tende a ter uma distribuição altamente distorcida para a direita. Isso se deve ao fato de que sua faixa é entre 0 e ∞, com o valor nulo igual a 1. Em teoria, para tamanhos amostrais extremamente grandes, a distribuição de $\widehat{OR}$ seria normal.

Infelizmente, esse requisito de tamanho amostral normalmente excede o da maioria dos estudos. Assim, as inferências são geralmente baseadas na distribuição amostral de $ln\, ln\,(\widehat{OR}) = \widehat{\beta_1}$, que tende a seguir uma distribuição normal para tamanhos amostrais muito menores. Obtemos um $100 \times (1 - \alpha)\%$ estimador de intervalo de confiança para a razão de chances, calculando primeiro os pontos finais de um estimador de intervalo de confiança para a relação log-odds $(i.e., \beta_1)$ e, em seguida, expondo os pontos finais deste intervalo. Em geral, os pontos finais são dados pela expressão:

$$exp\left[\hat{\beta}_1 \pm z_{1-\frac{\alpha}{2}} \times \widehat{SE}\left(\hat{\beta}_1\right)\right] \tag{3.24}$$

Devido à importância da razão de chances como medida de associação, muitos pacotes de software fornecem automaticamente estimativas de intervalo de ponto e confiança com base na exposição de cada coeficiente em um modelo de regressão logística encaixada. O usuário deve estar ciente de que essas quantidades relatadas automaticamente fornecem estimativas de razões de chances (odds ratios) de interesse em apenas alguns casos especiais (por exemplo, uma variável dicotômica codificada 0 e 1 que não está envolvida em nenhuma interação com outras variáveis), um ponto para o qual retornamos na próxima seção. Um dos principais objetivos deste capítulo é mostrar, usando as quatro etapas observadas anteriormente, que se pode obter estimativas de ponto e intervalo de confiança das razões de odds (chances), independentemente da complexidade do modelo montado.

Antes de concluir o caso variável dicotômica, é importante considerar o efeito que a codificação tem na computação do estimador das razões de odds. Na discussão anterior, observamos que o estimador é $\widehat{OR} = exp\left(\hat{\beta}_1\right)$ e que este está correto desde que se codifica a variável independente como 0 ou 1 (ou quaisquer dois valores que diferem por um). Qualquer outra codificação requer que se calcule o valor da diferença de logit para a codificação específica utilizada e, em seguida, exponize-se essa diferença, essencialmente seguindo os quatro passos, não apenas expondo cegamente o estimador do coeficiente.

Ilustramos a configuração da codificação alternativa em detalhes, pois ajuda a enfatizar os quatro passos do método geral para computar estimadores de razões de chances a partir de um modelo de regressão logística encaixada. Suponha que nossa covariada dicotômica é codificada usando valores a e b, e que, no Passo 1, gostaríamos de estimar a razão de chances para o covariado no nível a versus b. Em seguida, no Passo 2, substituímos os dois valores da covariada na equação, a fim de que o logit obtenha $\hat{g}(a) = \widehat{\beta_0} + \widehat{\beta_1}a$ e $\hat{g}(b) = \widehat{\beta_0} + \widehat{\beta_1}b$. Para o passo 3, calculamos a diferença nas duas equações e simplificamos algebricamente para obter a expressão para as log-odds como:

$$ln\left[\widehat{OR}(a,b)\right] = \hat{g}(x=a) - \hat{g}(x=b) = \left(\widehat{\beta_0} + \widehat{\beta_1}a\right) - \left(\widehat{\beta_0} + \widehat{\beta_1}b\right) = \widehat{\beta_1}(a \times b) \tag{3.25}$$

No Passo 4, exponenciamos a equação obtida no Passo 3, mostrado neste caso na equação (3.3), para obter nosso estimador da razão de chances, ou seja,

$$\widehat{OR}(a,b) = exp\,[\widehat{\beta_1}(a \times b)] \tag{3.26}$$

Nas equações (3.3) e (3.4), a notação $\widehat{OR}(a,b)$ denota a razão de chances específica

$$\widehat{OR}(a,b) = \pi(\hat{x} = a)[1 - \hat{\pi}\,(x = a)]/\,\pi(\hat{x} = b)[1 - \hat{\pi}\,(x = b)]$$

No caso usual, quando $a = 1$ e $b = 0$, nós suprimimos a e b e simplesmente usamos $\widehat{OR}$.

O método de codificação também influencia no cálculo dos pontos finais do intervalo de confiança. Para o exemplo, usando o desvio de meios de codificação, o erro padrão estimado necessário para a estimativa do intervalo de confiança é $\widehat{SE}\left(2\hat{\beta}_1\right) = \widehat{2\,SE}\left(\hat{\beta}_1\right)$. Assim, os pontos finais do intervalo de confiança são:

$$exp\,exp\left[2\widehat{\beta}_1 \pm z_{1-\frac{\alpha}{2}} \times 2\,\widehat{SE}\left(\widehat{\beta}_1\right)\right] \tag{3.26}$$

Em geral, os pontos finais do intervalo de confiança para a razão de chances dada na equação (3.5) são

$$exp\,exp\,[\widehat{\beta}_1(a-b) \pm z_{1-\frac{\alpha}{2}} \times |a - b|\,\widehat{SE}(\widehat{\beta}_1)] \tag{3.27}$$

em que $|a - b|$ é o valor absoluto de $(a - b)$. Isso é necessário, porque a pode ser menor que b. Como temos o controle de como codificamos nossas variáveis dicotômicas, recomendamos que, quando os juros se concentram na relação de probabilidades, elas sejam codificadas como 0 ou 1 para fins de análise.

Em resumo, para uma variável dicotômica, o parâmetro de interesse na maioria, se não todos, das configurações aplicadas é a razão de chances. Uma estimativa deste parâmetro pode ser obtida a partir de um modelo de regressão logística equipada, expondo o coeficiente estimado. Em uma configuração na qual a codificação não é 0 ou 1, a estimativa pode ser encontrada simplesmente seguindo as quatro etapas descritas

nesta seção. A relação entre o coeficiente de regressão logística e a razão de chances fornece a base para a nossa interpretação de todos os resultados de regressão logística.

Capítulo 4 – Aplicações do Modelo de Regressão Logística com uso da Linguagem R

4.1. Introdução

Após termos reunido, então, os conhecimentos e os métodos para se construir a Regressão Logística, vamos fazer uso disso para aplicar em diferentes exemplos. Tais exemplos têm diferenças significativas em suas naturezas, mostrando, assim, um largo uso e a maleabilidade que o Método de Regressão Logística tem. Inclusive, outros contextos que possam parecer desafiadores para serem compreendidos podem ser implementados com o uso de Regressão Logística.

Aliado ao que está sendo proposto, será utilizado a Linguagem de Programação R. Uma importante ferramenta na análise e interpretação de inúmeros problemas da Estatística. O seu poder computacional facilita a resolução dos problemas, utilizando corretamente suas implementações.

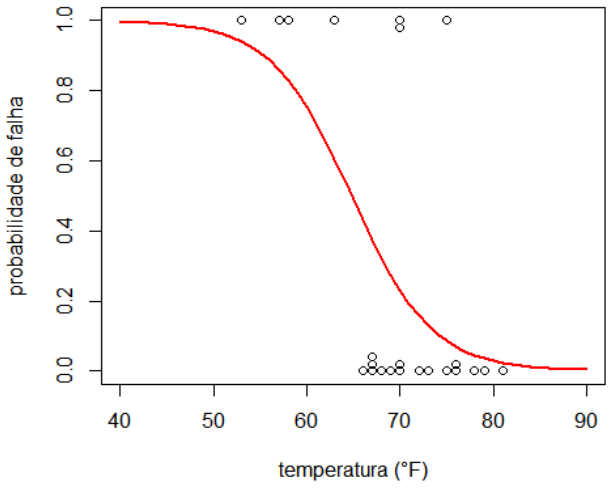
Exemplo 4.1. (CASELLA & BEGER, 2002). Um conjunto de dados de um experimento assustador é o das falhas nos anéis de vedação dos ônibus espaciais, os quais foram associadas à temperatura. Os dados na Tabela 4.1 fornecem as temperaturas na decolagem e se um anel de vedação falhou ou não. A partir da expressão (3.3), a solução numérica do sistema formado pelas equações $\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{0}$ produz os EMVs:$\widehat{\boldsymbol{\beta}}_{\mathbf{0}} = \mathbf{15,04}$ e $\widehat{\boldsymbol{\beta}}_{\mathbf{1}} = -\mathbf{0,232}$. A Figura 4.1 mostra a curva ajustada junto com os dados. O ônibus espacial Challenger explodiu durante a decolagem, matando os sete astronautas a bordo. A explosão foi o resultado de uma falha de um anel, que se acredita ter sido causada pelo clima frio incomum (31 °F) no momento do lançamento. A EMV da probabilidade de falha do anel em 31 °F é 0,9996.

Tabela 4.1 - Temperatura na hora do vôo (°F) e falha nos anéis (1= falha , 0 = sucesso).

Vôo nº	*Falha*	*Temperatura*
14	1	53
09	1	57
23	1	58
10	1	63
01	0	66
05	0	67
13	0	67
15	0	67
04	0	68
03	0	69
08	0	70
17	0	70
02	1	70
11	1	70
06	0	72
07	0	73
16	0	75
21	1	75
19	0	76
22	0	76
12	0	78
20	0	79
18	0	81

Fonte: Casella & Berger (2002).

Figura 4.1 - Dados da Tabela 4.1 e o gráfico do modelo logístico ajustado



Fonte: Elaborado pelo Autor (2023)

A matriz de informação observada dos parâmetros de interesse é dada por

$$I(\hat{\beta}_0, \hat{\beta}_1) = (3{,}15\ 214{,}75\ 214{,}75\ 14728{,}5\ ).$$

cuja inversa é

$$I(\hat{\beta}_0, \hat{\beta}_1)^{-1} = (54{,}44\ -0.80\ -0{,}80\ 0{,}012\ ).$$

As estimativas das variâncias $\left[\widehat{SE}(\hat{\beta}_0)\right]^2$ e $\left[\widehat{SE}(\hat{\beta}_1)\right]^2$são os elementos da diagonal da inversa da matriz $I(\hat{\beta}_0, \hat{\beta}_1)^{-1}$, I_{11}^{-1} e I_{22}^{-1}, respectivamente. A notação $\widehat{SE}(\hat{\beta}_0)$ representa o erro padrão de β_0.

De acordo com as equações (3.15) e (3.16) $\hat{\beta} \pm Z_{\frac{\alpha}{2}}\widehat{SE}(\widehat{\beta})$ é, para grandes amostras, um intervalo de confiança aproximado $100(1-\alpha)\%$ para β. Assim, o intervalo de confiança 95% para β_0 é dado por

$$15{,}043\ \pm\ 1{,}96 \times \sqrt{54{,}44} = [0.58\ ;\ 29.50]$$

e o intervalo de confiança 95%para para β_1 por

$-0{,}232\ \pm\ 1{,}96 \times \sqrt{0{,}012} = [-0.447\ ;\ -0.017]$, apoiando a conclusão que $\beta_1 < 0$.

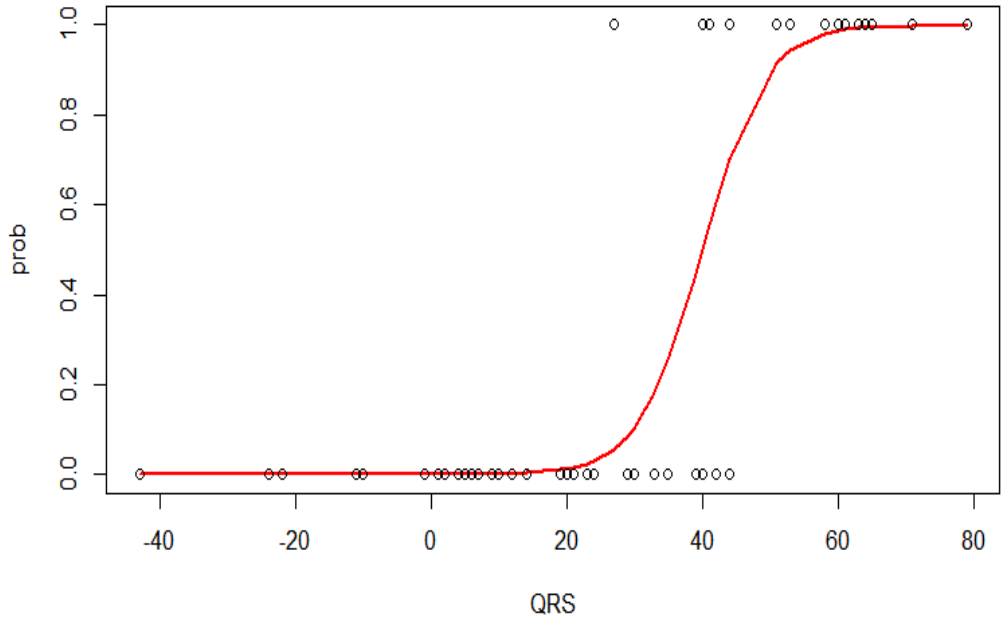
Às vezes, é mais comum neste modelo testar a hipótese $H_0\ :\ \beta = 0$, porque, como na regressão linear simples, essa hipótese afirma que não há relação entre a variável preditora e a variável de resposta. A estatística do teste de Wald, $Z = \hat{\beta}/\widehat{SE}(\hat{\beta})$, tem aproximadamente distribuição normal padrão, se H_0 for verdadeira e se o tamanho da amostra for grande. Portanto, H_0 de ser rejeitada se $|Z| \geq Z_{\frac{\alpha}{2}}$.

Exemplo 4.2. (LEE & WANG, 2003). Sabe-se que a adriamicina é eficaz no tratamento de certos tipos de câncer. Também é de conhecimento que a adriamicina é altamente tóxica. Alguns pacientes desenvolvem insuficiência cardíaca congestiva (ICC), mas outros que recebem uma dose semelhante de adriamicina não. Em uma tentativa de detectar fatores que aumentariam o risco de desenvolver cardiotoxicidade por adriamicina, várias características de 53 pacientes com câncer foram estudadas. Dezessete desses pacientes desenvolveram CHF e 36 pacientes não. Após uma investigação cuidadosa, descobriu-se que a dose total (z) e a diminuição percentual na voltagem do QRS eletrocardiográfico (z) estão mais intimamente relacionadas à ICC. A

Tabela 4.2. mostra os dados e algumas estatísticas amostrais. O seguinte modelo de regressão logística linear com variáveis transformadas.

O procedimento passo a passo (stepwise procedure) seleciona a diminuição percentual no QRS como a variável mais importante, seguida pela dose total (DT) e interação (DT x QRS). Os resultados da análise de regressão logística são apresentados. Os valores de log-verossimilhança passo a passo dados na última coluna indicam que apenas QRS é significativo desde $2(-10.185 + 33.254) = 46.138$, o que produz um p-valor menor que $0{,}001$. Nem a dose total, nem a interação são significativas.

Figura 4.2 - Dados e gráfico do modelo feito do exemplo 4.2.



Fonte: Elaborado pelo Autor (2023)

Para a obtenção das EMVs, intervalos de confiança assintóticos e testar hipóteses os códigos em linguagem R do Apêndice A, foram utilizados no Exemplo 4.2 e adaptados no Exemplo 4.1.

5. Considerações Finais

A regressão logística foi empregada para modelar situações em que a variável resposta é binária. Em geral, o objetivo é encontrar o melhor modelo para descrever a relação entre uma variável dependente e um conjunto de variáveis independentes, conhecidas como covariáveis.

Abordamos os principais conceitos e técnicas para o modelo de regressão logística. Podemos destacar a ideia principal para ajustar o modelo, a distinção entre o modelo de regressão logística e o modelo de regressão linear, a definição de variáveis independentes, a modelagem de relações entre variáveis explicativas categóricas ou quantitativas e a variável de resposta binária ou dicotômica, além das suposições e formas do modelo.

A aplicação da metodologia desenvolvida no Capítulo 4, explorada em dois exemplos iniciais, pode ser estendida para o modelo multivariado e para as demais categorias. Esses modelos também estão implementados em pacotes da linguagem *R* (R Core Team, 2020).

Apêndice A

Os Códigos em *R* para o Exemplo 4.2 são dados a seguir. Estes podem ser diretamente adaptados para o Exemplo 4.1.

```
### Exemplo 2
datalee=read.table("C:/...lee_ex_pp_391_2.txt", header=TRUE, strip.white=TRUE)
modelee2 <- glm(CHF ~datalee$QRS, data = datalee,
 family = binomial(link = "logit"))
summary(modelee2)
# Os dados (QRS,CHF) precisam estar ordenados
CHF<-datalee$CHF
QRS<-datalee$QRS
dt<-datalee$dt
qrs<-data.frame(QRS)
prob<-predict(modelee2, qrs, type='response')
plot(QRS, prob, xlab='QRS', type='l', col='red',lwd='2',
    ylab='prob');
points(QRS,CHF)
### Ou usando o ggplot2
library(ggplot2)
# load data from CSV
# Plot Predicted data and original data points
ggplot(datalee, aes(x=QRS, y=CHF)) +  geom_point() + coord_cartesian(xlim =c(-41, 79), ylim = c(0, 1))+
     stat_smooth(method="glm", color="red", se=FALSE,
            method.args = list(family=binomial))
```

Apêndice B

A Distribuição de t de Student

(MEYER,1983) Suponha-se que estimemos o, empregando a estimativa não - tendenciosa

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{2.1}$$

Nós consideraremos a variável aleatória

$$t = \frac{(\bar{X} - \mu)\sqrt{n}}{\hat{\sigma}} \tag{2.2}$$

É intuitivamente evidente que a distribuição de probabilidade da variável aleatória t deve ser muito mais complicada do que a de $Z = \left[\frac{(\bar{X}-\mu)}{\sigma}\right]\sqrt{n}$, porque na definição de t, ambos, numerador e denominador, são variáveis aleatórias enquanto Z é apenas uma função linear de $X_1, \dots, X_n$. Para obtermos a distribuição de t, levaremos em conta os seguintes fatos.

(a) $Z = \left[\frac{(\bar{X}-\mu)}{\sigma}\right]\sqrt{n}$ tem distribuição $N\ (0, 1)$.

(b) $V = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^2}$ tem uma distribuição de qui-quadrado, com $(n-1)$ graus de liberdade.

(c) $Z\ e\ V$ são variáveis aleatórias independentes. (Isto não é muito fácil de demonstrar, e aqui não o verificaremos.)

Com o auxílio do seguinte teorema, poderemos agora obter $fdp\ de\ t$:

Teorema 2.4 (MEYER,1983) Suponha-se que as variáveis aleatórias Z e V sejam independentes tenham, respectivamente, as distribuições $N\ (0, 1)\ x^2$. Defina-se

$$t = \frac{Z}{\sqrt{\frac{V}{k}}}$$

Então, a $fdp\ de\ t$ será dada por:

$$h_k(t) = \frac{\Gamma\left[\frac{(k+1)}{2}\right]}{\Gamma(\frac{k}{2})\sqrt{\pi k}}(1 + \frac{t}{k^2})^{\frac{-(k+1)}{2}}, -\infty < t < \infty \qquad (2.3)$$

Esta distribuição é conhecida como distribuição de t de Student, com graus de liberdade.

Voltaremos, agora, ao problema apresentado no início desta seção. Como obteremos um intervalo de confiança para a média de uma variável aleatória normalmente distribuída, se a variância for desconhecida?

Obteremos o seguinte intervalo de confiança para μ, com coeficiente de confiança (1 - α):

$$\left(\bar{X} - n^{-\frac{1}{2}}\hat{\sigma}t_{n-1,\frac{1-\alpha}{2}}, \bar{X} + n^{-\frac{1}{2}}\hat{\sigma}t_{n-1,\frac{1-\alpha}{2}}\right) \qquad (2.4)$$

Desse modo, o coeficiente de confiança acima apresenta a mesma estrutura que o anterior, com a importante diferença de que o valor conhecido de θ foi substituído pela sua estimativa $\hat{\theta}$e a constante $K_{\frac{1-\alpha}{2}}$, que anteriormente era obtida das tábuas da distribuição normal, foi substituída por $t_{n-1,\frac{1-\alpha}{2}}$, está obtida das tábuas da distribuição.

Concluímos que I não é uma constante, porque ele depende de $\hat{\theta}$, a qual por sua vez depende dos valores amostrais $(X_1, \ldots, X_n)$.

Método dos mínimos quadrados (M.M.Q.)

Definição 2.9 (MEYER,1983) Suponha-se que temos $E(Y) = \alpha X + \beta$, onde $\alpha, \beta \; e \; X$ são como explicadas acima. Seja $(x_1, Y_1), \ldots, (x_n, Y_n)$ uma amostra aleatória de Y. As estimativas de mínimos quadrados dos parâmetros $\alpha \; e \; \beta$ são aqueles valores de $\alpha \; e \; \beta$ que tornam mínima a expressão:

$$\sum_{i=1}^{n}[Y_i - (\alpha x_i + \beta)]^2 \qquad (2.5)$$

A fim de obter as estimativas desejadas para $\alpha\ e\ \beta$, procederemos da seguinte maneira: Façamos $S(\alpha,\beta)\ =\ \sum_{i=1}^{n}[Y_i-(\alpha x_i+\beta]^2$. Para tornar mínimo $S(\alpha,\beta)$, deveremos resolver as equações:

$$\frac{\partial S}{\partial \alpha}=0\ ;\ \frac{\partial S}{\partial \beta}=0 \qquad (2.6-7)$$

Derivando S, em relação a $\alpha\ e\ \beta$, obteremos:

$$\frac{\partial S}{\partial \alpha}=\ \sum_{i=1}^{n}2[Y_i-(\alpha x_i+\beta)](-x_i)=\ -2\ \sum_{i=1}^{n}[Y_i x_i-\alpha x_i^2-\beta x_i] \qquad (2.8)$$

$$\frac{\partial S}{\partial \beta}=\ \sum_{i=1}^{n}2[Y_i-(\alpha x_i+\beta)](-1)=\ -2\ \sum_{i=1}^{n}[Y_i-\alpha x_i-\beta] \qquad (2.9)$$

Por isso, $\frac{\partial S}{\partial \alpha}=0\ e\ \frac{\partial S}{\partial \beta}=0$ podem ser escritas, respectivamente:

$$\alpha\sum_{i=1}^{n}x_i^2+\ \beta\sum_{i=1}^{n}x_i=\ \sum_{i=1}^{n}Y_i x_i \qquad (2.10)$$

$$\alpha\sum_{i=1}^{n}x_i+n\beta=\ \sum_{i=1}^{n}Y_i \qquad (2.11)$$

Portanto, teremos duas equações lineares nas incógnitas $\alpha\ e\ \beta$. A solução poderá ser obtida da maneira usual, quer por eliminação direta, quer com o emprego de determinantes. Denotando as soluções por $\hat{\alpha}\ e\ \widehat{\beta}$, facilmente verificaremos que:

$$\hat{\alpha}=\frac{\sum_{i=1}^{n}\left(y_i-\underline{y}\right)\left(x_i-\underline{x}\right)}{\sum_{i=1}^{n}\left(x_i-\underline{x}\right)^2},\qquad onde\qquad \underline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i \qquad (2.12)$$

$$\hat{\beta}=\ \underline{Y}-\ \hat{\alpha}\underline{x},\qquad onde\quad \underline{Y}=\frac{1}{n}\sum_{i=1}^{n}Y_i \qquad (2.13)$$

As soluções acima serão sempre viáveis e únicas, desde que

$$\sum_{i=1}^{n}\left(x_i - \underline{x}\right)^2 \neq 0 \tag{2.14}$$

Porém, esta condição será satisfeita sempre que nem todos os x_i sejam iguais. A estimativa do parâmetro σ^2 não pode ser obtida pelos métodos acima. Vamos apenas afirmar que a estimativa usual de σ^2, em termos das estimativas de mínimos-quadrados $\hat{\alpha}\ e\ \widehat{\beta}$, é

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left[Y_i - (\hat{\alpha}x_i + \hat{\beta}\right]^2 \tag{2.15}$$

Intervalos de confiança.

(MEYER,1983) Suponha-se que X tenha distribuição $N(\mu;\sigma^2)$, onde se supõe σ^2 conhecido, enquanto é o parâmetro desconhecido. Seja $X_1, \ldots, X_n$uma amostra aleatória de X e seja $\underline{X}$a média amostral.

Sabemos que X tem distribuição $N(\mu;\frac{\sigma^2}{n})$ portanto, $Z = \left[\frac{(\underline{X}-\mu)}{\sigma}\right]\sqrt{n}$tem distribuição $N\ (0,1)$. Observe-se que, muito embora Z dependa de μ, sua distribuição de probabilidade não depende. Empregaremos este fato a nosso favor da seguinte maneira:

Considere-se

$$2\Phi(z) - 1 = P\left(-z < \frac{(\bar{X}-\mu)}{\sigma}\sqrt{n} < z\right) = P\left(\frac{-z\sigma}{\sqrt{n}} - \bar{X} \leq -\mu \leq \frac{z\sigma}{\sqrt{n}} - \bar{X}\right) =$$
$$= P\left(\bar{X} - \frac{z\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{z\sigma}{\sqrt{n}}\right) \tag{2.16}$$

Esta última expressão de probabilidade deve ser interpretada muito cuidadosamente. Ela não significa que a probabilidade do parâmetro μ cair dentro de um intervalo especificado seja igual a $2\Phi(z) - 1$; μ é um parâmetro, e ou está ou não está dentro do intervalo acima. De preferência, a expressão acima deve ser interpretada assim: $2\Phi(z) - 1$ é igual à probabilidade de que o intervalo $\left(\bar{X} - \frac{z\sigma}{\sqrt{n}}\ , \bar{X} + \frac{z\sigma}{\sqrt{n}}\right)$ contenha μ. Tal intervalo é denominado intervalo de confiança do parâmetro μ. Desde quez é

arbitrário, poderemos escolhe-lo de modo que a probabilidade acima seja igual, por exemplo, a $1 - \alpha$. Consequentemente, z ficará definido pela relação $\Phi(z) = \frac{1-\alpha}{2}$. Aquele valor de z, denotado por $K_{\frac{1-\alpha}{2}}$ pode ser obtido das tabuas da distribuição normal. Isto é, teremos $\Phi\left(K_{\frac{1-\alpha}{2}}\right) = \frac{1-\alpha}{2}$.

Em resumo: O intervalo $(\bar{X} - n^{\frac{1}{2}}\sigma K_{\frac{1-\alpha}{2}}, \bar{X} + n^{\frac{1}{2}}\sigma K_{\frac{1-\alpha}{2}})$ é um intervalo de confiança do parâmetro μ, com coeficiente de confiança $(1 - a)$, ou um intervalo de confiança $100\ (1 - a)$ por cento.

Referências Bibliográficas

BOLFARINE, Heleno; SANDOVAL, Mônica Carneiro. **Introdução à inferência estatística**. SBM, 2001.

CASELLA, George.; BERGER, Roger. L. **Statistical Inference**. 2002.

HOSMER JR, David W.; LEMESHOW, Stanley; STURDIVANT, Rodney X. **Applied logistic regression**. John Wiley & Sons, 2013.

LEE, Elisa T.; WANG, John. **Statistical methods for survival data analysis**. John Wiley & Sons, 2003.

MEYER, Paul L. **Probabilidade: aplicações à estatística**. Rio de Janeiro: Livros Técnicos e Científicos, 1983.

R Core Team. R: **A Language and Environment for Statistical Computing**. Vienna, Austria, 2020. Disponível em: <https://www.R-project.org/>