Revisão e Notas sobre Regressão Logística

Pedro Rupf Pereira Viana

5 de fevereiro de 2026

1 Introdução e Objetivos

Dentre as diversas técnicas de mineração de dados (isto é, o processo de descoberta automática de informações úteis em grandes depósitos de dados), a regressão logística difere das outras técnicas de mineração, principalmente pelo fato de sua variável dependente ser categórica; e mesmo quando ela não é dicotômica, é possível torná-la dicotômica, com a finalidade de aplicar esta técnica. Em relação as variáveis independentes, estas podem ser categóricas ou métricas.

Desta forma, a Regressão Logística é uma técnica que avalia a probabilidade de obtenção de uma das categorias da variável dependente, portanto, é capaz de obter a probabilidade de ocorrência de determinado evento, assim como a influência de cada variável independente no evento estudado.

2 Desenvolvimento

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, em função de uma ou mais variáveis independentes contínuas e/ou binárias. Então, a partir desse modelo gerado é possível calcular ou prever a probabilidade de um evento ocorrer, dado uma observação aleatória.

Suponha que queira-se analisar a ocorrência da apneia do sono, que é um distúrbio do sono potencialmente grave, em que a pessoa para de respirar, por alguns segundos, diversas vezes durante a noite. Existem vários fatores que podem influenciar nesse distúrbio, mas para este exemplo, vamos considerar apenas dois: *idade* e *peso*. Digamos que para esta análise, tenhamos uma amostra de cem indivíduos, contendo a idade, o peso e se ele tem apneia ou não, este é o nosso conjunto de observações. A variável dependente é a ocorrência ou não da apneia do sono, ter apneia é igual a 1, não ter apneia é igual a 0. As variáveis independentes são a idade e o peso. Para este exemplo, o que a regressão logística propõe é que, a partir dessas informações, é possível gerar um modelo logístico que possa prever a probabilidade de uma pessoa ter apneia do sono, baseando-se no peso e idade desta pessoa. Mas como veremos a seguir, este é apenas um dos objetivos da regressão logística.

O modelo de Regressão Logística permite:

- Modelar a probabilidade de um evento ocorrer dependendo dos valores das variáveis independentes, que podem ser categóricas ou contínuas;

  Então digamos que, a partir do modelo logístico gerado do problema da apneia do sono, queiramos saber qual a probabilidade de um indivíduo de 50 anos e 120 quilos, ter ou vir a desenvolver a apneia do sono. Ao inserir os dados no modelo, o resultado será um valor entre 0 e 1 que representa esta probabilidade. Suponhamos que o valor seja 0,75, assim uma pessoa de 50 anos e 120 quilos tem 75% de probabilidade de ter apneia do sono.

- Estimar a probabilidade de um evento ocorrer para uma observação selecionada aleatoriamente contra a probabilidade do evento não ocorrer;

  Se uma pessoa de 50 anos e 120 quilos tem probabilidade $p = 0,75$ de ter apneia. A probabilidade de não ter apneia é $1 - p$, logo, $1 - p = 0,25$. A probabilidade de um evento ocorrer, contra ele não ocorrer, é uma razão de probabilidades, $p/(1-p)$,

que é chamada de **chance**. Assim temos $0,75/0,25 = 3$, e isto significa que uma pessoa nessas características tem 3 vezes mais chance de ter apneia do sono do que de não ter.

- Prever o efeito do conjunto de variáveis sobre a variável dependente binária;

  Através da análise de regressão logística, pode-se concluir, por exemplo, que a variável peso é bastante significativa para o modelo de regressão, enquanto que a variável idade não contribui tanto para a eficácia do mesmo.

- classificar observações, estimando a probabilidade de uma observação estar em uma categoria determinada;

  A análise de regressão logística pode informar por exemplo, que indivíduos obesos, ou acima de uma determinada idade, podem ser mais propensos à esse distúrbio.

2.1 Função log it

Anteriormente, foi dito que a variável dependente na regressão logística segue a distribuição de Bernoulli, portanto é preciso conectar as variáveis independentes à distribuição Bernoulli presente na variável dependente e esse link é chamado de log it. Na regressão logística nós não conhecemos a probabilidade p, como é o padrão nos problemas de distribuição de Bernoulli. Logo, o objetivo do modelo logístico é estimar p para uma combinação linear das variáveis independentes. O p estimado é $\hat{p}$.

Para ligar a combinação linear de variáveis à distribuição de Bernoulli, é necessário uma função que as una, ou mapeie a combinação linear de variáveis que poderiam retornar qualquer valor em uma distribuição de probabilidades bernoulli com um domínio de 0 a 1. A razão de probabilidade é chamada de **chance**, e seu logaritmo natural, *log it*, é esta função representada na equação abaixo:

$$\ln\left(\textbf{chance}\right) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \tag{2.1}$$

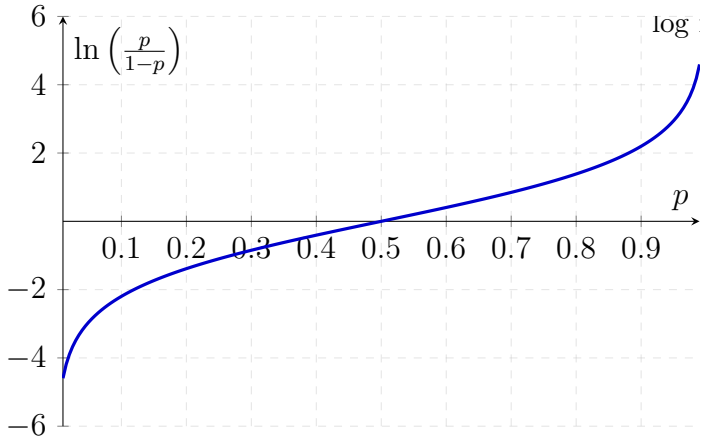

Figura 1: Função log it – relação entre probabilidade p e log-razão de chances

Pelo gráfico da função log it na Figura 1 acima, é possível compreendê-la melhor. A função vai a 0 mas não chega a tocar o eixo y, e o mesmo ocorre quando ela vai a 1. O que fica comprovado quando substituímos os valores na equação. Quando $p = 0$,

$\ln(0/1) = \ln(0) =$ indefinido. Quando $p = 1$, $ln(1/0) = ln(\infty) = \infty$. Ou seja, a função está dentro desse intervalo de 0 a 1, e quando estamos lidando com probabilidade, isto é algo muito útil, pois a probabilidade também é representada por valores dentro desse domínio. Deste modo, pela função logística, nunca poderá se obter uma probabilidade superior a 100% ou inferior a 0%.

Observando ainda a Figura 1, vejamos que quando p = 0.5 a função é 0. Substituindo o valor de p na função: $\ln(0,5/0,5) = \ln(1) = 0$. Isso significa que quando as probabilidades são iguais, a chance (isto é, a razão de probabilidades) é 1 e que o log it é 0.

No gráfico da função log it, os valores entre 0 e 1 percorreram o eixo x, mas se queremos que as probabilidades estejam no eixo y, isto pode ser obtido através da inversa da função log it. A partir da equação(2.1), temos que:

$$\log it^{-1}(\alpha) = \frac{1}{1 + e^{-\alpha}} = \frac{e^{\alpha}}{1 + e^{\alpha}} \tag{2.2}$$

onde $\alpha = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$. No modelo de regressão logística, α, será a combinação linear das variáveis e seus coeficientes. A inversa da função de log it retornará a probabilidade da variável dependente Y ser igual a 1.

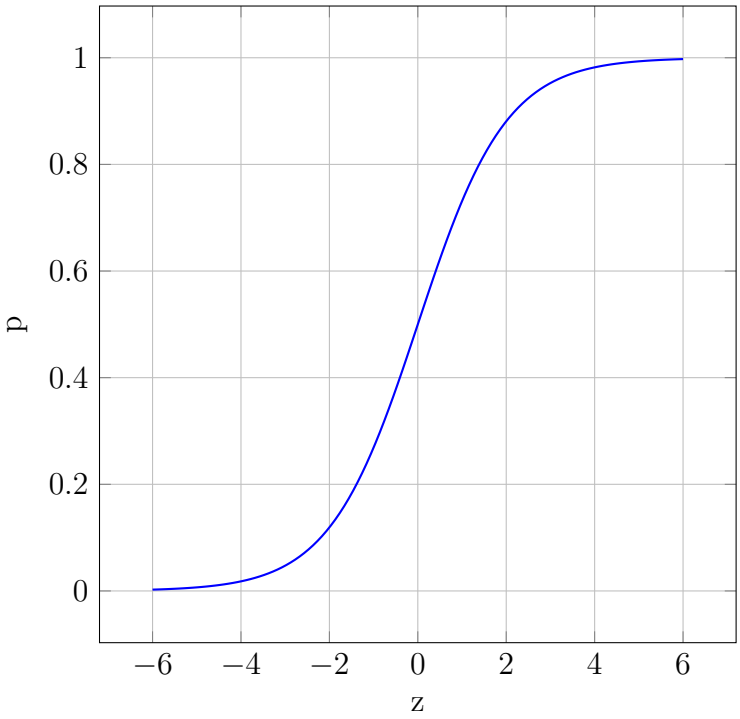

Figura 2: Função logística (inversa do logit)

Na Figura 2, observa-se que, o gráfico da inversa do log it é o mesmo do log it, apenas 90 graus invertido. Foi efetuada basicamente uma troca das coordenadas x e y, agora ao invés de ter o domínio da função de 0 a 1 no eixo x, temos o domínio de 0 a 1 no eixo y.

A representação gráfica da função inversa do log it na Figura 2 assume a forma parecida com um "S", também chamada de curva *sigmóide*, havendo áreas onde a mudança é acentuada e onde ela nem ocorre. As áreas onde pequenas variações nos valores de x causam grandes mudanças em valores de y representam áreas de maior probabilidade de mudança de estado da variável y em função de x.

3 Métodos de avaliação do modelo logístico

Após estimar os coeficientes, temos interesse em assegurar a significância das variáveis no modelo. Isto geralmente envolve formulação e teste de uma hipótese estatística para determinar se as variáveis independentes no modelo são significativamente relacionadas com a variável dependente. Para isto, há testes para avaliar o modelo logístico. Os testes mais utilizados são os testes da Razão da Verossimilhança, o teste de Wald e Pseudo R^2 de Cox e Snell.

3.1 Teste da Razão de Verossimilhança

Uma vez ajustado o modelo, é necessário testar a significância do modelo estimado, e isto pode ser feito através do teste da razão de verossimilhança. Esta medida testa simultaneamente se os coeficientes de regressão associados a β são todos nulos, com exceção de β_0. A comparação entre os valores observados e esperados usando a função de verossimilhança, é expressa da seguinte forma:

$$D = -2\ln\left(\frac{L(\beta)}{L(\beta_0)}\right) = -2\ln\left(\frac{L(\beta)}{L(\beta_0, \beta_1 = 0, ..., \beta_k = 0)}\right) \tag{3.1}$$

$$D = -2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\hat{\pi}_i}{p_i}\right) + (1 - y_i)\ln\left(\frac{1-\hat{\pi}_i}{1-p_i}\right)\right] \tag{3.2}$$

O modelo é dito saturado se contem todas as variáveis, enquanto o modelo ajustado corresponde ao modelo apenas com as variáveis desejadas para o estudo. Esta função D, também chamada de deviance (desvio), sempre é positiva e quanto menor, melhor é o ajuste do modelo. Considere a hipótese abaixo para testes:

$$H_0 \colon \beta_1 = \cdots = \beta_t = 0 \quad \text{versus} \quad H_1 \colon \exists j = 1, \ldots, p \ : \ \beta_j \neq 0$$

na hipótese nula H_0 a ser testada, os parâmetros do modelo serão igualados a 0. O modelo saturado que mantém o valor de seus coeficientes, representará a hipótese alternativa H_1.

Para estimar a significância de uma variável independente, comparam-se o valor de D com e sem variável independente na equação. A alteração no valor de D esperada pela inclusão da variável independente no modelo é obtida através de:

$$G = D\left(\frac{\text{modelo sem variável}}{\text{modelo com variável}}\right) \tag{3.3}$$

Ao rejeitar a hipótese nula, tem-se que a variável independente testada, é significativa para o modelo.

3.2 Teste de Wald

O teste de Wald é também utilizado na regressão logística para a determinação da significância dos coeficientes do modelo estimado, ele testa se cada coeficiente é significativamente diferente de zero. Deste modo, o teste de Wald verifica se uma determinada variável independente possui uma relação estatisticamente significativa com a variável dependente.

Se os coeficientes logísticos forem estatisticamente significativos, podemos interpretá-los em termos de seu impacto na probabilidade estimada, deste modo, na predição do objeto de estudo no grupo respectivo, isto é, no grupo do evento de interesse ($Y = 1$), ou no grupo da não ocorrência do evento ($Y = 0$).

O teste de Wald é obtido comparando a estimativa de máxima verossimilhança de um coeficiente e a estimativa do seu erro padrão:

$$W_j = \frac{\hat{\beta}_j}{\text{Var}(\hat{\beta}_j)} \tag{3.4}$$

3.3 Pseudo R^2 de Cox e Snell

A medida de ajuste do modelo logístico é o pseudo R^2 de Cox e Snell, que é uma extensão do coeficiente de determinação R^2 do modelo linear. Ele é definido como:

$$R^2_{CS} = 1 - \left(\frac{L(\beta_0)}{L(\beta)}\right)^{\frac{2}{n}} \tag{3.5}$$

onde $L(\beta_0)$ é a verossimilhança do modelo com apenas o termo constante e $L(\beta)$ é a verossimilhança do modelo completo. Este valor varia entre 0 e 1, sendo que valores mais altos indicam um melhor ajuste do modelo. No entanto, o valor máximo possível deste coeficiente não é 1, o que pode dificultar a interpretação em comparação com o R^2 linear.

O pseudo R^2 de Nagelkerke é uma versão normalizada deste coeficiente, definida como:

$$R^2_N = \frac{R^2_{CS}}{1 - L(\beta_0)^{\frac{2}{n}}} \tag{3.6}$$

que varia entre 0 e 1, permitindo uma interpretação mais direta em comparação com o R^2 linear.

4 Aplicações da Regressão Logística

A Regressão Logística tornou-se uma técnica padrão para análise de regressão de dados dicotômicos, principalmente nas ciências médicas. Mas ela apresentou um crescimento muito rápido, se expandindo para outras áreas além da saúde, sendo utilizada também no campo da econometria, administração, educação, ambiental e outros.

4.1 Óbito Neonatal

O período neonatal corresponde às quatro primeiras semanas de vida (0 a 28 dias incompletos). Denomina-se *período neonatal precoce* a primeira semana completa ou os sete primeiros dias de vida, e período neonatal tardio, as três semanas seguintes.

Óbito neonatal é o óbito que ocorre no período neonatal, isto é, entre 0 e 28 dias incompletos após o nascimento. A criança morta dentro deste período, dá-se o nome de *neomorto*.

A mortalidade infantil pode ser considerada um dos melhores indicadores da qualidade da assistência à saúde, bem como do nível socioeconômico de uma população. Tal índice compreende todos os óbitos de crianças com menos de um ano de idade, sendo formada

pelo óbito neonatal e o pós-neonatal, que abrange os óbitos ocorridos do 28º dia até um dia antes de se completar um ano de vida.

No Brasil, a taxa de mortalidade infantil teve redução de 50% entre 1990 e 2008. Mas apesar das taxas de mortalidade infantil estarem em queda, os dados indicam concentração dos óbitos no período neonatal, que ainda se mantém com valores elevados em comparação com as taxas de mortalidade pós-neonatal.

Para tentar reduzir estes valores elevados da mortalidade no período neonatal, técnicas como a regressão logística podem ser utilizadas para que se construa modelos que possibilitem identificar os fatores de risco para o óbito neonatal.

4.1.1 Regressão Logística no estudo do óbito neonatal

Modelos de análise clássicos pressupõem independência entre indivíduos e homogeneidade de variância, e desconsideram a hierarquia dos fatores preditores, isto é, não consideram que observações originadas de uma mesma unidade podem ser mais similares do que aquelas originadas de diferentes unidades. Isso pode levar à superestimação dos efeitos do agrupamento e induzir a conclusões imprecisas.

A análise de regressão logística é uma alternativa aos modelos clássicos ao considerar a variável dependente, em nível dicotômico, e as variáveis independentes, em qualquer nível, categórico ou contínuo.

Assim, para facilitar a identificação e a compreensão dos fatores associados ao óbito neonatal, pode-se dizer que a regressão logística é uma técnica apropriada. A variável dependente binária Y é a ocorrência ou não ocorrência do óbito em crianças com menos de 28 dias de vida.

$$Y = \begin{cases} 1 & \text{se ocorrência de óbito neonatal} \\ 0 & \text{caso contrário (não ocorrência de óbito neonatal)} \end{cases}$$

Características da mãe e da criança, assim como características socioeconômicas, são analisadas como determinantes da mortalidade infantil, algumas das mais utilizadas em trabalhos neste contexto são: peso ao nascer, se a criança é pré-termo, escore de *Apgar* no 1o e 5o minuto, idade da mãe, escolaridade materna, tipo de parto (normal ou cesárea), tipo de gestação (única ou múltipla), tabagismo, entre outros, renda familiar, número de pessoas que moram no mesmo domicílio, entre outros.

O uso de um modelo para avaliar os fatores de risco para a mortalidade infantil neonatal, compreendidos como indicadores de várias dimensões das condições de vida, é importante para compreender o quanto estes indicadores influenciam na ocorrência do óbito neonatal. Sendo assim possível, identificar grupos expostos a diferentes fatores de risco e detectar necessidades de saúde em diferentes subgrupos populacionais. Isto aumenta a esperança de que estes fatores possam ser minimizados e talvez até evitados.

4.1.2 Aplicação Exemplo

Martins e Velásquez-Meléndez (2004) testaram a associação de vários fatores com a mortalidade neonatal em Montes Claros, utilizando a técnica de regressão logística. Todos os dados e resultados apresentados nesta subseção foram obtidos pelos referidos autores.

De acordo com os autores, a população constituiu-se de 20.506 nascidos vivos na cidade de Montes Claros, MG, Brasil, entre o período de 1 de janeiro de 1997 a 31 de dezembro de 1999. Foi verificado banco de dados de óbitos e de nascimentos para identificar os

nascidos vivos que evoluíram para o óbito neonatal, no qual verificou-se 275 casos neste período. Após a verificação de registros com variáveis com valor omisso, foram excluídos 1491 registros, portanto a base amostral utilizada no estudo totalizou 19.015 registros.

A variável dependente do estudo foi a ocorrência de óbito neonatal (1 para a ocorrência do óbito, 0 para a não ocorrência do óbito), e as variáveis independentes estão relacionadas ao recém-nascido (sexo, peso ao nascer, escore de *Apgar* no 1º e 5º minutos de vida e idade gestacional), à gestação e parto (tipo de gravidez e parto, número de consultas de pré-natal e local do nascimento) e à mãe (grau de instrução, idade, filhos e abortos tidos). Segundo os autores, todas as associações entre os preditores e a variável dependente foram consideradas estatisticamente significantes.

Sobre o escore de *Apgar*, ele é um teste de avaliação de cinco sinais vitais do recém-nascido realizado no primeiro, quinto e décimo minuto após o nascimento. A pontuação varia de 0 a 10, e quanto mais próximo de 10, melhor.

Os autores apresentaram os resultados do estudo atráves dos valores de *odds ratio* (OR) obtidos pelo método da regressão logística. A Tabela 1 abaixo informa o OR de quatro das variáveis independentes inclusas neste estudo.

Tabela 1: Fatores associados ao óbito neonatal – Odds Ratio (OR) ajustado

Variáveis	OR
Peso ao nascer (1 = baixo peso; 0 = peso normal)	4,94
Idade gestacional ($\leq$ 37 sem. (1); $>$ 37 sem. (0))	5,68
Apgar 1º minuto (0 a 10)	0,75
Apgar 5º minuto (0 a 10)	0,76

O *odds ratio*, ou razão de chances, pode ser definido como a razão entre a chance de um evento ocorrer em um grupo e a chance de ocorrer em outro grupo. Em regressão logística, esta razão aparece diretamente relacionada aos coeficientes das variáveis independentes, favorecendo a interpretação dos resultados obtidos. Uma razão de chances de valor igual a 1 indica que o evento em estudo tem chances iguais de ocorrer nos dois grupos, maior que 1 indica chance maior de ocorrer no primeiro grupo, e entre 0 e 1 indica chance menor de ocorrer no primeiro grupo.

Trazendo este conceito para o estudo em questão, pela Tabela 1, tem-se que a variável "peso ao nascer" é binária (1 = baixo peso; 0 = peso normal), portanto, a leitura do valor do OR indica que recém-nascidos com baixo peso tem 4,94 mais chances de evoluir à óbito do que recém-nascidos com peso normal. Da mesma forma, crianças pré-termo (isto é, nascidas com menos de 37 semanas de gestação) tem 5,68 mais chances de evoluir à óbito em comparação às crianças a termo. As variáveis "Apgar 1º minuto" e "Apgar 5º minuto" são discretas e variam de unidade a unidade até 10, neste caso, a chance de um grupo é comparada à chance do grupo de unidade anterior. Observa-se ainda que elas possuem OR abaixo de 1, isto significa, por exemplo, que uma criança com score *Apgar* igual a 7 tem 0,75 menos chance de evoluir à óbito em comparacão com uma crianças com score *Apgar* igual a 6.

MARTINS, E. F.; VELÁSQUEZ-MELÉNDEZ, G. Determinantes da mortalidade neonatal a partir de uma coorte de nascidos vivos, montes claros, minas gerais, 1997–1999. *Rev. Bras. Saúde Matern. Infant.*, Recife, v. 4, n. 4, p. 405–412, 2004.