Regressão Logística

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. Os modelos de regressão constituem uma das ferramentas estatísticas mais importantes na análise estatística de dados quando se pretende modelar relações entre variáveis. Enquanto na regressão linear temos uma variável resposta contínua, na regressão logística a variável resposta é binária, 0 ou 1.

Tal como referido por Hosmer, Lemeshow & Sturdivant (2013), nos modelos de regressão linear simples ou múltipla, a variável dependente é uma variável aleatória de natureza contínua, sendo esta em alguns casos qualitativa e expressa em função de duas ou mais variáveis de natureza categórica, isto é, admite dois ou mais valores (Hosmer, Lemeshow, & Sturdivant, 2013). Assim, o que distingue o modelo de regressão logística do da regressão linear é que a variável resultado na regressão logística é binária (dicotómica) (Hosmer et al., 2013). Uma vez contabilizada esta diferença, os métodos utilizados numa análise utilizando a regressão logística seguem, mais ou menos, os mesmos princípios gerais utilizados na regressão linear (Hosmer et al., 2013).

As técnicas de regressão são versáteis na sua aplicação à investigação médica, pois permitem medir as associações, prever resultados, e controlar variáveis. Assim sendo, a regressão logística é uma forma eficiente e poderosa de analisar o efeito de um grupo de variáveis independentes sobre um resultado binário, quantificando a contribuição única de cada variável independente. Utilizando componentes da regressão linear refletidos na escala logítica, a regressão logística identifica iterativamente a combinação linear mais forte de variáveis com a maior probabilidade de detetar o resultado observado (Stoltzfus, 2011).

Foi usada a notação para representar a probabilidade condicionada em que Y se refere à variável resultado e x representa um valor especifico da variável independente. E lê-se a probabilidade de ocorrer o valor Y, dado o valor de x. Na regressão linear assumimos que esta média pode ser expressa como uma equação linear em x (ou alguma transformação de x ou Y ), tal como:

Esta expressão implica que para é possível tomar qualquer valor entre -e +.

A transformação do é o centro do estudo da regressão logística. Sendo esta transformação definida como:

A importância desta transformação é que a função g(x) passa a ter muitas das propriedades desejáveis de um modelo de regressão linear. A função logit pode ser linear no seus parâmetros, ser contínua e variar entre os valores de -a +, dependendo do domínio de x.

A quantidade é chamada o erro e expressa um desvio da observação em relação à média condicional. No caso de uma variável resposta dicotómica, o valor da variável resposta x é dada pela expressão:

em que o valor do erro, , pode assumir dois valores:

* se y = 1 então com a probabilidade ;
* se y=0 então com a probabilidade de .

Assim, apresenta uma distribuição com média zero e variância igual a ] Ou seja, a distribuição condicional da variável resultado segue uma distribuição binomial com probabilidade dada pela média condicional.

sendo *p* a probabilidade de ocorrência do evento e 1-p a probabilidade de não ocorrência.

Seja Y, uma variável dependente (resposta), e sejam (p-1) variáveis independentes. O objetivo é a construção de um modelo que relacione a variável aleatória Y

A função logit é mais interpretável do que a função logística comum. Portanto, esta função nada mais é do que uma função sigmóide que fornece os valores de “0” e “1”.

Para valores de X=0 e o valor de Y é como é possível verificar através do gráfico apresentado na figura X. Esta é a principal razão para o modelo logístico ser tão popular, pois possibilita a descrição da probabilidade de o paciente desenvolver ou não *delirium*.

Sendoe válido para um conjunto entre i = 1,2,..., p-1 até p-1 variáveis

A função logit pode ser linear no s parâmetros, ser contínuo e varia entre os valores de -a +, dependendo do domínio de X.

Os pressupostos básicos que devem ser cumpridos para a regressão logística incluem independência de erros, linearidade no logit para variáveis contínuas, ausência de multicolinearidade, e ausência de *outliers* fortemente influentes. Além disso, deve haver um número adequado de eventos por variável independente para evitar um modelo de “*overfiting*”. A adequação global do modelo de regressão logística resultante é avaliada utilizando várias medidas, sendo que os melhores resultados são caracterizados por uma menor diferença entre os valores observados e os valores previstos no modelo. Finalmente, os resultados para variáveis independentes são normalmente reportados como *odds ratios* (ORs) com intervalos de confiança de 95% (CIs) (Stoltzfus, 2011). A análise da curva ROC (Receiver Operating Characteristic) pode ser feita através de um gráfico que permite estudar a variação da sensibilidade e especificidade para cada valor de *cut-off.* O objetivo desta análise é identificar ou confirmar a qualidade do ajustamento do modelo. Aquando da observação do gráfico verifica-se que o ideal seria encontrar uma área sob a curva ROC (AUC-ROC) perto de 1, uma vez que, quanto mais próxima estiver a curva do canto superior esquerdo, mais verdadeiros positivos e menos falsos negativos serão obtidos.

Pode-se assim concluir que a regressão logística é uma forma eficiente e poderosa de avaliar contribuições variáveis independentes para um resultado binário, mas a sua exatidão depende em grande parte de uma cuidadosa seleção das variáveis com satisfação dos pressupostos básicos, bem como da escolha apropriada da estratégia de construção do modelo e validação dos resultados (Stoltzfus, 2011).

**Critérios de avaliação de modelos**

Dentre os critérios para seleção de modelos, os critérios baseados no máximo da função de verossimilhança (MFV) são os mais utilizados:

* Teste da Razão de Verossimilhança (TRV ou LR)
* Critério de Informação de Akaike (AIC)

O Akaike information criterion (AIC) é um critério amplamente utilizado na seleção do melhor modelo entre um conjunto de modelos candidatos (Yanagihara et al., 2012). Cada modelo é caracterizado por um valor de AIC, pelo que, a seleção do melhor modelo é seguida por aquele que apresentar o menor valor de AIC.

O critério AIC é definido pela seguinte equação:

onde 𝑘 representa o número de parâmetros independentes e 𝐿 representa o valor obtido da função de máxima verosimilhança do modelo.

* Critério Bayesiano de Schwarz (BIC).

O Bayesian Information Criterion (BIC), também é utilizado na seleção do melhor modelo, sendo que, aquele que apresentar o menor valor de BIC será o escolhido (Konishi & Kitagawa, 2008). O BIC é um critério baseado na probabilidade bayesiana e representa-se pela seguinte expressão:

**Tenho de melhorar a forma como está apresentado e ainda talvez inserir algumas das fórmulas mais importantes deste modelo.**