Regressão Logística

Mateus Fernandes de Souza

19 de março de 2022

1 Definição

A Regressão Logística (RegLog) é uma técnica tradicional estatística que utiliza observações indendentes para formar um modelo que possibilita predizer valores, normalmente de forma binária. Em sua forma, ela é muito parecida com a Regressão Linear, mas a grande diferença é que nessa a variável resposta pode ser vários valores numéricos, enquanto na logística ela assume valores entre 0 e 1, tratando assim de probabilidades, essa que significa qual é a probabilidade da variável resposta assuma o valor 1. Pelo padrão, todas as respostas com valores menos que 0,5 são consideradas com 0, enquanto as maiores ou iguals a 0,5 são consideradas 1, mas isso pode mudar de acordo com os parâmetros buscados por esse teste.

1.1 Fórmula

$p(X) = \frac{1}{1+e^{-(x'\beta)}}$, Sendo $(x'\beta) = \beta 0 + \beta 1Xi1 + \beta 2Xi2 + ... + \beta kXik$

1.2 Como encontrar β

Para encontrar os valores de β, é necessário fazer o seguinte cáculo: Considerando a variável resposta com distribuição de Bernoulli com função de probabilidade.

yi = 1, então $P(yi = 1) = \pi i$

yi = 0, então $P(yi = 0) = 1 - \pi i$

O Valor esperado da variável resposta é $E(yi) = x'\beta = \pi i$

$Odds = \frac{\pi}{1-\pi}$

Então, temos que $ln\frac{odds(xi+1)}{odds_x i} = \beta 1$

1.3 Interpretação dos parâmetros

$oddsration = \frac{odds(xi+1}{odds_x 1} = e^{\beta}1$

Exemplo: Varíavel resposta é do tipo morrer(1) e não morrer(0), e a variável preditora que está sendo analisada é a idade, então caso eu encontre, por exemplo, um valor para odds ration = 2, isso significa que a chance de morrer ao aumentar em 1 ano na idade aumenta em 2x em relação àquele que tem menos 1 ano de idade.

1.4 Como saber se uma variável é importante para o modelo

Após estimar os coeficientes β, temos interesse em assegurar a significância das variáveis do modelo. isto geralmente envolve a formulação e teste de uma hipótese estatística para determinar se a variável preditora no modelo é significativamente relacionada com a variável resposta. Os testes de hipóteses mais utilizados são os testes da Razão da Verossimilhança e Wald.

1.5 Medidas da qualidade do ajuste do modelo

O desempenho geral do modelo ajustado pode ser medido por diversos testes de qualidade de ajuste. Dois testes requerem dados replicados (múltiplas obsaervações com os mesmos valores para todos os

preditores): Qui Quadradado de Pearson e Deviance. O teste de Hosmer-Lemeshow é útil para conjuntos de dados não replicados ou que contêm apenas algumas observações replicadas (as observações são agrupadas com base em suas probabilidades estimadas)

1.6 Desempenho do modelo

Para avaliar o desempenho do modelo pode-se utilizar: Acurácia, Recall, Especifidade, Precisão e ROC-AUC. Essas são as principais formas de avaliação.

1.7 Gráfico