R para Ciência de Dados

Regressão Logística

Profa Carolina Paraíba e Prof Gilberto Sassi

Departamento de Estatística
Instituto de Matemática e Estatística
Universidade Federal da Bahia

Setembro de 2024

R para Ciência de Dados: Regressão Logística

R para Ciência de Dados: Regressão Logística

- Introdução
- Modelo de Regressão Logística
- Inferência
- Análise de Resíduos
- Diagnóstico de Influência
- Classificação
- Curva ROC
- Exemplos
- Regressão Logística no R

Modelos Lineares Generalizados

A classe de modelos lineares generalizados (MLG), proposta por Nelder e Wedderburn (1972), estende a classe dos modelos lineares normais.

A ideia básica por traz dos MLG consiste em permitir que a distribuição da variável resposta pertença à família exponencial de distribuições e flexibilizar a relação funcional entre a média da variável resposta $E(Y) = \mu$ e as covariáveis por meio de um preditor linear η e uma função de ligação g.

Modelos Lineares Generalizados

Os MLG envolvem três componentes:

1. **Componente aleatória:** representada por um conjunto de variáveis aleatórias (v.a.'s) independentes, $Y_1, \ldots, Y_n$, provenientes de uma mesma distribuição pertencente à família exponencial com

$$E(Y_i) = \mu_i; \quad i = 1, \ldots, n.$$

Modelos Lineares Generalizados

2. **Componente sistemática:** cada Y_i está associada a um conjunto de p variáveis explicativas, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$, que definem um preditor linear dado por

$$\eta_i = \boldsymbol{x}_i \boldsymbol{\beta}, \;\; i = 1, \ldots, n,$$

onde $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ é um vetor de parâmetros.

Modelos Lineares Generalizados

3. **Função de ligação:** uma função g monótona e diferenciável que relaciona a componente aleatória à componente sistemática, isto é, relaciona a média ao preditor linear,

$$g(\mu_i) = \eta_i.$$

A função de ligação que transforma a média μ_i no parâmetro natural θ_i é a *função de ligação canônica*. Para esta função de ligação, tem-se $g(\mu_i) = \eta_i = \theta_i$, isto é, o preditor linear modela diretamente o parâmetro canônico.

Modelo de Regressão Logística

Um caso particular de MLG é o modelo de regressão logística, que relaciona um conjunto de variáveis regressoras ao parâmetro binomial de variáveis binárias por meio da função de ligação logito.

Regressão Logística Binária: usada quando a resposta é binária (ou seja, tem dois resultados possíveis).

Regressão Logística Nominal: usada quando há três ou mais categorias sem ordenação natural dos níveis.

Regressão Logística Ordinal: usada quando existem três ou mais categorias com ordenação natural dos níveis, mas a ordenação dos níveis não significa necessariamente que os intervalos entre eles sejam iguais.

Variáveis binárias

Em muitas aplicações, a variável resposta de interesse representa um número fixo n de observações que podem assumir uma de duas possíveis categorias.

Por exemplo, a resposta pode ser "vivo" ou "morto", ou "presente" e "ausente".

Usualmente, usamos os termos sucesso e fracasso para cada uma duas categorias.

As duas categorias da variável podem ser representadas por uma variável indicadora binária assumindo os valores 0 e 1.

Variáveis binárias

A distribuição de Bernoulli é adequada para experimentos com 2 resultados possíveis, chamados de sucesso e fracasso, onde a probabilidade de sucesso é $\pi \in (0, 1)$ e a probabilidade de fracasso é $1 - \pi$.

A variável Y que atribui 1 ao sucesso e zero ao fracasso é chamada de variável de Bernoulli com parâmetro π: $Y \sim Bernoulli(\pi)$.

A função de probabilidade de $Y \sim Bernoulli(\pi)$ é dada por

$$f_Y(y|\pi) = \pi^y(1-\pi)^{1-y}, \; y = 0, 1; \quad 0 \leq \pi \leq 1. \tag{1}$$

A média e variância de uma variável Y com distribuição Bernoulli são:

$$E(Y) = \pi \quad \text{e} \quad Var(Y) = \pi(1-\pi). \tag{2}$$

Introdução

Variáveis binárias

1. Em uma análise sobre se as empresas comerciais têm ou não um departamento de relações industriais, de acordo com o tamanho da empresa (x), a variável resposta (Y) foi definida para ter os dois resultados possíveis: a empresa tem um departamento de relações industriais, a empresa não tem um departamento de relações industriais. Esses resultados podem ser codificados como 1 e 0, respectivamente.

2. Em um estudo de participação de mulheres na força de trabalho remunerada, como uma função de idade (x_1), número de filhos (x_2) e renda do marido (x_3), a variável resposta (Y) foi definida para ter os possíveis resultados: mulher casada na força de trabalho remunerada, mulher casada fora da força de trabalho remunerada. Esses resultados podem ser codificados como 1 e 0, respectivamente.

Variáveis binárias

Considere o modelo de regressão linear simples,

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i; \quad E(\epsilon_i) = 0; \quad Y_i = 0, 1. \tag{3}$$

Se a resposta é binária, $Y_i \sim Bernoulli(\pi_i)$, a resposta esperada tem um significado especial

$$E(Y_i) = \beta_0 + \beta_1 x_i = \pi_i. \tag{4}$$

A resposta média $E(Y_i) = \beta_0 + \beta_1 x_i$ é simplesmente a probabilidade de $Y_i = 1$ quando o valor da variável preditora é x_i.

Variáveis binárias

Ao considerar o modelo de regressão linear usual para respostas binárias, surgem alguns problemas:

1. O termo de erro do modelo não é normalmente distribuído. Segue que,

$$Y_i = 1: \quad \epsilon_i = 1 - \beta_0 - \beta_1 x_i; \tag{5}$$

$$Y_i = 0: \quad \epsilon_i = -\beta_0 - \beta_1 x_i. \tag{6}$$

2. A variância do termo de erro do modelo não é constante. Segue que,

$$Var\left(\epsilon_i\right) = \left(\beta_0 + \beta_1 x_i\right)\left(1 - \beta_0 - \beta_1 x_i\right). \tag{7}$$

Variáveis binárias

3. Restrições na função de resposta. Como a função de resposta representa probabilidades quando a variável resposta é uma variável binária 0, 1, as respostas médias devem ser restritas da seguinte forma:

$$0 \leq E(Y_i) = \pi_i \leq 1. \tag{8}$$

Usando a estrutura de MLG, vamos construir um *modelo de regressão logística* onde as respsostas Y_i são variáveis de Bernoulli independentes com média

$$E(Y_i) = \pi_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}. \tag{9}$$

Modelo Regressão Logística

Em muitos estudos, cada variável resposta Y_i pode estar associada a um vetor de covariáveis $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$, que são informações que influenciam a probabilidade binomial π_i.

O interesse estatístico é verificar a relação entre π_i e as covariáveis $\boldsymbol{x}_i$.

Para investigar esta relação é conveniente estabelecer um modelo formal. Como a distribuição Bernoulli pertence à família exponencial, esse problema pode ser visto como um caso particular de MLG.

Observação

Na prática, a construção do modelo necessita que algumas suposições sejam assumidas, por exemplo a independência entre as observações, linearidade da componente sistemática, entre outras. Essas suposições não podem ser garantidas, mas podem ser checadas.

Modelo Regressão Logística

Sob a estrutura de MLG, vamos supor que a dependência de π_i em $\boldsymbol{x}_i$ é dada pela combinação linear

$$\eta_i = \boldsymbol{x}_i\boldsymbol{\beta} = \sum_{j=1}^{p} x_{ij}\beta_j; \quad i = 1, \ldots, n. \tag{10}$$

A menos que restrições sejam impostas aos β_j's, temos que $-\infty < \eta_i < +\infty$.

Então, para expressar os π_i's como uma combinação linear dos $\boldsymbol{x}_i$'s, devemos usar uma função de ligação g que leve os valores do intervalo unitário $(0, 1)$ a valores no reais.

Para dados binários, é comum escolher uma transformação g que seja simétrica em torno de zero e que corresponda a uma função de distribuição acumulada (f.d.a.).

Modelo Regressão Logística

Quando g corresponde à função logística, temos

$$g(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \sum_{j=1}^{p} x_{ij}\beta_j; \quad i = 1, \ldots, n. \tag{11}$$

Assim, temos a formulação do modelo de regressão logística, onde

$$\pi_i = \frac{e^{\eta_i}}{1+e^{\eta_i}} = \frac{e^{\sum_{j=1}^{p} x_{ij}\beta_j}}{1+e^{\sum_{j=1}^{p} x_{ij}\beta_j}}; \quad i = 1, \ldots, n \tag{12}$$

ou

$$\mathit{logito}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i = \sum_{j=1}^{p} x_{ij}\beta_j; \quad i = 1, \ldots, n. \tag{13}$$

Modelo Regressão Logística

- π_i é a probabilidade de que uma observação esteja numa categoria especificada da variável binária Y_i, usualmente chamada de *probabilidade de sucesso*.
- Observe que o modelo descreve a *probabilidade de um evento* ocorrer em função de um conjunto de covariáveis $\boldsymbol{x}_i$.
- Com o modelo logístico, estimativas de π_i sempre estarão entre 0 e 1:
  - o numerador é positivo porque é a potência de um valor positivo (e);
  - o denominador é (1 + numerador), então o resultado sempre será menor do que 1.

Modelo Regressão Logística

Interpretação de β: efeito na probabilidade e na razão de chances

Suponha que $p = 3$ e que $x_{i1} = 1$ para todo i, então, o modelo pode ser escrito em termos do logaritmo da chance (odds) de resposta positiva,

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}. \tag{14}$$

Equivalentemente, o modelo pode ser escrito em termos da chance de resposta positiva,

$$\frac{\pi_i}{1-\pi_i} = \exp\{\beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}\}. \tag{15}$$

Modelo Regressão Logística

Interpretação de β: efeito na probabilidade e na razão de chances

Por fim, a probabilidade de resposta positiva é

$$\pi_i = g^{-1}(\eta_i) = \frac{\exp\{\beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}\}}{1 + \exp\{\beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}\}}. \qquad (16)$$

Modelo Regressão Logística

Interpretação de β: efeito na probabilidade e na razão de chances

Suponha que x_{i2} e x_{i3} não são funcionalmente relacionadas.

Para x_{i2} fixa, o modelo pode ser interpretado como segue: o efeito de uma unidade de mudança em x_{i3} é o aumento da chance por uma quantidade β_3; o efeito de uma unidade de mudança em x_{i3} é o aumento da chance de uma resposta positiva multiplicativamente pelo fator e^{β_3}.

Modelo Regressão Logística

Interpretação de β: efeito na probabilidade e na razão de chances

As interpretações na escala da probabilidade são mais complicadas pois o efeito em π_i de uma unidade de mudança em x_{i3} depende dos valores de x_{i2} e x_{i3}.

A derivada de π_i em relação a x_{i3} é

$$\frac{\partial \pi_i}{\partial x_{i3}} = \beta_3 \pi_i (1 - \pi_i).$$

Então, uma pequena mudança em x_{i3} tem um efeito maior se π_i é próximo de 0.5 do que se π_i é próximo de 0 ou 1 (como medida na escala de probabilidade).

Modelo Regressão Logística

Interpretação de β: efeito na probabilidade e na razão de chances

Para o caso geral em que $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$, temos que o parâmetro β_j refere-se ao efeito da j-ésima covariável no logaritmo da chance de resposta positiva (sendo que as outras covariáveis são mantidas fixas). Então e^{β_j} é o efeito multiplicativo do aumento de uma unidade na j-ésima covariável na chance de resposta positiva.

Modelo Regressão Logística

Respostas binomiais

Em alguns estudos, é possível termos observações repetidas em um mesmo nível da variável preditora.

Exemplo 1.

Num experimento de precificação, um novo produto foi apresentado a 1.000 consumidoras e, em seguida, foi perguntado a cada consumidora se ela compraria o produto em um determinado preço.

Cinco preços foram estudados e 200 consumidoras foram selecionadas aleatoriamente para cada nível de preço.

A variável resposta aqui é binária: compraria ou não compraria o produto.

A variável preditora é o preço e tem cinco níveis.

▪

Modelo Regressão Logística

Respostas binomiais

Assuma que m_i respostas binárias foram observadas no nível x_i, $i = 1, \ldots, n$.

Então, o i-ésimo valor da resposta binária em x_i pode ser denotado por Z_{ij}, tal que $Z_{ij} \sim Bernoulli(\pi_i)$, $i = 1, \ldots, n$ e $j = 1, \ldots, m_i$.

O número de 1's no nível x_i é denotado por Y_i:

$$Y_i = \sum_{j=1}^{m_i} Z_{ij}. \tag{17}$$

A porporção de 1's no nível x_i é π_i:

$$\pi_i = \frac{Y_i}{m_i}. \tag{18}$$

Modelo Regressão Logística

Respostas binomiais

A variável resposta Y_i tem distribuição binomial dada por

$$f_{Y_i}(Y_i|\pi_i) = \binom{m_i}{y_i} \pi_i{}^{y_i} (1-\pi_i)^{m_i - y_i}, \tag{19}$$

onde $\pi_i \in (0,1)$, $y_i = \{0, \ldots, m_i\}$.

A esperança e variância da variável resposta $Y_i \sim Binomial(m_i, \pi_i)$ são dadas por

$$E(Y_i) = m_i\pi_i \quad \text{e} \quad Var(Y_{.i}) = m_i\pi_i(1-\pi_i). \tag{20}$$

Modelo Regressão Logística

Respostas binomiais

Para o caso geral de n variáveis independentes $Y_1, \ldots, Y_n$ correspondentes ao número de sucessos em n diferentes subgrupos, as frequências de sucessos e fracassos são mostradas na Tabela 1.

Tabela 1: Frequências para n distribuições binomiais.

	Subgrupo		
	1	...	n
Sucesso	Y_1	...	Y_1
Fracasso	$m_1 - Y_1$	...	$m_n - Y_n$
Total	m_1	...	m_n

Note que para os tamanhos amostrais binomiais $\boldsymbol{m} = (m_1, \ldots, m_n)$, temos que o número total de observações binárias é $N = \sum_{i=1}^{n} m_i$.

Modelo Regressão Logística

Respostas binomiais

Quando $Y_i \sim Binomial(m_i, \pi_i)$, temos que,

$$E(Y_i) = \mu_i = m_i \pi_i, \ \ i = 1, \ldots, n. \tag{21}$$

Como m_i é considerado conhecido, modelar a média da variável resposta μ_i é equivalente a modelar a probabilidade binomial π_i.

Se $Y_i \sim Binomial(m_i, \pi_i)$, onde Y_i é o número de sucessos em m_i ensaios de Bernoulli, temos que a média da v.a. Y_i, $\mu_i = E(Y_i) = m_i \pi_i$, depende de m_i.

Então, vamos assumir que $y_1, \ldots, y_n$ são proporções binomiais tais que $m_i y_i \sim Binomial(m_i, \pi_i)$. Isto é, y_i é a proporção amostral de sucessos em m_i ensaios de Bernoulli e $E(Y_i) = \pi_i$ é independente de m_i.

Estimação

Em MLG, o método de estimação mais comumente utilizado é o de máxima verossimilhança (MV).

Para $\boldsymbol{y} = (y_1, \ldots, y_n)$ um vetor de n observações de $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$, em que cada $Y_i \sim Binomial(m_i, \pi_i)$ e os Y_i's são independentes, a função de log-verossimilhança de $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)$ é dada por

$$\ell(\boldsymbol{\pi}|\boldsymbol{y}) = \sum_{i=1}^{n} \left[y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + m_i \log(1-\pi_i) + \log\binom{m_i}{y_i} \right]. \quad (22)$$

Estimação

Temos que

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i = \sum_{j=1}^{p} x_{ij}\beta_j.$$

Então, a função de log-verossimilhança do modelo de regressão logística é dada por

$$\ell(\boldsymbol{\beta}|\boldsymbol{y}) \propto \sum_{i=1}^{n}\sum_{j=1}^{p} y_i x_{ij}\beta_j - \sum_{i=1}^{n} m_i \log\left[1+\exp\left(\sum_{j=1}^{p}\beta_j x_{ij}\right)\right]. \quad (23)$$

Inferência

Dstribuição assintótica de $\hat{\beta}$:

A distribuição assintótica de $\hat{\beta}$ é a base da construção de testes e intervalos de confiança, em amostras grandes, para os parâmetros dos MLG. Sob condições gerais de regularidade, para amostras grandes, tem-se

$$\hat{\beta} \overset{a}{\sim} N_p(\beta, \boldsymbol{I}^{-1}). \tag{24}$$

Testes de hipóteses (Teste de Wald)

Considere o teste de hipótese

$$H_0 : \beta_j = 0 \quad \text{contra} \quad H_1 : \beta_j \neq 0.$$

A estatística de Wald é definida por

$$Z = \frac{\hat{\beta}_j}{\sqrt{\widehat{Var}(\hat{\beta}_j)}} \overset{a}{\sim} N(0,1), \tag{25}$$

onde $\widehat{Var}(\hat{\beta}_j) = \widehat{Var}_{j,j}(\hat{\boldsymbol{\beta}}) = [(\boldsymbol{I}(\hat{\boldsymbol{\beta}}))^{-1}]_{j,j}$.

Assim, rejeita-se H_0 a um nível de $(1-\alpha)100\%$ de confiança se $|Z| > z_{1-\alpha/2}$.

Intervalos de confiança

De maneira geral, a estatística de Wald é a mais utilizada para construir intervalos de confiança $(1-\alpha)100\%$ assintóticos para cada um dos β_j's parâmetros.

Um intervalo de confiança $(1-\alpha)100\%$ para β_j é dado por

$$\hat{\beta}_j \pm z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{\beta}_j)}, \tag{26}$$

onde $Var(\hat{\beta}_j) = \widehat{Var}(\hat{\boldsymbol{\beta}})_{j,j} = [(\boldsymbol{I}(\hat{\boldsymbol{\beta}}))^{-1}]_{j,j}$.

Teste da Razão de Verossimilhanças (TRV)

Suponha que deseja-se comparar dois modelos aninhados M_0 (modelo simplificado correspondente a H_0) e M_1 (modelo mais completo correspondente a H_1).

Seja $\hat{\pi}_0$ os valores ajustados sob o modelo M_0 e $\hat{\pi}_1$ os valores ajustados sob o modelo M_1.

Teste da Razão de Verossimilhanças (TRV)

Então, a estatística do TRV para comparar M_0 e M_1 é dada por

$$\begin{aligned}\Lambda &= 2[\ell(\hat{\boldsymbol{\pi}}_1|\boldsymbol{y}) - \ell(\hat{\boldsymbol{\pi}}_0|\boldsymbol{y})] \\ &= 2\ell(\hat{\boldsymbol{\pi}}_1|\boldsymbol{y}) - 2\ell(\hat{\boldsymbol{\pi}}_0|\boldsymbol{y}) \\ &= D_{M_0}(\hat{\boldsymbol{\pi}}) - D_{M_1}(\hat{\boldsymbol{\pi}}). \qquad (27)\end{aligned}$$

Teste da Razão de Verossimilhanças (TRV)

A verossimilhança para um espaço menor M_0 não pode ser maior do que a verossimilhança sob um espaço maior M_1 isto é,

$$\ell(\hat{\boldsymbol{\pi}}_0|\boldsymbol{y}) \leq \ell(\hat{\boldsymbol{\pi}}_1|\boldsymbol{y}).$$

Então,

$$D_{M_1}(\hat{\boldsymbol{\pi}}) \leq D_{M_0}(\hat{\boldsymbol{\pi}}).$$

Assim, a estatística do TRV (27) é maior quando o modelo M_0 se ajusta mal aos dados quando comparado a M_1.

Análise de Resíduos

- **Resíduos Pearson:**

$$r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i(m_i - \hat{\mu}_i)/m_i}}, \tag{28}$$

onde $\hat{\mu}_i = m_i\hat{\pi}_i$ é o valor ajustado e
$\hat{\mu}_i(m_i - \hat{\mu}_i)/m_i = v(\hat{\mu}_i) = m_i\hat{\pi}_i(1 - \hat{\pi}_i) = \widehat{Var}(Y_i)$.

Análise de Resíduos

- **Resíduos Deviance:**

$$r_i^D = \text{sinal}(y_i - \hat{\mu}_i)\sqrt{d_i}, \tag{29}$$

onde d_i é a contribuição da i-ésima observação para o deviance $D(\hat{\pi})$.

Análise de Resíduos

- **Resíduos Estudentizados:**

$$r_i^S = \frac{r_i^P}{\sqrt{1-\hat{h}_{ii}}} = \frac{y_i - \hat{\mu}_i}{\sqrt{(1-\hat{h}_{ii})\hat{\mu}_i(m_i - \hat{\mu}_i)/m_i}}, \qquad (30)$$

onde $\hat{h}_{ii}$ é a alavancagem, dada pelo i-ésimo elemento da diagonal da matriz chapéu ponderada estimada $\boldsymbol{H}_{\widehat{\boldsymbol{W}}}$.

Análise de Resíduos

Como critério de avaliação desses três resíduos, costuma-se usar que observações com resíduos grande (em módulo) são possíveis observações atípicas.

Para um modelo bem ajustado, espera-se que os valores dos resíduos sejam aleatórios em torno de 0.

Análise de Resíduos

Para dados binários não agrupados e quando as variáveis explicativas são contínuas, cada $m_i = 1$. Então, y_i pode ser igual a apenas 0 ou 1, e um resíduo pode assumir apenas dois valores e geralmente não é informativo.

Nesses casos, gráficos de resíduos têm uso limitado, consistindo meramente de duas linhas paralelas de pontos.

Quando os dados podem ser agrupados em conjuntos de observações com valores preditores comuns, é melhor calcular resíduos para os dados agrupados.

Uma alternativa é usar os resíduos *binned* (agrupados), onde os dados são divididos em *bins* (grupos) com base em seus valores ajustados e calcula-se o resíduo médio e o valor médio ajustado para cada grupo.

Diagnóstico de Influência

As alavancagens de cada observação $\hat{h}_{ii}$ podem ser utilizadas no diagnóstico de observações influentes.

Outra medida de influência que também pode ser usada em regressão logística é a distância de Cook, que compara $\hat{\boldsymbol{\beta}}$ com $\hat{\boldsymbol{\beta}}_{(-i)}$, a estimativa de $\boldsymbol{\beta}$ com a i-ésima observação removida do conjunto de dados. Uma aproximação para a distância de Cook é dada por

$$D_i = (r_i^S)^2 \frac{\hat{h}_{ii}}{p(1-\hat{h}_{ii})}. \tag{31}$$

Como critério de avaliação, usa-se que observações com valores grandes de $\hat{h}_{ii}$ ou D_i são influentes.

Classificação

Quando a resposta é binária, ao final da estimação do modelo de regressão logística, obtemos as probabilidades estimadas de obter um sucesso ou um fracasso.

Porém, em muitas situações, nosso interesse é *classificar* uma observação como sucesso ou fracasso, com base no modelo (usando as covariáveis).

Classificação

Por exemplo, com base em característica de uma paciente, queremos classificá-la como diabética ou não diabética.

Isso pode ser feito, com base no modelo ajustado, definindo-se um valor de corte c, tal que

- se $\hat{\pi}_i > c$, então a i-ésima observação é classificada como sucesso;
- se $\hat{\pi}_i \leq c$, então a i-ésima observação é classificada como fracasso;

Algumas possibilidades: tomar $c = 0.5$; tomar c como a proporção amostral de 1's.

Classificação

Alguns questões que surgem:

- Como podemos verificar quão boa é essa regra de classificação?
- Quais indicadores podemos usar para fazer essa verificação?

Usualmente, nos baseamos no grau de acerto das classificações.

Classificação

- **Matriz de confusão:** tabela de classificação cruzada entre a classificação de acordo com o modelo estimado e a classificação observada na amostra.

Com base na matriz de confusão, podemos calcular algumas métricas:

- acurácia, precisão (sensibilidade e especificidade), recall, entre outras.

Classificação

- Acurácia: proporção de casos que são classificados corretamente.
- Sensibilidade: proporção de casos positivos (sucessos) que foram corretamente classificados como positivos.
- Especificidade: proporção de casos negativos (fracassos) que foram corretamente classificados como negativos.

Curva ROC

Uma forma de avaliar a performance de classificação de um modelo de regressão logística estimado é pela curva ROC (*Receiver Operating Characteristic*).

Uma curva ROC é um gráfico da sensibilidade em função de (1 - especificidade) para um possível valor de corte c.

A curva ROC pode ser mais informativa do que a matriz de confusão porque ela resume o poder preditivo do modelo para todos os valores de corte.

Curva ROC

Quando o valor de corte se aproxima de 0, quase todas as predições são 1: então a sensibilidade é próxima de 1 e a especificidade é próxima de 0 e o ponto para (1 - especificidade, sensibilidade) tem coordenada próxima de (1,1).

Quando o valor de corte se aproxima de 1, quase todas as predições são 0: então, a sensisbilidade é próxima de 0, a especificidade é próxima de 1 e o ponto para (1 - especificidade, sensibilidade) tem coordenada próxima de (0,0).

A curva ROC tem forma côncava conectando os pontos (0,0) e (1,1).

Curva ROC

Para uma dada especificidade, um melhor poder preditivo corresponde a uma maior sensibilidade. Porntanto, quanto melhor o poder preditivo, mais "alta" a curva.

Uma curva ROC de adivinhação aleatória será um linha diagonal.

A curva ROC de uma regra de classificação perfeita é um ponto no canto superior esquerdo do gráfico, onde a proporção verdadeiros positivos é 1 e a proporção de falsos positivos é zero.

Exemplos

Exemplo 2.

Uma gerente estudou o efeito da experiência em programação (em meses) na habilidade de completar uma tarefa de programação complexa dentro de um tempo determinado. 25 programadoras foram incluídas no estudo e todas receberam a mesma tarefa e a mesma quantidade de tempo para executá-la.

O conjunto de dados está disponível na planilha *programacao* do arquivo *dados_cdrl.xlsx*.

▪

Exemplos

Exemplo 3.

Considere o conjunto de dados para o estado do Maine, nos EUA, com informações sobre renda per capita média e a localização de cada condado do estado.

O conjunto de dados está disponível na planilha *maine* do arquivo *dados_cdrl.xlsx*.

▪

Exemplos

Exemplo 4.

A Tabela 2 mostra o número de besouros mortos após cinco horas de exposição a várias doses de dissulfeto de carbono gasoso.

O conjunto de dados está disponível na planilha *besouro* do arquivo *dados_cdrl.xlsx*.

▪

Exemplos

Tabela 2: Dados de mortalidade de besouros.

Dose (x_i em $log_{10} CS2 mgl^{-1}$)	Número de besouros (m_i)	Número de mortos (y_i)
1.6907	59	6
1.7242	60	13
1.7552	62	18
1.7842	56	28
1.8113	63	52
1.8369	59	53
1.8610	62	61
1.8839	60	60

Exemplos

Exemplo 5.

Uma pesquisadora está interessa em saber como as notas GRE (Graduate Record Exam scores), GPA (grade point average) e o prestígio da universidade onde candidatas/os cursaram a graduação afeta a admissão em um programa de pós-graduação.

O conjunto de dados está disponível na planilha *admissao* do arquivo *dados_cdrl.xlsx*.

▪

Exemplos

Exemplo 6.

Considere o conjunto de dados de $n = 27$ pacientes com leucemia disponível na planilha *leucemia* do arquivo *dados_cdrl.xlsx*.

A variável resposta é binária e indica se ocorreu remissão da leucemia (REMISS), que é dada por um 1. As variáveis preditoras são celularidade da seção do coágulo da medula (CELL), porcentagem diferencial de esfregaço de blastos (SMEAR), porcentagem de infiltrado absoluto de células de leucemia da medula (INFIL), índice de marcação percentual das células de leucemia da medula óssea (LI), número absoluto de blastos no sangue periférico (BLAST) e a temperatura mais alta antes do início do tratamento (TEMP).

▪

Exemplos

Exemplo 7.

Considere o conjunto de dados de disponível na planilha *diabetes* do arquivo *dados_cdrl.xlsx*.

Este conjunto de dados apresenta resultados de testes de diabetes coletados pelo Instituto Nacional de Diabetes e Doenças Digestivas e Renais dos EUA de uma amostra de mulheres com pelo menos 21 anos de idade, de herança indígena Pima e que viviam perto de Phoenix, Arizona. As variáveis observados foram: pregnant: número de vezes que engravidou; glucose: concentração plasmática de glicose (teste de tolerância à glicose); pressure: pressão arterial diastólica (mm Hg); triceps: espessura da prega cutânea do tríceps (mm); insulin: insulina sérica de 2 horas (mu U/ml); mass: índice de massa corporal (peso em kg/(altura em m)^2); pedigree: função de pedigree do diabetes; age: idade (anos); diabetes: fator que indica o resultado do teste de diabetes (neg/pos).

Exemplos

Exemplo 8.

O conjunto de dados disponível na planilha *inseto* do arquivo *dados_cdrl.xlsx* apresenta o resultado de um experimento realizado para testar testar o efeito de uma substância tóxica em insetos. Nesse experimento, em cada uma de seis doses da substância, 250 insetos foram expostos à substância e o número de insetos que morrem foi contado.

▪

Exemplos

Exemplo 9.

O conjunto de dados disponível na planilha *doenca* do arquivo *dados_cdrl.xlsx* apresenta dados de um estudo surto de uma doença propagada por mosquitos. No estudo, pessoas foram selecionada aleatoriamente em dois setores de uma cidade para determinar se tinham ou não a doença. A variável resposta foi codificada com 1, se a pessoa tinha a doença e 0 se não. Três variáveis preditoras foram incluídas no estudo: age: idade em anos; status socioeconômico com três níveis e representada por duas variáveis indicadoras (middle e lower); sector: variável categórica com dois níveis (setor 1 e setor 2).

▪

Exemplos

Exemplo 10.

O conjunto de dados disponível na planilha *cupom* do arquivo *dados_cdrl.xlsx* apresenta dados de uma pesquisa para avaliar o efeito de oferecer cupons de redução de preços de um determinado produto onde 1000 domicílios foram selecionados ao acaso para receber cupons de desconto variando na oferta de redução de preço.

▪

Exemplos

Exemplo 11.

O conjunto de dados disponível na planilha *pressao* do arquivo *dados_cdrl.xlsx* apresenta dados de uma pesquisa com indivíduos do sexo masculino em Framingham, Massachusetts. Nessa pesquisa, homens com idades entre 40 e 59 anos para os quais foram observada a pressão arterial e se desenvolveram doença cardíaca durante um período de acompanhamento de 6 anos.

▪

Regressão Logística no R

```
library(readxl)
library(ggthemes)
library(ROCR)
library(caret)
library(tidymodels)
library(tidyverse)
```

Regressão Logística no R

```
dados <- read_xlsx("../dados/dados_cdrl.xlsx",
                   sheet = "diabetes")

dados <- dados |>
  mutate(diabetes = factor(diabetes),
         diab_bin = ifelse(diabetes == "neg", 0, 1))
```

Regressão Logística no R

```
fit <- glm(diab_bin ~ glucose,
           family = binomial(link = "logit"),
           data = dados)

tidy(fit)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  -6.10     0.630       -9.68 3.71e-22
## 2 glucose       0.0424   0.00476      8.91 5.07e-19
```

Regressão Logística no R

```
summary(fit)
```

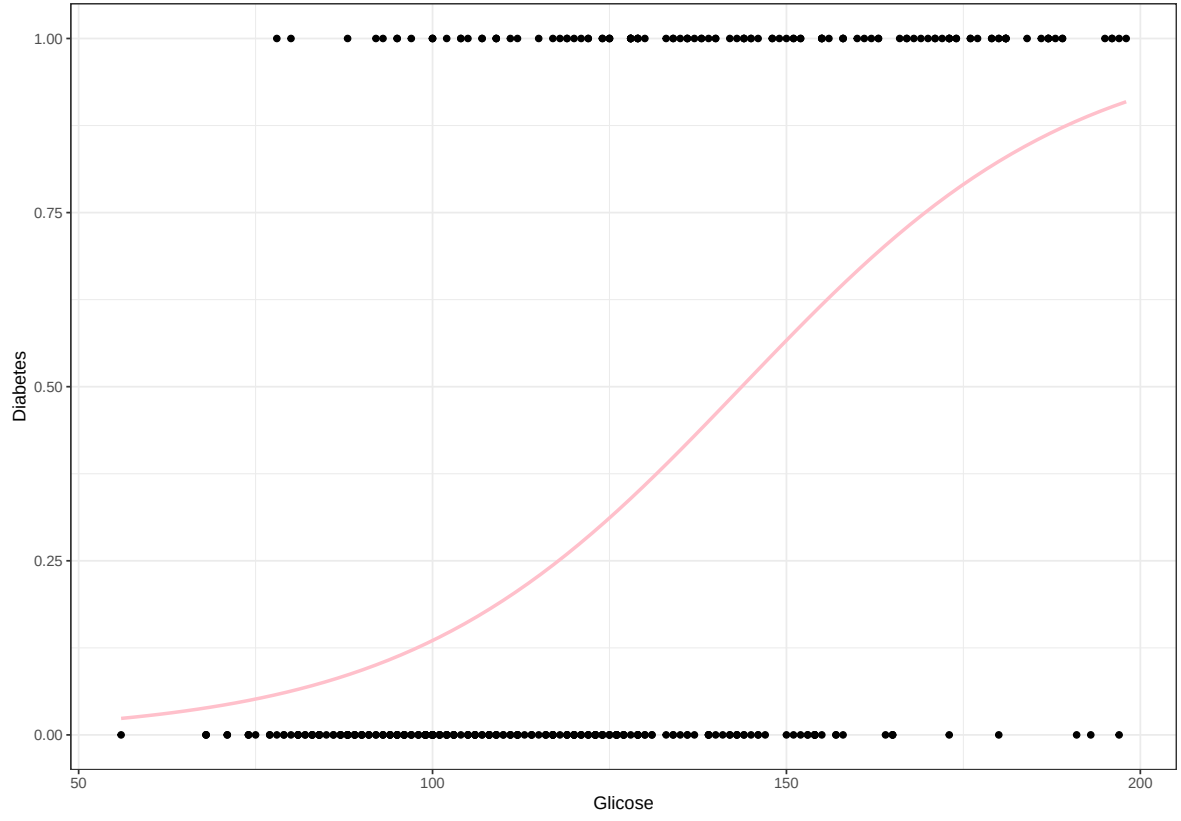
```
##
## Call:
## glm(formula = diab_bin ~ glucose, family = binomial(link = "logit"),
##     data = dados)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.095521   0.629787  -9.679   <2e-16 ***
## glucose      0.042421   0.004761   8.911   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 498.10  on 391  degrees of freedom
## Residual deviance: 386.67  on 390  degrees of freedom
## AIC: 390.67
##
## Number of Fisher Scoring iterations: 4
```

Regressão Logística no R

```
ggplot(dados, aes(x = glucose, y = diab_bin)) +
  geom_point() +
  stat_smooth(method = "glm", color = "pink",
              se = FALSE,
              method.args = list(family = binomial)) +
  labs(x = "Glicose", y = "Diabetes") +
  theme_bw()
```

Regressão Logística no R

```
## `geom_smooth()` using formula = 'y ~ x'
```



Regressão Logística no R

```r
# predições
probs <- predict(fit, type = "response")
pred_classe <- ifelse(probs < 0.5, 0, 1)

# acurácia
mean(pred_classe == dados$diab_bin)
```

```
## [1] 0.7678571
```

Regressão Logística no R

```
cats <- factor(ifelse(probs < 0.5, "neg", "pos"))

conf_mat <- confusionMatrix(
  data = relevel(cats, ref = "pos"),
  reference = relevel(dados$diabetes, ref = "pos"))
```