REGRESSÃO LOGÍSTICA

PROF. LETÍCIA RAPOSO
profleticiaraposo@gmail.com

TÓPICOS DA AULA

- Visão geral da classificação;
- Modelo de regressão logística;
- Interpretação dos parâmetros;
- Estimação dos parâmetros;
- Testes de significância;
- Seleção de variáveis;
- Avaliação da regressão;
- Desempenho dos modelos.

INTRODUÇÃO

- Em muitas situações, a variável resposta é *qualitativa (categórica)*.
- Abordagens para predizer respostas qualitativas: *classificação*.
- Muitos métodos usados para classificação prevêem primeiro a *probabilidade* de cada uma das categorias de uma variável qualitativa como base para fazer a classificação. Nesse sentido, eles também se comportam como métodos de regressão.
- Existem muitas técnicas de classificação possíveis que podem ser usadas para prever uma resposta qualitativa. Hoje falaremos de uma delas: a *regressão logística*.

VISÃO GERAL DA CLASSIFICAÇÃO

Exemplos de problemas de classificação:

1. Uma pessoa chega ao pronto-socorro com um conjunto de sintomas que podem ser atribuídos a uma das três condições médicas. Qual das três condições o indivíduo possui?

2. Um serviço bancário *on-line* deve ser capaz de determinar se uma transação que está sendo realizada no site é fraudulenta, com base no endereço IP do usuário, no histórico de transações anteriores e assim por diante.

3. Com base nos dados da sequência de DNA para um número de pacientes com e sem uma determinada doença, um biólogo gostaria de descobrir quais mutações de DNA são deletérias (causadoras de doenças) e quais não são.

VISÃO GERAL DA CLASSIFICAÇÃO

- Assim como na configuração de regressão, na classificação temos um conjunto de observações de treinamento $(x_1, y_1), \ldots (x_n, y_1)$, que podemos usar para construir um classificador.
- Desejamos que o classificador tenha um bom desempenho não apenas nos dados de treinamento, mas também em observações de teste que não foram usadas para treinar o classificador.

REGRESSÃO LOGÍSTICA

- Recomendada em situações em que a variável *resposta* é de natureza *dicotômica ou binária*.
- Quanto às *preditoras*, podem ser *categóricas ou não*.
- Assume que as *observações são independentes*; em outras palavras, que uma observação não afeta outra.
- Permite *estimar a probabilidade* associada à ocorrência de determinado evento em face de um conjunto de variáveis preditoras.

POR QUE NÃO A REGRESSÃO LINEAR?

- Suponha que estamos tentando prever a condição médica de um paciente na sala de emergência com base em seus sintomas. Existem três diagnósticos possíveis: _acidente vascular cerebral, overdose de drogas e convulsões epilépticas_.
- Poderíamos considerar a codificação desses valores como uma variável de resposta quantitativa, Y, da seguinte maneira:

$$Y = \begin{cases} 1, se\ AVC \\ 2, se\ overdose\ de\ drogas \\ 3, se\ convulsões\ epilépticas \end{cases}$$

Esta codificação implica uma _ordenação_ dos resultados, colocando uma overdose de drogas entre o derrame e a crise epiléptica, e que a diferença entre acidente vascular cerebral e overdose de drogas é a mesma que a diferença entre overdose e crise epléptica.

POR QUE NÃO A REGRESSÃO LINEAR?

- Se os valores da variável resposta assumissem um ordenamento natural, como leve, moderado e grave, e se considerássemos que a diferença entre leve e moderado é semelhante ao intervalo entre moderada e grave, então uma codificação de 1, 2, 3 seria razoável.
- Infelizmente, em geral, não há uma maneira natural de converter uma variável resposta qualitativa com *mais de dois níveis* em uma resposta quantitativa pronta para regressão linear.

POR QUE NÃO A REGRESSÃO LINEAR?

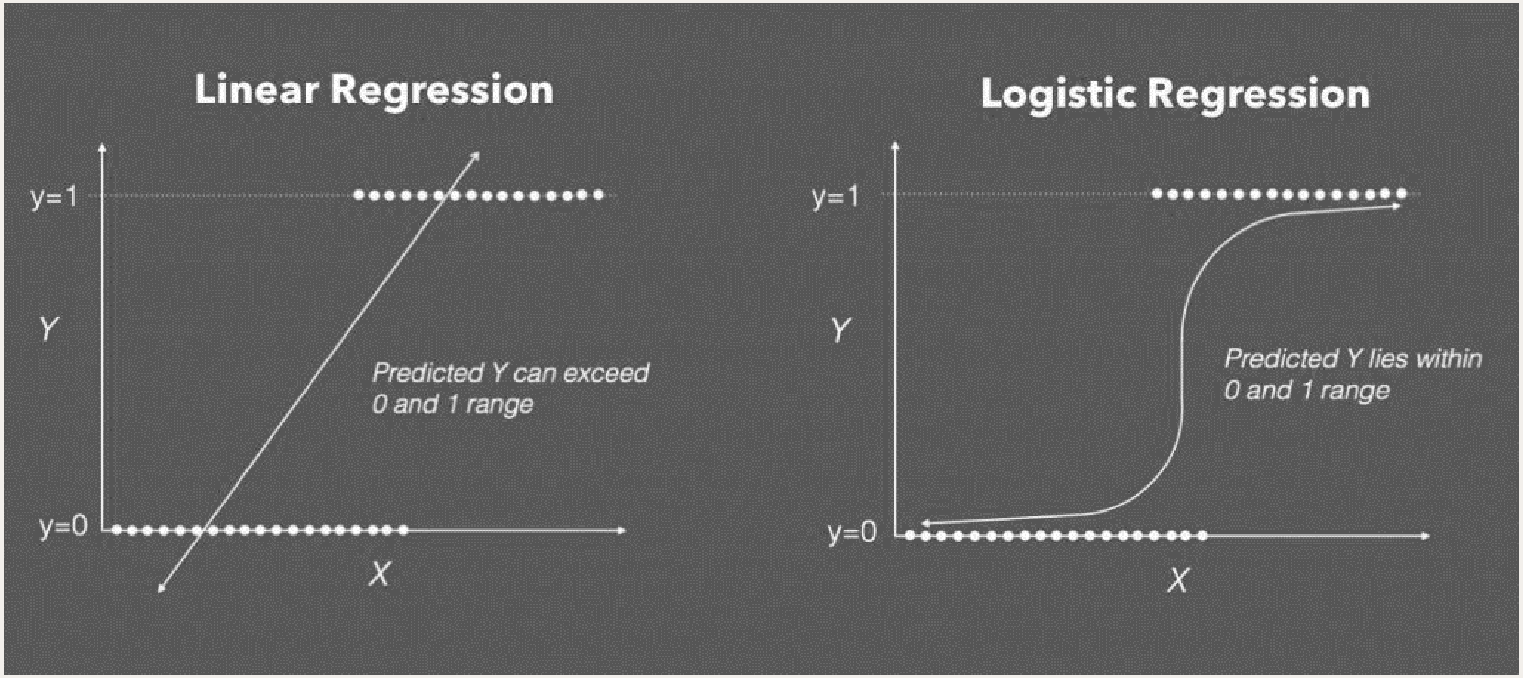
- Para uma resposta qualitativa binária (dois níveis), a situação é melhor. Por exemplo, apenas duas possibilidades para a condição médica do paciente: derrame e overdose de drogas.

$$Y = \begin{cases} 0, se\ AVC \\ 1, se\ overdose\ de\ drogas \end{cases}$$

Variável *dummy* para codificar a resposta

- Possibilidade: regressão linear para essa resposta binária e predizer a overdose de drogas se $\hat{y} > 0{,}5$ e o acidente vascular cerebral, caso contrário.
- No entanto, se usarmos regressão linear, algumas *estimativas podem estar fora do intervalo [0, 1]*, tornando-as difíceis de interpretar como probabilidades!

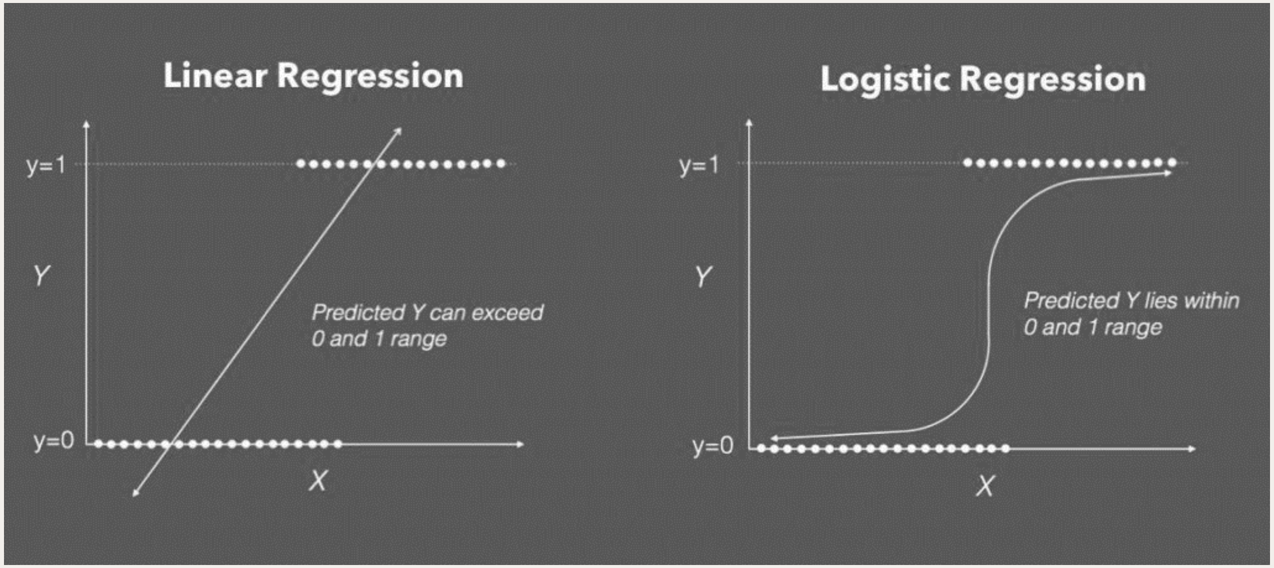
POR QUE NÃO A REGRESSÃO LINEAR?



https://www.machinelearningplus.com/machine-learning/logistic-regression-tutorial-examples-r/

Interpretando y como probabilidade, vemos que com a regressão linear essa probabilidade pode ser menor que 0 ou maior que 1.

POR QUE NÃO A REGRESSÃO LINEAR?



Queremos ser capazes de ter uma linha em forma de "s"para prever as probabilidades e descrever essa linha curva com os coeficientes da regressão linear.

REGRESSÃO LOGÍSTICA

- Suponha que o modelo tenha a forma

$$y_i = \mathbf{x}_i'\beta + \varepsilon_i$$

em que $\mathbf{x}_i' = [1, x_{i1}, x_{i2}, \cdots, x_{ik}]$, $\beta' = [\beta_0, \beta_1, \beta_2, \cdots, \beta_k]$ e a variável resposta tem valores entre 0 e 1.

- Assumiremos que a variável resposta é uma variável aleatória com distribuição de Bernoulli com função de probabilidade

y_i	Probabilidade
1	$P(y_i = 1) = \pi_i$
0	$P(y_i = 0) = 1 - \pi_i$

REGRESSÃO LOGÍSTICA

Uma vez que $E(\varepsilon_i) = 0$, o valor esperado da variável resposta é

$$E(y_i) = 1(\pi_i) + 0(1 - \pi_i) = \pi_i \qquad (1)$$

o que implica em

$$E(y_i) = x_i'\beta = \pi_i$$

Isso significa que a resposta esperada dada pela função resposta $E(y_i) = x_i'\beta$ é apenas a probabilidade de que a variável resposta assuma o valor 1.

REGRESSÃO LOGÍSTICA

- Há uma restrição na função de resposta, porque

$$0 \leq E(y_i) = \pi_i \leq 1$$

- Essa restrição pode causar problemas na escolha de uma função resposta linear, como assumimos inicialmente na Eq. (1). Isso porque desejamos que os *valores da função resposta fiquem entre 0 e 1*.
- Geralmente, quando a variável resposta é binária, há considerável evidência empírica indicando que a forma da *função resposta deva ser não-linear*.

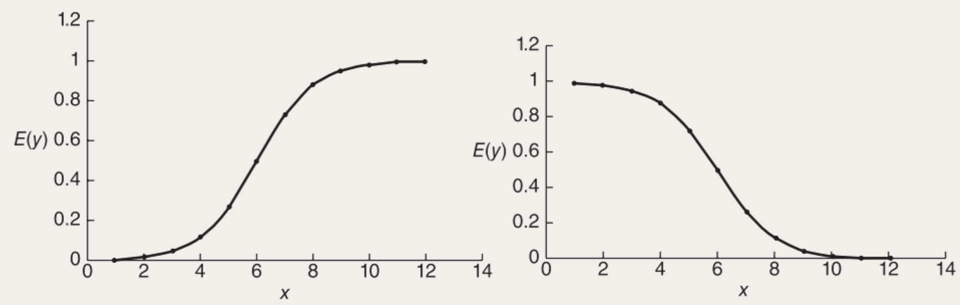
REGRESSÃO LOGÍSTICA

Usualmente é empregada uma função monotônica em formato de S. Esta função é chamada de *função logística* e tem a forma abaixo:

$$E(y) = \underbrace{\frac{e^{\mathbf{x}'\beta}}{1 + e^{\mathbf{x}'\beta}}} = \frac{1}{1 + e^{-\mathbf{x}'\beta}}$$

$$\frac{\oplus}{1+\oplus}$$

Valor entre 0 e 1



(a) (b)

Mas como transformar a linha em forma de "s" das probabilidades previstas em uma linha reta que pode ser descrita com os coeficientes?

REGRESSÃO LOGÍSTICA

A função logística pode ser facilmente *linearizada*.

Considere

$$\eta = \mathbf{x}'\beta$$

ser o preditor linear, onde η é definido pela transformação

$$\eta = \ln \frac{\pi}{1-\pi}$$

Esta transformação é frequentemente chamada de *transformação logit* da probabilidade π, e a razão $\frac{\pi}{1-\pi}$ é chamada de chance (*odds*).

Função de ligação que associa os valores esperados da resposta aos preditores lineares no modelo.

REGRESSÃO LOGÍSTICA

Sendo a resposta binária, os termos de erro ε_i só podem levar dois valores

$$\varepsilon_i = 1 - \mathbf{x}_i'\beta, \qquad y_i = 1$$

$$\varepsilon_i = -\mathbf{x}_i'\beta, \qquad y_i = 0$$

Consequentemente os erros não podem ser normais, e a variância dos erros não é constante.

$$\sigma_{yi}^2 = E\{y_i - E(y_i)\}^2 = (1-\pi_i)^2\pi_i + (0-\pi_i)^2(1-\pi_i) = \pi_i(1-\pi_i)$$

$$\sigma_{yi}^2 = E(y_i)[1 - E(y_i)]$$

ESTIMAÇÃO DOS PARÂMETROS

A forma geral de um modelo de regressão logística é

$$y_i = E(y_i) + \varepsilon_i$$

em que as observações são variáveis aleatórias independentes de Bernoulli com valores esperados

$$E(y_i) = \pi_i = \frac{\exp(\mathbf{x}_i'\beta)}{1 + \exp(\mathbf{x}_i'\beta)}$$

ESTIMAÇÃO DOS PARÂMETROS

- Utilizamos o método de máxima verossimilhança para estimar os parâmetros no preditor linear $\mathbf{x}_i'\beta$.
- Cada observação de amostra segue a distribuição de Bernoulli, então a distribuição de probabilidade de cada observação da amostra é

$$f_i(y_i) = \pi_i^{y_i}(1-\pi_i)^{\,1-y_i}, \qquad i = 1,2,\dots,n$$

e cada observação assume o valor 0 ou 1.

- Como as observações são independentes, a função de verossimilhança é

$$L(y_1, y_2, \dots, y_n, \beta) = \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{\,1-y_i}$$

Os valores são calculados computacionalmente.

INTERPRETAÇÃO DOS PARÂMETROS

- Considere o caso em que o preditor linear tem *apenas uma variável preditora*, de forma que o valor ajustado do preditor linear em um valor particular de x, x_i, é

$$\hat{\eta}(x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- O valor ajustado em $x_i + 1$ é

$$\hat{\eta}(x_i + 1) = \hat{\beta}_0 + \hat{\beta}_1 (x_i + 1)$$

e a diferença nos dois valores previstos é

$$\hat{\eta}(x_i + 1) - \hat{\eta}(x_i) = \hat{\beta}_1$$

$$\hat{\eta}(x_i + 1) - \hat{\eta}(x_i) = \ln\left(odds_{(x_i+1)}\right) - \ln\left(odds_{x_i}\right) = \ln\left(\frac{odds_{(x_i+1)}}{odds_{x_i}}\right) = \hat{\beta}_1$$

$$\hat{O}_R = \frac{odds_{(x_i+1)}}{odds_{x_i}} = e^{\hat{\beta}_1}$$

INTERPRETAÇÃO DOS PARÂMETROS

$$\boxed{\hat{O}_R = \frac{odds_{(x_i+1)}}{odds_{x_i}} = e^{\hat{\beta}_1}}$$

- Razão de chances de ocorrência do evento é igual a $\exp(\hat{\beta}_1)$ para variação de 1 unidade de x_i.
- Em geral, o aumento estimado na razão de chances associada a uma mudança de d unidades na variável de previsão é $\exp(d\hat{\beta}_1)$.

INTERPRETAÇÃO DOS PARÂMETROS

Quando a <u>*variável preditora é binária*</u>, podemos analisar da seguinte forma:

O preditor linear é $\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x$

Quando $x_1 = 0$ $\ln\left(\frac{\pi_0}{1-\pi_0}\right) = \beta_0$

Quando $x_1 = 1$ $\ln\left(\frac{\pi_1}{1-\pi_1}\right) = \beta_0 + \beta_1$

INTERPRETAÇÃO DOS PARÂMETROS

Dividindo o logit quando $x_i = 1$ pelo logit quando $x_i = 0$

$$\frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_0}{1-\pi_0}} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

$$\left.\begin{array}{l} e^{\beta_1} < 1 \rightarrow \text{exposição é fator protetor} \\ e^{\beta_1} > 1 \rightarrow \text{exposição é fator risco} \\ e^{\beta_1} = 1 \rightarrow \text{efeitos semelhantes (não há associação)} \end{array}\right\} \text{Para ocorrência do evento}$$

INFERÊNCIA ESTATÍSTICA NOS PARÂMETROS DO MODELO

- 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 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*.

O modelo que inclui a variável em questão nos diz mais sobre a variável resposta do que um modelo que não inclui essa variável?

TESTE DE RAZÃO DE VEROSSIMILHANÇA

- Deseja-se comparar os valores observados da variável resposta com os valores preditos obtidos dos modelos com e sem determinadas variáveis em questão.
- A comparação dos observados com os valores preditos é baseado no log da verossimilhança.
- É útil pensar em um valor observado da variável resposta também como sendo um valor predito resultante de um modelo saturado.

Deviance: para regressão logística, desempenha o mesmo papel que a soma dos quadrados residuais da regressão linear.

Necessário para obter uma quantidade cuja distribuição é conhecida e, portanto, pode ser usada para fins de teste de hipóteses.

$$D = -2\ln\left(\frac{\textit{verossimilhança do modelo ajustado}}{\textit{verossimilhança do modelo saturado}}\right)$$

Modelo saturado: aquele que contém tantos parâmetros quanto observações (o que se ajustaria perfeitamente).

TESTE DE RAZÃO DE VEROSSIMILHANÇA

- A *deviance* pode ser usada para comparar 2 modelos que não sejam saturados: modelo maior (com a variável avaliada) e modelo menor (sem a variável avaliada).

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

O modelo saturado é o mesmo para as 2 parcelas

$$G = -2 \ln\left(\frac{verossimilhança\ sem\ a\ variável}{verossimilhança\ com\ a\ variável}\right)$$

- Sob a hipótese nula, as variáveis omitidas não são significativas e a estatística G tem distribuição χ^2 com k graus de liberdade.

TESTE DE WALD

- O teste de Wald é obtido por comparação entre a estimativa de máxima verossimilhança do parâmetro $\hat{\beta}_j$ e a estimativa de seu erro padrão.
- Vamos considerar a seguinte hipótese

$$H_0: \beta_1 = 0$$
$$H_A: \beta_1 \neq 0$$

- A estatística do teste Wald para a regressão logística é

$$W_j = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$$

e tem distribuição normal padrão.

- Se não rejeitarmos H_0, temos que a variável x_j não explica a variável resposta.

RAZÃO DE VEROSSIMILHANÇA X WALD

- Hauck e Donner (1977) examinaram o desempenho do teste de Wald e descobriram que ele se comportou de maneira aberrante, muitas vezes falhando em rejeitar a hipótese nula quando o coeficiente era significativo. Assim, eles recomendaram que o teste da razão de verossimilhança seja o preferido.
- Na prática, a situação mais preocupante é quando os valores estão próximos e um teste tem $p < 0,05$ e o outro tem $p > 0,05$. Quando isso ocorre, usamos o valor p do teste da razão de verossimilhança.

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 observações com os mesmos valores para todos os preditores):
  - χ^2 de Pearson;
  - *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.

DEVIANCE

- Quando o modelo de regressão logística é adequado aos dados e o tamanho da amostra é grande, a *deviance* possui uma distribuição χ^2 com $n - k$ graus de liberdade, onde k é o número de parâmetros no modelo.
- *Pequenos valores de deviance (ou elevado valor p) implicam que o modelo fornece um ajuste satisfatório aos dados*, enquanto grandes valores de *deviance* implicam que o modelo atual não é adequado.
- Uma boa regra é dividir a *deviance* pelo graus de liberdade. Se a relação $\frac{D}{n-k} \gg 1$, o modelo atual não é adequado aos dados.

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

χ^2 DE PEARSON

- Compara as probabilidades de sucesso e fracasso observadas e esperadas em cada grupo de observações.
- O número esperado de sucessos $n_i\hat{\pi}_i$ e o número esperado de fracassos é $n_i(1-\hat{\pi}_i), i = 1,2,\dots,n$.
- A estatística χ^2 de Pearson é

$$\chi^2_{n-k} = \sum_{i=1}^{n}\left\{\frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i} + \frac{[(n_i - y_i) - n_i(1-\hat{\pi}_i)]^2}{n_i(1-n_i\hat{\pi}_i)}\right\} = \sum_{i=1}^{n}\frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1-\hat{\pi}_i)}$$

- Valores pequenos da estatística (ou um *grande valor p*) implicam que o *modelo fornece um ajuste satisfatório aos dados*.

HOSMER-LEMESHOW

- Quando não há réplicas nas variáveis preditoras, as observações podem ser agrupadas para realizar o teste de Hosmer - Lemeshow.
- Neste procedimento, as observações são classificadas em g grupos com base nas probabilidades estimadas de sucesso.
- Geralmente, são utilizados cerca de 10 grupos e os números observados de sucessos O_j e fracassos $N_j - O_j$ são comparados com as frequências esperadas em cada grupo, $N_j\bar{\pi}_j$ e $N_j(1-\bar{\pi}_j)$, onde N_j é o número de observações no j-ésimo grupo e a probabilidade média de sucesso estimada no j-ésimo grupo é $\bar{\pi}_j = \sum_{i \in grupo\ j} \hat{\pi}_i / N_j$.

HOMER-LEMESHOW

- A estatística de Homer-Lemeshow é

$$HL = \sum_{j=1}^{n} \frac{\left(O_j - N_j\bar{\pi}_j\right)^2}{N_j\bar{\pi}_j\left(1 - \bar{\pi}_j\right)}$$

- Se o modelo de regressão logística está correto, a estatística HL segue uma distribuição χ^2 com $g - 2$ graus de liberdade quando o tamanho da amostra é grande.
- A H0 sustenta que o modelo se ajusta aos dados. *Grandes valores da estatística HL (valor p <0,05) implicam que o modelo não é adequado aos dados*. Também é útil calcular a razão da estatística de HL para o número de graus de liberdade $g - k$ com valores próximos à unidade implicando em um ajuste adequado.

INTERVALO DE CONFIANÇA PARA OS PARÂMETROS E RAZÃO DE CHANCES

- A inferência de Wald é utilizada para construir intervalos de confiança na regressão logística.
- Considere primeiro encontrar intervalos de confiança em coeficientes de regressão individuais no preditor linear.
- Um intervalo de confiança de aproximadamente $100\,(1-\alpha)\%$ no j-ésimo coeficiente do modelo é

$$\hat{\beta}_i - Z_{\alpha/2}SE(\hat{\beta}_i) \leq \beta_i \leq \hat{\beta}_i + Z_{\alpha/2}SE(\hat{\beta}_i)$$

- Como $\hat{O}_R = \exp(\hat{\beta}_i)$

$$\exp[\hat{\beta}_i - Z_{\alpha/2}SE(\hat{\beta}_i)] \leq O_R \leq \exp[\hat{\beta}_i + Z_{\alpha/2}SE(\hat{\beta}_i)]$$

SELEÇÃO DE MODELOS

- A estatística de Homer-Lemeshow é

- Se o modelo de regressão logística está correto, a estatística HL segue uma distribuição χ^2 com $g - 2$ graus de liberdade quando o tamanho da amostra é grande.
- A H0 sustenta que o modelo se ajusta aos dados. Grandes valores da estatística HL (valor $p < 0{,}05$) implicam que o modelo não é adequado aos dados. Também é útil calcular a razão da estatística de HL para o número de graus de liberdade $g - k$ com valores próximos à unidade implicando em um ajuste adequado.

O caminho tradicional da estatística é procurar por um *modelo parcimonioso, estável e que descreva o fenômeno estudado*. Alguns passos devem ser seguidos para a construção do modelo:

1. Realizar análises univariadas cuidadosas;

2. Identificar variáveis com potencial impacto;

3. Estudar as inter-relações entre as diferentes variáveis.

4. Decidir qual, ou quais técnicas para seleção de variáveis serão empregadas.

FORWARD SELECTION

1. A variável que apresenta menor valor-p no teste da RV é escolhida;
2. Escolhe-se uma segunda variável que produza o maior aumento na RV quando adicionada ao modelo;
3. Aplica-se o teste da RV para verificar se a contribuição desta nova variável é significativa;
4. O processo continua até que nenhuma variável acrescida no modelo cause aumento significativo na RV.

BACKWARD SELECTION

1. Ajuste de todas as variáveis preditoras candidatas a ficar no modelo;
2. Compara-se a *deviance* do modelo com todas as variáveis com a *deviance* dos modelos que resultam da exclusão individual de cada variável. Se valor p do teste da RV for significativo, a variável fica no modelo e o procedimento se encerra; se não for, ela sai.
3. Escolhe-se a próxima variável que menos contribui e testa-se a sua significância. Se for significativa, ela fica no modelo e processo se encerra, caso contrário, ela sai e processo continua.

SELEÇÃO DE VARIÁVEIS

- A estatística de Homer-Lemeshow é

- Se o modelo de regressão logística está correto, a estatística HL segue uma distribuição χ^2 com $g - 2$ graus de liberdade quando o tamanho da amostra é grande.
- A H0 sustenta que o modelo se ajusta aos dados. Grandes valores da estatística HL (valor p <0,05) implicam que o modelo não é adequado aos dados. Também é útil calcular a razão da estatística de HL para o número de graus de liberdade $g - k$ com valores próximos à unidade implicando em um ajuste adequado.

Stepwise: combinação das duas outras seleções.

- Começa com *forward*, mas após entrada da 2ª variável, o teste da RV é realizado para verificar se a 1ª permanece no modelo.
- Caso permaneça, uma 3ª variável é selecionada (*forward*).
- Se uma 3ª variável entra no modelo, testa-se para verificar se as duas primeiras continuam. (Pode acontecer que uma delas ou as duas sejam eliminadas).
- Tenta-se então a inclusão de uma nova variável. Caso entre, tenta-se a eliminação das que já estão no modelo.
- O procedimento acaba quando não se consegue nem adicionar, nem eliminar variáveis;

SELEÇÃO DE VARIÁVEIS

- A estatística de Homer-Lemeshow é

- Se o modelo de regressão logística está correto, a estatística HL segue uma distribuição χ^2 com $g - 2$ graus de liberdade quando o tamanho da amostra é grande.
- A H0 sustenta que o modelo se ajusta aos dados. Grandes valores da estatística HL (valor p <0,05) implicam que o modelo não é adequado aos dados. Também é útil calcular a razão da estatística de HL para o número de graus de liberdade $g - k$ com valores próximos à unidade implicando em um ajuste adequado.

- Medidas, como AIC e BIC, vistas na regressão linear múltipla, podem ser usadas para selecionar o melhor conjunto de variáveis preditoras para o modelo.
- O AIC é um indicador importante do ajuste do modelo.
  - Quanto menor o AIC, melhor.
  - AIC penaliza o aumento do número de coeficientes no modelo.
  - Observar o valor do AIC de um único modelo não ajuda muito. É mais útil na comparação de modelos. O modelo com o menor AIC será relativamente melhor.

VERIFICAÇÃO DIAGNÓSTICA NA REGRESSÃO LOGÍSTICA

- Etapa importante para verificar a dependência do modelo estatístico em relação às várias observações que foram coletadas.
- Para isso são usadas as técnicas de diagnóstico, que verificam se as suposições do modelo estão satisfeitas e identificam características dos dados, como observações influentes, que causem alguma mudança nas estimativas dos coeficientes, levando a problemas nas conclusões geradas pelo modelo.
- Resíduos podem ser usados para verificação diagnóstica e investigação da adequação do modelo de regressão logística.
- Os resíduos comuns são definidos como

$$e_i = y_i - \hat{y}_i = y_i - n_i\hat{\pi}_i, \qquad i = 1,2,\dots,n$$

DEVIANCE

Na regressão logística, a quantidade análoga à soma dos quadrados dos resíduos da regressão linear é a *deviance*.

Isso leva a uma *deviance* residual, definida como

$$d_i = \pm\left\{2\left[y_i \ln\left(\frac{y_i}{n_i\hat{\pi}_i}\right) + (n_i - y_i)\ln\left(\frac{n_i - y_i}{n_i(1-\hat{\pi}_i)}\right)\right]\right\}^{1/2}, i = 1,2,\ldots,n$$

Mesmo sinal do resíduo correspondente

$$y_i = 0, d_i = -\sqrt{-2n\ln(1-\hat{\pi}_i)}$$

$$y_i = 1, d_i = \sqrt{-2n\ln(\hat{\pi}_i)}$$

RESÍDUO DE PEARSON

Da mesma forma, podemos definir um resíduo de Pearson como

$$r_i = \frac{(y_i - n_i\hat{\pi}_i)}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}}, \qquad i = 1,2,\dots,n$$

VERIFICAÇÃO DIAGNÓSTICA NA REGRESSÃO LOGÍSTICA

- A *deviance* e os resíduos de Pearson são os mais apropriados para a realização de verificações de adequação do modelo.
- As plotagens desses *resíduos versus a probabilidade estimada* e um *gráfico de probabilidade normal dos resíduos de desvio* são úteis para verificar o ajuste do modelo em pontos de dados individuais e verificar possíveis desvios.
- A distância de Cook, vista nos modelos de regressão linear, também pode ser utilizada na determinação de *pontos influentes*.
- Assim como no modelo linear, uma métrica para diagnosticar outliers é a *leverage*.

DESEMPENHO DO MODELO

A *matriz de confusão* é muito utilizada para avaliar os modelos de classificação.

	Verdadeiro	
Predito	Positivo	Negativo
Positivo	VP	FP
Negativo	FN	VN

- Acurácia: $\frac{VP + VN}{VP + FP + VN + FN}$
- Sensibilidade (revocação - *recall*): $\frac{VP}{VP + FN}$
- Especificidade: $\frac{VN}{VN + FP}$
- Precisão, *precision:* $\frac{VP}{VP+FP}$

DESEMPENHO DO MODELO

- Valor preditivo positivo

$$\frac{S \times P}{(S \times P) + (1 - E) \times (1 - P)}$$

- Valor preditivo negativo

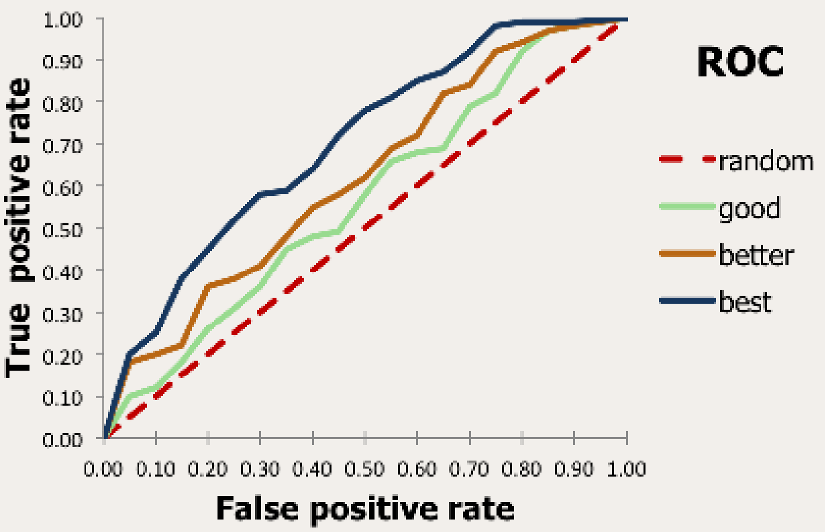
$$\frac{E \times (1 - P)}{(1 - S) \times P + E \times (1 - P)}$$

- Medida F: média harmônica de precisão e revocação. Fica entre 0 e 1. Quanto maior o valor, melhor o modelo. É formulado como

$$2\,\frac{revocação\ \times\ precisão}{precisão\ +\ revocação}$$

DESEMPENHO DO MODELO

- A curva ROC (*Receiver Operator Characteristic*) determina o desempenho de um modelo de classificação em um valor limite definido pelo usuário usando a área sob a curva ROC (*Area Under Curve*, AUC).
- Quanto maior a área, melhor o modelo.



REFERÊNCIAS

- JAMES, Gareth et al. **An introduction to statistical learning**. New York: springer, 2013.
- MONTGOMERY, Douglas C.; PECK, Elizabeth A.; VINING, G. Geoffrey. **Introduction to linear regression analysis**. John Wiley & Sons, 2012.
- DANCEY, Christine P.; REIDY, John G.; ROWE, Richard. **Estatística Sem Matemática para as Ciências da Saúde**. Penso Editora, 2017.
- McDonald, J.H. 2014. **Handbook of Biological Statistics** (3rd ed.). Sparky House Publishing, Baltimore, Maryland.