

Introdução à Regressão Logística

(Vilma Mayumi Tachibana)

Alguns Problemas

- Capacidade de prever o que poderá acontecer e tomar as decisões escolhendo as poss. alternativas.
- Uso da reg. logística para estimar a prob. de não pagamento de dívidas.
- Avaliações de risco do crédito. Instituição Financeira (banco de dados amplo e consistente) para previsão de solvência. Ex: Minussi et al (2002) com 323 clientes (empresas de setor industrial) e 49 ind. financeiros para prever prob. de um pg. ocorrer.
- Um componente de máquina resistirá à alta temperatura?
- Modelagem da fidelidade dos clientes

com a academia em que praticam esportes - Costa, G.G.O - SEGET

• Satisfação no trabalho de servidores de uma instituição educacional (Medeiros, et.al, 2007) - SEGET.

• Estudo sobre a participação do esposo no mercado de trabalho como função da idade da esposa, número de filhos e rend. do marido.

• Utilização da regressão logística para a classificação de famílias quanto à condição de pobreza nas RMs do Rio de Janeiro e Recife nos anos de 70, 80 e 91.

(Kondo, S. 2001, ENEP, ABEP)

1º Votar no
pluriprincipio

• Geomarketing : o uso de regressão logística múltipla para o mapeamento de regiões geográficas de alto potencial mercadológico . Rodeski, 2010 , TCC , UFRGS . Jonhyra Fachel.

• Um entomologista deseja estudar

a resistência de uma espécie de besouro submetida a diferentes dosagens de uma substância nociva.

- Um médico realizar uma pesquisa buscando avaliar a eficiácia de um medicamento (Figueira, 2006, Silvia Lopez).

O que tais exemplos têm em comum?

Relação entre variável resposta e variáveis explicativas.

$$Y = f(X_1, X_2, \dots, X_n)$$

Exemplo mais comum: regressão

Análise de Regressão - agrupa informações de uma ou de várias variáveis que visam explicar um fenômeno de interesse (encontrar relacionamento entre variáveis)

domingo segunda terça quarta quinta sexta sábado _____ / _____ / _____

variáveis) apresentando como resultado uma função de expressão matemática, mentre como as variáveis podem explicar as alterações no evento de estudo.

Nos nossos exemplos - variável resposta é dicotómica (binária) ou poli-tómica.

Régressão Linear Simples

O modelo de regressão linear descreve a variável Y como uma soma de uma quantidade determinística e uma quantidade aleatória.

$$\text{obs} = \text{previsível} + \text{aleatório}$$

$$Y = f(X) + \dots = f(X_1, X_2, \dots, X_n) + \varepsilon$$

Dados n pares de dados via X_i, Y_i , ($i=1, 2, \dots, n$)

domingo segunda terça quarta quinta sexta sábado _____ / _____ / _____

se admitirmos que Y é uma função linear de X , então:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon$$

β_0 é o valor de Y , quando X for nulo

β_1 é o coeficiente angular, o quanto varia a média de Y para o aumento de uma unidade da var. X .

• Para estimarão dos parâmetros são exigidas várias suposições:

- os erros tem média 0, mesma variabilidade em todos os níveis da variável auxiliar X (homocedásticos) e não correlacionados.

• para testar hipóteses e construir I.C. dos parâmetros, $\text{ENN}(0, \sigma^2)$

Então, no modelo simples em que a relação entre X e Y é expressão por uma linha reta:

$$E(Y|x) = \beta_0 + \beta_1 x$$

e

$$\text{Var}(Y|x) = \text{Var}(\beta_0 + \beta_1 x + \varepsilon) = \sigma^2$$

Regressão Logística Simples

Relacionar Y com X

Y é uma v. qualitativa binária bivariada -
tómica - tem duas prob.

Variável Indicadora 0 ou 1, -> out

Esta diferença com a regressão linear
vai implicar na escolha do modelo paramétrico e

Applied Logistic Regression - David W.
Hosmer e Stanley Lemeshow (2000)

Exemplo do Livro:

Idade e presença ou ausência de problema

domingo segunda terça quarta quinta sexta sábado _____ / _____ / _____

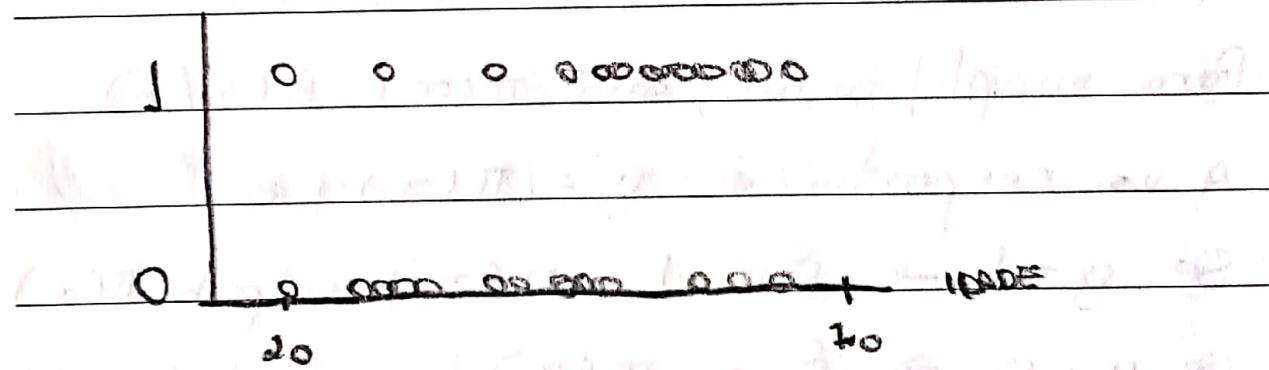
cardíaco (PCor)

ID. G-ET. IDADE. PCOR. T. Z.

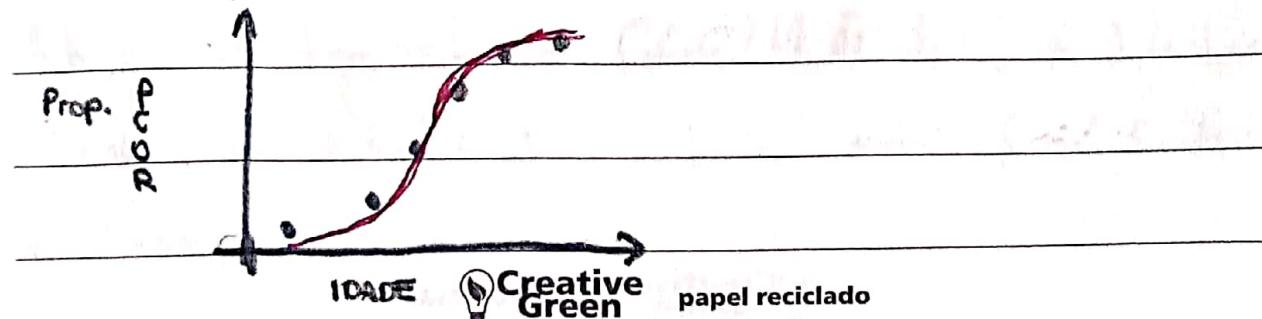
↓

grupo etário

Gráficos simples



Agrupar os dados e verificar a proporção de casos no respectivo intervalo (grupo).



Reg. Logística

$$0 \leq E(Y/x) \leq 1$$

A mudança em $E(Y/x)$ para cada unidade de x torna-se progressivamente menor, quando a média fica muito próxima de 0 ou de 1.

Para simplificação, seja $\pi(x) = E(Y/x)$

A v.a. resposta é $y = \pi(x) + \epsilon$

Se $y=1 \rightarrow \epsilon = 1 - \pi(x)$, com prob. $\pi(x)$

Se $y=0 \rightarrow \epsilon = -\pi(x)$, com prob $1-\pi(x)$.

Na reg linear, (diferenças)

$$-\infty < \mu < \infty$$

$$\epsilon \sim N(0,1)$$

domingo segunda terça quarta quinta sexta sábado

Na res. logística

$$\varepsilon = \begin{cases} 1 - \pi(x) & \text{if } \pi(x) < \pi \\ 0 & \text{otherwise} \end{cases}$$

$$p(\varepsilon) = \begin{cases} \pi(x) & \text{if } \varepsilon = 1 \\ 1 - \pi(x) & \text{if } \varepsilon = 0 \end{cases}$$

↳

$$E(\varepsilon) = 0$$

$$V(\varepsilon) = \pi(x)(1 - \pi(x))$$

for $\varepsilon \sim \text{Bernoulli}(\pi(x))$

A forma de relação - curva S

Função de distribuição acumulada de uma v.a.

Algumas transformações: $[0, 1] \rightarrow [-\infty, +\infty[$,
para $p(x) \in [0, 1]$, faz com que $g(p(x))$
 $\in [-\infty, +\infty[$

mais utilizadas

• Probabilidade

De Valor Extremo

Modelo logístico com a função de ligação logit

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \quad \text{ou} \quad \pi(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

$$1 - \pi(x) = \frac{1}{1 + e^{\beta_0 + \beta_1 x}}$$

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 - e^{-\beta_0 - \beta_1 x}}$$

$$1 - \pi(x)$$

domingo segunda terça quarta quinta sexta sábado domingo

$$\ln(e^{\beta_0 + \beta_1 x}) = \beta_0 + \beta_1 x$$

$$g(x) = \ln\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \beta_0 + \beta_1 x$$

$g(x)$ tem muitas propriedades de um modelo de regressão linear. É linear em seus parâmetros, pode ser contínua e variar de $-\infty$ a $+\infty$ dependendo da amplitude de x .

Estimação dos coeficientes desconhecidos β_0 e β_1

Procedimento: método de máxima verossimilhança (que produz valores para os parâmetros desconhecidos que

domingo segunda terça quarta quinta sexta sábado

maximizam a probabilidade de obter o conjunto de dados observados.

$$\xi(x_i) = \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i}$$

Como os y_i são independentes

$$l(\beta) = \prod_{i=1}^n \xi(x_i) = \prod_{i=1}^n \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i}$$

$$L(\beta) = \ln(l(\beta)) = \sum \{ y_i \ln[\pi(x_i)] + (1-y_i) \ln[1-\pi(x_i)] \}$$

Resultado interesse: $\sum_{i=1}^n y_i = \sum_{i=1}^n \hat{\pi}(x_i)$

$\hat{\pi}(x_i)$ é o ~~mpn~~ V de $\pi(x_i)$

Para o exemplo

$$\hat{\beta}_0 = -5,31 \quad e \quad \hat{\beta}_1 = 0,111$$

$$\hat{\pi}(x) = e^{-5,31 + 0,111 \cdot \text{Idade}}$$

$$1 + e^{-5,31 + 0,111 \cdot \text{Idade}}$$

É uma estimativa do logito dado pela equação

$$\hat{g}(x) = -5,31 + 0,111 \cdot \text{Idade}$$

Idade(x)	g(x)	$\pi(x)$
30	-2,01	0,118
45	-0,315	0,442
60	1,35	0,794
72	2,682	0,936

Verificações da qualidade do modelo ajustado

Ver se as variáveis são relevantes

- Comparação entre os valores preditos e saturados

$$D = -2 \ln$$

Verossi. modelo atual

Veras. modelo antigo

48

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

em que $\hat{\pi}_i = \hat{\pi}(x_i)$

A estatística D é denominada Deviance.

A mudança em D devido à inclusão da variável independente no modelo é dado por

$$\Delta D = D(\text{modelo sem variável}) - D(\text{modelo com variável})$$

- domingo segunda terça quarta quinta sexta sábado _____ / _____ / _____

Tal teste tem o mesmo papel que o númerador do teste F no R. L.

faz o caso de uma única v. indep.

$$G = -2 \left\{ \sum [y_i \ln(\hat{\pi}_i) + (1-y_i) \ln(1-\hat{\pi}_i)] \right\}$$

$$= [n_1 \ln(n_1) + n_0 \ln(n_0) - n \ln(n)]$$

Neste caso,

$$G = -2 [53,677 - [43 \ln(43) + 57 \ln(57) - 100 \ln(100)]]$$

$$= 29,31$$

$$P(X^2(1) > 29,31) < 0,001$$

Temos evidência que a idade é uma v.a significativa.



domingo segunda terça quarta quinta sexta sábado

Outros testes:

Teste de Wald

$$W = \frac{\hat{\beta}_1}{\text{EPI}(\hat{\beta}_1)}$$

$$W = 0,111 / 0,024 = 4,61$$

$$P(|z| > 4,61) = 2,01 \times 10^{-6}$$

Regressão Logística Múltipla

No modelo substituimos $\beta_0 + \beta_1 X$ por

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

Nesse caso:

$$\pi(x) = e^{g(x)}$$

$$1 + e^{g(x)}$$

domingo segunda terça quarta quinta sexta sábado / / / / /

Se alguma var. for discreta, nominal, tratar com var. de planejamento (dummy)

Ex:

Classe Social

Alta

D₁

D₂

0

0

Média

)

0

Baixa

0

1

$$g(x) = \beta_0 + \beta_1 x + \dots + \sum_{u=1}^{K_u-1} \beta_{ju} D_{ju} + \beta_p x^p$$

raio
↓
[D₁ D₂ D₃]

$$x_1 = \begin{cases} -1 & \text{ausente} \\ 1 & \text{presente} \end{cases}$$

D₁

D₂

D₃

branca

-1

-1

-1

→ efeitos

negra

)

0

0

hispano

0

)

0

Outros

0

0

1



Creative
Green

papel reciclado

Procedimento análogo ao anterior (caso univariado)

No entanto, agora temos $p+1$ equações de máxima verossimilhança.

Exemplo

Estudo sobre fatores de risco associados ao baixo peso dos recém nascidos (BP). O objetivo era identificar fatores associados ao nascimento de bebês com menos de 2500 gramos.

189 mulheres $n_1 = 59$ e $n_0 = 130$

Resultados

$$\hat{g}(x) = 1,295 + 0,024 P_{Dade} - 0,014 P_M + 1,004 P_{S1} + 0,433 P_{S2} + 0,049 NVM$$

domingo segunda terça quarta quinta sexta sábado _____ / _____ / _____

A) 19 182 2 0

B) 28 120 3 0

C) 28 95 1 2

D) 22 125 1 2

Raca 1. Branca (0 0)

2. Negra (1 0)

3. Outras (0 1)

4 -

Mulher g(x) $\pi(x)$ peso classif

A -0,705 0,33 2530 0

B -0,629 0,35 709 1

C -0,805 0,31 2466 1

D -1,032 0,26 3696 0

Verificar se as var. são importantes no modelo:



Creative
Green

papel reciclado

$$\left\{ \begin{array}{l} H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0 \\ H_1: \text{pelo menos um } \beta_i \neq 0 \end{array} \right.$$

Mesma estatística G, só que agora o modelo tem $p+1$ parâmetros.

Log da veros. com todas as var = -111,286

Log .. " sem nenhuma var. = -117,336

$$G = -2[-117,336 + 111,286]$$

$$G = -2 \left[\sum [y_i \ln(\hat{\pi}_i) + (1-y_i) \ln(1-\hat{\pi}_i)] - [n_1 \ln(n_1) + n_0 \ln(n_0) - n \ln(n)] \right]$$

$$G = -2 \{ [-111,286] - [59 \ln(59) + 130 \ln(130) - 189 \ln(189)] \} = 12,099$$

$$G \sim \chi^2_6$$



domingo segunda terça quarta quinta sexta sábado _____ / _____ / _____

$$P[\chi^2 > 12,099] = 0,034$$

A hipótese nula ~~não~~ deve ser rejeitada.

NVM - Nº de vezes que visito o médico

Para verificar a significância de cada variável, temos.

teste de Wald:

$$W_j = \frac{\hat{\beta}_j}{\hat{S.E}(\hat{\beta}_j)} \quad j=0,1,\dots,p$$

Outro método - retira-se as var. da interesse em estudo e observa-se o comportamento de

Teste de significância dos coeficientes

$$W = \hat{\beta}^T [\sum (\hat{\beta})^{-1}] \hat{\beta} = \hat{\beta}^T \hat{\beta}$$

Interpretação dos Modelos

a diferença no logito para uma pessoa com $\Sigma = 1$ e $\Sigma = 0$ é

$$g(1) - g(0) = [\beta_0 + \beta_1] - [\beta_0] = \beta_1$$

Para interpretar o resultado uma medida é razão de chances (odds ratios)

$$\psi = \frac{\pi(1)}{[1-\pi(1)]} = \frac{\pi(0)}{[1-\pi(0)]}$$

$$= \frac{\pi(1)[1-\pi(0)]}{\pi(0)[1-\pi(0)]} = OR$$

Creative
Green] papel reciclado

$$\psi = \frac{e^{\beta_0 + \beta_1}}{1 + e^{\beta_0 + \beta_1}} \quad \text{if } \beta_1 > 0$$
$$\psi = \frac{1}{1 + e^{\beta_0 + \beta_1}} \quad \text{if } \beta_1 < 0$$

γ = pres. câncer X : fumante

Se $e^{OR} = 2$, o câncer de pulmão ocorre

2 vezes mais com mais freqüências em que fuma do que nos não fumantes, na pop. de estudo.

X = exercícios físicos

Se γ - PCard e $OR = 0,5$, a ocorrência de PCor é metade entre aqueles que praticam exercícios físicos do que entre aqueles que não praticam.

SS(1)

SS(0)

Presente

21

22

Ausente

6

51

domingo segunda terça quarta quinta sexta sábado _____ / _____ / _____

$$OR = \frac{21}{6} \quad SI = 8,1$$

$$22 \quad " \quad "$$

Ou,

$$OR = e^{2,094} = 8,1$$

J

Gef. Estimado Desvio Padrão

Idade	2,094	0,529	3,96
-------	-------	-------	------

Constante	-0,841	0,255	-3,20
-----------	--------	-------	-------

domingo segunda terça quarta quinta sexta sábado

_____ / _____ / _____

Documento Dor. - Bloco de Notas

Copiar datos

Tratamentos

P, A, B

Acceso Trat Trat

Sexo

Ideas

tempo sentindo a dor

Dor passou depois do tratamento

Nesse modelo,

$$g(x) = 15,57 + 3,18 T_{r1} + 3,7 T_{r2} + 1,83 \text{ genero} \\ - 0,2621 \text{ idade} + 0,00586 \text{ tempo}$$

$$P(Y=1) = \pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}}$$

domingo segunda terça quarta quinta sexta sábado _____ / _____ / _____

$\pi(x) = 0$ S - não evento
 N - evento

Dados uti

Solución

Analyze

Interactive data analysis



Creative
Green

papel reciclado