Capítulo 4

Regressão Logística

Introdução

Assim como na Regressão Linear, o objetivo da Regressão Logística é encontrar um modelo *razoável, preciso* e *parcimonioso* que descreva a relação entre:

- uma varíavel de saída (*variável dependente / resposta*); e
- um conjunto de variáveis independentes (*variáveis preditoras / explicativas*)

O que distingue um modelo de regressão logística de um modelo de regressão linear é que no primeiro a variável resposta é *binária* (*dicotômica*).
Tal diferença entre regressão logística e linear se reflete tanto na escolha do modelo paramétrico como nas premissas. Porém, os princípios gerais são os mesmos em ambas as abordagens.

Regressão Logística Simples

Exemplo: Estudo de doenças arteriais coronarianas

- Para a exposição que segue, apresentamos como motivação um exemplo no qual o interesse é explorar a associação entre idade e a presença/ausência de doença arterial coronariana (*coronary heart disease*) em uma certa população.
- Dados: amostra de 100 indivíduos contendo a idade (AGE) e a presença (1) ou ausência (0) de doença arterial coronariana (CHD) para cada indivíduo.
- Análise preliminar sugere que a freqüência de CHD entre indivíduos mais velhos é maior do que entre indivíduos mais jovens (vide tabela e gráfico de dispersão).

ID	AGE	CHD
1	20	0
2	23	0
3	24	0
4	25	0
5	25	1
6	26	0
7	26	0
8	28	0
9	28	0
10	29	0
11	30	0
12	30	0
13	30	0
14	30	0
15	30	0
16	30	1
17	32	0
18	32	0
19	33	0
20	33	0

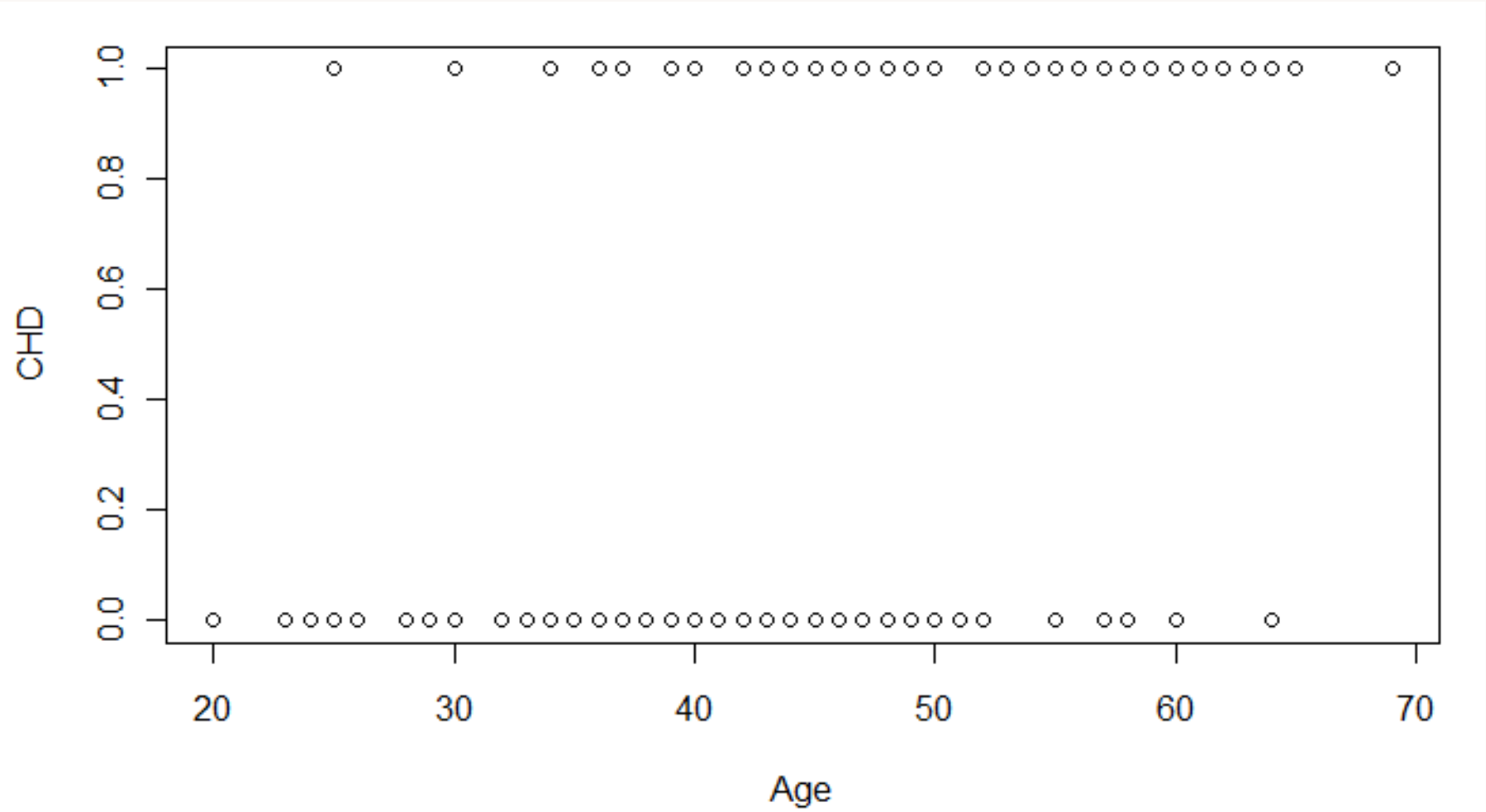
ID	AGE	CHD
21	34	0
22	34	0
23	34	1
24	34	0
25	34	0
26	35	0
27	35	0
28	36	0
29	36	1
30	36	0
31	37	0
32	37	1
33	37	0
34	38	0
35	38	0
36	39	0
37	39	1
38	40	0
39	40	1
40	41	0

ID	AGE	CHD
41	41	0
42	42	0
43	42	0
44	42	0
45	42	1
46	43	0
47	43	0
48	43	1
49	44	0
50	44	0
51	44	1
52	44	1
53	45	0
54	45	1
55	46	0
56	46	1
57	47	0
58	47	0
59	47	1
60	48	0

ID	AGE	CHD
61	48	1
62	48	1
63	49	0
64	49	0
65	49	1
66	50	0
67	50	1
68	51	0
69	52	0
70	52	1
71	53	1
72	53	1
73	54	1
74	55	0
75	55	1
76	55	1
77	56	1
78	56	1
79	56	1
80	57	0

ID	AGE	CHD
81	57	0
82	57	1
83	57	1
84	57	1
85	57	1
86	58	0
87	58	1
88	58	1
89	59	1
90	59	1
91	60	0
92	60	1
93	61	1
94	62	1
95	62	1
96	63	1
97	64	0
98	64	1
99	65	1
100	69	1

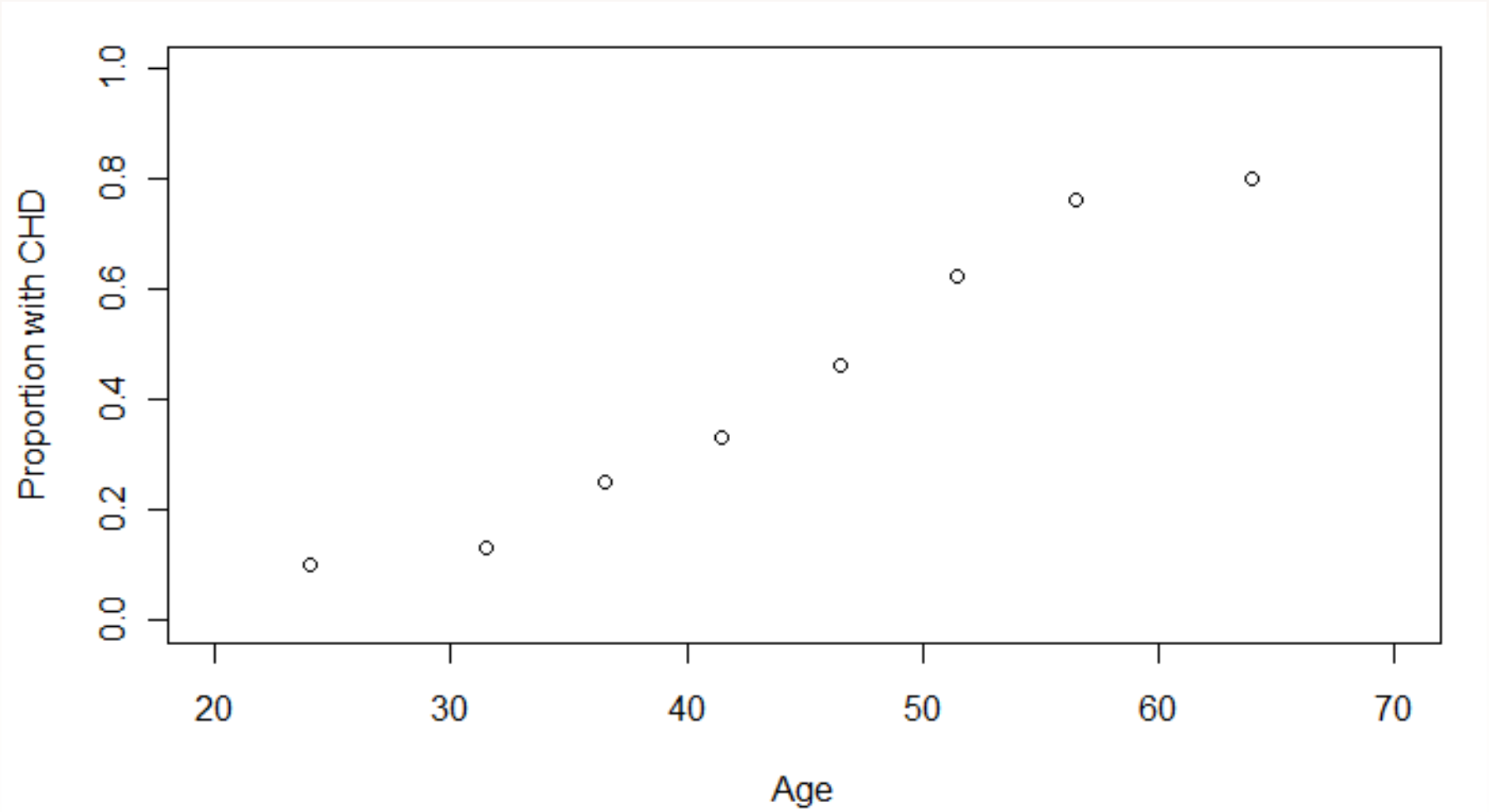
Gráfico 1: Dados individuais de idade e ocorrência de CHD



Agrupando-se os dados em faixas etárias, é possível visualizar melhor a relação entre CHD e AGE.

Age Group	*n*	CHD		Mean (Proportion)
		Absent	Present	
20 - 29	10	9	1	0.10
30 - 34	15	13	2	0.13
35 - 39	12	9	3	0.25
40 - 44	15	10	5	0.33
45 - 49	13	7	6	0.46
50 - 54	8	3	5	0.63
55 - 59	17	4	13	0.76
60 - 69	10	2	8	0.80
Total	100	57	43	0.43

Gráfico 2: Proporções de ocorrência de CHD por faixa etária



Regressão linear $\times$ logística

Em um problema de regressão, a quantidade chave é o valor médio da variável resposta, dado o valor da variável independente.
Esta quantidade é denominada *média condicional* e é expressa como $E(Y|x)$, onde Y denota a variável resposta e x denota um valor da variável independente.
$E(Y|x)$: "Valor esperado de Y, dado o valor x".

Na Regressão Linear, assume-se que esta média pode ser expressa como uma equação linear em x (ou alguma transformação de x ou de Y):

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

Isto significa que é possível para $E(Y|x)$ assumir qualquer valor para x variando na reta real $(-\infty, +\infty)$.

Características da média condicional $E(Y|x)$ quando Y é dicotômica:

- $0 \leq E(Y|x) \leq 1$
- $E(Y|x)$ se aproxima de 0 e de 1 de forma gradual:
  A mudança em $E(Y|x)$ por unidade de variação em x se torna progressivamente menor à medida em que $E(Y|x)$ se aproxima de 0 ou de 1.
- Perfil semelhante ao de uma distribuição acumulada de uma variável aleatória contínua.

É usual utilizar distribuições acumuladas conhecidas como modelos para $E(Y|x)$ quando Y é dicotômica. Na Regressão Logística, a distribuição adotada é a *Distribuição Logística*.

Função Logística / Logito

Função logística

Para simplificar a notação, denotaremos $\pi(x) = E(Y|x)$ a média condicional de Y quando a distribuição logística é utilizada.
A forma específica do modelo de regressão utilizado é

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

Transformação logito*

Transformação chave em regressão logística: inversa da função logística:

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

* *logit, em inglês*

$g(x)$ é linear nos parâmetros β_0, β_1, pode ser contínua e pode variar de $-\infty$ a $+\infty$, dependendo do valor de x.

Distribuição do erro na regressão linear

Na regressão linear assume-se que uma observação da variável resposta pode ser expressa como $y = E(Y|x) + \epsilon$. A quantidade ϵ é denominada *erro* e expressa o desvio de uma observação da respectiva média condicional. Em regressão linear é usual assumir que $\epsilon \sim Normal(0, \sigma^2)$.

Distribuição do erro na regressão logística

Para variáveis resposta dicotômicas, podemos expressar o valor de Y dado x como $y = \pi(x) + \epsilon$.
A quantidade ϵ pode assumir dois valores:

- Com probabilidade $\pi(x)$, $y = 1$ e portanto $\epsilon = 1 - \pi(x)$;
- Com probabilidade $1 - \pi(x)$, $y = 0$ e portanto $\epsilon = -\pi(x)$.

Logo, ϵ tem uma distribuição com média 0 e variância igual a $\pi(x)[1 - \pi(x)]$. Ou seja, Y segue uma distribuição binomial com probabilidade dada pela média condicional, $\pi(x)$.

Ajuste do Modelo de Regressão Logística

- Suponha que tenhamos uma amostra de n observações independentes de pares $(x_i, y_i), i = 1, 2, \ldots, n$, onde $y_i \in \{0, 1\}$ denota a ausência (0) ou presença (1) de uma característica dicotômica, e x_i denota o valor da variável independente para o $i-$ésimo indivíduo.
- Ajustar o modelo de regressão logística da equação (1) sobre um conjunto de dados requer que estimemos os valores de β_0 e β_1, os parâmetros desconhecidos.
- Usualmente, na regressão logística utiliza-se o método de *máxima verossimilhança*.
- Notação: denotaremos por $\boldsymbol{\beta} = [\beta_0, \beta_1]$ o vetor de parâmetros.

Função de Verossimilhança

- Se $Y \in \{0, 1\}$, a expressão para $\pi(x)$ na equação (1) fornece, para um dado valor de β, a probabilidade condicional $P(Y = 1|x)$.
- Logo, $1 - \pi(x)$ fornece a probabilidade condicional $P(Y = 0|x)$.
- Assim, para os pares (x_i, y_i) onde $y_i = 1$, a contribuição para a função de verossimilhança é $\pi(x_i)$, e para aqueles pares onde $y_i = 0$, a contribuição para a função de verossimilhança é $1 - \pi(x_i)$.
- Uma forma conveniente de expressar a contribuição do par (x_i, y_i) para a função de verossimilhança é através da expressão

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

Função de Verossimilhança

- Assumindo-se que as observações sejam independentes, a função de verossimilhança é expressa por

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

- Otimizar a *log verossimilhança* é mais fácil matematicamente:

$$L(\beta) = \ln l(\beta) = \sum_{i=1}^{n} \{y_i \ln \pi(x_i) + (1 - y_i) \ln[1 - \pi(x_i)]\} \ .$$

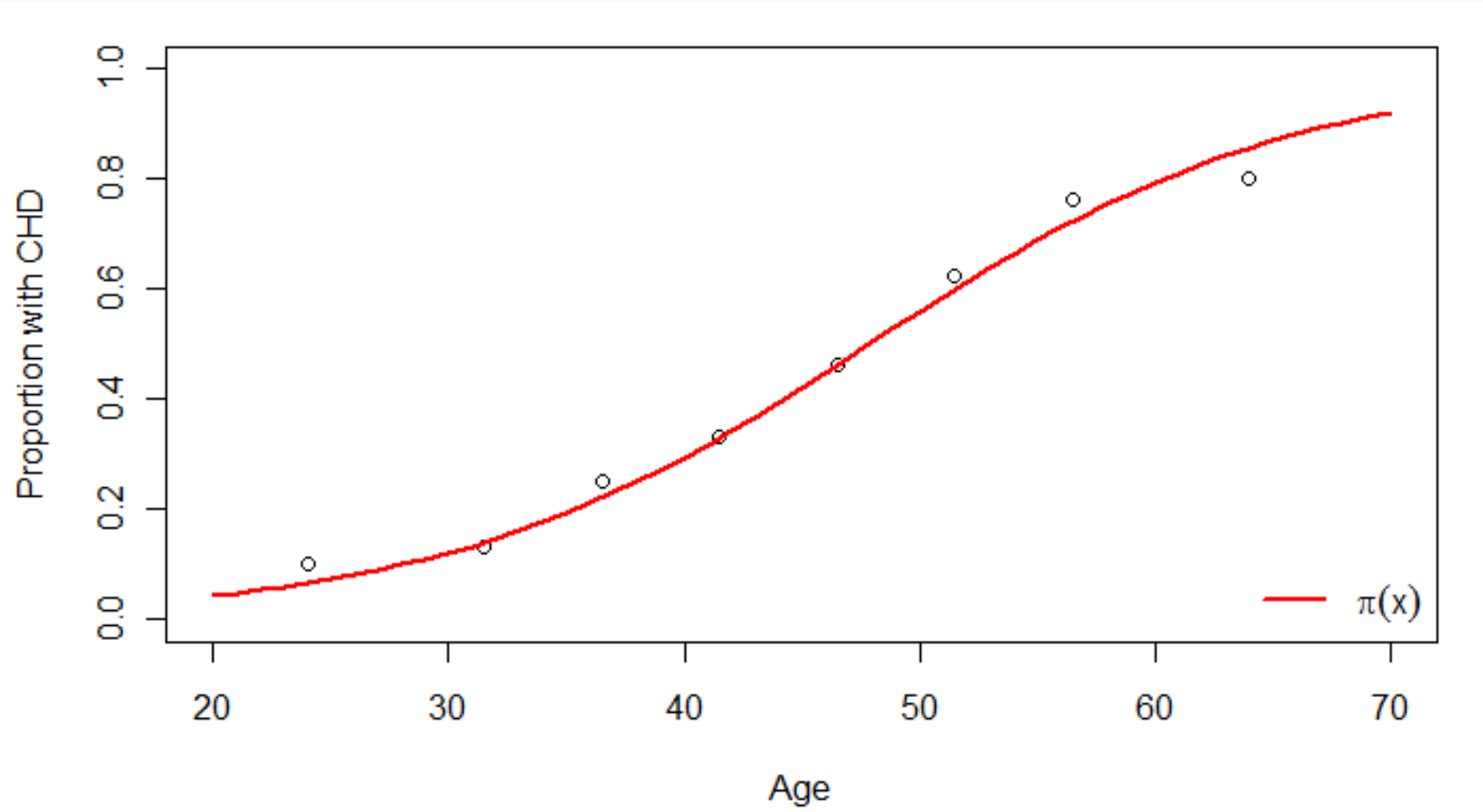
- Para encontrar o valor de β que maximiza $L(\beta)$ derivamos $L(\beta)$ com respeito a β_0 e β_1 e igualamos as expressões a zero (Cramer, 2003):

$$\sum_i [y_i - \pi(x_i)] = 0 \ , \quad \sum_i x_i [y_i - \pi(x_i)] = 0 \ .$$

- Notação:
  $\hat{\beta}$: estimativa de máxima verossimilhança para β.
  $\hat{\pi}(x_i)$: estimativa de máxima verossimilhança para $\pi(x_i)$.

No exemplo anterior:

- $\hat{\beta}_0 = -5.309$, $\hat{\beta}_1 = 0.111$
- $\hat{g}(age) = -5.309 + 0.111 age$



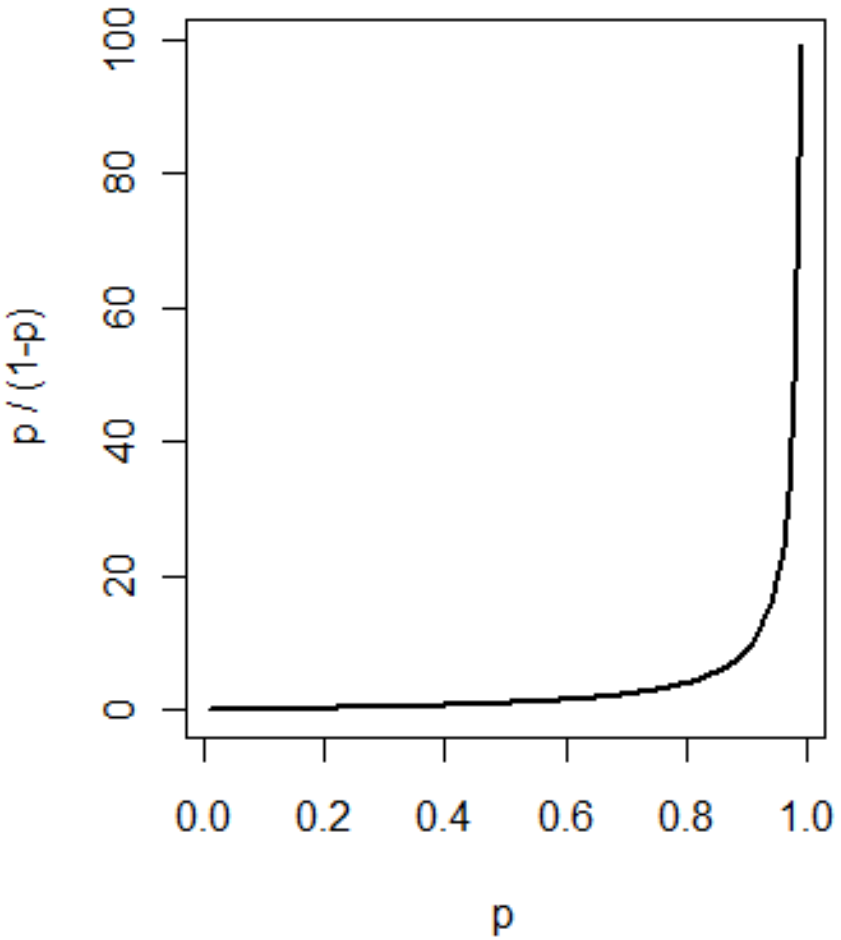
Interpretação dos parâmetros

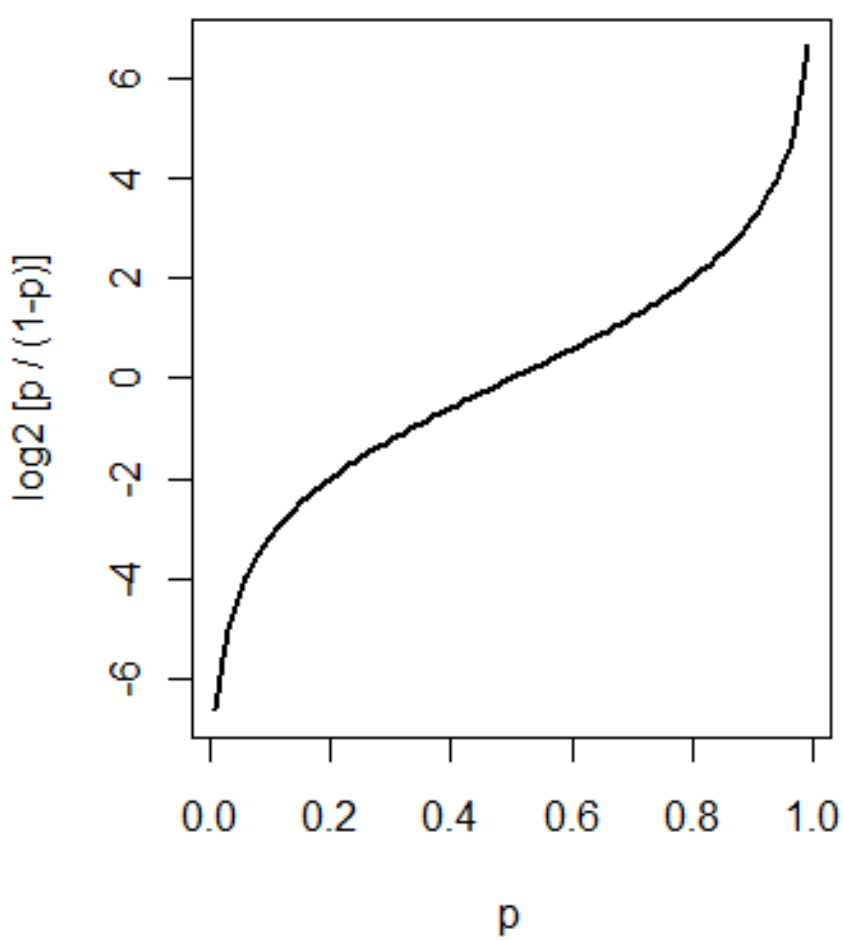
Chances (*Odds*)

- Em certos contextos (p.ex. apostas) é comum se referir à probabilidade de um evento em termos de *chances* (*odds*, em inglês).
- Seja p a probabilidade de um evento ocorrer. As *chances* a favor desse evento são dadas pela razão entre a probabilidade de sucesso e a probabilidade de fracasso: $p/(1-p)$
- P.ex. dizer que um indivíduo tem probabilidade 0.8 de desenvolver uma doença arterial coronária (CHD) é equivalente a dizer que a chance a favor da CHD $0.8/0.2$, ou ainda 4:1.
- Um inconveniente é que essa medida é assimétrica: No exemplo anterior, o indivíduo tem chance $0.8/0.2 = 4$ de desenvolver CHD, e chance $0.2/0.8 = 0.25$ de não desenvolver CHD. Essa assimetria pode ser contornada aplicando-se o logaritmo (em qualquer base > 1):

$$\log_2(0.8/0.2) = 2, \quad \log_2(0.2/0.8) = -2.$$

Ou seja, $\forall\, b > 1$, $\log_b[p/(1-p)] = -log_b[(1-p)/p]$.


p / (1-p)
0
20
40
60
80
100
0.0
0.2
0.4
0.6
0.8
1.0
p


log2 [p / (1-p)]
-6
-4
-2
0
2
4
6
0.0
0.2
0.4
0.6
0.8
1.0
p

Relação entre coeficientes e chances:

- Uma vez que $\beta_0 + \beta_1 x = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right]$, temos:

$$\begin{aligned} g(x+1) - g(x) &= \ln\left[\frac{\pi(x+1)}{1-\pi(x+1)}\right] - \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] \\ &= \beta_0 + \beta_1(x+1) - \beta_0 - \beta_1 x = \beta_1. \end{aligned}$$

- Logo, β_1 representa o incremento esperado no log das chances a favor do evento Y por unidade de incremento na variável X.
- Outra interpretação: β_1 representa o log da razão de chances entre $P(Y = 1|x+1)$ e $P(Y = 1|x)$.

Regressão Logística Múltipla

- Considere uma coleção de p variáveis independentes denotada pelo vetor $\mathbf{x}' = (x_1, x_2, \ldots, x_p)$.
- Como anteriormente, denotemos a probabilidade condicional $\pi(\mathbf{x}) = P(Y = 1|\mathbf{x})$. O logito do modelo de regressão logística múltipla é dada pela equação

$$g(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p,$$

e portanto o modelo de regressão logística é

$$\pi(\mathbf{x}) = \frac{e^{g(\mathbf{x})}}{1 + e^{g(\mathbf{x})}} = \frac{1}{1 + e^{-g(\mathbf{x})}}$$

Ajuste do modelo de regressão logística múltipla

- Assuma que tenhamos uma amostra de n observações independentes $(\mathbf{x}_i, y_i), i = 1, 2, \ldots n$. Denotemos por $\boldsymbol{\beta}' = (\beta_0, \beta_1, \ldots, \beta_p)$ o vetor dos parâmetros a serem ajustados.
- A função de log-verossimilhança é trivialmente extendida, assim como suas respectivas derivadas:

$$\sum_i [y_i - \pi(\mathbf{x}_i)] = 0 \ , \quad \sum_i x_{ij}[y_i - \pi(\mathbf{x}_i)] = 0 \ , \ j = 1, 2 \ldots, p \ .$$

Covariância dos Parâmetros

- Através de intervalos de confiança para os parâmetros e testes de significância, é possível avaliar a precisão do ajuste e definir as variáveis preditoras importantes no modelo.
- Essas etapas são baseadas na matriz de covariância dos parâmetros.
- O método usualmente adotado é oriundo da teoria de estimação de máxima verossimilhança.
- Essa teoria estabelece que, sob certas condições, os estimadores das variâncias e covariâncias podem ser obtidos a partir da matrix das derivadas parciais de 2^a ordem da função de log-verossimilhança.

- Na regressão logística, essas derivadas parciais possuem a seguinte forma geral:

$$\frac{\partial^2 L(\beta)}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij}^2 \pi_i (1 - \pi_i)$$

e

$$\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_l} == -\sum_{i=1}^{n} x_{ij} x_{il} \pi_i (1 - \pi_i)$$

para $j, j = 0, 1, 2, \ldots, p$ onde π_i denota $\pi(\mathbf{x}_i)$.

- A matriz $(p+1) \times (p+1)$ obtida pelas equações anteriores é usualmente denominada *Hessiana* da função de log-verossimilhança, aqui denotada por $\mathbf{H}(\beta)$. A matriz $\mathbf{I}(\beta) = -\mathbf{H}(\beta)$ a *matriz de informação observada de Fisher*.
- A inversa da matriz de informação fornece um limitante inferior para a matriz de covariância dos parâmetros, $\text{Var}(\beta) = \mathbf{I}^{-1}(\beta)$.

Notação

- $\hat{\text{Var}}(\hat{\boldsymbol{\beta}})$ denota a matriz obtida pela avaliação de $\text{Var}(\boldsymbol{\beta})$ sobre $\hat{\boldsymbol{\beta}}$.
- $\hat{\text{Var}}(\hat{\beta}_j)$ denota o j-ésimo elemento da diagonal de $\hat{\text{Var}}(\hat{\boldsymbol{\beta}})$.
- $\hat{\text{Cov}}(\hat{\beta}_i, \hat{\beta}_j)$ denota o elemento na posição (i, j), $i \neq j$, de $\hat{\text{Var}}(\hat{\boldsymbol{\beta}})$.

Erros-padrão dos parâmetros

Os desvios-padrão dos coeficientes estimados, usualmente denominados *erros-padrão*, são estimados por

$$\hat{\text{SE}}(\hat{\beta}_j) = \left[\hat{\text{Var}}(\hat{\beta}_j)\right]^{1/2}.$$

Notação Matricial para a Matriz de Informação

Uma formulação matricial da matriz de informação é dada por
$\hat{\mathbf{I}}(\hat{\beta}) = \mathbf{X}'\mathbf{V}\mathbf{X}$
onde $\mathbf{X}$ é uma matriz $n \times (p+1)$ contendo os dados de cada indivíduo, e $\mathbf{V}$ é uma matriz diagonal $n \times n$ com os elementos $\hat{\pi}_j(1-\hat{\pi}_j)$:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

$$\mathbf{V} = \begin{bmatrix} \hat{\pi}_1(1-\hat{\pi}_1) & 0 & \dots & 0 \\ 0 & \hat{\pi}_2(1-\hat{\pi}_2) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \hat{\pi}_n(1-\hat{\pi}_n) \end{bmatrix}$$

Intervalos de Confiança

- Como usual, o intervalo de confiança para os parâmetros é obtido a partir de suas estimativas e de seus erros padrões:

$$IC_{(1-\alpha)}(\beta_j) = \hat{\beta}_j \pm z_{1-\alpha/2}\, \hat{\text{SE}}(\hat{\beta}_j)\ ,$$

onde $z_{1-\alpha/2}$ corresponde ao quantil $(1-\alpha)$ da distribuição normal padrão.

- No exemplo do estudo de doenças arteriais coronarianas,

$$IC_{0.95}(\beta_0) = (-7.531, -3.087), IC_{0.95}(\beta_1) = (0.064, 0.158),$$

sugerindo que o incremento em ano de idade resulta no aumento de 0.064 a 0.158 no log da chance de CHD.

- Outro aspecto de interesse é obter o intervalo de confiança para a previsão $g(x)$.
- Na regressão logística simples, a formulação é direta: Considerando que

$$\hat{g}(x) = \hat{\beta}_0 + \hat{\beta}_1 x \ ,$$

temos

$$\hat{\text{Var}}[\hat{g}(x)] = \hat{\text{Var}}(\beta_0) + x^2 \hat{\text{Var}}(\beta_1) + 2x\hat{\text{Cov}}(\hat{\beta}_0, \hat{\beta}_1).$$

- A formulação geral incluindo a regressão múltipla pode ser apresentada em notação matricial:

$$\hat{g}(\mathbf{x}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p = \mathbf{x}'\hat{\boldsymbol{\beta}}$$

onde $\mathbf{x}' = (1, x_1, x_2, \ldots x_p)$. Logo,

$$\hat{\text{Var}}[\hat{g}(\mathbf{x})] = \sum_{j=0}^{p} x_j^2 \hat{\text{Var}}(\hat{\beta}_j) + \sum_{j=0}^{p-1} \sum_{k=j+1}^{p} 2x_j x_k \hat{\text{Cov}}(\hat{\beta}_j, \hat{\beta}_k).$$

- Recordando a notação $\hat{\mathbf{I}}(\hat{\beta}) = \mathbf{X}'\mathbf{V}\mathbf{X}$, temos

$$\hat{\text{Var}}(\hat{\beta}) = (\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}.$$

- Assim, $\hat{\text{Var}}[\hat{g}(\mathbf{x})]$ pode ser expressa como

$$\hat{\text{Var}}[\hat{g}(\mathbf{x})] = \mathbf{x}'\hat{\text{Var}}(\hat{\beta})\mathbf{x} = \mathbf{x}'(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{x}.$$

Teste de Significância dos Coeficientes

Uma vez ajustado o modelo de regressão logística, é necessário confirmar a associação entre as variáveis independentes e a variável resposta. Usualmente, utilizam-se *testes de significância*.

Teste da razão de verossimilhança

- O *teste da razão de verossimilhança* é um testes bastante difundido, por ter uma formulação geral, intuitiva e de fácil implementação.
- Para verificar a significância de coeficientes individuais β_j, o teste utiliza a estatística

$$G = -2 \ln \left[\frac{\text{verossimilhança do modelo } \textit{sem} \text{ a variável } (\beta_j = 0)}{\text{verossimilhança do modelo } \textit{com} \text{ a variável } (\beta_j \neq 0)} \right].$$

- Sob a hipótese de $\beta_j = 0$, a estatística G possui uma distribuição chi-quadrado com p graus de liberdade (note que o modelo completo possui $p + 1$ coeficientes, pois inclui β_0).

Teste de Wald

- O *teste de Wald* compara a estimativa de máxima verossimilhança $\hat{\beta}_i$ com o valor testado ($\beta_i = 0$), assumindo que a diferença entre essas medidas é aproximadamente normal.
- Mais especificamente, o teste é obtido comparando-se a estimativa de máxima verossimilhança do parâmetro de interesse, $\hat{\beta}_j$, com a estimativa de seu erro padrão:

$$W = \frac{\hat{\beta}_j}{\widehat{\text{SE}}(\hat{\beta}_j)} \quad .$$

- Sob a hipótese $\beta_j = 0$, a estatística W segue distribuição normal padrão. Assim, considerando-se a hipótese alternativa $\beta_j \neq 0$, o p-valor corresponde à probabilidade
$P(|z| > W) = 2P(z > W)$,
onde z denota uma variável com distribuição normal padrão.

Referências

- D.W.Hosmer, S.Lemeshow (2000). *Applied Logistic Regression*, 2nd ed, Wiley.
- J. S. Cramer (2003). *The origins and development of the logit model*. Cambridge UP. http://www.cambridge.org/resources/0521815886/1208_default.pdf