Aula 5

Estudos com o tempo de duração

Modelo de Regressão logística

Quando não é possível estabelecer o número de participantes , devido a uma limitação de ocorrência (incidência). O tempo de duração pode ser preestabelecido.

Neste caso N_{ij} , i,j =1,2 são contagens aleatórias, com N_{ij} e também N, conhecidas somente após o término da coleta dos dados.

		Categoria da Resposta Y		
Categoria da Variável X		j	j	Totais
		1	2	
i	1	n11	n12	n1.
i	2	n21	n22	n2.
Totais		n.1	n.2	n

Óbito HIV segundo sexo e Região em 2011.

	Sexo		
Região	FEM	MASC	Total
Sudeste	1078	270	1348
Sul	319	65	384
Total	1397	335	1732

Boletim de Epidemiologia 2012

	Bacteria		
Placa	TipoI	TipoII	Total
A	246	17	263
B	458	32	490
Total	704	49	753

A distribuição de Poisson

A distribuição de Poisson pode ser vista como um limite "especial" da distribuição binomial: Suponha-se que n aumenta e que p diminui de tal forma que $E(X) = np$ se mantém constante, seja λ essa constante, então tem-se

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} =$$

$$= \frac{n(n-1)\cdots(n-x+1)}{x!}\left(\frac{np}{n}\right)^x\left(1-\frac{np}{n}\right)^n\left(1-\frac{np}{n}\right)^{-x} =$$

$$= \frac{n(n-1)\cdots(n-x+1)}{x!}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-x} =$$

$$= \frac{n(n-1)\cdots(n-x+1)}{n^x}\frac{\lambda^x}{x!}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-x} =$$

$$= \frac{n}{n}\frac{n-1}{n}\cdots\frac{n-x+1}{n}\frac{\lambda^x}{x!}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-x}$$

$$\downarrow \quad \downarrow \qquad \downarrow \qquad \downarrow \qquad \downarrow$$

$$1 \quad 1 \qquad 1 \qquad e^{-\lambda} \qquad 1$$

logo

$$\lim_{n\to\infty, np=\lambda} P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0,1,2,\ldots \ (\lambda > 0)$$

Definição: Dado um intervalo de números reais suponha-se que certas ocorrências surgem aleatoriamente ao longo do intervalo. Se o intervalo puder ser subdividido em subintervalos de comprimento suficientemente pequeno de modo a que se verifique:

1) A probabilidade de mais do que uma ocorrência num subintervalo é zero.

2) A probabilidade duma ocorrência num subintervalo é constante e proporcional ao comprimento do subintervalo.

3) As ocorrências nos diversos subintervalos são independentes.

- Assumimos que Nij tem uma distribuição binomial com parâmetro μij=tλij,
- λij a taxa média por unidade de tempo
- t a duração do experimento ou observação
- Sendo Nij independentes temos o produto de Poisson independentes.

$$P(\mathbf{N}=\mathbf{n})=\prod_{i=1}^{2}\prod_{j=1}^{2}P(N_{ij}=n_{ij})=\prod_{i=1}^{2}\prod_{j=1}^{2}\frac{e^{-\mu_{ij}}(\mu_{ij})^{n_{ij}}}{(n_{ij})!},\quad \mu_{ij}>0$$

$$\text{com } (\mathbf{N}=\mathbf{n})=(N_{11}=n_{11},N_{12}=n_{12},N_{21}=n_{21},N_{22}=n_{22}).$$

com $\mu_{ij}=\hat{n}_{ij}$

As hipóteses neste modelo serão:

$$H_0: \frac{\mu_{1j}}{\mu_{1\bullet}} = \frac{\mu_{2j}}{\mu_{2\bullet}} \left(= \frac{\mu_{\bullet j}}{\mu} \right), \quad \text{para j} = 1,2.$$

$$H_A: \frac{\mu_{1j}}{\mu_{1\bullet}} \neq \frac{\mu_{2j}}{\mu_{2\bullet}}$$

$$H_0 : \mu_{ij} = \frac{(\mu_{i\bullet})\ (\mu_{\bullet j})}{\mu}, \text{ para i, j} = 1,2.$$

$$H_A: \mu_{ij} \neq \frac{(\mu_{i\bullet})\ (\mu_{\bullet j})}{\mu}, \text{ para ao menos um par } ij.$$

Pode-se utilizar as estatiscas já estudadas para verificar a H0
Tp, Tn e TL

$$E(N_{ij}) = \mu_{ij} \qquad E(N_{ij}) = \frac{(\mu_{i\bullet})(\mu_{\bullet j})}{\mu}$$

$$e_{ij} = \frac{(n_{i\bullet})(n_{\bullet j})}{n}$$

Usualmente utiliza-se a distribuição multinomial nestes casos

Justificativa: a distribuição de probabilidades do vetor $(N_{11}, \ldots, N_{22})$, N_{ij} Poisson independentes, condicional à soma $N = \sum_{i,j} N_{ij}$, segue distribuição Multinomial (N, **p**), com **p** = $(p_{11}, \ldots, p_{22})$, em que $p_{ij} = \frac{\mu_{ij}}{\sum_{i,j} \mu_{ij}}$, para i,j = 1, 2.

Esquema Estudos Epidemiologicos

Quadro 7-1 Esquema para Análise de Estudos Epidemiológicos

Tipo de Estudo	Medida de Ocorrência	Medidas de Associação		Medida de Significância Estatística
		Proporcionalidade	Diferença	
Ecológico	Médias/freqüências	Razão de médias/correlação	–	Teste de diferença de médias (Z e t) Teste de significância da correlação
Seccional	Prevalência	Razão de prevalência	Diferença de prevalência (DP)	Teste de diferença de proporções (Z e t) Teste de qui-quadrado (x^2)
Coorte	Incidência	Risco relativo (RR)	Risco atribuível (RA, RAP%)	Teste de qui-quadrado (x^2)
Caso-controle	–	*Odds ratio* (OR)	RA de Levin (RAP%)	Teste de qui-quadrado (x^2) Mantel-Haenszel (MH x^2)

Modelos para dados onde a variável resposta é dicotômica

- Y -> {0 ou 1}
- $X_1, X_2, X_3, \ldots X_k$

$$E(Y|x) = \text{soma } (Y * P(Y|x)$$

Exemplos

Tabela 0.1: Doença Coronária

	Doença Coronária			
Idade	Nao	Sim	Total	$P(Y = 1\|x)$
20-29	9	1	10	0,10
30-34	13	2	15	0,13
35-39	9	3	12	0,25
40-44	10	4	14	0,29
45-49	7	5	12	0,42
50-54	3	6	9	0,67
55-59	4	13	17	0,76
60-69	2	8	10	0,80
Total	57	43	100	0,43

Exemplos

Tabela 0.2: Ocorrência de Violência

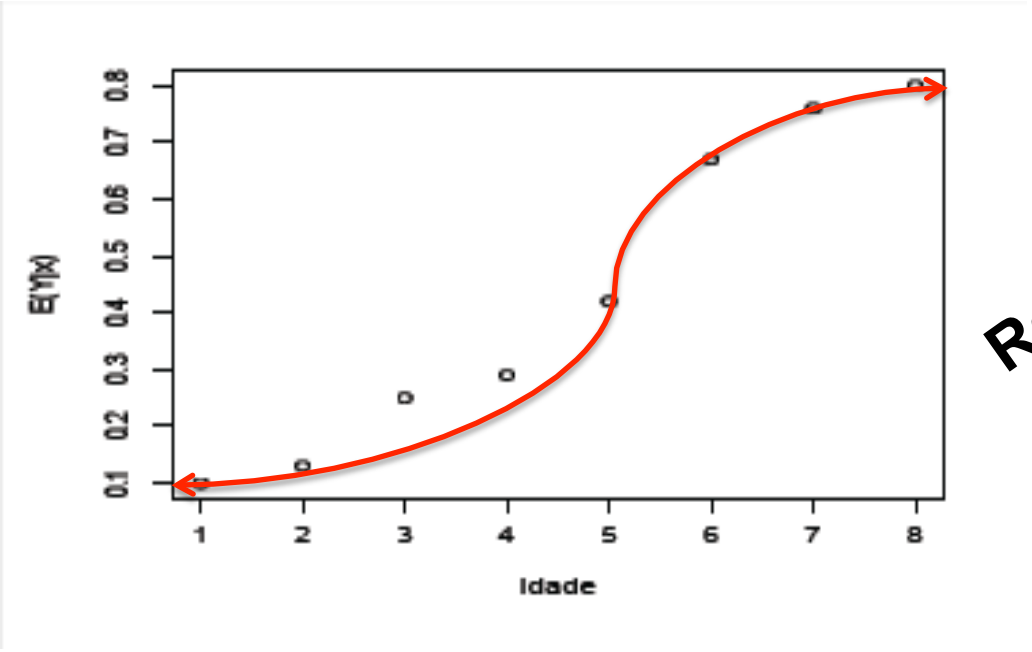
	Violencia Fisica			
Idade	0	1	Total	$P(Y = 1\|x)$
15 - 25 anos	95	47	142	0,33
26 - 35 anos	86	45	131	0,34
36 - 45 anos	88	58	146	0,40
46 - 49 anos	55	21	76	0,28
Total	324	171	495	0,35

Quantidade de interesse

$$E(Y|x) = \text{soma}\,(Y * P(Y|x)$$

$$E(Y|x) = 1 * P(Y=1|x) + 0 * P(Y=0|x)$$

$$= P(Y=1|x)$$



Relação não é linear

Função não linear

$$\underbrace{E(Y \mid \mathbf{x}) = P(Y = 1 \mid \mathbf{x})} = \theta(\mathbf{x}) = \frac{\exp\left\{\beta_0 + \sum_{k=1}^{p} \beta_k \mathrm{x}_k\right\}}{1 + \exp\left\{\beta_0 + \sum_{k=1}^{p} \beta_k \mathrm{x}_k\right\}}$$

$$1 - \theta(\mathbf{x}) = \frac{1}{1 + \exp\left\{\beta_0 + \sum_{k=1}^{p} \beta_k \mathrm{x}_k\right\}}.$$

sendo $\mathbf{x} = (\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_p)$ = valores observados das variáveis $\mathbf{X}$,
β_0 = constante e β_k $(k = 1, \ldots, p)$ os p parâmetros de regressão.

Função linearizável

$$\theta(\mathbf{x}) = \frac{\exp\{\boldsymbol{\beta}'\mathbf{x}\}}{1+\exp\{\boldsymbol{\beta}'\mathbf{x}\}} \quad \text{e} \quad 1-\theta(\mathbf{x}) = \frac{1}{1+\exp\{\boldsymbol{\beta}'\mathbf{x}\}}$$

$$\ln\left(\frac{\theta(\mathbf{x})}{1-\theta(\mathbf{x})}\right) = \beta_0 + \sum_{k=1}^{p} \beta_k \mathbf{x}_k = \boldsymbol{\beta}'\mathbf{x}.$$

Esta transformação é denominada **logito**.

A razão entre $\theta(\mathbf{x})$ e $1-\theta(\mathbf{x})$ $\Rightarrow$ definição de *odds*

$\Downarrow$

$$odds = \frac{\theta(\mathbf{x})}{1-\theta(\mathbf{x})} = \exp\{\boldsymbol{\beta}'\mathbf{x}\}.$$

Como encontrar quantidades estimadas

Estimação de $\beta \Rightarrow$ Método da Máxima Verossimilhança

$$L(\beta) = \prod_{\ell=1}^{n} P(Y = y_\ell \mid \mathbf{x}_\ell) = \prod_{\ell=1}^{n} \left(\theta(\mathbf{x}_\ell)\right)^{y_\ell} \left(1 - \theta(\mathbf{x}_\ell)\right)^{1-y_\ell}$$

- $y_\ell = 1$, se indivíduo l apresentou a resposta e $y_\ell = 0$, c.c.
- Valores de β que maximizam $\ln L(\beta) \Rightarrow \widehat{\beta}$.
- Distribuição assintótica de $\widehat{\beta} \Rightarrow$ Normal

Estimação da matriz de variâncias-covariâncias de $\widehat{\beta}$

$\Sigma(\beta) = [I(\beta)]^{-1} =$ matriz de variâncias-covariâncias

- $I(\beta) =$ matriz contendo o negativo das derivadas parciais de 2^{a} ordem de $\ln L(\beta)$.
- Estimadores são obtidos por avaliar $\Sigma(\beta)$ em $\widehat{\beta}$.

Qual variável explicativa é importante?

- Testar hipóteses relativas aos parâmetros β_k $(k = 1, \ldots, p)$

1. Teste da Razão de Verossimilhanças (TRV)

$$TRV = -2\ln\left[\frac{L_S}{L_C}\right] = \underbrace{2\ln(L_C) - 2\ln(L_S)}_{\text{equivalente à diferença de } \textit{deviances}} \sim \chi^2_{(q)}$$

- L_S: função de verossimilhança associada ao modelo sem a(s) variável(is) sob investigação
- L_C: função de verossimilhança associada ao modelo com a(s) variável(is) sob investigação
- q = diferença de parâmetros entre os dois modelos.

- Considere que modelos encaixados sejam ajustados aos dados de um estudo em que Y é binária e X_1 e X_2 são categóricas com duas categorias cada.

Tabela de Análise de *Deviances* (ANODEV).

Modelos	g.l.	*Deviances*	*TRV*	$\neq$ g.l.
Nulo	gl_N	D_N		
X_1	$gl_N - 1$	D_1	$D_N - D_1$	1
$X_2 \mid X_1$	$gl_N - 2$	D_2	$D_1 - D_2$	1
$X_1 * X_2 \mid X_1, X_2$	$gl_N - 3$	D_3	$D_2 - D_3$	1

gl_N = g.l. do modelo nulo = número de subpopulações $-$ 1

- Obs: na presença de dados faltantes, o tamanho amostral nos modelos sequenciais dependerá das variáveis X_k que os compõem $\Rightarrow$ *TRV* apresentará problemas.

Ainda escolhendo variáveis

2. Teste de Wald (Wald, 1943)

i) Para testar hipóteses relativas a um parâmetro

$$H_0 : \beta_k = 0,\ k = 1, \ldots, p$$

$$W = \frac{(\widehat{\beta}_j)^2}{\widehat{Var}(\widehat{\beta}_j)} \sim \chi^2_1$$

ii) Para hipóteses relativas a $q \geq 2$ parâmetros

$$H_0 : \beta^* = \mathbf{0}\ (\beta^* = \text{vetor } q \times 1)$$

$$W = (\widehat{\beta}^*)' [\widehat{\Sigma}(\widehat{\beta}^*)]^{-1} (\widehat{\beta}^*) \sim \chi^2_q$$

- Sob a hipótese H_0: modelo ajustado é satisfatório, faz-se uso de estatisticas que resumem a concordância entre os valores observados e os preditos pelo modelo.

$$Q_P = \sum_{i,j} \frac{\left(n_{ij} - e_{ij}\right)^2}{e_{ij}} \sim \chi^2_m$$

$$Q_L = 2\sum_{i,j} n_{ij} \ln\left(\frac{n_{ij}}{e_{ij}}\right) \sim \chi^2_m$$

$$e_{ij} = n_{i+}\,\widehat{\theta}(\mathbf{x}_i),\ j = 1 \quad \text{e} \quad e_{ij} = n_{i+}\,(1 - \widehat{\theta}(\mathbf{x}_i)),\ j = 2.$$

n_{i+} = sujeitos na i-ésima subpopulação da tabela de dados $s \times 2$.

$\widehat{\theta}(\mathbf{x}_i)$ = probabilidade $P(Y = 1 \mid \mathbf{x}_i)$ predita pelo modelo ajustado.

e_{ij} = frequências esperadas sob o modelo ajustado.

m = nº subpopulações − nº parâmetros do modelo ajustado.

- Na presença de variáveis contínuas $\Rightarrow$ frequências muito pequenas para a grande maioria das s subpopulações.

$\Downarrow$

inviabiliza o uso de Q_L e Q_P

$\Downarrow$

Hosmer e Lemeshow (1989) propuseram **uma estatística alternativa**, Q_{HL}, que é obtida calculando-se a estatística qui-quadrado de Pearson a partir de uma tabela $g \times 2$ de frequências observadas e preditas

Estatística de Hosmer e Lemeshow

- Inicialmente, as n observações são ordenadas em ordem crescente das probabilidades $\theta(\mathbf{x})$ preditas pelo modelo.
- Tais observações são, então, divididas em g grupos ($g = 10$, por exemplo). No $1^{\underline{0}}$ grupo ficam as n_1 observações com probabilidades estimadas $< 0{,}1$ e, no último, as n_g observações com probabilidades $\geq 0{,}9$.

$$Q_{HL} = \sum_{i=1}^{g} \frac{(o_i - n_i \overline{\theta}(\mathbf{x}_i))^2}{n_i \overline{\theta}(\mathbf{x}_i)(1 - \overline{\theta}(\mathbf{x}_i))} \sim \chi^2_{(g-2)}$$

$n_i =$ frequência de observações no grupo i

$o_i =$ frequência de resposta $Y = 1$ no grupo i

$\overline{\theta}(\mathbf{x}_i) =$ probab. média estimada de resposta $Y = 1$ no grupo i.

Voltamos ao exemplo

Tabela 0.1: Doença Coronária

	Doença Coronária			
Idade	Nao	Sim	Total	$P(Y=1\|x)$
20-29	9	1	10	0,10
30-34	13	2	15	0,13
35-39	9	3	12	0,25
40-44	10	4	14	0,29
45-49	7	5	12	0,42
50-54	3	6	9	0,67
55-59	4	13	17	0,76
60-69	2	8	10	0,80
Total	57	43	100	0,43

$\widehat{\beta}_0 = -5{,}123$ (e.p. $= 1{,}11$) e $\widehat{\beta}_1 = 0{,}1058$ (e.p. $= 0{,}023$).

$$\widehat{\theta}(\mathbf{x}) = \frac{\exp\{-5,123 + 0,1058\,\mathbf{x}\}}{1 + \exp\{-5,123 + 0,1058\,\mathbf{x}\}}$$

$$\ln\left(\frac{\widehat{\theta}(\mathbf{x})}{1-\widehat{\theta}(\mathbf{x})}\right) = -5,123 + 0,1058\,\mathbf{x}$$

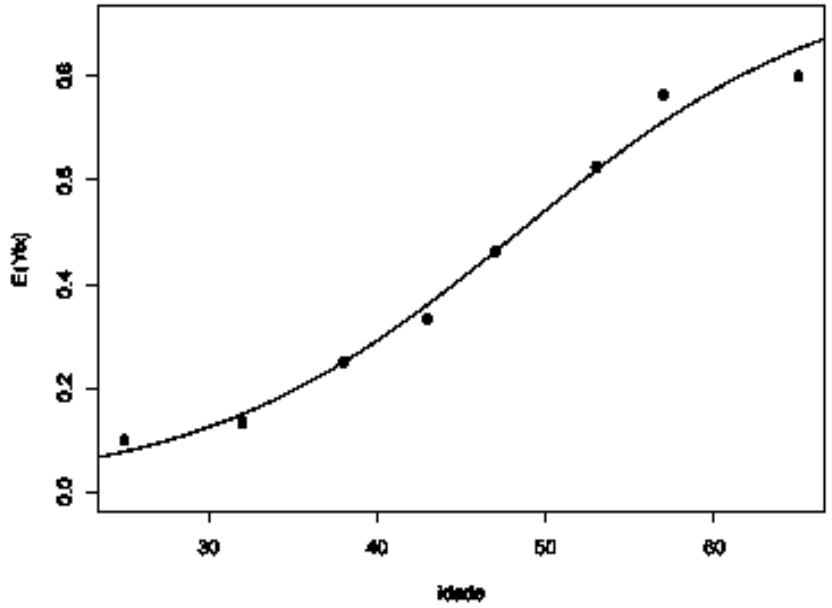
Tabela 1. Diferenças de *deviances*

Modelos	g.l.	*Deviances*	Diferenças	$\neq$ g.l.
Nulo	7	28,7015		
X: idade	6	0,5838	28,1177	1

Tabela 2. Análise de *Deviance* (ANODEV)

Fonte de variação	g.l.	*Deviances*	*TRV*	valor p
Regressão	1	28,1177	28,1177	< 0,00001
Deviance residual	6	0,5838		
Deviance total	7	28,7015		

- Evidências de associação entre idade e doença coronária.
- Ainda, teste de Wald $\Rrightarrow W = 20,49$ (g.l. = 1, $p < 0,00001$)



A partir do modelo ajustado tem-se, por exemplo:

x_i	$\widehat{\theta}(x_i)$	$1-\widehat{\theta}(x_i)$	$\frac{\widehat{\theta}(x_i)}{1-\widehat{\theta}(x_i)} = \exp\{\widehat{\beta}_0+\widehat{\beta}_1 x_i\}$
26	0,0853	0,9147	$\exp\{\widehat{\beta}_0+\widehat{\beta}_1 * 26\} = 0{,}093$
27	0,0939	0,9061	$\exp\{\widehat{\beta}_0+\widehat{\beta}_1 * 27\} = 0{,}103$
65	0,8524	0,1476	$\exp\{\widehat{\beta}_0+\widehat{\beta}_1 * 65\} = 5{,}774$

- Relembrando que $\frac{\theta(\mathbf{x}_i)}{1-\theta(\mathbf{x}_i)} = odds$, segue que:

$$\widehat{OR} = \frac{odds_{(27)}}{odds_{(26)}} = \exp\{\widehat{\beta}_1(27-26)\} = \exp\{\widehat{\beta}_1\} \approx 1,11$$

$$\widehat{OR} = \frac{odds_{(65)}}{odds_{(26)}} = \exp\{\widehat{\beta}_1(65-26)\} = \exp\{\widehat{\beta}_1 * 39\} \approx 62$$

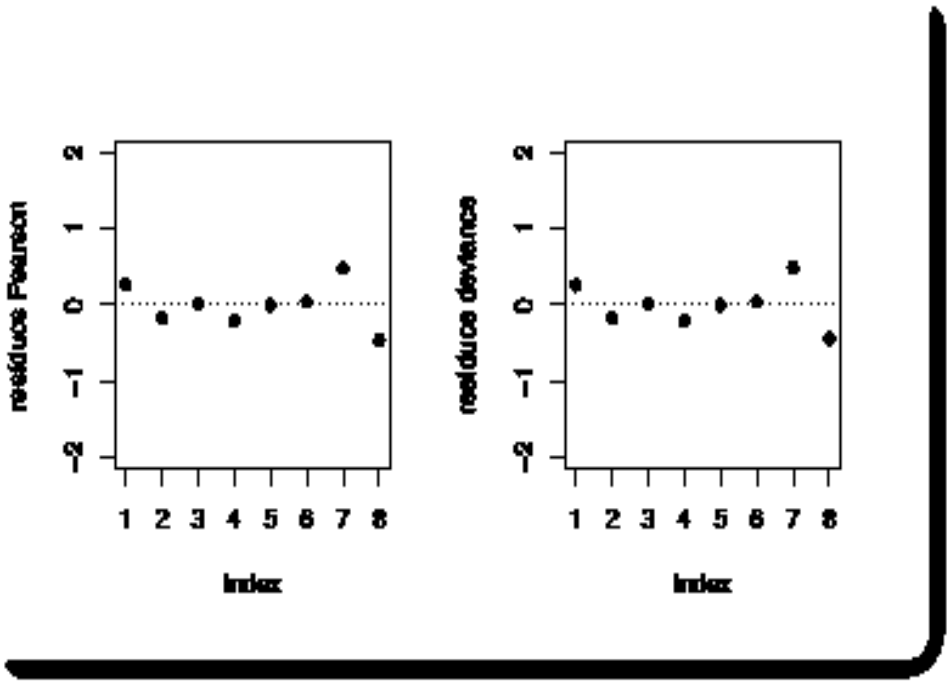
- A *odds* de doença coronária entre indivíduos com 65 anos de idade é $\approx$ 62 vezes a dos indivíduos com 26 anos.

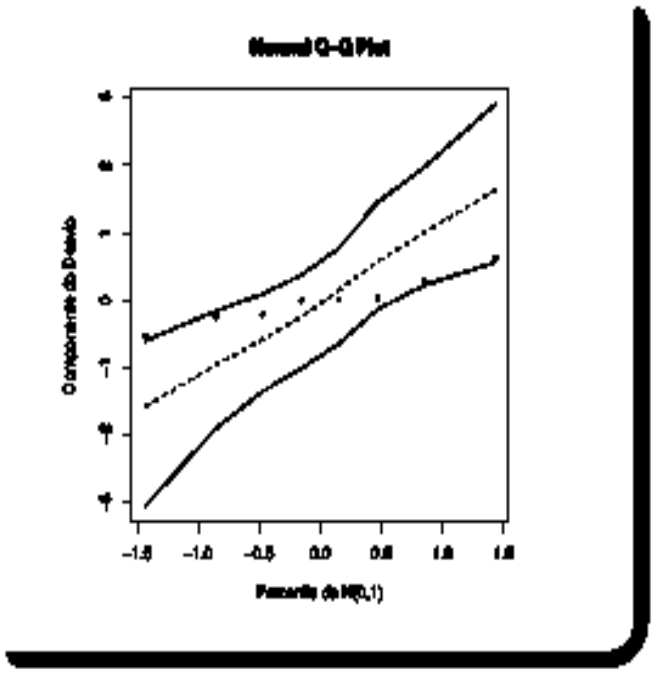
Obs: OR nos modelos de regressão logística são denominadas OR ajustadas, uma vez que o efeito β_k associado à covariável k é estimado na presença dos demais no modelo.

Ajuste bom????

Qual a qualidade deste ajuste????

- $Q_p = 0{,}59$ ($p = 0{,}9965$) e $Q_L = 0{,}58$ ($p = 0{,}9967$), g.l.= 6.
- Resíduos c_i e d_i entre $-2{,}5$ e 2,5.





- Área abaixo da curva ROC: AUC $\approx 0{,}79$.
- Evidências favoráveis ao modelo ajustado.

- Limitação das estatísticas Q_p e $Q_L \Rightarrow$ único valor é utilizado para resumir uma quantidade considerável de informação.
- Pregibon (1981) estendeu os métodos de diagnóstico de regressão linear para a regressão logística, fazendo uso dos componentes individuais das estatísticas Q_p e Q_L.

$$c_i = \underbrace{\frac{n_{i1} - (n_{i+})\,\widehat{\theta}(\mathbf{x}_i)}{\sqrt{(n_{i+})\,\widehat{\theta}(\mathbf{x}_i)\,(1-\widehat{\theta}(\mathbf{x}_i))}}}_{\text{resíduos de Pearson}}, i = 1, \cdots, s.$$

- Componentes c_i são denominados **resíduos de Pearson**, pois a soma deles ao quadrado resulta em Q_P, i.e.,

$$Q_P = \sum_{i=1}^{s} (c_i)^2$$

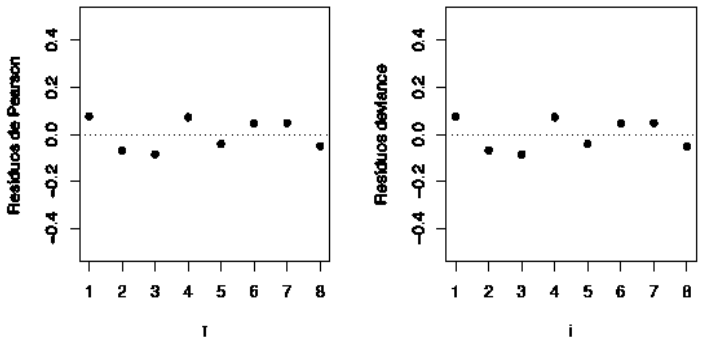
- Limitação das estatísticas Q_p e $Q_L \Rightarrow$ único valor é utilizado para resumir uma quantidade considerável de informação.
- Pregibon (1981) estendeu os métodos de diagnóstico de regressão linear para a regressão logística, fazendo uso dos componentes individuais das estatísticas Q_p e Q_L.

$$c_i = \underbrace{\frac{n_{i1} - (n_{i+})\,\widehat{\theta}(\mathbf{x}_i)}{\sqrt{(n_{i+})\,\widehat{\theta}(\mathbf{x}_i)\,(1-\widehat{\theta}(\mathbf{x}_i))}}}_{\text{resíduos de Pearson}}, i = 1, \cdots, s.$$

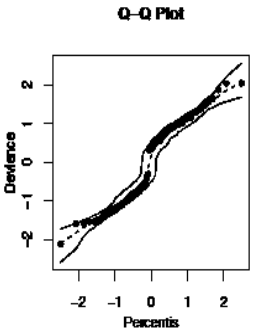
- Componentes c_i são denominados **resíduos de Pearson**, pois a soma deles ao quadrado resulta em Q_P, i.e.,

$$Q_P = \sum_{i=1}^{s} (c_i)^2$$

- Distribuição aproximada dos resíduos c_i e $d_i \sim N(0,1)$.
- Resíduos excedendo $\pm$ 2,5 pode indicar
  - possível falta de ajuste do modelo
  - presença de *outliers*
  - padrões sistemáticos de variação.



- Assumindo que os resíduos d_i seguem distribuição aproximada normal $\Rightarrow$ construir *normal Q-Q plot* com envelope simulado (Davison e Gigli, 1989).



- Se os resíduos estiverem dentro do envelope simulado $\Rightarrow$ evidências favoráveis ao modelo ajustado.

Ben e Yohai (2004) argumentam, contudo, que para alguns MLG, tal distribuição pode estar distante da normalidade.

Assim, propuseram uma estimativa da distribuição dos resíduos d_i, de modo que no *Q-Q plot* tais resíduos são graficado versus os quantis da distribuição estimada.

Para avaliar o poder preditivo do modelo é necessário estabelecer um *ponto de corte* ($0 < \text{pc} < 1$), tal que:

a) Probabilidades preditas pelo modelo $\geq$ pc $\Rightarrow$ Y = 1
b) Probabilidades preditas pelo modelo $<$ pc $\Rightarrow$ Y = 0.

Resposta Observada	Resposta Predita pelo Modelo		Totais
	Y = 1 (+)	Y = 0 (−)	
Y = 1 (+)	a	b	(a + b)
Y = 0 (−)	c	d	(c + d)
Totais	(a + c)	(b + d)	n

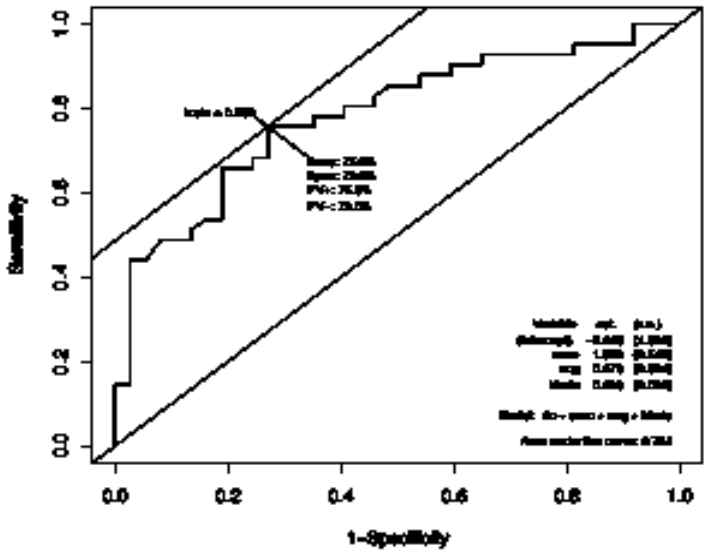
$\Rightarrow$ **Sensibilidade** $= \frac{a}{a+b}$ = taxa de verdadeiros +

$\Rightarrow$ **Especificidade** $= \frac{d}{c+d}$ = taxa de verdadeiros −

$\Rightarrow$ **Valor Preditivo** $= \frac{a+d}{n}$ = proporção geral de acertos

Para diversos pontos de corte $\Rightarrow$ **Curva ROC**

- Pares (x, y) = (1 − especificidade, sensibilidade).
- Modelo com discriminação perfeita $\Rightarrow$ (x, y) = (0,1).
- Pontos de corte próximos ao canto superior esquerdo, produzirão os maiores % de acertos (V+ e V−).
- Quanto mais próxima de 1 for a área abaixo da curva, melhor o poder de predição do modelo.



NO R

```
resim<-c(1,2,3,5,6,5,13,8)
resnao<-c(9,13,9,10,7,3,4,2)
idade<-c(25,32,38,43,47,53,57,65)
dados<-as.data.frame(cbind(resim,resnao,idade))
attach(dados)
ajust<-glm(as.matrix(dados[,c(1,2)])~idade,
       family=binomial(link="logit"),data=dados)
anova(ajust,test="Chisq")
summary(ajust)
ajust$y
ajust$fitted.values
dev<-residuals(ajust,type='deviance')
QL<-sum(dev^2)
pl<-1-pchisq(QL,6)
cbind(QL,pl)
plot(dev,ylim=c(-2,2),ylab="residuos deviance",pch=16)
abline(h=0, lty=3)
```

```
rpears<-residuals(ajust,type='pearson')
rpears
QP<-sum(rpears^2)
p2<-1-pchisq(QP,6)
cbind(QP,p2)

plot(rpears,ylim=c(-2,2),ylab="residuos Pearson",pch=16)
abline(h=0,lty=3)

theta<-resim/(resim+resnao)
plot(idade,theta,ylim=range(0,0.9),xlab="idade",
     ylab="E(Y|x)",pch=16)
idade<-20:70
modajust<-(exp(-5.123+0.1058*idade))/(1+ exp(-5.123+
          0.1058*idade))
lines(idade,modajust)
```

```
ntot<-c(10,15,12,15,13,8,17,10)
fit.model<-ajust
source("http://www.ime.usp.br/~giapaula/envelr_bino")

dados1<-read.table("coronaria.txt",h=T)
# dados 1 = arquivo com 1 indivíduo por linha (100 x 2)
attach(dados1)
dados1[1:3,]
y idade
1 25
0 25
0 25
...
require(Epi)
ROC(form=y~idade,plot="ROC")
```

FUNÇÃO LIGAÇÃO

Alguns *links* para dados com resposta binária.

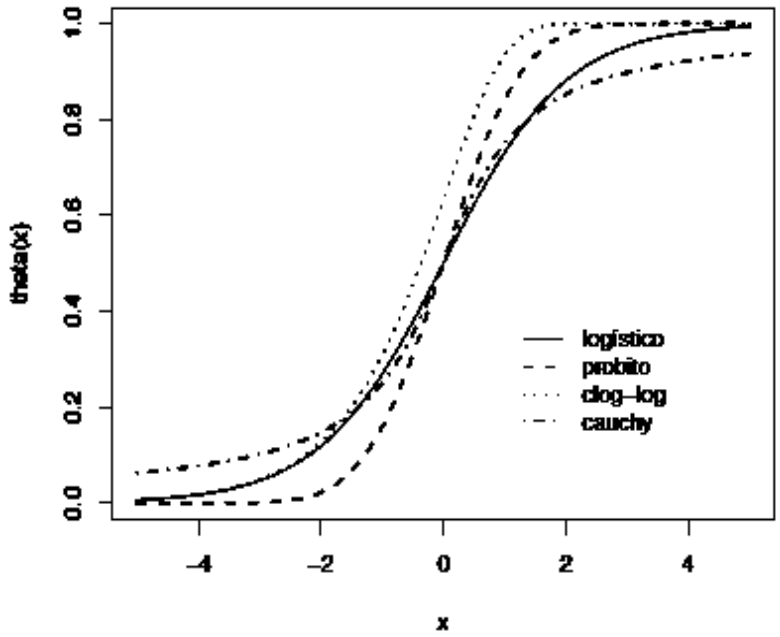
$\theta(\mathbf{x}) = F(\beta'\mathbf{x})$	*links* paramétricos alternativos	
$\frac{\exp\{\beta'\mathbf{x}\}}{1+\exp\{\beta'\mathbf{x}\}}$	logito	$\Rightarrow \ln\left(\frac{\theta(\mathbf{x})}{1-\theta(\mathbf{x})}\right)$
$\Phi(\beta'\mathbf{x})$	probito	$\Rightarrow \Phi^{-1}(\theta(\mathbf{x}))$
$1-\exp\{-\exp\{\beta'\mathbf{x}\}\}$	clog-log	$\Rightarrow \ln(-\ln(1-\theta(\mathbf{x})))$
$\frac{1}{2}+\frac{\text{arctg}(\beta'\mathbf{x})}{\pi}$	cauchy	$\Rightarrow F^{-1}(\theta(\mathbf{x}))$

$\Phi(\cdot)$ denota a função de distribuição da N(0,1), arctg = arco tangente

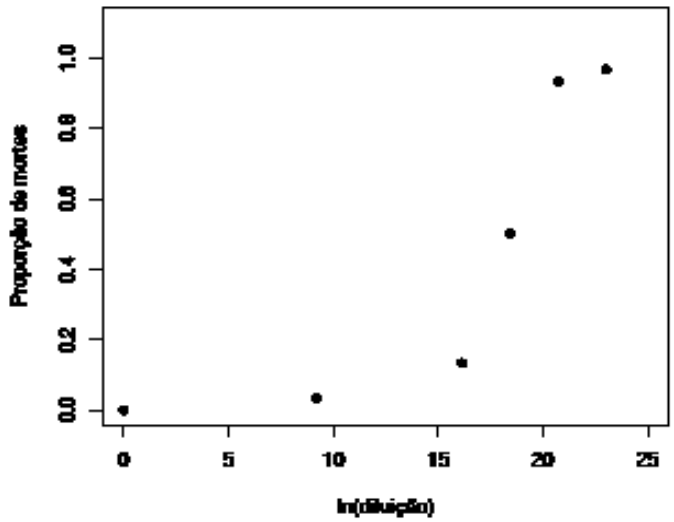
$F(\cdot)$denota a função de distribuição da Cauchy(0,1) $\sim$ t-Student$_{(1g.l.)}$

$\Rrightarrow$ Simétricos: logístico, probito e cauchy

$\Rrightarrow$ Assimétricos: complemento log-log.



- Procedimentos de estimação, qualidade e diagnóstico são análogos aos do modelo logístico.
- Interpretação dos parâmetros difere da apresentada para o modelo logístico.



Logístico $\Rightarrow \quad \theta(x_i) = \dfrac{\exp\{\beta_0 + \beta_1 x_i\}}{1 + \exp\{\beta_0 + \beta_1 x_i\}}$

Probito $\Rightarrow \quad \theta(x_i) = \Phi(\beta_0 + \beta_1 x_i)$

Clog-log $\Rightarrow \quad \theta(x_i) = 1 - \exp\{-\exp\{\beta_0 + \beta_1 x_i\}\}$

Cauchy $\Rightarrow \quad \theta(x_i) = \dfrac{1}{2} + \dfrac{\arctan(\beta_0 + \beta_1 x_i)}{\pi}$

EXEMPLO

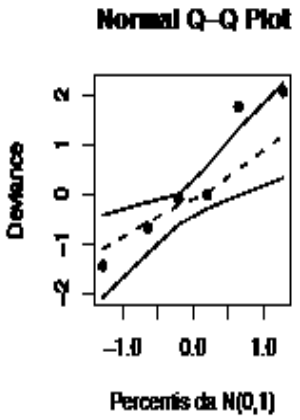
- Bioensaio conduzido em laboratório por Machado (2006).
- Objetivo: concentração ideal de uma suspensão viral.

Diluições (CPI/ml)	Mortes		Totais
	Sim	Não	
Testemunha	0	30	30
10^3	1	29	30
10^6	4	26	30
10^7	15	15	30
10^8	28	2	30
10^9	29	1	30

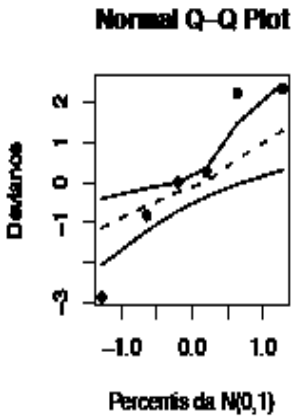
- $x_i =$ logaritmo neperiano das diluições.

Estatística *deviance* de qualidade de ajuste

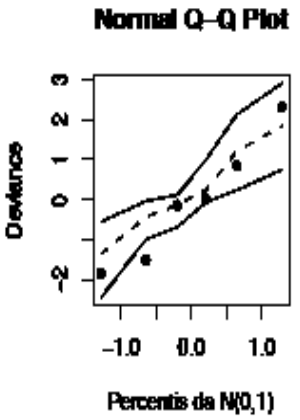
	Logito	Probito	Clog-log	Cauchy
Q_L	6,59	10,99	6,18	1,72
p-valor	0,158	0,027	0,186	0,787



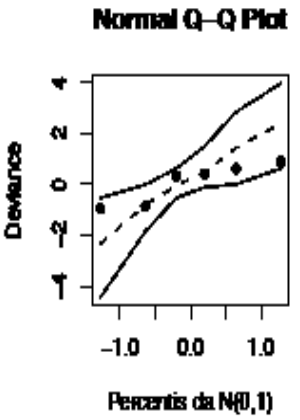
a) logito



b) probito



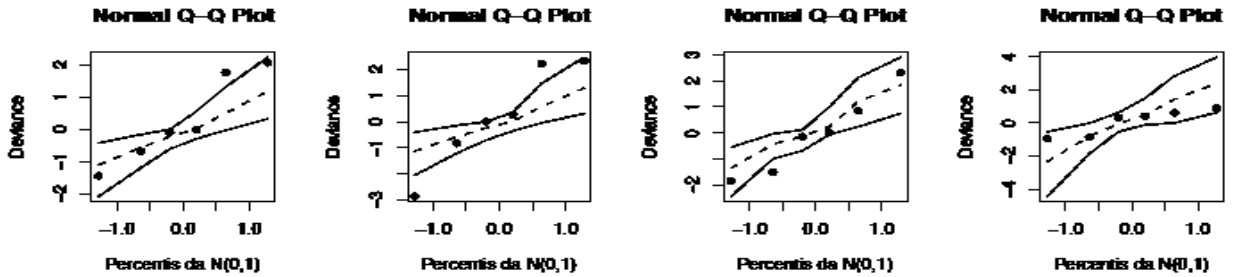
c) clog-log



d) Cauchy

Estatística *deviance* de qualidade de ajuste

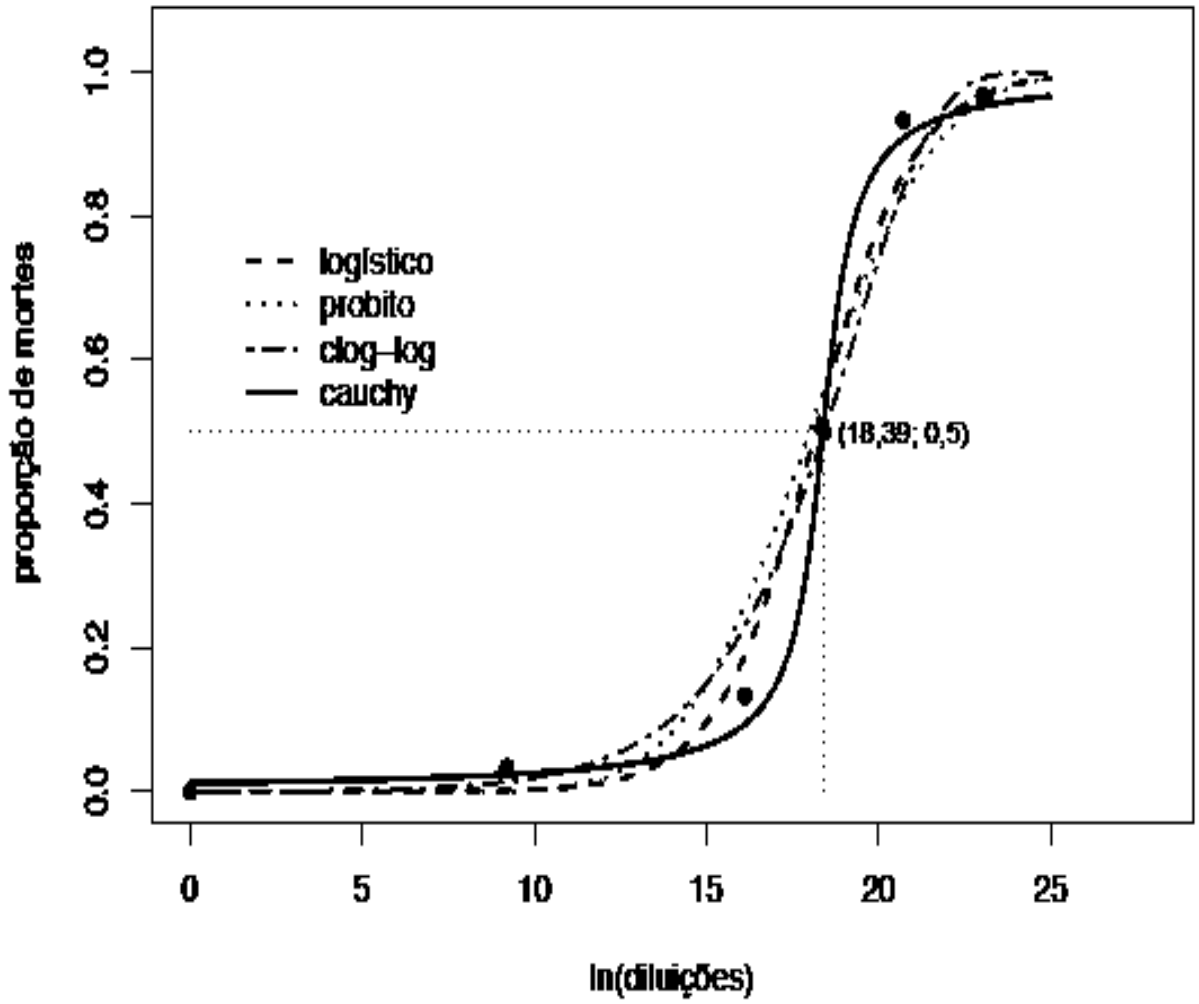
	Logito	Probito	Clog-log	Cauchy
Q_L	6,59	10,99	6,18	1,72
p-valor	0,158	0,027	0,186	0,787



a) logito b) probito c) clog-log d) Cauchy

		$x_{50} = \ln(\widehat{LD}_{50})$	$\widehat{LD}_{50}$
Logito	$\ln\left(\frac{0,50}{1-0,50}\right) = \widehat{\beta}_0 + \widehat{\beta}_1 x$	$-\frac{\widehat{\beta}_0}{\widehat{\beta}_1} \approx 18,17$	$(7,7)^7$
Probito	$\Phi^{-1}(0,50) = \widehat{\beta}_0 + \widehat{\beta}_1 x$	$-\frac{\widehat{\beta}_0}{\widehat{\beta}_1} \approx 18,00$	$(6,6)^7$
Clog-log	$\ln(-\ln(1-0,50)) = \widehat{\beta}_0 + \widehat{\beta}_1 x$	$\frac{-0,3665-\widehat{\beta}_0}{\widehat{\beta}_1} \approx 18,43$	$(10)^7$
Cauchy	$F^{-1}(0,50) = \widehat{\beta}_0 + \widehat{\beta}_1 x$	$-\frac{\widehat{\beta}_0}{\widehat{\beta}_1} \approx 18,39$	$(9,7)^7$

Obs: para os modelos com *links* simétricos $\Rightarrow x_{50} = -\frac{\widehat{\beta}_0}{\widehat{\beta}_1}$.


proporção de mortes
0.0
0.2
0.4
0.6
0.8
1.0
logístico
probito
clog–log
cauchy
(18,39; 0,5)
0
5
10
15
20
25
ln(diluições)

```
resim<-c(0,1,4,15,28,29)
resnao<-c(30,29,26,15,2,1)
lnd<-c(0,9.21,16.12,18.42,20.72,23.02)
dados<-as.data.frame(cbind(resim,resnao,lnd))
attach(dados)

ajuste4<-glm(as.matrix(dados[,c(1,2)])~lnd,
          family=binomial(link="cauchit"),data=dados)
ajuste4
anova(ajuste4,test="Chisq")
summary(ajuste4)
ntot<-c(30,30,30,30,30,30)
fit.model<-ajuste4
source("http://www.ime.usp.br/~giapaula/envelr_bino")

x<-seq(0,25,0.1)
m4<- pcauchy(-26.678+1.451*x)

plot(lnd,resim/(resim+resnao),pch=16, ylab="proporção de
mortes",xlab="ln(diluições)",xlim=c(0,28),ylim=c(0,1.05))
lines(x,m4,lty=1,lwd=2,col=1)
legend(1,0.8,lty=c(1),col=c(1),lwd=2,c("cauchy"),bty="n")

lines(c(18.386,18.386),c(0,0.50),lty=3)
lines(c(0,18.386),c(0.50,0.50),lty=3)
legend(17.7,0.55,c("(18.386, 0.5)"),bty="n",cex=0.8)
```