Introdução aos Modelos Lineares Generalizados Com Aplicações

Guatemala

Prof. Dr. Tiago Almeida de Oliveira

Departamento de Estatística
Universidade Estadual da Paraíba

Sumário

Modelagem Estatística

Modelagem estatística é um processo de descobrimento.

- O que é um modelo estatístico?

Modelo Estatístico

Modelagem Estatística

Modelagem estatística é um processo de descobrimento.

- O que é um modelo estatístico?

Modelo Estatístico

É a equação que descreve o fenômeno mais flutuações devido ao acaso

Modelagem Estatística

Modelagem estatística é um processo de descobrimento.

- O que é um modelo estatístico?

Modelo Estatístico

É a equação que descreve o fenômeno mais flutuações devido ao acaso

(Modelo estatístico = Modelo matemático + incerteza)

Introdução

Modelagem

- Modelo é uma versão simplificada de alguns aspectos do mundo real;

Introdução

Modelagem

- Modelo é uma versão simplificada de alguns aspectos do mundo real;
- Podemos dizer que modelo é uma representação em pequena escala de entidades físicas;

Introdução

Modelagem

- Modelo é uma versão simplificada de alguns aspectos do mundo real;
- Podemos dizer que modelo é uma representação em pequena escala de entidades físicas;
- A construção de modelos implica numa compreensão dos dados Dados disponíveis que são um subconjunto dos dados que poderiam ser coletados;

Introdução

Modelagem

- Modelo é uma versão simplificada de alguns aspectos do mundo real;
- Podemos dizer que modelo é uma representação em pequena escala de entidades físicas;
- A construção de modelos implica numa compreensão dos dados Dados disponíveis que são um subconjunto dos dados que poderiam ser coletados;
- O modelo serve para obter inferências para um grupo maior ou para obter compreensão do mecanismo (sistema) gerador dos dados observados;

Introdução

Modelagem

- Modelo é uma versão simplificada de alguns aspectos do mundo real;
- Podemos dizer que modelo é uma representação em pequena escala de entidades físicas;
- A construção de modelos implica numa compreensão dos dados Dados disponíveis que são um subconjunto dos dados que poderiam ser coletados;
- O modelo serve para obter inferências para um grupo maior ou para obter compreensão do mecanismo (sistema) gerador dos dados observados;
- Os modelos variam de acordo com a acurácia da sua representação;

Introdução

Modelagem

- Modelo é uma versão simplificada de alguns aspectos do mundo real;
- Podemos dizer que modelo é uma representação em pequena escala de entidades físicas;
- A construção de modelos implica numa compreensão dos dados Dados disponíveis que são um subconjunto dos dados que poderiam ser coletados;
- O modelo serve para obter inferências para um grupo maior ou para obter compreensão do mecanismo (sistema) gerador dos dados observados;
- Os modelos variam de acordo com a acurácia da sua representação;
- O ponto chave da modelagem está nesta acurácia que varia de acordo com o objetivo da análise.

Tipos de Modelos

Objetivos de um modelo Modelo Explicativo ou Descritivo

- Estudar a associação entre fatores de risco e desfecho (outcome). Exemplos:
- Avaliar a magnitude de associação de uma exposição e um desfecho ajustada pelo efeitos de possíveis fatores de confundimento ou de interação
- Investigar fatores determinantes de uma doença em plantas, ex, avaliar o efeito de um determinado fator de risco na ocorrência de uma doença controlado por fatores de confundimento e considerando possíveis fatores modificadores de efeito da associação principal em questão
- Acurácia do modelo não precisa ser perfeita

Tipos de Modelos - continuação

Objetivos de um modelo Modelo Preditivo

- Modelo em que o objetivo central é fazer predição do desfecho. Exemplos:
- Predição de um desfecho para ajudar na tomada de decisão de um tratamento
- Desenvolvimento de classificação de doença ou estagiamento (elaboração de um score)
- Identificação de fatores biológicos que podem ajudar elucidar a patologia da doença
- Acurácia do modelo é importante

Construção de um modelo

Passos envolvidos na construção de um modelo estatístico

- Formulação dos modelos
- Especificar uma expressão matemática para descrever o comportamento geral de acordo com as crenças do analista/investigador. Esta expressão também é conhecida como componente sistemático do modelo.
- Incorporar, na parte sistemática do modelo, uma certa quantidade de flutuações da variável resposta, denominada componente aleatório do modelo
- Especificar como combinar os componentes sistemático e aleatório

Construção de um modelo - continuação

Passos envolvidos no desenvolvimento de um modelo estatístico - continuação

- Inferência dos parâmetros do modelo (estimação e testes de hipóteses)
- Avaliação dos modelos
- avaliar premissas dos modelos
- avaliar o ajuste global do modelo que poderá depender do objetivo do modelo
- Reformulação (se necessário)

Motivação 1: variável resposta dicotômica

Hosmer e Lemeshow, em Applied Logistic Regression (Wiley, 1989), porém adaptado aqui para o contexto de agronomia.

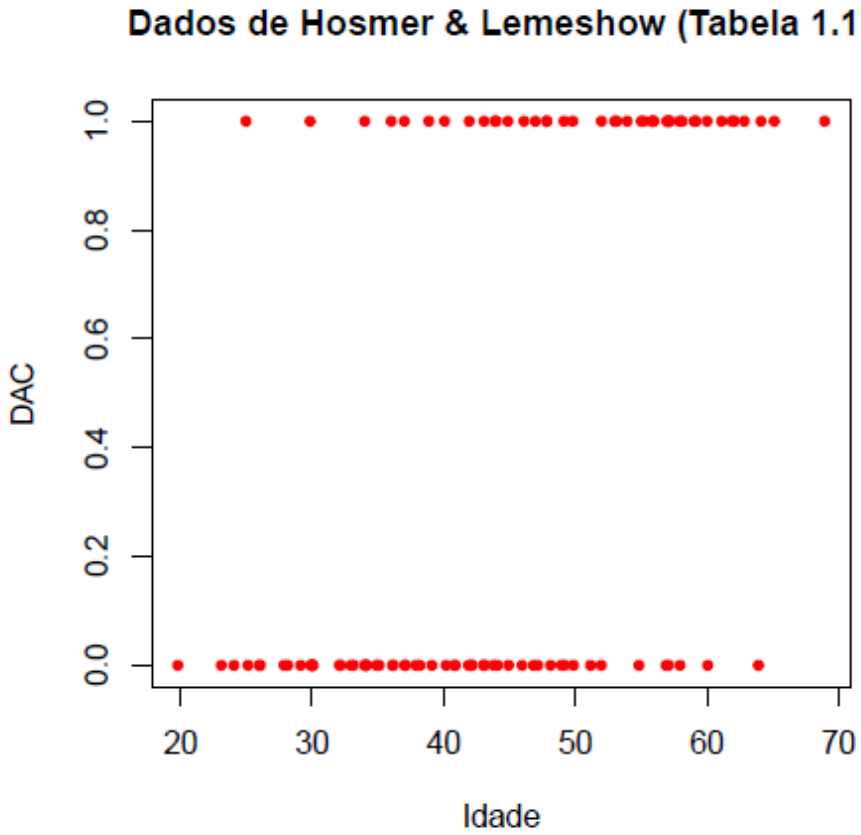
- Se têm dados sobre n = 100 observações, com variáveis:
- idade – numérica;
- doença em plantas – variável dicotômica (sim/não; 1/0).

Exemplo motivador: variável resposta dicotômica

Eis os primeiros seis valores da *data frame* correspondente:

```
head(HosLem.tudo)
Idade DAP
1 20 0
2 23 0
3 24 0
4 25 0
5 25 1
6 26 0
```

Quer-se relacionar a existência de DAP (variável resposta Y) com a idade (preditor X). O gráfico Y vs. X é pouco animador.



NOTA: Foi usado o comando `jitter` na variável idade:

```
> plot(DAP ~ jitter(Idade), data=HosLem.tudo, cex=0.8, col="red",pc
+ xlab="Idade",main="Dados de Hosmer \& Lemeshow (Tabela 1.1)")
```

Estudar a tendência do número de mortes

Objetivo: estudar a tendência no número de mortes (y_i) na Austrália a cada três meses de 1983 à 1986 (t_i) - Adaptado Andreozzi

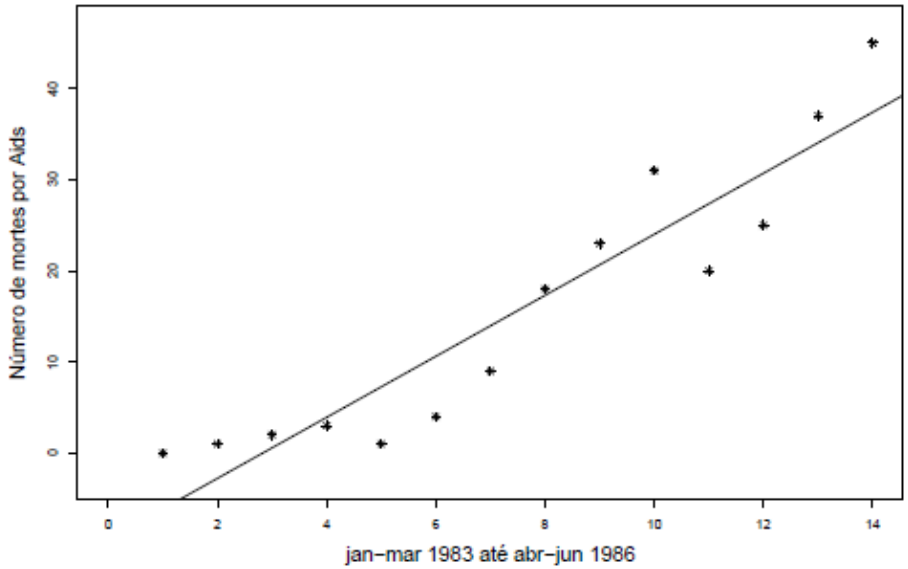


Figura: Fonte: Andreozzi

A Reta de regressão, parece razoável mas fornece valores esperados negativos para os períodos 1 e 2, assim:

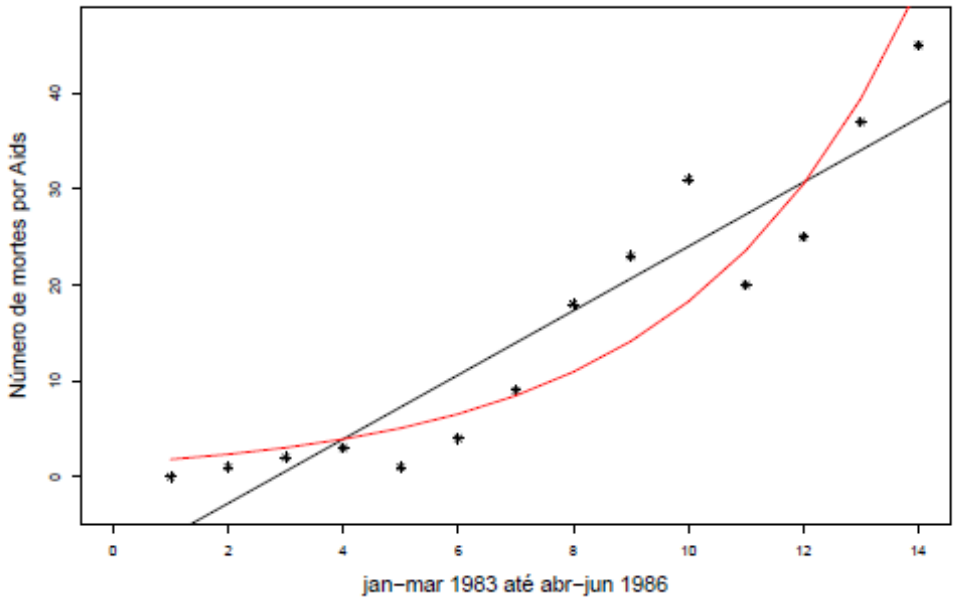


Figura: Fonte: Andreozzi

$$Y_i \sim Poi(\mu_i)$$
$$E(Y_i) = Var(Y_i) = \mu_i$$
$$ln(\mu_i) = \beta_0 + \beta_1 t_i$$

O Modelo Linear

Podemos recordar que o modelo linear relaciona

- um variável resposta numérica Y com
- preditores $X_1, X_2, ..., X_p$ (numéricos e/ou fatores)

por meio da equação, para n observações independentes Y_i:

$$Y_i = \beta_0 + \beta_1 x_1(i) + \beta_2 x_2(i) + ... + \beta_p x_p(i) + \epsilon_i$$

com $\epsilon_i \cap N(0, \sigma^2)(i = 1, 2, ..., n)$.

isto é, tal que $E[Y_i|X_1 = x_1(i), ..., X_p = x_p(i)]$ é dada por:

$$E[Y_i] = \beta_0 + \beta_1 x_1(i) + \beta_2 x_2(i) + ... + \beta_p x_p(i),$$

Y_i independentes, com distribuição Normal.

A generalização do modelo linear

Modelo Linear

- $E[Y_i] = \beta_0 + \beta_1 x_1(i) + \beta_2 x_2(i) + ... + \beta_p x_p(i)$,
- Y_i com distribuição Normal.

Modelo Linear Generalizado (MLG)

- $g(E[Y_i]) = \beta_0 + \beta_1 x_1(i) + \beta_2 x_2(i) + ... + \beta_p x_p(i)$, com g uma função invertível chamada função de ligação.
- Y_i com distribuição pertente a família exponencial de distribuições.

Assim, um MLG modela o valor esperado de uma variável resposta com distribuição na família exponencial, através da equação:

$$\mu_i = E[Y_i] = g^{-1}(\beta_0 + \beta_1 x_1(i) + \beta_2 x_2(i) + ... + \beta_p x_p(i)).$$

`Nota`: O Modelo Linear é caso particular de MLG: a Normal pertence à familia exponencial de distribuições e a função de ligação é a identidade: $g(x) = x, \forall$ x.

Modelo Linear Generalizado (MLG)

- Teoria unificadora de modelos lineares para variáveis resposta contínua e discreta introduzida por Nelder e Wedderburn em 1972;
- Modela o valor esperado da variável resposta;
- É considerado uma extensão do modelo linear clássico;
- Extensão da distribuição considerada e da função que relaciona o valor esperado e as covariáveis;
- Distribuição da variável resposta Família exponencial (Normal, Binomial, Bernoulli, Poisson, Exponencial, Gama, Binomial Negativa, Multinomial).

Os MLG são caracterizados pela seguinte estrutura:

1. Componente Aleatório
2. Componente Sistemático (ou estrutural)
3. Função de ligação

As três componentes de um MLG

Na definição de McCullagh e Nelder (1989), um modelo linear generalizado é definido sobre três componentes fundamentais:

Primeiro Componente

1. A variável resposta que se pretende modelar, trata-se de uma:

- variável aleatória;
- da qual se recolhem n observações independentes; e
- cuja distribuição de probabilidade faz parte da família exponencial de distribuições (FE);

Nota: a distribuição de probabilidades da variável - resposta aleatória Y não se restringe a Normal, podendo ser qualquer uma pertecente a FE.

As três componentes de um MLG - cont.

2. Componente Sistemática: consiste em uma combinação linear de variáveis preditoras.

Havendo p variáveis preditoras e n observações:

$$\beta_0 + \beta_1 x_1(i) + \beta_2 x_2(i) + ... + \beta_p x_p(i), \qquad \forall i \in \{1, ..., n\}.$$

as variáveis preditoras podem ser numéricas ou fatores.

Define-se a matriz do modelo identicamente ao modelo $X_{n\times(p+1)}$: a primeira coluna de 1's (associada a constante aditiva) e p colunas representando as variáveis explicativas.

As três componentes de um MLG - cont.

Forma Matricial

$$X = \begin{bmatrix} 1 & x_1(1) & x_2(1) & \cdots & x_p(1) \\ 1 & x_1(2) & x_2(2) & \cdots & x_p(2) \\ 1 & x_1(3) & x_2(3) & \cdots & x_p(3) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_1(n) & x_2(n) & \cdots & x_p(n) \end{bmatrix}$$

Componente sistemática do modelo é dada por:

$$\vec{\eta} = X\vec{\beta},$$

sendo $\beta = (\beta_0, \beta_1, \beta_2, \cdots, \beta_p)$ o vetor de coeficientes que define as n combinações lineares das variáveis preditoras, dadas em $\vec{\eta}$.

As três componentes de um MLG - cont.

3. Função de ligação: Uma função de ligação diferenciável e monótona g que associa as componentes aleatória e sistemática:

$$\begin{aligned} g(\mu) &= g(E([\vec{Y}]) = X\beta \\ \Leftrightarrow g(\mu_i) &= g(E[Y_i]) = \vec{\mathbf{x}}_{[i]}^t\beta \\ &= \beta_0\beta_1 x_1(i) + \beta_2 x_2(i) +, \cdots, +\beta_p x_p(i) (\forall i = 1:n) \end{aligned}$$

sendo:

- $\vec{Y}$ o vetor com as n observações $\{Y_i\}_{i=1}^n$.
- $\mu = E[\vec{Y}] = (\mu_1, \mu_2, ..., \mu_n)^t$ o vetor esperado de $\vec{Y}$;
- $x_{[i]}^{\vec{t}}$ a i-ésima linha da matriz $\mathbf{X}$, com os valores dos preditores na i-ésima observação da variável resposta.

As três componentes de um MLG (cont.)

Ou seja, e nas palavras de Agresti (1990, p.81):

Um MLG é um modelo linear para uma transformação da esperança um variável aleatória cuja distribuição pertence à família exponencial.

Nota: ao contrário do Modelo Linear, aqui não são explicitados erros aleatórios aditivos. A flutuação aleatória da variável-resposta é dada diretamente pela sua distribuição de probabilidades.

Caso a função g seja invertível (o que sucede se a monotonia acima exigida for estrita), pode se escrever:

$$g(\mu) = g(E[\vec{Y}]) = X\beta \quad \Leftrightarrow \quad \mu = E[\vec{Y}] = g^{-1}(X\beta)$$

$$g(\mu_i) = \vec{\mathbf{x}}_{[i]}^t\beta = \beta_0 + \sum_{j=1}^{p}\beta_j x_j(i) \quad \Leftrightarrow \quad \mu_i = g^{-1}(\vec{\mathbf{x}}_{[i]}^t\beta) = g^{-1}\left(\beta_0 + \sum_{j=1}^{p}\beta_j x_j(i)\right)$$

A família exponencial de distribuições inclui, entre outras:

- a Normal;
- a Poisson (para variáveis de contagem);
- a Bernoulli (para variáveis dicotomicas);
- a “Binomial/n"(para proporções de exitos em n provas de Bernoulli);
- a Gama (distribuição contínua assimétrica), inclui a Exponencial como caso particular;
- a Gaussiana inversa (distribuição contínua assimétrica).

A família exponencial de distribuições

Diz-se que uma variável aleatória (componente sistemático) Y tem distribuição na família exponencial (bi-paramétrica) usada por McCullagh & Nelder (1989), se a sua função densidade (caso Y contínua) ou de massa probabilística (se Y discreta) se for escrita na forma:

$$f(y|\theta,\phi) = \exp\left\{a(\phi)^{-1}\left(y\theta - b(\theta)\right)\right\} + c(y,\phi)$$

em que,

- θ e ϕ são parâmetros (escalares reais); e
- $a(\cdot), b(\cdot)$ e $c(\cdot)$ são funções reais conhecidas.

os parâmetros designam-se:

- θ - parâmetro natural;e
- ϕ - parâmetro de dispersão.

A Normal

A familía exponencial inclui a distribuição Normal:

$$f(y|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} = e^{\frac{y\mu - \frac{\mu^2}{2}}{\sigma^2}} + ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{y^2}{2\sigma^2}$$

é da forma indicada, com:

- $theta = \mu$
-
- $b(\theta) = \frac{\theta^2}{2} = \frac{\mu^2}{2}$
- $a(\phi) = \phi = \sigma^2$
- $c(y,\phi) = ln\left(\frac{1}{\sqrt{2\pi\phi}}\right) - \frac{y^2}{2\phi} = ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{y^2}{2\sigma^2}$

A Poisson

Uma variável aleatória discreta tem distribuição de Poisson se toma valores em N_0 com função masssa de probabilidade

$$p[y = k] = \frac{\lambda^k}{k!} e^{-\lambda}$$

Para os valores y in {0,1,2,...}, podemos escrever a função de massa de probabilidade de uma Poisson como:

$$f(y|\lambda) = e^{-\lambda} \frac{lambda^y}{y!} = e^{-lambda + yln(\lambda) - ln(y!)} \quad (1)$$

que é da família exponencial com:

- $\theta = ln(\lambda)$
- $\phi = 1$
- $b(\theta) = e^{\theta} = \lambda$
- $a(\phi) = 1$
- $c(y, \phi) = -ln(y!)$

A Bernoulli

A variável aleatória dicotomica − ou seja, binária − Y segue distribuição Bernoulli com parâmetro p, se toma valor 1 com probabilidade p e valor 0 com probabilidade $1 - p$.

Para valores $y = 0$ ou $y = 1$, a função de massa de probabilidade de uma distribuição Bernoulli pode ser escrita como:

$$f(y|p) = p^y(1-p)^{1-y} = e^{ln(1-p)+yln\left(\frac{p}{1-p}\right)}$$

que é da familia exponencial com:

- $\theta = ln\left(\frac{p}{1-p}\right)$
- $\phi = 1$
- $b(\theta) = ln(1 + e^{\theta}) = -ln(1-p)$
- $a(\phi) = 1$
- $c(y, \phi) = 0$

A Binomial

A Distribuição Binomial não pertence à família exponencial de distribuições. Mas se $X \cap B(n,p)$, então $Y = \frac{1}{n}X$ pertence à família exponencial.

Tem-se que $P[Y = y] = P[X = ny]$. A função de massa de probabilidade de Y pode ser escrita da seguinte forma, para $y \in F = \{0, \frac{1}{n}, \frac{2}{n}, ..., 1\}$:

$$f(y|p) = p^{ny}(1-p)^{n(1-y)} = e^{\frac{yln\left(\frac{p}{1-p}\right)+ln(1-p)}{\frac{1}{n}}+ln[bin]}$$

que é da familia exponencial com:

- $\theta = ln\left(\frac{p}{1-p}\right)$
- $\phi = \frac{1}{n}$
- $b(\theta) = ln(1+e^{\theta}) = -ln(1-p)$
- $a(\phi) = \frac{1}{n}$
- $c(y,\phi) = ln$

A Gama

Uma variável aleatória Y tem distribuição Gama com parâmetros μ e ν se toma valores em R^+, com função densidade da forma:

$$f(y|\mu,\nu) = \frac{\nu^\nu}{\mu^\nu \Gamma(\nu)} y^{\nu-1} e^{-\frac{\nu y}{\mu}} = e^{\frac{\left(-\frac{1}{\mu}\right)y + ln\left(\frac{1}{\mu}\right)}{\frac{1}{\nu}} + \nu ln\nu - ln\Gamma(\nu) + (\nu-1)lny}$$

que é da família exponencial com:

- $\theta = -\frac{1}{\mu}$
- $\phi = \frac{1}{\nu}$
- $b(\theta) = -ln\left(\frac{1}{\mu}\right) = -ln(-\theta)$
- $a(\phi) = \phi = \frac{1}{\nu}$
- $c(y,\phi) = \nu ln\nu - ln\Gamma(\nu) + (\nu-1)lny$

A família das distribuição Gama inclui como caso particular a distribuição Qui-Quadrado (χ^2_n se $\nu = \frac{n}{2}$ e $\mu = \nu$) e também a distribuição Exponencial ($\nu = 1$).

Funções de Ligação

A mais simples é a ligação identidade: $g(\mu) = \mu$.
Essa é a função de ligação utilizada no modelo linear.
As mais importantes funções de ligação tornam , para cada distribuição da família exponencial, o valor esperado da variável-resposta igual ao parâmetro natural, θ.
Em um modelo linear generalizado, a função $g(\cdot)$, diz-se uma função canônica para a variável resposta Y, se $g(E[Y]) = \theta$. Existe uma função de ligação canônica associada a cada distribuição da variável resposta.
As funções de ligação canônica são úteis porque simplificam de forma assinalável o estudo do Modelo. A ligação canônica representa de alguma forma uma função de ligação “natural"para o respectivo tipo de distribuição da variável-resposta.

Algumas funções de ligação canônicas

Distribuição	$E[Y]$	Ligação canónica	ϕ	$V[Y]$
Normal	μ	Identidade: $g(\mu) = \mu = \Theta$	σ^2	ϕ
Poisson	λ	Log: $g(\lambda) = \ln(\lambda) = \Theta$	1	λ
Bernoulli	p	Logit: $g(p) = \ln\left(\frac{p}{1-p}\right) = \Theta$	1	$p(1-p)$
Binomial/n	p	Logit: $g(p) = \ln\left(\frac{p}{1-p}\right) = \Theta$	$\frac{1}{n}$	$\frac{p(1-p)}{n}$
Gama	μ	Recíproco: $g(\mu) = \frac{1}{\mu} = -\Theta$	$\frac{1}{\nu}$	$\frac{\mu^2}{\nu}$

O Modelo Linear como um MLG

Ei alguns exemplos de MLGs:

1. O Modelo Linear.

O modelo linear é um caso particular de MLG, em que:

- cada uma das n observações da variável resposta Y tem distribuição Normal, com variância constante σ^2;
- a função de ligação é a função identidade.

A função de ligação identidade é a ligação canônica para a distribuição Normal.

MLGs para variáveis respostas dicotômicas

Considera-se um Modelo com variável resposta dicotômica (binária), ex., que apenas toma dois possíveis valores: 0 e 1, e cuja distribuição é Bernoulli, com probabilidades p (para 1) e $1 - p$ (para 0).
Adimite-se que o parâmetro p varia nas n observações de Y, e o valor esperado da i-ésima observação de Y é dado por:

$$E[y_i] = 1 \cdot p_i + 0 \cdot (1 - p_i) = p_i$$

Uma função de ligação vai relacionar este valores esperado p_i da variável-resposta com uma combinação linear dos preditores:

$$g(p(\vec{x})) = \vec{x}^t \beta \Leftrightarrow p(\vec{x}) = g^{-1}\left(\vec{x}^t \beta\right) \tag{2}$$

A Regressão Logística

1. A regressão Logística;

A função de ligação canônica transforma p no parâmetro natural θ da distribuição Bernoulli: $\theta = ln\left(\frac{p}{1-p}\right)$. Logo, a função de ligação canônica para a variável resposta de Bernoulli é a função *logit*:

$$g(p) = ln\left(fracp1 - p\right) \tag{3}$$

Com estas opções, o MLG é conhecido por Regressão Logística.
A função de ligação *logit* é o logaritmo do quociente entre a probabilidade de Y tomar o valor 1 (“êxito") e a probabilidade de tomar o valor 0 (“fracasso").
Esse quociente é conhecido na literatura anglo- saxônica por *odds ratio*.

A regressão Logística (cont.)

Consideremos que os *logits* dos valores esperados p_i são combinações lineares das variáveis preditoras $X_0, X_1, ..., X_p$. Concretamente, dado um conjunto $\vec{x} = (x_1, x_2, ..., x_p)$ de observações nas variáveis preditoras, tem-se:

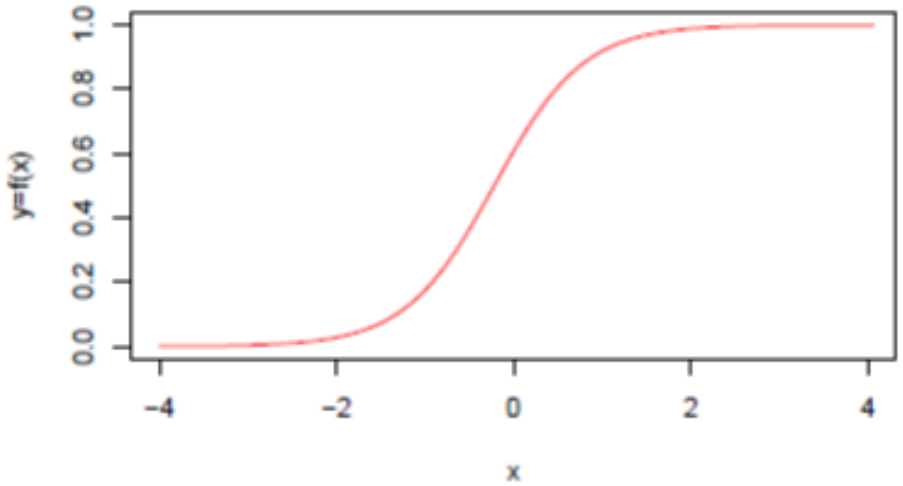
$$g(p) = ln\left(\frac{p}{1-p}\right) = \vec{x}\beta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Logo, a relação entre o valor esperado de Y_i (a probabilidade de êxito de Y) e o vetor de valores das variáveis preditoras, $\vec{x}$, é:

$$p(\vec{x}\beta) = g^{-1}(\vec{x}\beta) = \frac{1}{1+e^{-\vec{x}\beta}} = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p)}} \quad (4)$$

No caso de uma única variável preditora quantitativa, a relação entre Y e X é uma curva logística, que origina o nome Regressão Logística.

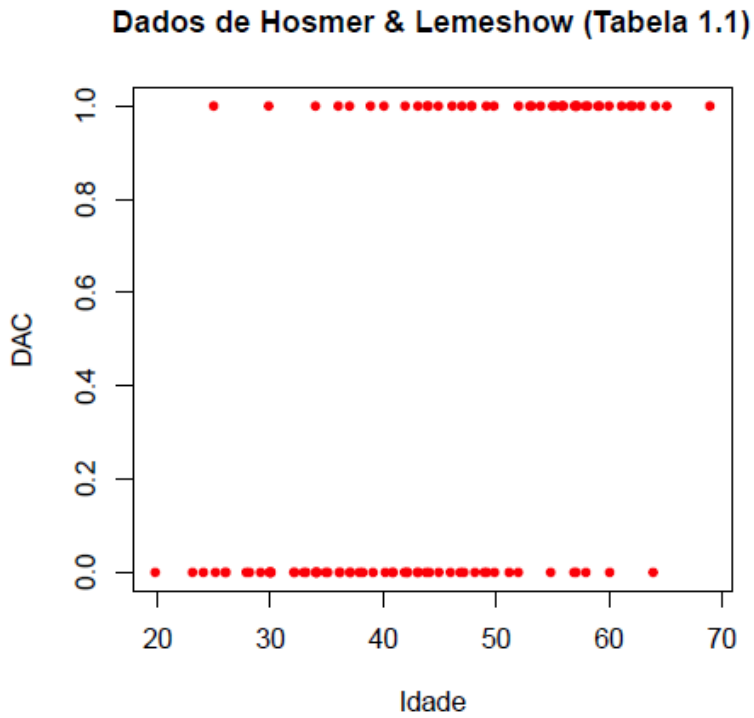
$$p(x) = g^{-1}(\beta_0 + \beta_1 x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \tag{5}$$



É uma função crescente, caso $\beta_1 > 0$, e descrescente caso $\beta_1 < 0$.
Quando há vários preditores, Y tem relação logística com a parte sistemática $\eta = \beta_0 + beta_1 x_1 +, \cdots, \beta_p x_p$.

Novamente o exemplo DAP - adaptado

- idade - numérica;
- doença - variável dicotômica (sim/não;1/0);



A variável resposta é dicotômica (binária): aplica-se uma regressão logística? Será preciso relacionar $p = E[Y]$, a probabilidade de ter doença arterial coronária, com a idade X.

Exemplo: A função de ligação

Para procurar uma função de ligação adequada, é necessário visualizar a relação entre idade e probabilidade de DAP.

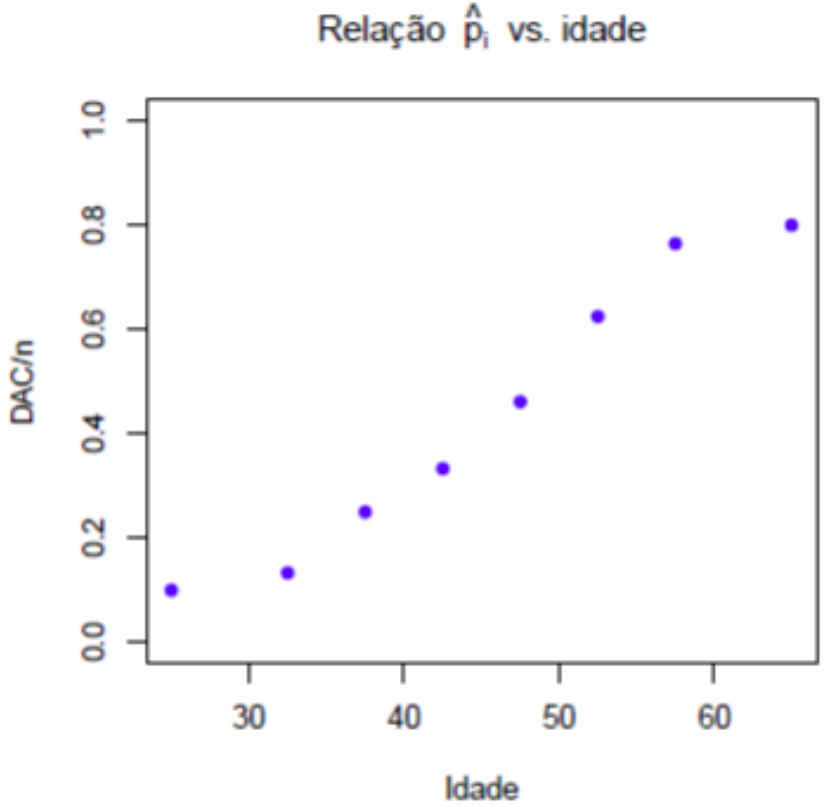
- Havendo repetições para cada idade, pode estimar-se p_i a partir da frequência relativa DAP na i-ésima idade;
- Havendo poucas repetições em cada idade, pode-se agrupar as observações em classes de idade.

Classe	n_i	DAC	$\hat{p}_i$
20-30-	10	1	0.100
30-35-	15	2	0.133
35-40-	12	3	0.250
40-45-	15	5	0.333
45-50-	13	6	0.462
50-55-	8	5	0.635
55-60-	17	13	0.765
60-70-	10	8	0.800

Exemplo: os dados tabelados

```
HosLem <-data.frame(Idade=c(25,32.5,37.5,42.5,47.5,52.5,57.5,65),
n=c(10,15,12,15,13,8,17,10),DAC=c(1,2,3,5,6,5,13,8))
rownames(HosLem) <- c("20-30-","30-35-","35-40-","40-45-","45-50-",
"50-55-","55-60-","60-70-")
HosLem
       Idade  n DAC
20-30-  25.0 10   1
30-35-  32.5 15   2
35-40-  37.5 12   3
40-45-  42.5 15   5
45-50-  47.5 13   6
50-55-  52.5  8   5
55-60-  57.5 17  13
60-70-  65.0 10   8

plot(DAC/n ~ Idade, data=HosLem, ylim=c(0,1),
+ main=expression(paste("Relação ",hat(p)[i], "vs. idade")),
 pch=16, col="blue")
```

Exemplo: $\hat{p}_i$ vs. idade



Temos uma relação sigmóide, talvez logística.

Resposta dicotômica e Binomial

A tabela anterior, resultante de agrupar as idades em classes, transformou a variável resposta Y_i, Bernoulli (1/0), em uma variável resposta Y_j que conta, em cada classe j, o número de “êxitos"(uns) nas n_j provas de Bernoulli dessa classe.

Para observações independentes, Y_j tem distribuição Binomial: $Y_j \cap B(n_j, p_j)$, em que p_j é a probabilidade de “êxito"na classe j.

Já vimos que a Binomial não pertence à família de distribuições exponenciais. Mas se $Y \cap B(n, p)$, então a distribuição da proporção de êxitos $W = \frac{1}{n}Y$ pertence à família exponencial.

A distribuição “Binomial/n"

Existem ligações íntimas, no contexto de MLGs, entre considerar que:

- temos n variáveis resposta Bernoulli, com parâmetros p_i; ou
- temos m variáveis resposta $Y_i \cap B(n_i, p_i)$.

O tratamento destas opções alternativas é igual, desde que transforme as Binomias Y_i em proporções de êxitos, i.e., desde que se considere novas v.a.s resposta $W_i = Y_i/n_i$, cujas distribuições pertencem à família exponencial de distribuições.

Bernoulli e “Binomial/n"podem ser vistas como essencialmente a mesma coisa, apresentada de forma diferente. A ligação canônica, que dar Bernoulli, quer da Binomial/n é a função logit:

$$g(p) = ln\left(\frac{p}{1-p}\right) \tag{6}$$

GLMs no R

No R, o comando crucial para o ajustamento de Modelos Lineares Generalizados é o comando glm.

Dos numerosos argumentos desta função, dois são cruciais:

formula: indica, de forma análoga à usada no modelo linear, qual a componente aleatória (à esquerda de um “ ") e quais os preditores (à direita, e separados por sinais de soma):

```
y~x1+x2+x3+...+xp
```

family: indica simultaneamente a distribuição de probabilidades da componente aleatória Y e a função de ligação do modelo.

A indicação da distribuição de probabilidades de Y faz-se por meio de uma palavra-chave, que se segue ao nome do argumento.
Por exemplo, um modelo com componente aleatória Bernoulli ou Binomial/n, indica-se assim:

```
family=binomial
```

Por omissão, é usada a função de ligação canônica dessa distribuição.
Caso se deseje outra função de ligação (implementada) acrescenta-se ao nome da distribuição, entre parentes, o argumento `link` com a especificação da função de ligação.
Por exemplo, um modelo probit pode ser indicado da seguinte forma:

```
family=binomial(link="probit")
```

Assim, ajusta-se um MLG no R invocando o comando glm com três argumentos:

```
glm(formula, family, data)
```

Em uma Regressão Logística,

- family=binomial. Não é necessário especificar a função de ligação: por omissão é usada a ligação canônica da distribuição especificada.
- podem usar-se dados em uma de 2 formas:
- observações dicotômicas individuais (como o *data frame* `Hoslem.tudo`;
- observações tabeladas para valores repetidos do(s) preditor(es)(como o *data frame* `Hoslem`).

As formulas para a Regressão Logística

As fórmulas do comando `glm` são semelhantes às do Modelo Linear:

```
y~x1+x2+,\cdots,xp
```

Mas em uma Regressão Logística, aos dois tipos de dados correspondem a objetos y de natureza diferente:

- Se dados contêm observações individuais, y é vetor de 0s e 1s:

```
glm(DAP~idade, family=binomial, data=HosLem.tudo)
```

- Se os dados tabelados, y deve ser uma matriz de duas colunas: uma com o número de “sim"s e outra com os número de “não"s, para cada valor do(s) preditor(es):

```
glm(cbind(DAP,n-DAP)~idade, family=binomial, data=HosLem)
```

Exemplo: ajustamento do modelo

Ajustar o modelo com base nas observações dicotômicas tabeladas:

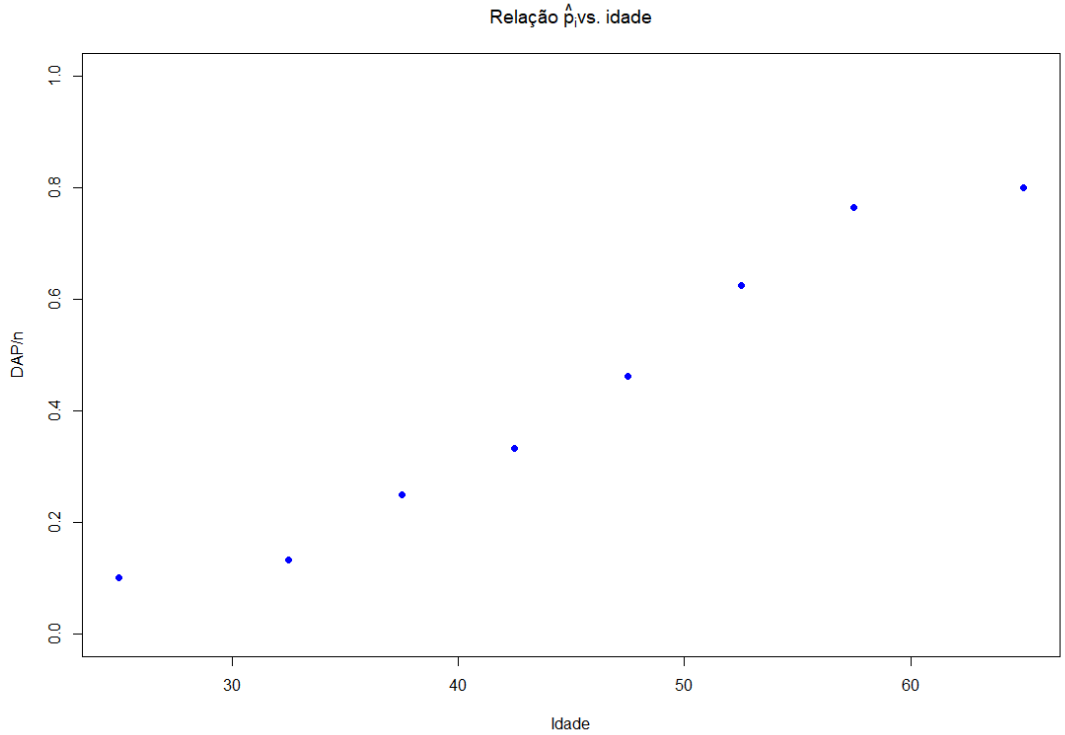
```
glm(cbind(DAP,n-DAP)~idade, family=binomial, data=HosLem)
```

```
Call:  glm(formula = cbind(DAC, n - DAC) ~ Idade, family = binomial,
data = HosLem)

Coefficients:
(Intercept)        Idade
-5.091             0.105

Degrees of Freedom: 7 Total (i.e. Null);  6 Residual
Null Deviance:     28.7
Residual Deviance: 0.5242  AIC: 25.66
```

Exemplo: ajustamento do modelo (cont.)

Sobrepondo a logística ajustada ao gráfico dos $\hat{p}_i$ vs. idade:

```
logistica <-function(b0,b1,x){
1/(1+exp(-(b0+b1*x)))
}
curve(logistica(b0=-5.3095, b1=0.1109, x), from=20, to=70,
col="blue", add=TRUE)
```



O resultado da função glm

Tal como o comando `lm`, também o comando `glm` produz uma `list`. Nas componentes dessa lista há informação sobre o ajustamento.

```
HosLem.glm <- glm(cbind(DAC,n-DAC) ~ Idade , family=binomial, data=
names(HosLem.glm)
```

```
[1] "coefficients"      "residuals"         "fitted.values"
[4] "effects"           "R"                 "rank"
[7] "qr"                "family"            "linear.predictors"
[10] "deviance"          "aic"               "null.deviance"
[13] "iter"              "weights"           "prior.weights"
[16] "df.residual"       "df.null"           "y"
[19] "converged"         "boundary"          "model"
[22] "call"              "formula"           "terms"
[25] "data"              "offset"            "control"
[28] "method"            "contrasts"         "xlevels"
```

Para aprofundar cada componente consultar: `help(glm)`
Para invocar uma componente usa-se a referenciação usual de listas:

A função coef

Tal como para os Modelos Lineares, existem funções para facilitar a extração de informação em um ajustamento de MLG. Algumas funções iniciais:
`coef` - devolve um vetor com os valores estimados dos parâmetros $\beta_0, \beta_1, \cdots, \beta_p$, ou seja, com os valores $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$:

```
HosLem.tudo.glm <- glm(DAC ~ idade, family=binomial, data=HosLem.tu
coef(HosLem.tudo.glm)
(Intercept)   idade
-5.3094534   0.1109211
```

A função predict

`predict` - por omissão, devolve vetor com os valores da combinação linear estimada dos preditores usados no ajustamento, ou seja, da componente sistemática $\hat{\beta}_0 + \hat{\beta}_1 x_1(i) + \cdots + \hat{\beta}_p x_p(i)$:

```
predict(HosLem.glm)
   20-30-      30-35-      35-40-      40-45-      45-50-      50-55-
-2.4652550 -1.6776115 -1.1525158 -0.6274202 -0.1023245  0.4227711
  55-60-      60-70-
0.9478668  1.7355102
```

A função predict (cont.)

Pode também estimar a combinação linear de valores não usados no ajustamento.
Os novos valores são dados numa data frame com nomes iguais aos usados nos dados originais:

```
> predict(HosLem.tudo.glm, newdata=data.frame(Idade=26))
     1
-2.425504
predict(HosLem.glm, newdata=data.frame(Idade=c(26,53,74)))
      1          2          3
-2.3602358 0.4752807 2.6806824
```

A função fitted

`fitted` – devolve um vector com os valores ajustados do valor esperado de Y_i , ou seja, de $\hat{p}_i = g^{-1}(\hat{\beta}_0 + \hat{\beta}_1 x_1(i) + \cdots + \hat{\beta}_p x_p(i))$.

```
fitted(HosLem.glm)
  20-30-     30-35-     35-40-     40-45-     45-50-     50-55-
0.07833012 0.15741201 0.24002985 0.34809573 0.47444116 0.60414616 0.
  60-70-
0.85011588
```

Um resultado análogo pode ser obtido através da função `predict`, indicando a opção `type=‘‘response"`:

```
predict(HosLem.glm, type="response")
  20-30-     30-35-     35-40-     40-45-     45-50-     50-55-
0.07833012 0.15741201 0.24002985 0.34809573 0.47444116 0.60414616 0.
  60-70-
0.85011588
```

Notas sobre a Regressão Logística

- A função logística tem boas propriedades para representar uma probabilidade: para qualquer valor da componente sistemática,

$$p(x_1, x_2, \cdots, x_p) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + , \cdots + \beta_p x_p)}} \tag{7}$$

toma valores entre 0 e 1. O mesmo não acontece com uma relação linear $p(x_1, \cdot, x_p) = \beta_0 + \beta_1 x_1 + \cdot + \beta_p x_p$, que pode tomar valores em toda a reta real R

- No caso de haver uma única variável preditora quantitativa, trocando os acontecimentos que dão à variável aleatória Y os valores 0 e 1, uma função decrescente para $p = P[Y = 1]$ transforma-se em uma função crescente.

Mais notas sobre a Regressão Logística

No caso de haver uma única variável preditora quantitativa, o parâmetro β_1 tem a seguinte interpretação:

- como,

$$\frac{p(x)}{1-p(x)} = e^{\beta_0} \cdot e^{\beta_1 x}, \tag{8}$$

cada aumento de uma unidade na variável preditora X traduz-se em um efeito multiplicativo sobre o *odds ratio*, de e^{β_1}:

$$\frac{p(x+1)}{1-p(x+1)} = e^{\beta_0} \cdot e^{\beta_1 (x+1)} = e^{\beta_0} \cdot e^{\beta_1 x} \cdot e^{\beta_1} = \frac{p(x)}{1-p(x)} \cdot e^{\beta_1}. \tag{9}$$

- o que é o mesmo que dizer que se traduz em um efeito aditivo, em β_1 unidades, sobre o *log-odds ratio*:

$$log\left[\frac{p(x+1)}{1-p(x+1)}\right] = log\left[\frac{p(x)}{1-p(x)}\right] + \beta_1. \tag{10}$$

Mais notas sobre a Regressão Logística

Quando há mais do que uma variável preditora quantitativa:

- a função de ligação *logit* gera uma relação logística para a probabilidade de êxito p, como função dos valores da parte sistemática η (combinação linear das variáveis preditoras).
- a interpretação dos coeficientes β_j generaliza-se: um aumento de uma unidade na variável preditora j (mantendo as restantes constantes) traduz-se numa multiplicação do *odds ratio* por um fator e^{β_j} .

Para preditores categóricos (fatores),

- seja ζ_j uma variável indicatriz. O correspondente parâmetro β_j indica o incremento no *log-odds ratio* resultante de uma observação passar a pertencer à categoria de que ζ_j é indicatriz.

A Regressão Logística (cont.)

O modelo de regressão logística é uma opção a considerar sempre que a variável-resposta Y assinala qual de duas categorias de classificação se verifica e se pretende relacionar a probabilidade do acontecimento associado ao valor 1 com um conjunto de variáveis preditoras.

A função logística revela rigidez estrutural, com um ponto de inflexão associado à probabilidade $p = 0.5$, em torno do qual há simetria da curva.

A função de ligação g pode ser substituída por outras funções, que já não serão funções de ligação canónicas para uma distribuição Bernoulli. Nesse caso, já não se fala em regressão logística.

Estimação de parâmetros em MLGs

A estimação de parâmetros em Modelos Lineares Generalizados é feita pelo Método da Máxima Verosimilhança. A estimação incide em primeiro lugar sobre os parâmetros β_j da parte sistemática do modelo.

O fato da estimação se basear na função verosimilhança significa que, ao contrário do que acontece com o Modelo Linear, em GLMs as hipóteses distribucionais são cruciais para a estimação dos parâmetros.

O fato das distribuições consideradas em MLGs pertencerem à família exponencial de distribuições gera algumas particularidades na estimação.

Verossimilhança na família exponencial

A função de verossimilhança para n observações independentes $y_1, y_2, \cdots, y_n$ em uma distribuição qualquer da família exponencial, é:

$$L(\theta, \phi; y_1, y_2, \cdots, y_n) = \prod_{i=1}^{n} f(y_i; \theta_i, \phi_i) = e^{\sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)} - c(y_i, \phi_i)}$$

Maximizar a verossimilhança é maximizar a log-verossimilhança $l(\theta, \phi; y_1, y_2, \cdots, y_n) = log(L(\theta, \phi; y_1, y_2, \cdots, y_n))$:

$$l(\theta, \phi; y_1, y_2, \cdots, y_n) = \sum_{i=1}^{n} \left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)} - c(y_i, \phi_i) \right]$$

Máxima Verossimilhança em MLGs

Um MLG, a componente sistemática e o valor esperado da variável resposta estão relacionados por $g(E[Y]) = \vec{x}^t\beta$. No caso de uma função de ligação canônica tem-se $\theta = \vec{x}^t\beta$.

Em geral, pode escrever-se a log-verossimilhança como função dos parâmetros desconhecidos β.

Estimar os parâmetros pelo método da máxima verossimilhança consiste em escolher o vetor β que torne máxima a função de log-verossimilhança $l(\beta)$.

Máxima Verossimilhança em MLGs (cont.)

A maximização da função de $p+1$ variáveis $l(\beta)$ tem como condição necessária:

$$\frac{\partial l(\beta)}{\partial \beta_j} = 0, \qquad \forall \quad j = 0 : p$$

Admite-se que as funções $a(\cdot), b(\cdot)$ e $c(\cdot)$ são suficientemente regulares para que as operações envolvidas estejam bem definidas.

No caso de um Modelo Linear Generalizado genérico, não existe a garantia de que haja máximo desta função log-verossimilhança(pelo menos para os valores admissíveis dos parâmetros β), nem que, existindo máximo, este seja único.

Nos casos concretos abordados, a situação será de máximo único.

Exemplo: o Caso da Regressão Logística

No Modelo de Regressão Logística, as n observações independentes referem-se a uma variável aleatória com distribuição Bernoulli.
A função de verossimilhança é dada por:

$$L(\vec{p}, \vec{y}) = \prod_{i=1}^{n} e^{ln(1-p_i)+y_i ln(\frac{p_i}{1-p_i})}$$

e a log-verossimilhança por:

$$l(\vec{p}, \vec{y}) = \sum_{i=1}^{n} \left(ln(1-p_i) + y_i ln(\frac{p_i}{1-p_i}) \right)$$

Uma vez que a função de ligação é dada por $g(p) = ln\left(\frac{p}{1-p}\right) = \vec{x}\beta$, tem-se a seguinte expressão para a log-verossimilhança como função dos parâmetros β:

$$l(\beta) = \sum_{i=1}^{n} \left(-ln\left(1 + e^{\vec{x}\beta}\right) + y_i \vec{x}\beta \right)$$

Estimação na Regressão Logística (cont.)

tem-se que:

$$l(\beta) = \sum_{i=1}^{n} \left(\beta_0 y_i + \sum_{k=1}^{p} y_i x_{k(i)} \beta_k \right) - \sum_{i=1}^{n} ln \left(1 + e^{\beta_0 + \sum_{k=1}^{p} x_{x(i)} \beta_k} \right)$$

Condição necessária para a existência de extremo da log-verossimilhança no ponto $\vec{\beta} = \vec{\hat{\beta}}$ é que:

$$\begin{cases} \frac{\partial l(\beta)}{\partial \beta_0} = & \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \frac{e^{\hat{\beta}_0 + \sum_{k=1}^{p} x_{k(i)} \hat{\beta}_k}}{1 + e^{\hat{\beta}_0 + \sum_{k=1}^{p} x_{k(i)} \hat{\beta}_k}} = 0 \\ \frac{\partial l(\beta)}{\partial \beta_j} = & \sum_{i=1}^{n} y_i x_{j(i)} - \sum_{i=1}^{n} \frac{e^{\hat{\beta}_0 + \sum_{k=1}^{p} x_{k(i)} \hat{\beta}_k}}{1 + e^{\hat{\beta}_0 + \sum_{k=1}^{p} x_{k(i)} \hat{\beta}_k}} \cdot x_{j(i)} = 0 \qquad \forall j = 1 : p \end{cases}$$

Estas $p + 1$ equações normais formam um sistema não linear de equações nas p+1 incógnitas $\hat{\beta}_j (j = 0, ..., p)$.

Estimação na Regressão Logística (cont.)

A não linearidade nos parâmetros β não permite explicitar uma solução $\hat{\beta}$ do sistema de equações.
Mas existe uma notação mnemónica, definindo o vetor $\vec{p}$ de probabilidades estimadas, cuja i-ésima componente é dada por:

$$\hat{p}_i = \frac{e^{\hat{\beta}_0 + \sum_{k=1}^{p} x_{k(i)} \hat{\beta}_k}}{1 + e^{\hat{\beta}_0 + \sum_{k=1}^{p} x_{k(i)} \hat{\beta}_k}}$$

e uma matriz $\mathbf{X}$ que tal como no Modelo Linear) tem uma primeira coluna de n uns e em cada uma de p colunas adicionais tem as n observações de uma das p variáveis preditoras. Com esta notação, o sistema de p+1 equações toma a forma:

$$X^t \vec{y} = X^t \vec{\hat{p}} \tag{11}$$

Sendo um sistema não linear, a sua solução exigirá métodos numéricos que serão considerados mais adiante.

Algoritmos de estimação

Foi visto que, em geral, o sistema $p+1$ equações normais associado à maximização da função de log-verossimilhança em um modelo linear generalizado é um sistema não linear:

$$\frac{\partial l(\beta)}{\partial \beta_j} = 0, \qquad j = 0 : p \tag{12}$$

Um algoritmo numérico de resolução utilizado no contexto de MLGs é uma modificação do algoritmo de Newton Raphson, conhecido por vários nomes: Método Iterativo de Mínimos Quadrados Ponderados (IWLS) ou (Re)ponderados (IRLS), ou ainda *Método de Fisher (Fisher Scoring Method*, em inglês)

O método de Newton-Raphson trabalha com uma aproximação de segunda ordem da função log-verossimilhança (fórmula de Taylor),c om desenvolvimento em torno de uma estimativa inicial do vetor β.

Algoritmos de estimação (cont.)

Designado por:

- $\beta^{[0]}$; a solução inicial para β;
- $\frac{\partial l(\beta)}{\partial \beta}$ o vetor gradiente de $l(\beta)$ calculado no ponto β;
- $\mathbf{H}_\beta$ a matriz Hessiana das segundas derivadas parciais da função $l(\cdot)$, nesse mesmo ponto,

tem-se a aproximação:

$$l(\beta) \propto l_0(\beta) = l\left(\beta^{[0]}\right) + \left(\frac{\partial l(\beta^{[0]})}{\partial \beta}\right)^t \left(\beta - \beta^{[0]}\right) + \frac{1}{2}\left(\beta - \beta^{[0]}\right)^t \mathbf{H}_\beta^{[0]} \left(\beta - \beta^{[0]}\right)$$

Em vez de maximizar $l(\beta)$, maximiza-se a aproximação $l_0(\beta)$.

Algoritmos de estimação (cont.)

O cálculo do vetor gradiente é simples para produtos internos ou formas quadráticas, como já se viu no estudo do modelo linear:

Se $h(\vec{x}) = \vec{a}^t\vec{x}$, tem-se $\frac{\partial h(\vec{x})}{\partial \vec{x}} = \frac{\partial(\vec{a}^t\vec{x})}{\partial \vec{x}} = \vec{a}$.

Se $h(\vec{x}) = \vec{x}^t A\vec{x}$, tem-se $\frac{\partial h(\vec{x})}{\partial \vec{x}} = \frac{\partial(\vec{x}^t A\vec{x})}{\partial \vec{x}} = 2A\vec{x}$. Assim,

$$\frac{\partial l_0\beta}{\partial \beta} = \frac{\partial l_0\beta^{[0]}}{\partial \beta} + H_\beta^{[0]}\left(\beta - \beta^{[0]}\right).$$

adimintindo a invertibilidade de $H_\beta^{[0]}$, tem-se:

$$\frac{\partial l_0(\beta)}{\partial \beta} = 0 \Leftrightarrow \beta = \beta^{[0]} - H_{\beta^{[0]}}^{-1}\left(\frac{\partial l\beta^{[0]}}{\partial \beta}\right)$$

O algoritmo Newton-Raphson itera esta relação.

Algoritmos de estimação (cont.)

tome-se:

$$\beta^{[i+1]} = \beta^{[i]} - H^{-1}_{\beta^{[i]}} \left(\frac{\partial l(\beta^{[i]})}{\partial \beta} \right)$$

Notas:

- A possibilidade de aplicar com êxito este algoritmo exige a existência e invertibilidade das matrizes Hessianas de l nos sucessivos pontos $\beta^{[i]}$;
- Não está garantida a convergência do algoritmo a partir de qualquer ponto inicial $\beta^{[0]}$, mesmo quando existe e é único o máximo da função de log-verossimilhança;
- Dada a existência e unicidade do máximo, a convergência é tanto melhor quanto mais próximo $\beta^{[0]}$estiver do máximo.

ALgoritmos de estimação (cont.)

O cálculo da matriz Hessiana da log-versossimilhança nos pontos $\beta^{[}i]$ é computacionamente exigente.

O algoritmo de Fisher é uma modificação do algoritmo de Newton-Raphson, que subistitui a matriz Hessiana pela matriz de informação de Fisher, definida como o simétrico da esperança da matriz Hessiana:

$$\iota_{\beta^{[i]}} = -E\left[H_{\beta^{[i]}}\right]$$

Assim, a iteração que está na base do Algoritmo de Fisher é:

$$\beta^{[i+1]} = \beta^{[i]} - \iota^{-1}_{\beta^{[i]}} \left(\frac{\partial l(\beta^{[i]})}{\partial \beta}\right)$$

Material Complementar: Algoritmos (cont.)

Quando se considera um MLG com a função de ligação canonica, a matriz Hessiana da log-verossimilhança não depende da variável resposta Y, pelo que a Hessiana e sou valor esperado coincidem.

Logo, neste caso os métodos de Fisher e Newton-Raphson coincidem.

Esta é uma das razões que confere às ligações canônicas a sua importância.

MC: Algoritmos de estimação (cont.)

O algoritmo de Fisher é também conhecido por Método Iterativo de Mínimos Quadrados Ponderados (IWS) ou (Re)ponderados (IRLS) porque é, em geral, possível (re)escrever a expressão anterior para $\beta^{[i+1]}$ na forma:

$$\beta^{[i+1]} = \left(X^t W^{[i]} X\right)^{-1} X^t W^{[i]} \vec{z}^{[i]}$$

em que:

- $\vec{z}^{[i]}$ é uma linearização da função de ligação g(y), escrita como função dos parâmetros β; e
- $W^{[i]}$ é uma matriz diagonal.

Para alguns modelos, as expressões concretas de $\vec{z}^{[i]} e W^{[i]}$ serão vistas adiante.

MC: Algoritmos de estimação (cont.)

A expressão anterior significa que o algoritmo de Fisher está associado a uma projeção não ortogonal, em que, quer o vetor $\vec{z}^{[i]}$, quer os subespaços envolvidos na projeção, são redefinidos em cada iteração do algoritmo.
A matriz $X(X^t W^{[}i]X)^{-1} X^{[t]} W^{[i]}$ é idempotente
Não é, em geral, simétrica, a não ser que a matriz diagonal
$W^{[i]} verifique X^t W^{[i]} = X^{[t]}$.
O método de Fisher baseia-se em ideias de Mínimos Quadrados em sentido generalizado, isto é, envolvendo projeções não ortogonais.

PULA

MC: IRLS para a Regressão Logística (cont.)

A matriz Hessiana da função de log-verossimilhança l, nos pontos correspondentes às iterações $\beta^{[i]}$, é constituída pelos valores destas derivadas parciais de segunda ordem .

Como acontece sempre quando se trabalha com Modelos que utilizam a função de ligação canônica, estes elementos das matrizes Hessianas não dependem dos valores observados da variável resposta Y, pelo que a Hessiana e o seu valor esperado coincidem (os Métodos de Newton-Raphson e de Fisher coincidem)

Defina-se a matriz $n \times n$ diagonal $\mathbf{W}$, cujos elementos diagonais são dados pelos n valores $p_i(1 - p_i)$

PULA

A regressão Probit

Outro exemplo de MLG é o modelo probit de Bliss (1935), muito frequente em Toxicologia.

Tal como na Regressão Logística, tem-se:

- variável resposta dicotômica (com distribuição Bernoulli).
- componente sistemática, dada por uma combinação linear de variáveis preditoras

Diferente da Regressão logística é a função de ligação.

Na regressão Logística, a função de ligação exprime p como uma função logística da componente sistemática $\eta = \vec{x}\beta$.

Aqui, escolhe-se uma outra relação sigmóide: a função de distribuição cumulativa (f.d.c) de uma Normal reduzida, Φ.

$$p(\vec{x}^t\beta) = g^{-1}\left(\vec{x}^t\beta\right)$$

em que, Φ indica a fdc de uma N(0,1)
Esta opção significa considerar como função de ligação a inversa da f.d.c em uma Normal reduzida, ou seja, $g = \Phi^{-1}$:

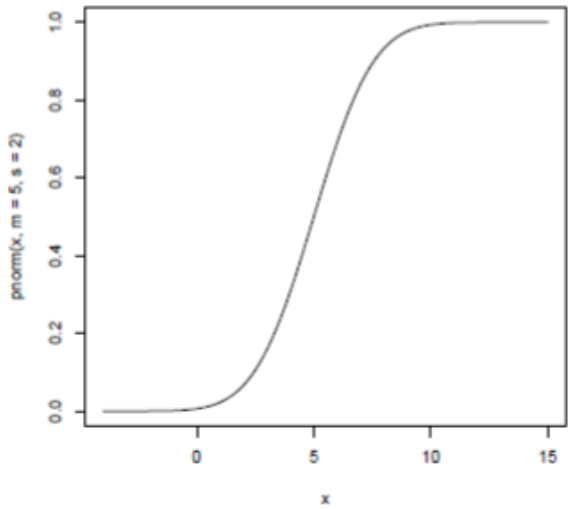
$$\vec{x}^t\beta = g\left(p(\vec{x}^t\beta)\right) = \Phi^{-1}\left(p(x^t\beta)\right)$$

A regressão Probit (cont.)

No caso de haver uma única variável preditora, tem-se:

$$p(x;\beta_0,\beta_1) = g^{-1}(\beta_0+\beta_1 x) = \Phi(\beta_0+\beta_1 x) = \Phi\left(\frac{X-\mu}{\sigma}\right),$$

em que, $\beta_0 = -\frac{\mu}{\sigma}$ e $\beta_1 = \frac{1}{\sigma}$. ex., a probabilidade de êxito p relaciona-se com a variável preditora X através da fdc de uma $N(\mu,\sigma^2)$, com $\sigma = \frac{1}{\beta_1}$ e $\mu = -\frac{\beta_0}{\beta_1}$.



A Regressão Probit (cont.)

Em geral, para qualquer número de variáveis preditoras, a probabilidade de êxito $p = P[Y = 1]$ é dada, no Modelo Probit, por uma função cujo comportamento é muito semelhante ao do Modelo Logit:

- função estritamente crescente,
- com um único ponto de inflexão quando o preditor linear $\vec{x}^t\beta = 0$,
- a que corresponde uma probabilidade de êxito $p(0) = 0.5$;
- com simetria em torno do ponto de inflexão, isto é, $p(-\eta) = 1 - p(\eta)$, para qualquer η.

Inconvenientes:

- não há interpretação fácil do significado dos parâmetros β_j;
- a função de ligação não é canônica.

A Regressão Probit em toxicologia

No contexto toxicológico, é frequente:

- existir uma variável preditora X que indica a dosagem (ou log-dosagem) de um determinado produto tóxico;
- para cada dosagem há um nível de tolerância t: o limiar acima do qual o produto tóxico provoca a morte do índivíduo;
- esse nível de tolerância varia entre indivíduos e pode ser representado por uma v.a. T.

Definindo a v.a. binária Y:

$$y = \begin{cases} 1, & \text{Indivíduo morre;} \\ 0, & \text{individuo sobrevive.} \end{cases}$$

A Regressão Probit em toxicologia (cont.)

Tem-se:

$$P[Y = 1|x] = P[T \leq x] = p(x)$$

Admitindo que a tolerância T segue uma distribuição $N(\mu, \sigma^2)$,

$$p(x) = \Phi\left(\frac{X - \mu}{\sigma}\right).$$

tem-se o Modelo Probit com X como única variável preditora.
Os coeficientes verificam $\beta_0 = -\frac{\mu}{\sigma}$ e $\beta_1 = \frac{1}{\sigma}$, estando pois associados aos parâmetros da distribuição de T.
Ilustremos a aplicação de uma Regressão Probit, no `R`, aos dados do exemplo adaptado de DAP, já considerado antes.

Regressão Probit no R

Em uma regressão probit, há que especificar a respectiva função de ligação, como opção do argumento `family`, da seguinte forma:

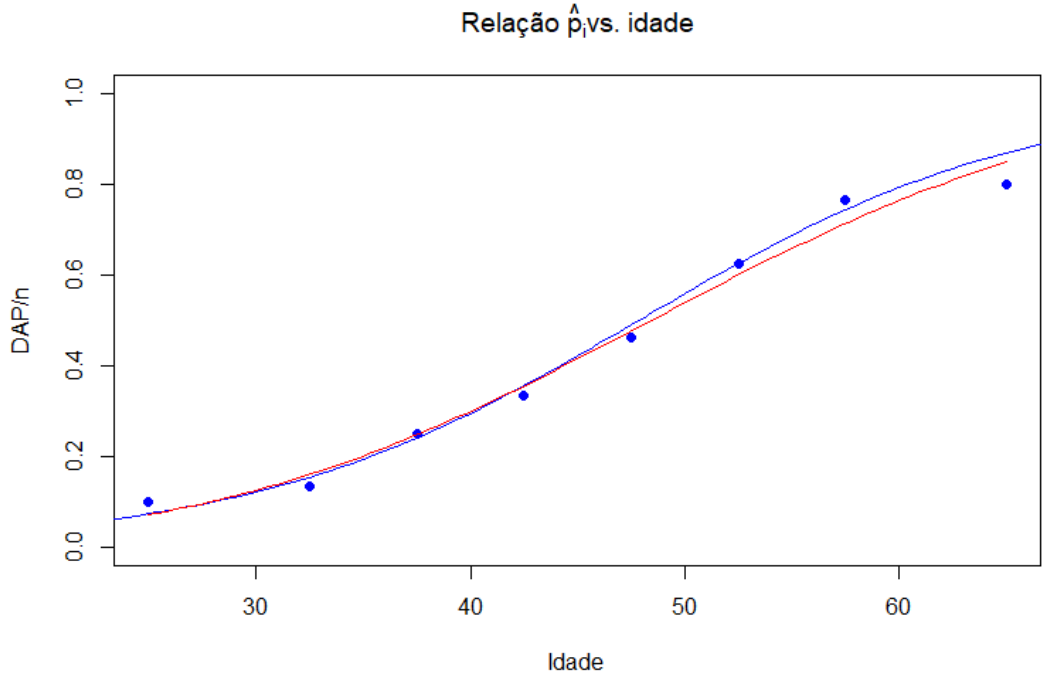
```
glm(cbind(DAC,n-DAC)~Idade, family=binomial(link=probit), data=HosLe

Call:  glm(formula = cbind(DAC, n - DAC) ~ Idade, family = binomial(
data = HosLem)

Coefficients:
(Intercept)        Idade
-3.0245           0.0624

Degrees of Freedom: 7 Total (i.e. Null);  6 Residual
Null Deviance:      28.7
Residual Deviance: 0.6529  AIC: 25.79
```

Tal como no caso da Regressão Logística, a variável resposta pode ser explicitada sob a forma de uma matriz de duas colunas, indicando o número de “êxitos"e o número de “fracassos"(como acima) ou, alternativamente, como

Regressão Probit no R (cont.)

A curva ajustada de probabilidade de DAP sobre idade (x), tem equação : $p(x) = \Phi(-3.0245 + 0.0624x)$. Eis a curva, sobreposta à nuvem de pontos (vermelho) (a em azul é a curva logística ajustada).

```
curve(pnorm(-3.0245+0.0624*x), add=T, col="red")
```



O modelo complemento log log

No mesmo contexto de variável resposta dicotômica Y, outra escolha frequente de função de ligação, com tradição histórica desde 1922 no estudo de organismos infecciosos consiste em tomar para probabilidade de êxito ($Y = 1$):

$$p(x) = g^{-1}\left(x^t\beta\right) = 1 - e^{-e^{x^t\beta}}$$

A função p é a diferença entre uma curva de Gompertz com valor assintótico $\alpha = 1$ e esse mesmo valor assintótico. O fato de se fixar o valor assintótico em 1 é natural, uma vez que a função p descreve *probabilidades*.
O contradomínio da função agora definida é o intervalo (0,1).

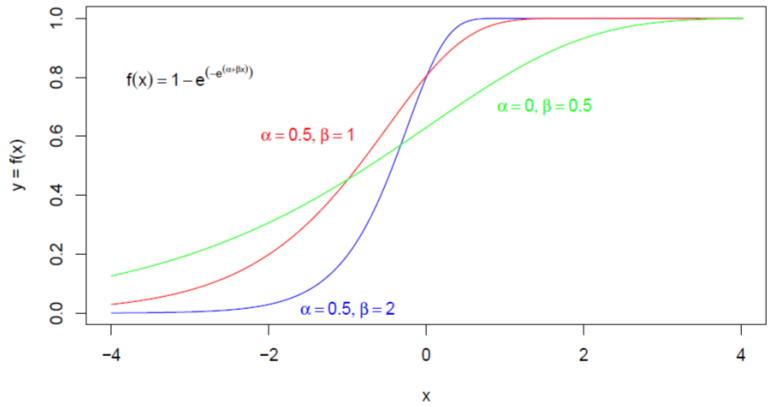
O modelo Complemento log log (cont.)

A função de ligação será, neste caso, da forma:

$$\vec{x}^t\beta = g\left(p(\vec{x}^t\beta)\right) = ln\left(-ln(1-p(\vec{x}^t\beta))\right)$$

de onde a designição do modelo que usa esta função de ligação.

No caso de haver uma única variável preditora X, a função $p(x)$ é a função distribuição cumulativa da distribuição Gumbel:



(13)

O modelo Complemento log log (cont.)

Esta função para p tem analogias e diferenças de comportamento em relação aos Modelos Logit e Probit:

- é igualmente estritamente monótona; tem igualmente um único ponto de inflexão, quando $\eta = 0$;
- mas o valor de probabilidade associado já não se encontra a meio caminho na escala de probabilidades, sendo $p(0) = 1 - \frac{1}{e}$;
- isso significa que a “fase de aceleração"da curva de probabilidades decorre até um valor superior da probabilidade (1-1/e $\approx$ 0,632) do que nas Regressões *Logit* e *Probit*.

Tal como no caso do Modelo *Probit*, os coeficientes β_j da componente sistemática não têm um significado tão facilmente interpretável como em uma Regressão Logística.

Complemento Log Log no R

Ajustar o modelo com função de ligação complemento log log faz-se especificando o valor *cloglog*

```
glm(cbind(DAC,n-DAC)~Idade, family=binomial(link=cloglog), data=HosL

Call:  glm(formula = cbind(DAC, n - DAC) ~ Idade, family = binomial(
data = HosLem)

Coefficients:
(Intercept)        Idade
-4.00470      0.07311

Degrees of Freedom: 7 Total (i.e. Null);  6 Residual
Null Deviance:     28.7
Residual Deviance: 1.148  AIC: 26.29
```

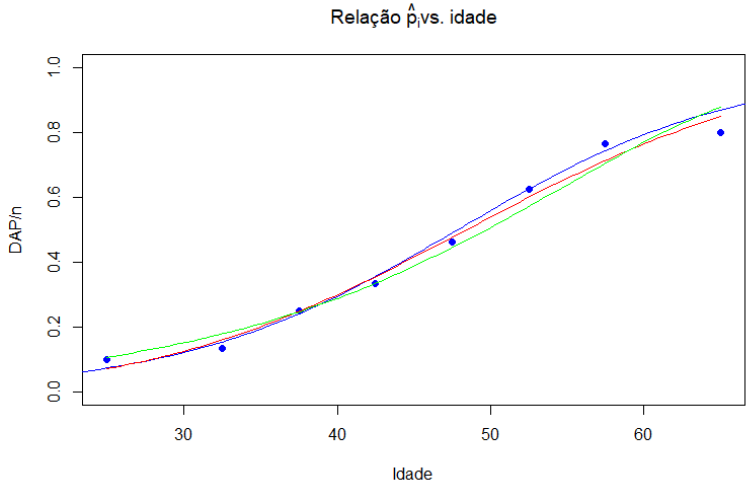
A curva ajustada é:

$$p(x) = 1 - e^{-e^{-4,00470+0,07311x}}.$$

Complemento log log no R (cont.)

A curva ajustada, sobreposta à nuvem de pontos do exemplo adaptado de DAP, é,

$$p(x) = 1 - e^{-e^{-4,00470+0,07311x}}.$$



Outras funções de ligação para respostas binárias

Foram consideradas três funções de ligação em modelo de resposta Bernoulli, cujas inversas são sigmóides. Em dois casos, usaram-se inversas de funções de distribuições cumulativas:

- f.d.c. de um Normal reduzida, no Modelo Probit;
- f.d.c. de uma Gumbel, no Modelo Complemento log log.

Uma generalização óbvia consiste em utilizar outra f.d.c. de uma variável aleatória contínua, gerando novos MLGs de resposta dicotômica.

No `R`, além das opções acima referidas, pode usar-se uma f.d.c. da distribuilção Cauchy.

Outras funções de ligação (cont.)

Outra possível generalização das funções de ligação para dados binários consiste em considerar a seguinte família de funções de ligação, que depende de um parâmetro, δ:

$$g(p;\delta) = ln\left[\frac{(1/(1-p))^{\delta}-1}{\delta}\right]$$

A função de ligação logit correspondente a tomar $\delta = 1$. A função de ligação complemento log log corresponde ao limite quando $\delta \to 0$.

Inferência: propriedades dos estimadores MV

Quaisquer estimadores β de máxima verossimilhança são:

- assintoticamente multinormais
- assintoticamente centrados ($E[\beta] \to \beta$)
- assintoticamente de matriz de variâncias-covariâncias $\iota_{\beta^{[i]}}$, em que,

$$\iota_\beta = -E[H_\beta]$$

é a matriz de Informação de Fisher, sendo H_β a matriz Hessiana da log-verossimilhança l, no ponto β, cujo elemento (j,m) é:

$$(H_\beta)_{(j,m)} = \frac{\partial^2 l}{\partial \beta_j \partial \beta_m}$$

Conclusão: Pode se fazer inferência (assintótica) em MLGs!

Inferência em MLGs

Aplicando estes resultados gerais aos estimaores β, obtém-se, assintoticamente:

$$\beta \sim N_{(p+1)}(\beta, \iota_\beta^{-1})$$

em que, ι_β é a matriz de informação de Fisher da log-verossimilhança da amostra, calculada no ponto β.

A dimensão da amostra tem uma importância grande para garantir a fiabilidade destes resultados.

Repara-se na semelhança com o resultado distribucional que serve de base à inferência em um modelo linear. As mesmas propriedades da Multinormal podem ser usadas para obter resultados análogos.

Inferência em MLGs (cont.)

Teorema

Dada um MLG (e admitindo certas condições de regularidade), os estiamdores de Máxima Verossimilhança β verificam, assintoticamente:

- Dado um vetor não aleatório a_{p+1} : $\frac{\vec{a}\hat{\beta}-\vec{a}\beta}{\sqrt{\vec{a}^t \iota_\beta}}$

O teorema permite obter intervalos de confiança e teste de hipóteses (aproximados) para combinações lineares dos parâmetros β.

Inferência em MLGs (cont.)

A derivação de resultados para combinações lineares dos parâmetros inclui como casos particulares importantes, resultados sobre parâmetros individuais e sobre somas ou diferenças de parâmetros.
Na expressão que serve de base aos ICs e Testes de Hipóteses surge a inversa da matriz de informação no ponto desconhecido β. Essa matriz desconhecida é substituída por outra, conhecida: a matriz de informação calculada para a estimativa β.
Esta substituição reforça a necessidade de grandes amostras para se possa confiar nos resultados.

Inferência em MLGs (cont.)

Intervalos de Confiança (assintóticos)

Um intervalo assintótico a $(1-\alpha)\times 100\%$ de confiança para a combinação linear $\vec{a}^t\beta$ é dado por:

$$\left(\vec{a}\hat{\beta} - z_{\frac{\alpha}{2}} \cdot \sqrt{\vec{a}\iota^{-1}_{\hat{\beta}}\vec{a}}; \vec{a}\hat{\beta} + z_{\frac{\alpha}{2}} \cdot \sqrt{\vec{a}\iota^{-1}_{\hat{\beta}}\vec{a}}\right)$$

sendo ι^{-1}_{β} a inversa da matriz de informação de Fisher da log-verossimilhança, calculada no ponto β.

Inferência em MLGs (cont.)

Teste de Hipóteses (assintótico)

Em um MLG, um teste de hipóteses (assintótico) bilateral a uma combinação linear dos β é:

- Hipóteses:

$$H_0 : \vec{a}\beta = c \qquad vs. \qquad H_1 : \vec{a}\beta \neq c$$

- Estatística de teste:

$$Z = \frac{\vec{a}\hat{\beta} - \vec{a}\beta_{|H_0}}{\sqrt{\vec{a}\iota_{\hat{\beta}}^{-1}\vec{a}}} \sim N(0,1),$$

- Região Crítica: Bilateral. Rejeitar H_0 se $|Z_calc| > Z_{\frac{\alpha}{2}}$ de teste:

Defini-se testes unilaterais, com hipóteses e RCs análogas às do modelo linear.

A função `summary` tem método para MLGs, gerando resultados análogos aos de modelos lineares.

A tabela de `Coefficients` tem colunas análogas:

- `Estimate` - valores estimados dos parâmetros β_j;
- `Std. Error` - os respectivos desvios padrão estimados, σ_{β_j}, ex., as raízes quadradas dos elementos diagonais da matriz $\iota_{\beta[i]}$;
- `z value` - o valor calculado da estatística $Z = \frac{\beta_j}{\sigma_{\beta_j}}$, para um teste às hipóteses $H_0 : \beta_j = 0 \qquad$ vs. $\qquad H_1 : \beta_j \neq 0$;
- `Pr(>|z|)` - O $p - value$ (bilateral) da estatística da coluna anterior (calculado em uma N(0,1)).

O teste referido pode servir para determinar a dispensabilidade de algum preditor.

Inferência no R

Na listagem do comando **summary** tem-se a informação fundamental para construir ICs ou Testes a parâmetros, em um MLG.

```
DAP.logistica <- glm(cbind(DAP,n-DAP) ~ Idade,
family=binomial, data=HosLem)
summary(DAP.logistica)

Call:
glm(formula = cbind(DAP, n - DAP) ~ Idade, family = binomial,
data = HosLem)

Deviance Residuals:
 Min            1Q    Median            3Q       Max
-0.42620  -0.15577  -0.00636   0.15210   0.41168

Coefficients:
Estimate     Std. Error     z      value Pr(>|z|)
(Intercept) -5.09073    1.09753  -4.638 3.51e-06 ***
Idade        0.10502    0.02308   4.551 5.35e-06 ***
```

A matriz de covariâncias dos estimadores no R

O comando `vcov` devolve a matriz de (co)variâncias dos estimadores, $\hat{\beta}$, ou seja, a inversa da matriz de informaçãoo de Fisher, $\iota_{\beta[i]}$:

```
 vcov(DAP.logistica)
             (Intercept)         Idade
(Intercept)  1.20457613 -0.0247424726
Idade       -0.02474247  0.0005325726
```

Esta é a matriz usada para construir um intervalo assintótico a $(1-\alpha)\times 100\%$ de confiança para a combinação linear $\vec{a}^t\beta$:

$$\left(\vec{a}\hat{\beta} - z_{\frac{\alpha}{2}} \cdot \sqrt{\vec{a}\iota_{\hat{\beta}}^{-1}\vec{a}}; \vec{a}\hat{\beta} + z_{\frac{\alpha}{2}} \cdot \sqrt{\vec{a}\iota_{\hat{\beta}}^{-1}\vec{a}}\right)$$

Intervalos de confiança para β_j no R

Os intervalos de confiança para os parâmetros individuais β_j são dados pela função `confint.default`:

```
confint.default(DAP.logistica)
                2.5 \%     97.5 \%
(Intercept) -7.24185609 -2.9396103
Idade        0.05978799  0.1502503
```

Venables & Ripley, no módulo MASS, disponibilizam um método alternativo (computacionalmente mais exigente) de construir intervalos de confiança em MLGs, denominado *profiling*. É automaticamente invocado, pela função confint:

```
confint(DAP.logistica)
Waiting for profiling to be done...
                   2.5 %     97.5 %
(Intercept) -7.42548805 -3.0887956
Idade        0.06276942  0.1539715
```

MLGs para variáveis resposta de Poisson

Consideremos agora modelos em que a componente aleatória Y tem distribuição de Poisson.

A distribuição de Poisson surge com muita frequência, associada à contagem de acontecimentos aleatórios (quando se pode admitir que não há acontecimentos simultâneos).

Se Y tem distribuição de Poisson, toma valores em N_0 com probabilidades $P[Y = k] = \frac{\lambda^k}{k!} e^{-lambda}$, para $\lambda > 0$.

Esta distribuição não é indicada para situações em que seja fixado à partida o número máximo de observações ou realizações do fenômeno, como sucede com uma Binomial.

Funções de ligação e ligação canônica

O valor esperado de $Y \Cap Po(\lambda)$ é o parâmetro λ.

Uma função de ligação será uma função $g(\cdot)$ tal que:

$$g(\lambda) = \vec{x}\beta,$$

em que $\vec{x}^t\beta$ é a componente sistemática do Modelo.

O parâmetro natural da distribuição de Poisson é $\theta = ln(\lambda)$.
Assim, a função de ligação canônica para uma componente aleatória com distribuição de Poisson é a função de ligação logarítimica:

$$g(\lambda) = ln(\lambda) = \vec{x}^t\beta \rightleftarrows \lambda = g^{-1}(\vec{x}^t\beta) = e^{\vec{x}^t\beta} \tag{14}$$

Um Modelo assim definido designa-se um Modelo Log Linear.

Modelos Log Lineares

São modelos com:

- componente aleatória de Poisson;
- função de ligação logaritmo natural, que é a ligação canônica para a Poisson.

Nota: a ligação apenas permite valores positivos do parâmetro λ, o que está estruturalmente de acordo com as características do parâmetro λ de uma distribuição Poisson.

Interpretação dos parâmetros β_j

No caso de haver uma única variável preditora X, a relação entre o parâmetro λ da distribuição Poisson e o preditor fica:

$$\lambda(x) = e_0^{\beta} \cdot e^{\beta_1 X}$$

O aumento de uma unidade no valor do preditor multiplica o valor esperado da variável resposta por e^{β_j}.

A interpretação generaliza-se para mais do que uma variável preditora. Com p variáveis preditoras tem-se:

$$\lambda(x) = e_0^{\beta} \cdot e^{\beta_1 X_1} \cdot e^{\beta_2 X_2} \ldots e^{\beta_p X_p}.$$

Um aumento de uma unidade no valor da variável preditora X_j, mantendo as restantes variáveis preditoras constantes, multiplica o valor esperado de Y por e^{β_j}.

Fatores preditores e tabelas de contingência

No caso de uma variável indicadora X_j, tem-se que a pertinença à categoria assinalada pela variável indicadora X_j multiplicada pelo parâmetro λ da distribuição de Poisson por e^{β_j}.

Os modelos log lineares têm grande importância no estudo de tabelas de contigência, cujas margens correspondem a diferentes fatores e cujo recheio corresponde a contagens de observações nos cruzamentos de níveis correspondentes.

Tal como nos casos anteriores, outras funções de ligação são concebíveis para variáveis resposta com distribuição de Poisson.

Exemplo: Modelos Log lineares

Em um modelo Log linear, as n observações independentes são de uma variável aleatória com distribuição de Poisson.
A função de verossimilhança destas n observações é dada por:

$$L(\lambda; \vec{y}) = \prod_{i=1}^{n} e^{-\lambda_i} \frac{\lambda_i^{y_i}}{y_i!}$$

E a log-verossimilhança por:

$$L(\lambda; \vec{y}) = \sum_{i=1}^{n} (-\lambda_i + y_i ln lambda_i - ln y_i!)$$

A função de ligação é dada por $g(\lambda) = ln(\lambda) = \vec{x}^t \beta$. Eis a expressão para a log-verossimilhança como função dos parâmetros β:

$$l(\beta) = \sum_{i=1}^{n} (-e^{\vec{x}^t \beta} + y_i \vec{x}^t \beta - ln y_i!)$$

Estimação em modelos log lineares (cont.)

Deixando cair a última parcela, que é a constante nos parâmetros β_j, logo dispensável na identificação dos máximos:

$$l(\beta) = \sum_{i=1}^{n} \left(-e \sum_{k=0}^{p} x_{k(i)} \beta_k + y_i \sum_{k=0}^{p} x_{k(i)} \beta_k \right) \tag{15}$$

(com a conversão $x_{0(i)} = 1, \forall i$). Condição necessária para a existência de extremo da extremo da log-verossimilhança no ponto $\beta = \hat{\beta}$ é que

$$\frac{\partial l(\hat{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} x_{j(i)} \left[y_i - e^{\sum_{k=0}^{p} X_{k(i)} \hat{\beta}_k} \right] = 0, \qquad \forall j = 0 : p$$

Estimação em modelos log-lineares (cont.)

Tal como no caso anterior, estas p+1 equações formam um sistema não-linear de equações nas $p+1$ incógnitas $\hat{\beta}_j$, j=0:p.

De novo, embora o sistema de equações seja não linear, é possível utilizar uma notação mnemónica matricial, definindo o vetor λ de probabilidade estimadas, cuja $i-$ésima componente é dada por:

$$\lambda = e^{\sum_{k=0}^{p} X_{k(i)} \hat{\beta}_k}$$

Com esta notação, o sistema de p+1 equações toma a forma:

$$X^t \vec{y} = X^t \lambda$$

A não linearidade do sistema exige métodos númericos.

IRLS para modelos log-lineares

No contexto do Modelo log-linear, as derivadas parciais de primeira ordem da log-verossimilhança são:

$$\frac{\partial l(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} x_j(i) \left[y_i - e^{\sum_{k=0}^{p} x_{k(i)} \beta_k} \right], \forall j = 0 : p$$

$$\leftrightarrows \frac{\partial l(\beta)}{\partial \beta} = X^t \vec{y} - X^t \lambda$$

Assim, as derivadas parciais de segunda ordem são:

$$\frac{\partial^2 l(\beta)}{\partial \beta_l \partial \beta_j} = -\sum_{i=1}^{n} x_{j(i)} x_{l(i)} e^{\sum_{k=0}^{p} x_{k(i)} \beta_k}$$

$$\frac{\partial^2 l(\beta)}{\partial \beta_l \partial \beta_j} = -\sum_{i=1}^{n} x_{j(i)} x_{l(i)} \lambda_i$$

IRLS para modelos log lineares (cont.)

De novo, a função de ligação é canônica e os elementos da matriz Hessiana não dependem de Y, pelo que Hessiana e seu valor esperado são iguais, ex., os métodos de Newton-Raphson e Fisher coincidem.
Defina-se a matriz $n \times n$ diagonal W, cujos elementos diagonais são dados pelos n valores λ_i. A matriz Hessiana e a correspondente matriz de informação de Fisher podem escrever-se como:

$$H = X^t W X \qquad \iota = X^t W X$$

A equação que define a iteração dos vetores β no algoritmo IRLS é:

$$\beta[i+1] = \beta[i] + \left(X^t W^{[i]}]X\right)^{-1} X^t \left(\vec{y} - \lambda^{[i]}\right)$$

Definindo o vetor

$$\vec{z}^{[i]} = X\beta^{[i]} + (W^{[i]})^{-1}\left(\vec{y} - \lambda^{[i]}\right)$$

tem se uma expressão de transição idêntica à do Modelo Logit:

Avaliação da qualidade de um MLG

Conceito importante na avaliação da qualidade de um MLG é o conceito de Desvio de um Modelo (*deviance* em inglês).

O desvio desempenha nos MLGs um papel análogo ao da Soma de Quadrados Residual nos Modelos Lineares.

No estudo do Modelo Linear foi introduzida a noção de Modelo Nulo: um Modelo em que o preditor linear é constituído apenas por uma constante e toda a variação nos valores observados é variação residual, não explicada pelo Modelo.

No estudo de Modelos Lineares Generalizados é de utilidade um Modelo que ocupa o extremo oposto na gama de possíveis modelos: O Modelo Saturado que tem tantos parâmetros quantas as observações de Y disponíveis.

Modelo Nulo e Modelo Saturado (cont.)

Em um modelo Saturado, o ajustamento é “perfeito", mas inútil: a estimativa de cada valor esperado de Y coincide totalmente com o valor observado de Y correspondente, isto é $E[\hat{Y}_i] = Y_i$.

Recorde-se que, quer no Modelo Logístico, quer no Modelo Log Linear, o sistema de equações normais resultante da condição necessária para a existência de máximo da log-verossimilhança toma a forma $X\vec{y} = X\hat{\mu}$, em que $\hat{\mu}$ indica o vetor estimado de $E[Y_i]$ para as n observações.
Em um modelo saturado, com tantos parâmetros quantas observações, X é de tipo $n \times n$ e, em geral, invertível. Nesse caso, $\hat{\mu} = \vec{y}$.

Desvios

Assim, um modelo saturado ocupa o polo oposto em relação ao Modelo Nulo: enquanto que neste último tudo é variação residual, não explicada pelo modelo, em um modelo saturado tudo é “explicado"pelo modelo, não havendo lugar a variação residual.

Um tal ajustamento “total"dos dados ao modelo é, em geral, ilusório. Mas é de utilidade como termo de comparação para medir o grau de ajustamento de um conjunto de dados a um MLG, medindo-se o afastamento em relação a este ajustamento “ideal".
É nessa ideia que se baseia a definição do conceito de Desvio ou *Deviance*.

Desvios (cont.)

Considere-se um Modelo Linear Generalizado baseado em n observações independentes da variável resposta Y.
seja $\hat{\beta}_M$ o vetor estimado dos seus parâmetros e $l_m(\hat{\beta}_M)$ a respectiva log-verossimilhança (máxima).
Considere um modelo saturado com n parâmetros. Designe-se por $l_T(\hat{\beta}_T)$ a log verossimilhança correspondente (isto é, a log-verossimilhança obtida substituindo cada parâmetro estimado $\hat{\mu}_i$ pela observação correspondente y_i).
Defini-se o desvio como sendo:

$$D^* = -2\left(l_m(\hat{\beta}_M) - l_T(\hat{\beta}_T)\right)$$

$$l(\theta, \phi) = \sum_{i=1}^{n} \left[\frac{y_i \theta_i - b(\theta_i)}{a(\phi_i)} - c(y_i, \phi_i) \right]$$

O desvio correspondente, indicado pelas letras M e T os estimadores associados ao parâmetro natural θ, e admitindo conhecidos os parâmetros de dispersão, vem:

$$D^* = -2 \left(l_m(\hat{\beta}_M) - l_T(\hat{\beta}_T) \right) = 2 \sum_{i=1}^{n} \left[\frac{y_i(\theta_i^T - \theta_i^M) - [b(\theta_i^T) - b(\theta_i^M)]}{a(\phi_i)} \right]$$

Na expressão do desvio surge o parâmetro de dispersão ϕ.

As expressões para os desvios são mais simples caso o parâmetro de dispersão seja uma constante, que não exige estimação. É o caso das distribuições de Poisson, Bernoulli e ou Binomial/n:

- $\phi = 1$ na Poisson;
- $\phi = 1$ na Bernoulli;
- $\phi = \frac{1}{n}$ na Binomial/n.

Mas, para distribuições bi paramétricas da família exponencial em que o parâmetro ϕ não é conhecido, ϕ tem de ser estimado a partir dos dados para se poder calcular o desvio.

Desvios na Poisson e Binomial/n

Substituindo as expressões já citadas na definição geral do desvio, tem-se as seguintes expressões para MLGs em que T tem:

- distribuição de Poisson;

$$\begin{aligned} D^* &= 2\sum_{i=1}^{n}[y_i\left(ln(y_i) - ln(\hat{\lambda}_i)\right) - y_i + \hat{\lambda}_i] \\ \Leftrightarrow D^* &= 2\sum_{i=1}^{n}\left[y_i\left(\frac{ln(y_i)}{\hat{\lambda}_i}\right) - (y_i - \hat{\lambda}_i)\right] \end{aligned}$$

- Distribuição Binomial/n:

$$D^* = 2\sum_{i=1}^{n} n_i\left\{y_i\left(\frac{ln(y_i)}{\hat{p}_i}\right) - (1-y_i)ln\left(\frac{1-y_i}{1-\hat{p}_i}\right)\right\}$$

Desvios e devios reduzidos (cont.)

Para distribuições em que seja necessário estimar ϕ, é hábito definir um conceito alternativo de Desvio. Admitindo que

$$a(\phi) = \frac{\phi}{w_i},$$

para constantes w_i conhecidas e ϕ comum a todas as observações:

$$D^* = -2(l(\hat{\theta}^M) - l(\hat{\theta}^T)) = 2\sum_{i=1}^{n} \frac{w_i}{\phi}\left[y_i(\hat{\theta}_i^T - \hat{\theta}_i^M) - [b(\hat{\theta}_i^T) - b(\hat{\theta}_i^M)]\right] \quad (16)$$

É usual chamar a D^* o desvio reduzido (*scaled deviance*) e reservar a expressão desvio (*deviance*) para D, definido tal que:

$$D^* = \frac{D}{\phi},$$

$$\Leftrightarrow D = 2\sum_{i=1}^{n} w_i\left[y_i(\hat{\theta}_i^T - \hat{\theta}_i^M) - [b(\hat{\theta}_i^T) - b(\hat{\theta}_i^M)]\right]$$

Desvio e desvio reduzido na Normal

O desvio reduzido na Normal, (admitindo a variância σ_i^2 de cada observação conhecida e escrevendo $\hat{\mu}^M$ apenas como $\hat{\mu}_i$) é:

$$D^* = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\sigma_i^2} \tag{17}$$

com a hipótese usual do Modelo Linear de que $\sigma_i^2 = \sigma^2 = \phi$ para todas as observações, o desvio da Normal vem:

$$D = \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2 = SQRE$$

ou seja, o devio e a tradicional Soma de Quadrados Residual, coincidem.

O AIC em MLGs

O critério de Informação de Akaike (AIC) defini-se, em um MLG com p preditores (e constante aditiva), como

$$AIC = -2 \cdot l(\hat{\beta}, Y) = 2(p+1)$$

- Quanto menor o valor do AIC (para igual variável resposta, Y), melhor o ajustamento do modelo.
- O AIC pode ser usado como critério de comparação de modelos e submodelos;
- Nota-se a relação entre o desvio reduzido D^* de um MLG e o seu AIC: ambos definidos a custa da log-verossimilhança.

A razão de verossimilhanças

Um teste à admissibilidade de um Submodelo pode ser obtido com base em um resultado mais geral: o teste à razão de verossimilhanças.
Seja $(Y_1, Y_2, \cdots, Y_n)$ um amostra aleatória. Seja $L(\theta, X)$ a sua função verossimilhança, em que θ designa um vetor de parâmetros. Sejam Θ_0 e Θ_1 dois conjuntos alternativos de condições sobre os valores dos parâmetros θ.
Designa-se razão de verossimilhanças a:

$$R_n(X) = \frac{1}{1}$$

Em alguns contextos, a transformação $\Lambda = -2ln(R_n)$ pode ser utilizada como estatística de um teste às hipóteses:

$$H_0 : \theta \in Theta_0 vs H_1 : \theta \in \Theta_1.$$

Teorema de Wilks

O teorema de Wilks garante que, sob H_0 (e com certas condições de regularidade da função de verossimilhança) $\Lambda = -2ln(R_n)$ tem distribuição assintótica χ^2_q, onde q indica o número de restrições impostas aos parâmetros em H_0:

$$\Lambda = -2(l(\theta; X) - l(\theta; X)) \sim \chi^2_q$$

No contexto de comparação de modelos e submodelos em um MLG,

- q é a diferença entre o número de parâmetros do modelo completo $(\Theta_0 \Cap \Theta_1)$ e do submodelo $(\Theta_0) : q = p - k$.
- $\Lambda = D^*_S - D^*_M$ é a diferença dos desvios:
- do submodelo, $D^*_S = -2(l(\hat{\theta}^S) - l(\hat{\theta}^T))$.
- do submodelo, $D^*_M = -2(l(\hat{\theta}^M) - l(\hat{\theta}^T))$.

Teste de Wilks a Submodelos

No contexto de um Modelo Linear Generalizado, os parâmetros θ são os $p+1$ coeficientes β_j da combinação linear que constitui a componente sistemática do Modelo.

Sejam Θ_0 os valores resultantes de impor a restrição $\beta_{\vec{S}} = \vec{0}$.

Por Θ_1 indica-se a condição complementar: $\beta_{\vec{S}} \neq \vec{0}$.

O máximo da função log-verossimilhança para $\theta \in \Theta_0 \Cap \Theta_1$ correspondente às estimativas MV do Modelo Completo.

O máximo da função log-verossimilhança para $\theta \in \Theta_0$ são as estimativas $\hat{\beta}_S$ de Máxima Verossimilhança do Submodelo.

Teste de Wilks a Submodelos

Assim, a estatística do Teste de Wilks a modelos encaixados é a diferença dos Desvios de Modelo e Submodelo.

Teste de Wilk a Submodelos Encaixados

Hipóteses:

$$\begin{aligned} H_0 : \beta_j &= = 0, \forall j \notin S vs. H_1 : \beta_j \neq 0 t.q. \exists j \notin S \\ \Leftrightarrow H_0 : \beta_S &= 0(\text{submodelo OK}) vs. H_1 : \beta_S \neq 0(\text{modelo Melhor}) \end{aligned}$$

Estatística do Teste:

$$\Lambda = D_S^* - D_M^* \sim \chi^2_{p-k},$$

Região Crítica:

Unilateral direito. Rejeitar H_0 se $\Lambda_{calc} > \chi^2_{\alpha;(p-k)}$.
No caso do parâmetro de dispersão ϕ não ser conhecido, o cálculo de D (que envolve ϕ) fica condicionado. São necessários testes alternativos, ou trabalhar

Teste de Wilks ao Ajustamento Global

Para MLGs cuja componente sistemática inclui uma parcela aditiva constante, o conceito de ajustamento global do Modelo pode ser semelhante ao usado no estudo do Modelo Linear: compare-se o ajustamento do Modelo e do Submodelo Nulo, que se obtém sem qualquer variável preditora (apenas com a constante).
No Submodelo Nulo tem-se:

$$g(E[Y_i]) = \beta_0 \Leftrightarrow E[Y_i] = g^{-1}(\beta_0), \forall i = 1 : n. \tag{18}$$

Ou seja, a variação de $E[Y]$ não depende de variáveis preditoras. Se esse Submodelo Nulo não se ajustar de forma significativamente diferente do Modelo sob estudo, conclui-se pela inutilidade do Modelo.

Teste de Wilks ao Ajustamento Global

beginblockTeste de Wilks ao Ajustamento Global
Hipóteses:

$$H_0 \beta_j = 0 (\text{Modelo inutil}), \forall j = 1 : p vs H_1 : \beta_j \neq 0 \text{Modelo útil} t.q. \exists j = 1 : p$$

Estatística do Teste:

$$\Lambda = D_N^* - D_M^* \sim \chi_p^2,$$

Região Crítica:

Unilateral direito. Rejeitar H_0 se $\Lambda_{calc} > \chi^2_{\alpha;(p)}$.
D_N^* - desvio do modelo nulo.

Exemplo Do livro de Venables e Ripley

Uma experiência estuda a resistência da larva do tabaco heliothis virescens a doses de uma substância tóxica.

Lotes de 20 traças de cada sexo foram expostas, durante 3 dias, a doses da referida substância. Registou-se o número de indivíduos de cada lote que morria até ao fim desse período de exposição. Os resultados são sintetizados na seguinte tabela (doses em μ g).

	Dose					
Sexo	1	2	4	8	16	32
Machos	1	4	9	13	18	20
Fêmeas	0	2	6	10	12	16

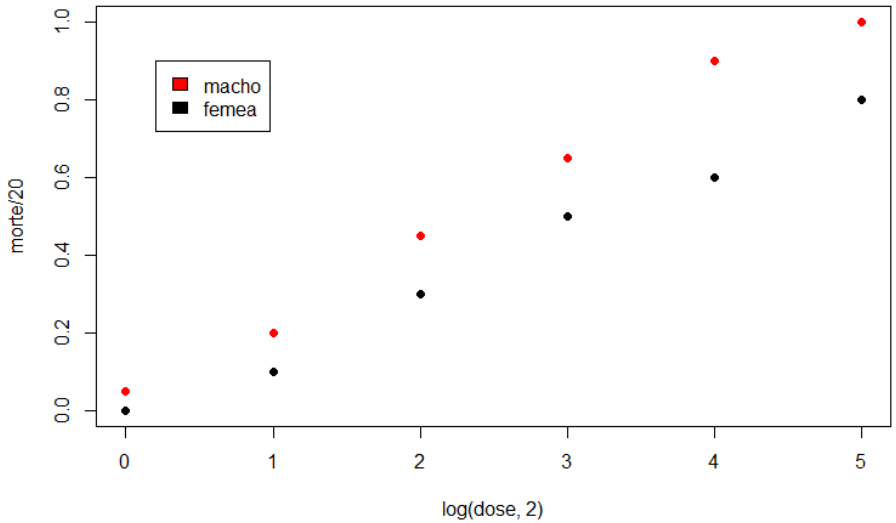
Trata-se de dados com a variável resposta Binomial (número de mortes em $n = 20 \times 12 = 240$ larvas expostas ao tóxico).

Exemplo MLG (cont.

```
############## Bill venables

morte <- c(1,4,9,13,18,20,0,2,6,10,12,16)
sexo <- factor(rep(c("macho","femea"),c(6,6)))
dose <- rep(2^(0:5),2)
tabaco <- data.frame(morte,sexo,dose)
tabaco

morte  sexo dose
1      1 macho    1
2      4 macho    2
3      9 macho    4
4     13 macho    8
5     18 macho   16
6     20 macho   32
7      0 femea    1
8      2 femea    2
9      6 femea    4
```

Exercicio 2

```
attach(tabaco)
The following objects are masked _by_ .GlobalEnv:

dose, morte, sexo

plot(log(dose,2),morte/20,col=as.numeric(sexo),pch=16)
legend(0.2,0.9,legend=c("macho","femea"), fill=c("red","black"))
```



Embora uma relação linear parece adequada, uma relação sigmoide é

Exercício 2 no R (cont.)

Para ajustar uma Regressão Probit, utiliza-se a opção link=“probit"na definição do argumento family:

```
glm(cbind(morte,20-morte) ~ log(dose,2),family=binomial(link="probit
```

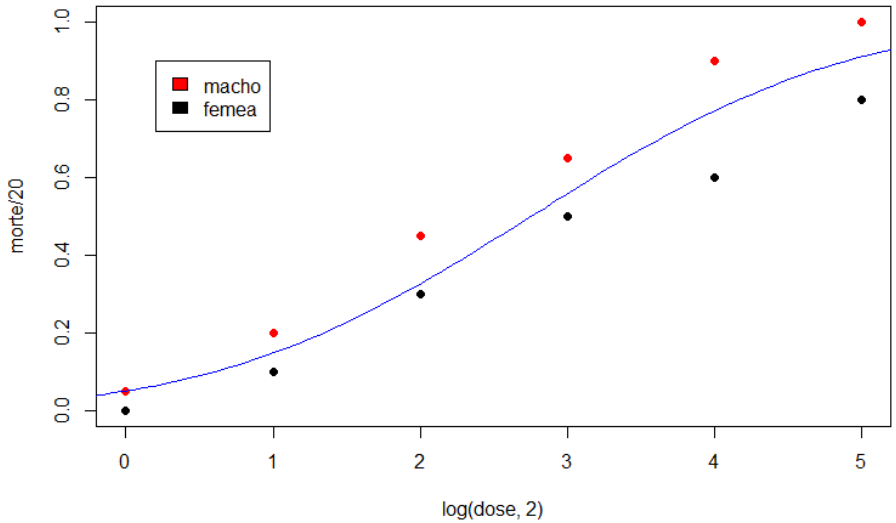
```
Call:  glm(formula = cbind(morte, 20 - morte) ~ log(dose, 2), family
data = tabaco)

Coefficients:
(Intercept)  log(dose, 2)
-1.6431         0.5966

Degrees of Freedom: 11 Total (i.e. Null);  10 Residual
Null Deviance:      124.9
Residual Deviance: 16.41  AIC: 50.52
```

A relação estimada é : $p(x) = \Phi(-1,6431 + 0,5966 log_2(x))$

Exercicio 2 no R (cont.)

Sobrepõe se a curva ajustada à nuvem de pontos, com o comando:

```
curve(pnorm(-1.6431+0.5966*x), from=-1, to=6, col="blue", add=TRUE)
```



Teste de ajustamento Global no R

No R, um teste de Wilks, comparando um modelo MLG com o modelo nulo corresponde, pode ser feito utilizando o comando `anova`, com o argumento `test="Chisq"`.

```
tabaco.glm <- glm(cbind(morte,20-morte) ~ log(dose,2),
+                   family=binomial(link="probit"), data=tabaco)
> anova(tabaco.glm, test="Chisq")
Analysis of Deviance Table
Model: binomial, link: probit
Response: cbind(morte, 20 - morte)
Terms added sequentially (first to last)
Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                             11    124.876
log(dose, 2)  1   108.46         10     16.414 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Como previsível, o modelo ajusta-se significativamente melhor do que um modelo nulo, sem preditores.

Exercicio 13 no R

Também é possível cruzar fatores com preditores numéricos, como numa Análise de Covariância.

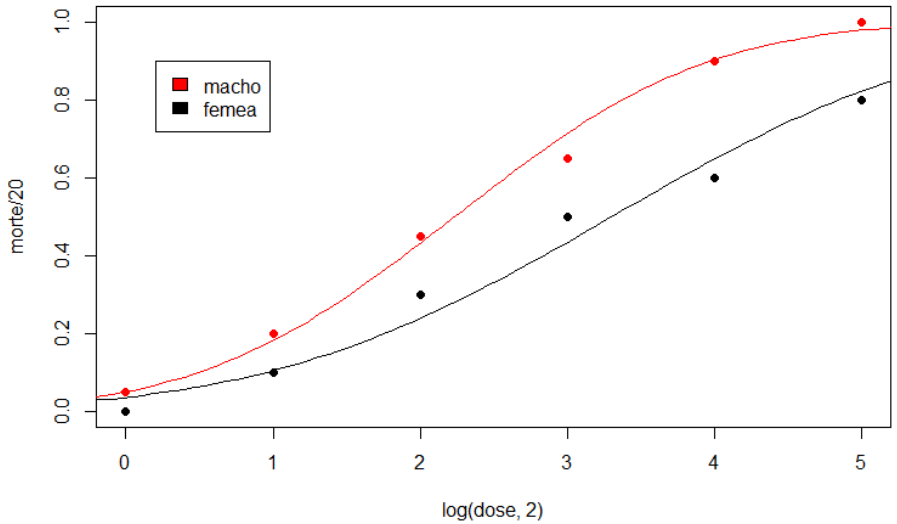
```
summary(glm(cbind(vomial(link="probit"), data=tabaco))
Call:
glm(formula = cbind(morte, 20 - morte) ~ log(dose, 2) * sexo,
family = binomial(link = "probit"), data = tabaco)

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept)            -1.80072    0.29832  -6.036 1.58e-09 ***
log(dose, 2)            0.54523    0.09138   5.966 2.43e-09 ***
sexomacho               0.15479    0.41635   0.372    0.710
log(dose, 2):sexomacho  0.19165    0.14259   1.344    0.179
---

Null deviance: 124.876  on 11  degrees of freedom
Residual deviance:   3.768  on  8  degrees of freedom
AIC: 41.878
```

As relações estimadas são:

- $p(x) = \Phi(-1,80072 + 0,54532 log_2(x))$ nas fêmeas; e
- p(x)=$\Phi((-1,80072 + 0,15479) + (0,54532 + 0,19165) log_2(x))$ nos machos.

Veja como o desvio baixou de 16,41 para apenas 3,768.

```
plot(morte/20 ~ log(dose,2),col=sexo, data=tabaco,pch=16)
curve(pnorm(-1.80072+0.54523*x), from=-1, to=6, col="black",add=TRUE
curve(pnorm((-1.80072+0.15479)+(0.54523+0.19165)*x), from=-1, to=6,
+ col="red", add=TRUE)
legend(0.2,0.9,legend=c("macho","femea"), fill=c("red","black"))
```



(19)

Para saber se há vantagem em considerar modelos diferentes para cada sexo, comparam-se os modelos, usando o teste de Wilks.

```
tabaco.glmS <- glm(cbind(morte,20-morte) ~ log(dose,2) * sexo, famil
anova(tabaco.glm, tabaco.glmS, test="Chisq")
Analysis of Deviance Table

Model 1: cbind(morte, 20 - morte) ~ log(dose, 2)
Model 2: cbind(morte, 20 - morte) ~ log(dose, 2) * sexo
Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        10     16.414
2         8      3.768  2   12.646 0.001795 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Há vantagens evidentes na distinção por sexo (previsível pelo ajustamento gráfico).

Seleção de Submodelos

Tal como no Modelo Linear, a escolha de um submodelo adequado pode ser determinado por considerações de diversa ordem.

Caso não haja um submodelo proposto, a pesquisa completa dos $2^p - 2$ possíveis submodelos coloca as mesmas dificuldades computacionais já consideradas no estudo do Modelo Linear. A função `eleaps` do módulo R `subselect` permite efetuar pesquisas completas para submodelos MLG ótimos de uma dada cardinalidade.

Alternativamente, é possível usar algoritmos de exclusão ou inclusão sequenciais, semelhantes aos usados no Modelo Linear, mas adotando como critério para a inclusão/exclusão de variáveis a maior/menor redução (significativa) que geram no Desvio.

Algoritmos sequenciais no R

No R,

- o comando `anova` fornece a informação básica para efetuar um Teste de razão de verosimilhanças a Submodelos encaixados (indicando os submodelos como argumentos do comando); e
- os comandos `drop1` e `add1` fornecem a informação básica para proceder aos algoritmos de exclusão/inclusão sequenciais de variáveis preditoras, na escolha de Submodelos.
- o comando `step` automatiza os algoritmos de seleção sequencial com base no AIC. É respeitada a natureza dos preditores categóricos e a hierarquia dos tipos de efeitos que lhe estão associados.

Algoritmos de seleção de preditores no R

Para ilustrar a aplicação do algoritmo de exclusão sequencial, vejamos o exemplo já considerado, associado ao Exercício 13:

```
 step(tabaco.glmS)
Start:  AIC=41.88
cbind(morte, 20 - morte) ~ log(dose, 2) * sexo

Df Deviance    AIC
- log(dose, 2):sexo  1    5.566 41.676
<none>                    3.768 41.878

Step:  AIC=41.68
cbind(morte, 20 - morte) ~ log(dose, 2) + sexo

Df Deviance     AIC
<none>               5.566  41.676
- sexo           1   16.414  50.524
- log(dose, 2)   1  118.799 152.909
```

A distribuição Gama na família exponencial

Uma variável aleatória Y tem distribuição Gama com parâmetros μ e ν se toma valores em R^+, com função densidade da forma

$$f(y|\mu,\nu) = \frac{\nu^\nu}{\mu^\nu \Gamma(\nu)} y^{\nu-1} e^{-\frac{\nu y}{\mu}} = e^{\frac{\left(-\frac{1}{\mu}\right)y + ln\left(\frac{1}{\mu}\right)}{\frac{1}{\nu}} + \nu ln\nu - ln\Gamma(\nu) + (\nu-1)lny}$$

que é da familia exponencial com:

- $\theta = -\frac{1}{\mu}$
- $= \frac{1}{\nu}$
- $b(\theta) = -ln(\frac{1}{\mu}) = -ln(-\theta)$
- $a(\phi) = \phi = \frac{1}{\nu}$
- $c(y,\phi) = \nu ln\nu - ln\Gamma(\nu) + (\nu-1)lny$

A família das distribuições Gama inclui como caso particular a distribuição Qui-quadrado (χ^2_n se $\nu = \frac{n}{2}$ e $\mu = n$) e a distribuição Exponencial ($\nu = 1$)

Vejamos agora um exemplo de MLG com variável resposta contínua, não Normal. Consideremos uma componente aleatória Y com distribuição Gama (que, como sabemos, inclui como casos particulares uma Exponencial ou uma Qui-quadrado).
Se $Y \Cap G(\mu, \nu)$, tem-se:

$$E[Y] = \mu \qquad e \qquad V[Y] = \frac{\mu^2}{\nu}$$

Assim, na distribuição Gama a variância é proporcional ao quadrado da média. Esta propriedade sugere que MLGs com componente aleatória Gama podem ser úteis em situações onde a variância dos dados não seja constante, mas proporcional ao quadrado da média.

Funções de ligação e ligação canônica na Gama

Uma vez que para $Y \Cap G(\mu, \nu)$ se verifica $E[Y] = \mu$, as funções de ligação g em um MLG com variável resposta Gama relacionam a média μ com as combinações lineares das variáveis preditoras:

$$g(\mu) = x^t\beta = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$$

A função de ligação canônica para modelos com distribuição Gama será a função g que transforma o valor esperado de Y no parâmetro natural $\theta = -\frac{1}{\mu}$. Como o sinal negativo não é relevante na discussão, é hábito definir a função de ligação canônica para modelos com variável resposta Gama apenas como a função recíproca:

$$g(\mu) = \frac{1}{\mu}$$

Um único preditor na Gama

O modelo fica completo equacionando a parte sistemática a esta transformação canônica do valor esperado de Y:

$$g(\mu) = \frac{1}{\mu} = x^t\beta \qquad \leftrightarrow \qquad \mu(x^t\beta) = g^{-1}\left(x^t\beta\right) = \frac{1}{x^t\beta}$$

No caso particular de haver uma única variável preditora, a relação que acabamos de estabelecer diz que o valor médio de Y é dado por uma curva de tipo hiperbólico,

$$E[Y] = \frac{1}{\beta_0 + \beta_1 x}.$$

Esta função tem sido usada em Agronomia para modelar curvas de rendimento por planta (Y), em função da densidade da cultura (X).

Um preditor transformado

Caso se opte por trabalhar com os recíprocos de um único preditor, ou seja com a transformação $X^* = \frac{1}{x}$, o valor esperado fica,

$$E[Y] = \frac{1}{\beta_0 + \beta_1/x} = \frac{x}{x\beta_0 + \beta_1},$$

pelo que o valor esperado de Y será dado pela curva de Michaelis-Menten (com a parametrização de Shinozaki-Kira).
Nota: embora o valor esperado da variável resposta Y tenha de ser positivo (uma vez que uma variável Y com distribuição Gama só toma valores positivos), na relação estabelecida o valor esperado pode ser negativo para alguns valores da(s) variável(is) preditora(s).
Assim, e ao contrário de modelos anteriores, não existe uma "garantia estrutural"de que os valores de μ estimados façam sentido.

Desvio e desvio reduzido na Gama

Tem-se, a partir das expressões para D^* e para D e tendo em conta que $\theta = \frac{1}{\mu}, b(\theta) = -ln(-\theta) = ln(\mu), \phi = \frac{1}{\nu} e a(\phi) = \phi = \frac{1}{\nu}$:

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

Admitindo que $a(\phi) = \frac{\phi}{w_i}$, para algum conjunto de constantes w_i, o desvio não vem muito diferente (apenas substituindo ν_i por w_i).
Com a hipótese da igualdade de parâmetros de dispersão nas n observações, fica-se com uma expressão mais simples para o desvio:

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

Resíduos e Validação do Modelo

O conceito de resíduos, $e_i = y_i - \hat{y}_i$, usado no Modelo Linear como ferramenta para a validação das hipóteses subjacentes ao Modelo, tem de ser adaptado nos MLGs, onde, diversamente do que acontecia nos Modelos Lineares, não se contempla a existência de erros aleatórios aditivos.
Em Modelos Lineares Generalizados utilizam-se diversos conceitos de resíduos, sendo os principais os

- resíduos de Pearson; e os
- resíduos do desvio.

Referências I