# Modelos Matemáticos e Aplicações
# Generalized Linear Models

Jorge Cadima

Matemática (DCEB), Instituto Superior de Agronomia (ULisboa)

2022-23

# Bibliography


- Agresti, A. (2015) *Foundations of Linear and Generalized Linear Models*, Wiley.
- Dobson, A.J. & Barnett, A.G. (2008) *An Introduction to Generalized Linear Models*, 3rd ed., CRC Press.
- McCullaugh, P. & Nelder, J. (1989) *Generalized Linear Models*, Chapman & Hall.
- McCulloch, C. & Searle, S. (2001) *Generalized, Linear, and Mixed Models*, John Wiley & Sons. **Mat 600-62**.
- Agresti, A. (1990) *Categorical Data Analysis*, John Wiley & Sons. **Mat 401-62**.
- Hosmer, D.W. & Lemeshow, S. (1989) *Applied Logistic Regression*, John Wiley & Sons. **Mat 258-62**.


GLMs in  :


- Faraway, J.J. (2006) *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*, Chapman & Hall/CRC.
- Fox, J. & Weisberg, S. (2011) *An R Companion to Applied Regression, 2d Ed*, Sage Publications (`R` *package*: `car`).
- Venables & Ripley (2002). *Modern Applied Statistics with S* (4a. edição), Springer. (`R` *package*: `MASS`).

# Generalized Linear Models

Generalized Linear Models, GLMs:

- are a very vast family of models;
- extend the Linear Model;
- encompass many previously known and used models, sometimes over many decades. Among them:
  - the *probit* model;
  - the *logit* (or Logistic Regression) model;
  - log-linear models
  - the linear model itself.
- the "umbrella" for GLMs was introduced and formalized by McCullagh and Nelder (1989);

# Motivating example: binary response variable

## Hosmer & Lemeshow example

Hosmer and Lemeshow, in *Applied Logistic Regression* (Wiley, 1989) has data on $n = 100$ patients, with variables:

- age (`Idade`) – numeric;
- coronary heart disease (DAC) – binary variable (yes/no; 1/0).

Here are the first six row of the corresponding HL *data frame*:

```
> head(HL)
```

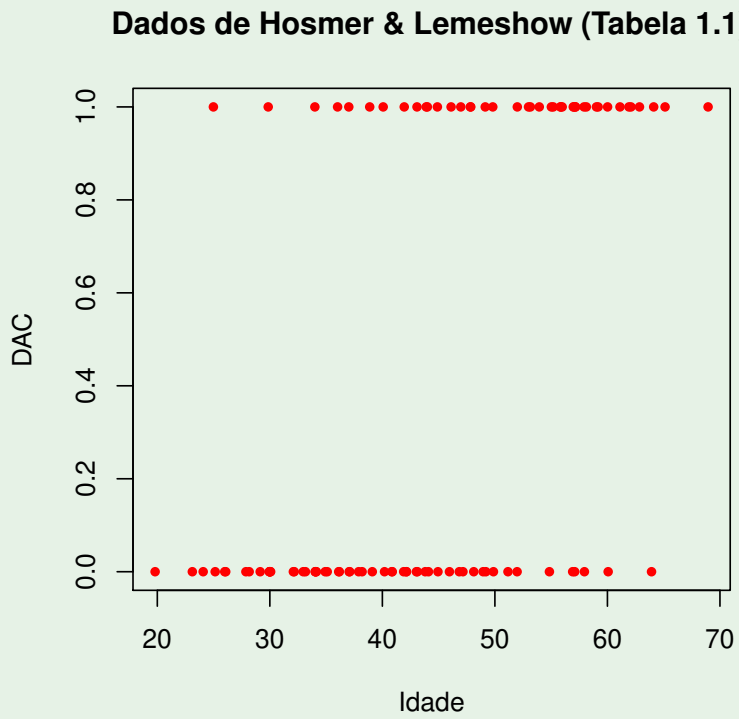
```
  Idade  DAC
1    20    0
2    23    0
3    24    0
4    25    0
5    25    1
6    26    0
```

We wish to relate the existence of coronary heart disease (DAC) (response $Y$) with age (Idade, predictor $X$). The scatterplot of $Y$ vs. $X$ is not promising.

# Example 1: DAC *vs.* Idade

## Hosmer & Lemeshow Example

```
> plot(DAC ~ Idade , data=HL , cex=0.8 , col="red" , pch=16 ,
+ xlab="Idade" , main="Dados de Hosmer & Lemeshow (Tabela 1.1)")
```



# The Linear Model

Recall that the linear model relates

- a numerical response variable $Y$ with
- predictors $X_1$, $X_2$, ..., $X_p$,

by the equation, using $n$ independent observations $Y_i$:

$$Y_i = \beta_0 + \beta_1 X_{1_{(i)}} + \beta_2 X_{2_{(i)}} + ... \beta_p X_{p_{(i)}} + \varepsilon_i \ ,$$

with $\varepsilon_i \frown \mathcal{N}(0, \sigma^2)$ $(i = 1, 2, ..., n)$.

That is, in the Linear Model:

- $E[Y_i | X_1 = x_{1_{(i)}}, ..., X_p = x_{p_{(i)}}] = \beta_0 + \beta_1 x_{1_{(i)}} + \beta_2 x_{2_{(i)}} + ... \beta_p x_{p_{(i)}}$ ,
- $Y_i$ independent, with Normal distributions (and equal variances).

# The generalization of the linear model

Linear Model

- $E[Y_i] = \beta_0 + \beta_1 x_{1_{(i)}} + \beta_2 x_{2_{(i)}} + \ldots \beta_p x_{p_{(i)}}$ ,
- $Y_i$ with Normal distribution.

In a Generalized Linear Model there are two extensions:

- $g(E[Y_i]) = \beta_0 + \beta_1 x_{1_{(i)}} + \beta_2 x_{2_{(i)}} + \ldots \beta_p x_{p_{(i)}}$ ,
  with $g$ an invertible function called link function.
- $Y_i$ with distribution in the exponential family of distributions.

Hence, a GLM models the expected value of a response variable whose distribution belongs to the exponential family, through the equation:

$$\mu_i = E[Y_i] = g^{-1}(\beta_0 + \beta_1 x_{1_{(i)}} + \beta_2 x_{2_{(i)}} + \ldots \beta_p x_{p_{(i)}}) .$$

Note: Linear Models are specific instances of GLMs: the Normal distribution is in the exponential family of distributions and the link function is the identity: $g(x) = x, \forall\, x$.

# The three components of a GLM (cont.)

In the words of Agresti (1990, p.81):

> *a GLM is a linear model for a transformation of the expected value of a random variable whose distribution belongs to the exponential family.*

Note: unlike in Linear Models, in GLMs there are no explicit additive random errors. The random fluctuation of the response variable is introduced when we specify its probability distribution.

# The exponential family of distributions

The exponential family of distributions includes, among others:

- the Normal
- the Poisson (for count variables)
- the Bernoulli (for binary variables)
- the "Binomial/n" (for proportions of successes in $n$ Bernoulli trials)
- the Gamma (asymmetric continuous distribution);
  which has the Exponential as a special case.
- the inverse Gaussian (asymmetric continuous distribution).

Note: Notice how the exponential family includes distributions of both continuous and discrete random variables.

# The three components of a GLM

In the definition by McCullagh and Nelder (1989), a Generalized Linear Model has three fundamental components:

1) Random Component:
The response variable $Y$ that we wish to model. It is:

- random variable;
- of which we collect $n$ independent observations; and
- with probability distribution in the exponential family (defined later);

Note: the probability distribution of the random response variable $Y$ is no longer restricted to be Normal. It can be any distribution in the exponential family of distributions. Some generalizations of GLMs admit distributions that are not in the exponential family.

# The three components of a GLM (cont.)

2) Systematic Component:
Is a linear combination of predictors, assumed to non-random.

There are $p$ predictors and $n$ observations:

$$\beta_0 + \beta_1 x_{1(i)} + \beta_2 x_{2(i)} + \beta_3 x_{3(i)} + ... + \beta_p x_{p(i)} \quad , \qquad \forall\, i \in \{1, ..., n\} \ .$$

The predictors may be numerical variables, factors or a mixture of both, as in the Linear Model.

We define the model matrix $\mathbf{X}_{n\times(p+1)}$ as in the Linear Model: a first column of ones (associated with the additive constant) and $p$ additional columns given by the observations of each predictor (indicator variables, in the case of factors).

# The three components of a GLM (cont.)

$$\mathbf{X} = \begin{bmatrix} 1 & x_{1_{(1)}} & x_{2_{(1)}} & \cdots & x_{p_{(1)}} \\ 1 & x_{1_{(2)}} & x_{2_{(2)}} & \cdots & x_{p_{(2)}} \\ 1 & x_{1_{(3)}} & x_{2_{(3)}} & \cdots & x_{p_{(3)}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1_{(n)}} & x_{2_{(n)}} & \cdots & x_{p_{(n)}} \end{bmatrix}$$

The vector of the model's systematic component is given by:

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

with $\vec{\boldsymbol{\beta}} = (\beta_0, \beta_1, \beta_2, ..., \beta_p)$ the vector of coefficients defining the $n$ linear (affine) combinations of the predictors, given in $\vec{\boldsymbol{\eta}}$.

# The three components of a GLM (cont.)

3) Link function:
a differentiable and strictly monotone function $g$ which associates the random and systematic components as follows:

$$\begin{aligned} g(\mu_i) &= g(E[Y_i]) = \vec{\mathbf{x}}_{[i]}^t \vec{\boldsymbol{\beta}} \\ &= \beta_0 + \beta_1 x_{1(i)} + \beta_2 x_{2(i)} + \ldots + \beta_p x_{p(i)} \qquad (\forall\, i = 1:n) \end{aligned}$$

where $\vec{\mathbf{x}}_{[i]}^t$ is the $i$-th row of matrix $\mathbf{X}$, containing the values of the predictors for the $i$-th observation.

The expected value of $Y$, given the predictors, is thus:

$$\mu_i = g^{-1}\left(\vec{\mathbf{x}}_{[i]}^t \vec{\boldsymbol{\beta}}\right) = g^{-1}\left(\beta_0 + \beta_1 x_{1(i)} + \beta_2 x_{2(i)} + \ldots + \beta_p x_{p(i)}\right)$$

# The exponential family of distributions

McCullagh & Nelder (1989) define a random variable $Y$ has having distribution in the (2-parameter) exponential family, if its density function (for continuous $Y$) or probability mass function (if $Y$ discrete) can be written as:

$$f(\ y \mid \theta, \phi\ ) \quad = \quad \mathrm{e}^{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)}$$

where

- $\theta$ and $\phi$ are parameters (real scalars); and
- $a(\cdot), b(\cdot)$ and $c(\cdot)$ are known real functions.

The parameters are called:

- $\theta$ – natural parameter; and
- $\phi$ – dispersion parameter.

# Normal (Gaussian)

The exponential family includes the Normal distribution, with density function:

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

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

which is of the form $f(y\,|\,\theta,\phi) = e^{\frac{y\theta-b(\theta)}{a(\phi)}+c(y,\phi)}$, with:

- $\theta = \mu$ (natural parameter)
- $\phi = \sigma^2$ (dispersion parameter)
- $b(\theta) = \frac{\theta^2}{2} = \frac{\mu^2}{2}$
- $a(\phi) = \phi = \sigma^2$
- $c(y,\phi) = -\ln(\sqrt{2\pi\phi}) - \frac{y^2}{2\phi} = -\ln(\sigma\sqrt{2\pi}) - \frac{y^2}{2\sigma^2}$

# Bernoulli

The (binary) random variable $Y$ is said to be a Bernoulli with parameter $p$, if it takes value 1 with probability $p$ and value 0 with probability $1-p$.

For $y=0$ or $y=1$, the Bernoulli probability mass function can be written as:

$$\begin{aligned} f(y|p) &= p^y(1-p)^{1-y} = e^{\ln p^y}\cdot e^{\ln(1-p)^{(1-y)}} = e^{y\ln p}\cdot e^{(1-y)\ln(1-p)} \\ &= e^{y\ln p+\ln(1-p)-y\ln(1-p)} = e^{y\ln\left(\frac{p}{1-p}\right)+\ln(1-p)} \end{aligned}$$

which is in the exponential family $f(y\,|\,\theta,\phi) = e^{\frac{y\theta-b(\theta)}{a(\phi)}+c(y,\phi)}$, with:

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

# GLMs for binary response variables

Consider a Model with binary response variable, i.e., $Y$ can only take two possible values: 0 and 1, with a Bernoulli distribution:

$$Y = \begin{cases} 1 & , \quad p \\ 0 & , \quad 1-p \end{cases}$$

We assume that the parameter $p$ varies in the $n$ observations of $Y$.

The expected value of the $i$-th observation of $Y$ is (also) the probability of success:

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

Any link function relates the probability of success $p_i$ with a linear combination of the predictors:

$$g(E[Y_i]) = g(p_i) = \vec{\mathbf{x}}_{[i]}^t \vec{\boldsymbol{\beta}} \qquad \Longleftrightarrow \qquad E[Y_i] = p_i = g^{-1}\left(\vec{\mathbf{x}}_{[i]}^t \vec{\boldsymbol{\beta}}\right) ,$$

com $\vec{\mathbf{x}}_{[i]}^t \vec{\boldsymbol{\beta}} = \beta_0 + \beta_1 x_{1_i} + \beta_2 x_{2_i} + \ldots + \beta_p x_{p_i}$.

# Link functions

The simplest link function is the identity link: $g(\mu) = \mu$, which is the link function used in the Linear Model.

Each distribution in the exponential family, has a link function that turns the expected value of $Y$ into the natural parameter, $\theta$.

## Canonical link function

In a Generalized Linear Model, the function $g(\cdot)$ is said to be the canonical link function for response variable $Y$, if $g(E[Y]) = \theta$.

Canonical link functions are useful because they simplify the study of the Model. Canonical links are in a sense the "natural" link function for each type of distribution of the response variable.

# The Linear Model as a GLM

Here are some examples of GLMs:

## The Linear Model

The Linear Model is a specific instance of a GLM, for which:

- each of the $n$ observations of the response variable $Y$ has a Normal distribution, with common variance $\sigma^2$;
- the link function is the identity function, since we assume:

$$\mu = E[Y] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p \ .$$

The identity link function is the canonical link function for a Normal distribution, since $\theta = \mu = E[Y]$.

# Logistic Regression

## Logistic Regression

The canonical link function of a Bernoulli distribution is the *logit* function, that transforms $p = E[Y]$ into the natural parameter $\theta = \ln\left(\frac{p}{1-p}\right)$:

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

A GLM for binary response variables, with the *logit* link function is called a Logistic Regression.

The *logit* link function is the logarithm of the ratio of the probability of $Y$ taking value 1 ("success") and the probability of value 0 ("failure"). This ratio is called the *odds ratio*. The *logit* link function is called the log-odds ratio.

# Logistic Regression (cont.)

Given the values $\vec{\mathbf{x}} = (x_1, x_2, ..., x_p)$ for the predictors, this model assumes:

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

$$\Leftrightarrow \quad \frac{p}{1-p} = \mathrm{e}^{\vec{\mathbf{x}}^t \vec{\boldsymbol{\beta}}} \quad \Leftrightarrow \quad \frac{1-p}{p} = \mathrm{e}^{-\vec{\mathbf{x}}^t \vec{\boldsymbol{\beta}}} \quad \Leftrightarrow \quad \frac{1}{p} = 1 + \mathrm{e}^{-\vec{\mathbf{x}}^t \vec{\boldsymbol{\beta}}} \quad \Leftrightarrow \quad p = \frac{1}{1 + \mathrm{e}^{-\vec{\mathbf{x}}^t \vec{\boldsymbol{\beta}}}}.$$

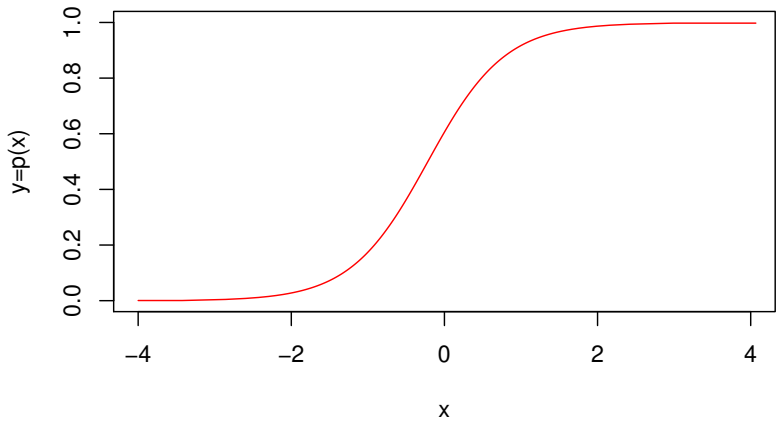
In a Logistic Regression, the relation between the probability of success, $p = E[Y]$, and the values of the predictors, $\vec{\mathbf{x}}$, is:

$$p(\vec{\mathbf{x}}) = g^{-1}\left(\vec{\mathbf{x}}^t \vec{\boldsymbol{\beta}}\right) = \frac{1}{1 + \mathrm{e}^{-\vec{\mathbf{x}}^t \boldsymbol{\beta}}} = \frac{1}{1 + \mathrm{e}^{-(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p)}}$$

# Logistic Regression (cont.)

If we have a single numerical predictor, the relation between $Y$ and $X$ is a logistic curve, which gives rise to the name Logistic Regression.

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



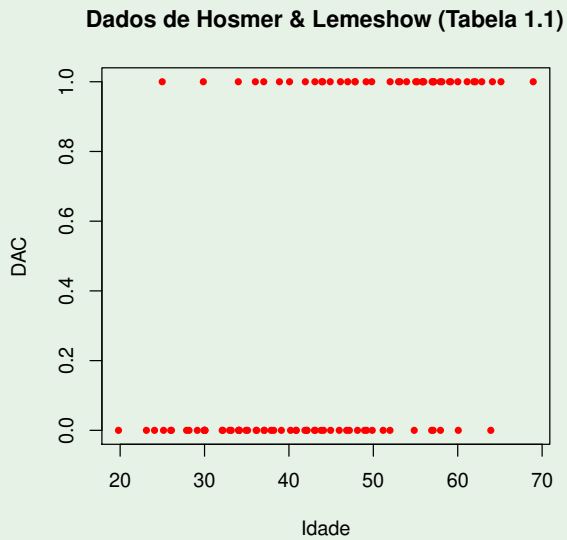

It is an increasing function, if $\beta_1 > 0$ and a decreasing function if $\beta_1 < 0$.

With more than one predictor, $p(\vec{\mathbf{x}}) = \frac{1}{1+e^{-(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)}}$ defines a hypersurface in $\mathbb{R}^{p+1}$.

# Again the DAC example

## Hosmer & Lemeshow data

- age – numerical;
- coronary heart disease – binary variable (yes/no; 1/0).




The response variable is binary. We must relate $p = E[Y]$, the probability of having coronary heart disease, with age $X$.

# Example: The link function

## Hosmer & Lemeshow Data (cont.)

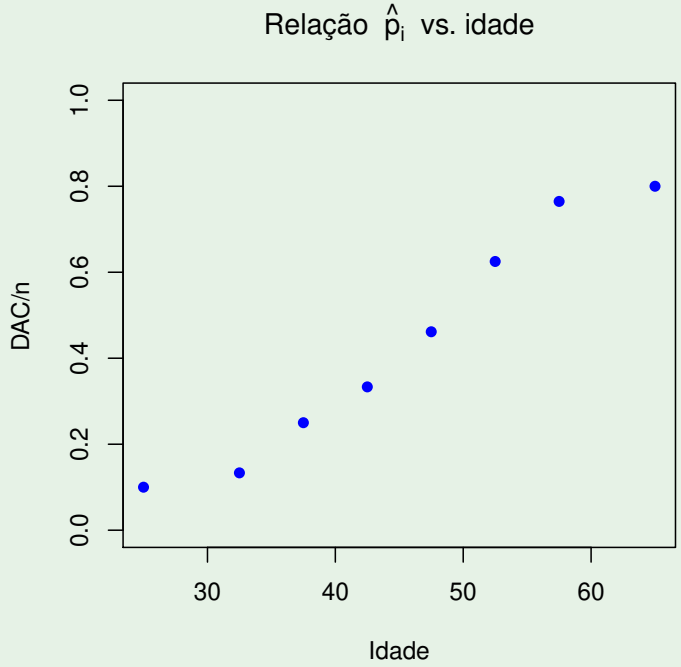
In order to find a suitable link function, we need a different plot, to see the relation between age and the probability of DAC.

- With repetitions for each age, we can estimate $p_i$ from the relative frequency of DAC at age $i$;
- With no (or few) repetitions at each age, we can group the observations in age classes.

| Class | $n_i$ | 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 |

# Example: $\hat{p}_i$ *vs.* age

This is the plot of estimated probabilities vs. age:




Note: Here “age” is the midpoint of each age class.

We have a sigmoidal relation. Maybe a logistic?

# Example 1: creating the frequency table in 

To create the table on slide 24, we used function `hist`, that allows the grouping of ages in age classes. With argument `plot=FALSE`, instead of drawing a histogram, the command returns the information to be used. Argument `right=FALSE` provides class intervals open on the right.

We begin by defining the class borders.

```
> frclass <- c(20,30,35,40,45,50,55,60,70)
> hist(HL$Idade, breaks=frclass, plot=FALSE, right=FALSE)
$breaks     <--  class borders
[1] 20 30 35 40 45 50 55 60 70
$counts     <-- absolute frequencies
[1] 10 15 12 15 13  8 17 10
$density    <-- histogram rectangle heights
[1] 0.010 0.030 0.024 0.030 0.026 0.016 0.034 0.010
$mids       <-- midpoints for each class
[1] 25.0 32.5 37.5 42.5 47.5 52.5 57.5 65.0
[...]
```

# Example 1: the frequency table (cont.)

Select components `counts` and `mids` to create the table columns:

```
> info <- hist(HL$Idade, breaks=frclass, plot=FALSE, right=FALSE)
> HL.tab <- data.frame(idade=info$mids, nobs=info$counts)
> HL.tab
  idade nobs
1  25.0   10
2  32.5   15
3  37.5   12
4  42.5   15
5  47.5   13
6  52.5    8
7  57.5   17
8  65.0   10
```

Still missing is the column with the number of patients with coronary heart disease (DAC) within each age class. Create it by repeating commands, but selecting only the rows of `HL` where `DAC` is 1.

# Example 1: the grouped data

The remaining table columns, including row names:

```
> HL.tab$DAC <-        <-- creats a new column called DAC
+     hist(HL$Idade[HL$DAC==1], breaks=frclass, plot=F, right=F)$counts
> rownames(HL.tab) <-    <-- gives names to table rows
+           paste("[" , frclass[-9] , "," , frclass[-1] , "[" , sep="")
> HL.tab
```

| | idade | nobs | 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 |

The plot on slide 25 was created with the command:

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

# Binary response and Binomial

Grouping in age classes transforms the Bernoulli response variable $Y_i$ (1/0) into a response variable $Y_j$ that counts, in age class $j$, the number of "successes" (ones) in the $n_j$ Bernoulli trials for that class.

For independent observations, $Y_j$ has a Binomial distribution: $Y_j \frown B(n_j, p_j)$, where $p_j$ is the probability of "success" in class $j$.

The Binomial distribution does not belong to the exponential family. But the proportion of successes in $n$ Bernoulli trials, $W = Y/n$ (with known $n$) does.

There are close connections, in GLMs, between having:

- $n$ observations of Bernoulli response variables, with parameters $p_i$; and
- $m$ observations of 'Binomial/n' response variables $W_j = \frac{Y_j}{n_j}$, with $Y_j \frown B(n_j, p_j)$.

Bernoulli and "Binomial/n" can be seen as alternative formulations of similar situations. The canonical link function for both Bernoulli and Binomial/n distributions is the *logit* function:

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

# 'Binomial/n'

$X \frown B(n,p)$, if $P[X=x] = \binom{n}{x} p^x (1-p)^{n-x}$.

$Y = \frac{1}{n}X$ has distribution in the exponential family:

We have $P[Y=y] = P[X=ny]$. The probability mass function of $Y$ can be re-written as follows, for $y \in \{0, \frac{1}{n}, \frac{2}{n}, ..., 1\}$:

$$\begin{aligned} f(y|p) &= \binom{n}{ny} p^{ny}(1-p)^{n-ny} = e^{\ln\left[\binom{n}{ny}\right]} \cdot e^{ny \ln p} \cdot e^{(n-ny)\ln(1-p)} \\ &= e^{\ln\left[\binom{n}{ny}\right] + ny \ln p + n\ln(1-p) - ny\ln(1-p)} = e^{\frac{y\ln\left(\frac{p}{1-p}\right)+\ln(1-p)}{\frac{1}{n}} + \ln\left[\binom{n}{ny}\right]} \end{aligned}$$

It belongs to the exponential family $f(y\,|\,\theta,\phi) = e^{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)}$, with:

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

# GLMs in R

In `R`, the command to fit Generalized Linear Models is `glm`.

There are three fundamental arguments for this command:

formula indicates the random component (response variable) and the systematic component (predictors), in the same way as for linear models:

$$y \sim x1 + x2 + x3 + ... + xp$$

family indicates both the probability distribution of the random component $Y$ and the model's link function.

data indicates the *data frame* where the variables can be found.

# GLMs in R (cont.)

The specification of the probability distribution of $Y$ is done with a keyword, following the name of the argument `family`.

For example, a model with Bernoulli or Binomial/n random component, is specified as:

```
family = binomial
```

By default, the canonical link function of the distribution is used.

To use a different link function (among those implemented) we must add, between parentheses, the argument `link` and the specification of the link function.

For example, a probit model may be indicated as follows:

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

# Example: fitting the model

Thus, a GLM is fitted in R by invoking the command `glm` with three arguments:

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

In a Logistic Regression,

- `family=binomial`.

  It is not necessary to specify the link function: by default, the canonical link of the specified distribution is used.

- we can supply the data in one of 2 formats:
  - individual binary observations (as in the *data frame* `HL`);
  - observations grouped by repeated predictor values (as in the *data frame* `HL.tab`).

# The formula for a Logistic Regression

The formula in the `glm` command is similar to that in a Linear Model:

$$y \quad \sim \quad x1 + x2 + ... + xp$$

But in a Logistic Regression, the two types of dataset correspond to two different *y* objects:

- If the data is provided as individual observations, *y* is a vector of 0/1s:

  > glm(DAC ∼ Idade , family=binomial , data=HL)

- If the dataset has tabled data, *y* must be a 2-column matrix: one column containing the number of "yes"s and another the number of "no"s, for each predictor value (or class midpoint):

  > glm(cbind(DAC,nobs-DAC) ∼ idade , family=binomial, data=HL.tab)

# The Hosmer & Lemeshow example

## Hosmer & Lemeshow data (cont.)

Fitting the model using the individual binary observations:

> glm(DAC ~ Idade , family=binomial, data=HL)

```
Call:  glm(formula = DAC ~ Idade , family = binomial , data = HL)
Coefficients:
(Intercept)         Idade
    -5.3095        0.1109          <---- estimated parameters
```

The fitted logistic equation is:

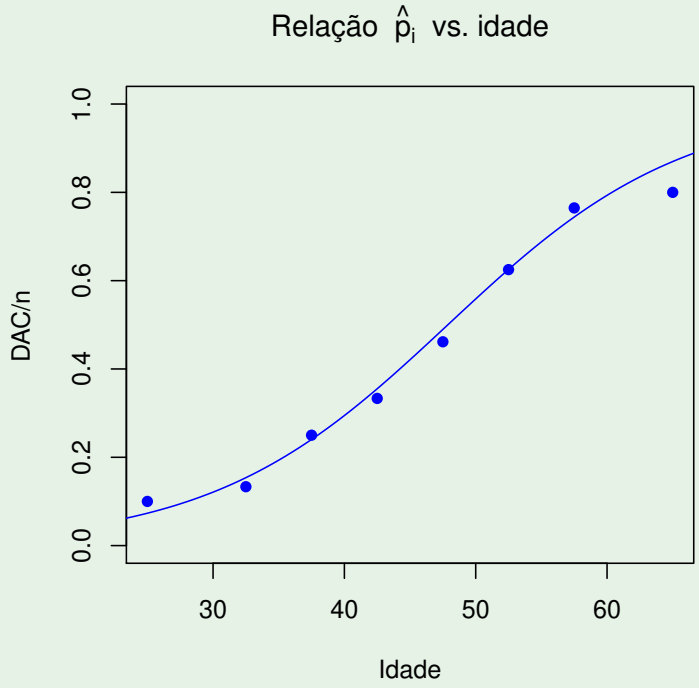
$$y = \frac{1}{1+\mathrm{e}^{-(b_0+b_1 x)}} = \frac{1}{1+\mathrm{e}^{-(-5.3095+0.1109x)}}$$

# Example (cont.)

## Hosmer & Lemeshow data (cont.)

Drawing the fitted logistic on the plot of $\hat{p}_i$ *vs.* age:

```
> 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)
```



# Example: fitting the model (cont.)

## Hosmer & Lemeshow data (cont.)

Fitting the model using the tabled (grouped) data:

> glm( cbind(DAC,nobs-DAC) ~ idade , family=binomial, data=HL.tab)

```
Call: glm(formula=cbind(DAC,nobs-DAC)~idade, family=binomial,data=HL.tab)
Coefficients:
(Intercept)        idade
     -5.091        0.105          <----  estimated parameters
```

The fitted logistic equation is:

$$y = \frac{1}{1+\mathrm{e}^{-(b_0+b_1 x)}} = \frac{1}{1+\mathrm{e}^{-(-5.091+0.105x)}}$$

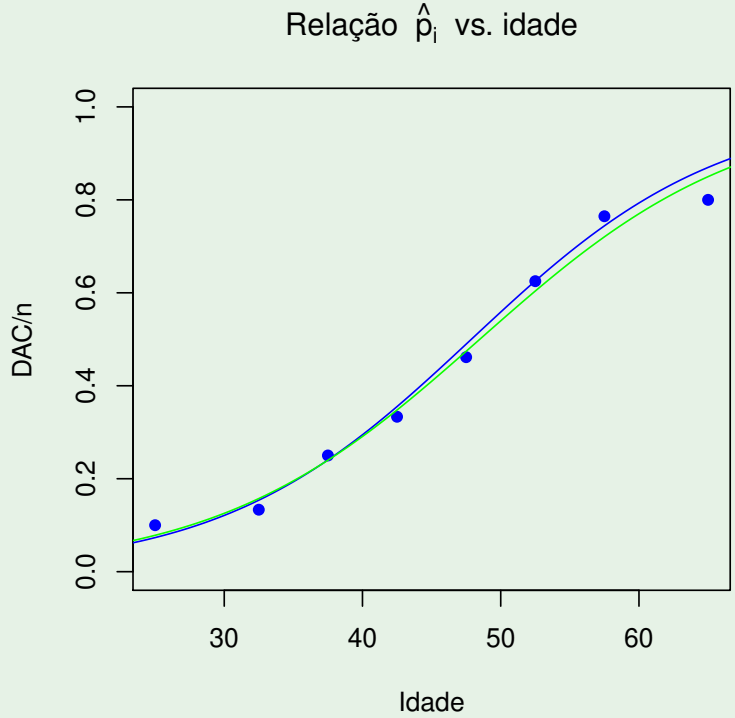
Note: The small differences with the results of the previous fit are due to the grouping in age classes: the data are different.

# Example: fitting the model (cont.)

## Hosmer & Lemeshow data (cont.)

Drawing the fitted logistic on the $\hat{p}_i$ *vs.* age plot (retaining the previous curve):

```
> curve(logistica(b0=-5.091, b1=0.105, x), from=20, to=70, col="green", add=TRUE)
```



# The `glm` command output

As with command `lm`, so too the `glm` command produces a *list*. The list components contain information regarding the fit.

```
> HLtab.glm <- glm(cbind(DAC,nobs-DAC)~idade,family=binomial,data=HL.tab)
> names(HLtab.glm)

 [1] "coefficients"      "residuals"         "fitted.values"     "effects"
 [5] "R"                 "rank"              "qr"                "family"
 [9] "linear.predictors" "deviance"          "aic"               "null.deviance"
[13] "iter"              "weights"           "prior.weights"     "df.residual"
[17] "df.null"           "y"                 "converged"         "boundary"
[21] "model"             "call"              "formula"           "terms"
[25] "data"              "offset"            "control"           "method"
[29] "contrasts"         "xlevels"
```

For more information on each component, check: `help(glm)`

We can invoke a component in the usual way:

## Hosmer & Lemeshow data (cont.)

```
> HLtab.glm$coef

(Intercept)       idade
 -5.0907332   0.1050191
```

# The command `coef`

As for Linear Models, there are commands to extract information from a fitted GLM. Among them:

`coef` – produces a vector with the estimated values of the parameters $\beta_0$, $\beta_1$, ...,$\beta_p$, i.e., values $b_0$, $b_1$, ...., $b_p$:

## Hosmer & Lemeshow data (cont.)

```
> HL.glm <- glm(DAC ~ Idade, family=binomial, data=HL)
> coef(HL.glm)

 (Intercept)        idade
  -5.3094534    0.1109211
```

# The command `predict`

`predict` – by default, the command returns the values of the estimated linear combination of the predictor values used to fit the data, that is, the systematic component values $b_0 + b_1 x_{1_{(i)}} + \ldots + b_p x_{p_{(i)}}$.

## Hosmer & Lemeshow data (cont.)

> predict(HLtab.glm)

```
   [20,30[    [30,35[    [35,40[    [40,45[    [45,50[    [50,55[    [55,60[    [60,70[
-2.4652550 -1.6776115 -1.1525158 -0.6274202 -0.1023245  0.4227711  0.9478668  1.7355102
```

> predict(HL.glm)

```
         1          2          3          4          5          6          7
-3.09103053 -2.75826710 -2.64734596 -2.53642482 -2.53642482 -2.42550368 -2.42550368
.....
        99         100
 1.90042087  2.34410544
```

# The command `predict` (cont.)

`predict` can also be used to estimate a linear combination of values not used to fit the model.

The new values are given in a *data frame*, with names equal to those used in the original dataset.

## Hosmer & Lemeshow data (cont.)

```
> predict(HL.glm, newdata=data.frame(Idade=26))

        1
-2.425504

> predict(HLtab.glm, newdata=data.frame(idade=c(26,53,74)))

         1          2          3
-2.3602358  0.4752807  2.6806824
```

# The command `fitted`

`fitted` – returns the fitted expected values of $Y_i$, that is, $\hat{p}_i = g^{-1}(b_0 + b_1 x_{1_{(i)}} + ... + b_p x_{p_{(i)}})$.

```
> fitted(HLtab.glm)

   [20,30[    [30,35[    [35,40[    [40,45[    [45,50[    [50,55[    [55,60[    [60,70[
0.07833012 0.15741201 0.24002985 0.34809573 0.47444116 0.60414616 0.72068596 0.85011588
```

A similar result is obtained with the `predict` command, using argument `type="response"`:

```
> predict(HLtab.glm, type="response")

   [20,30[    [30,35[    [35,40[    [40,45[    [45,50[    [50,55[    [55,60[    [60,70[
0.07833012 0.15741201 0.24002985 0.34809573 0.47444116 0.60414616 0.72068596 0.85011588
```

Thus, we can estimate $\hat{p}$ for new predictor values.

```
> predict(HLtab.glm, newdata=data.frame(idade=c(26,53,74)),type="response")

        1         2         3
0.0862556 0.6166329 0.9358771
```

# Comments about Logistic Regression

- The logistic function has good properties to represent a probability: for *any* value of the systematic component, the values of the logistic function

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

  are between 0 and 1. The same does not happen with a linear relation $p(x_1, ..., x_p) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$, whose values are in $\mathbb{R}$.

- In the case of a single numerical predictor, exchanging the events associated with values 0 and 1, a decreasing function for $p = P[Y = 1]$ becomes an increasing function.

# More comments about Logistic Regressions

In the case of a single numerical predictor, parameter $\beta_1$ has the following interpretation:

- since

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

each increase by one unit in the predictor variable $X$ translates into a multiplicative effect on the odds ratio, of $e^{\beta_1}$:

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

- equivalently, it induces and additive effect of $\beta_1$ units on the log-odds ratio:

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

# More comments on a Logistic Regression

When there is more than one numerical predictor:

- the *logit* link function produces a logistic relation for the probability of success $p$, as a function of the values of the systematic component $\eta$ (linear combination of the predictors).
- the interpretation of the coefficients $\beta_j$ can be generalised: an increase of one unit in predictor $j$ (with other predictors fixed) is associated with a multiplication of the odds ratio by the factor $e^{\beta_j}$.

For categorical predictors (factors),

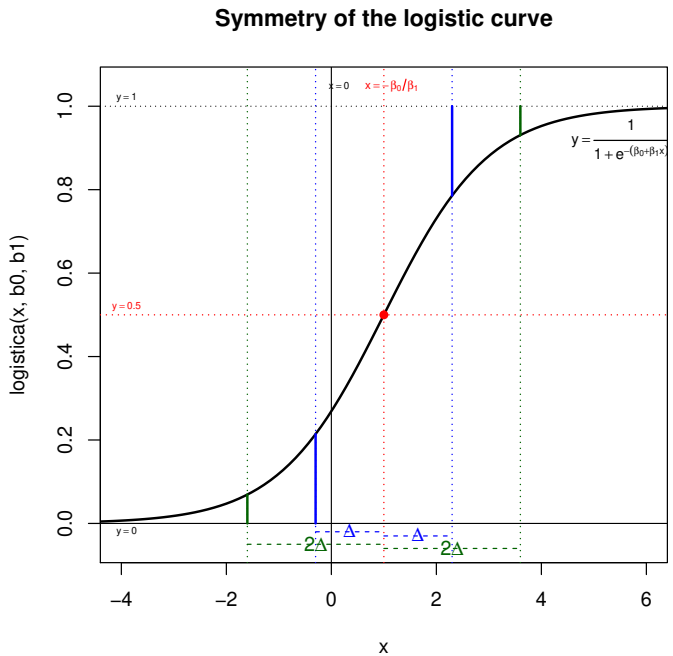
- Let $\vec{\mathscr{I}}_j$ be an indicator variable. The corresponding parameter $\beta_j$ indicates the increase in the log-odds ratio that results from an observation belonging to the category associated with the indicator $\vec{\mathscr{I}}_j$.

# Characteristics of a logistic curve

The logistic may be too rigid. With a single numerical predictor, the logistic $f(x) = \left[1 + \mathrm{e}^{-(\beta_0 + \beta_1 x)}\right]^{-1}$ has a point of inflection corresponding to probability $p = 0.5$, the curve being symmetric around this point:

$$f''(x) = 0 \quad \Leftrightarrow \quad \mathrm{e}^{-(\beta_0+\beta_1 x)} = 1 \quad \Leftrightarrow \quad x = -\frac{\beta_0}{\beta_1} \quad \Rightarrow \quad f\left(-\frac{\beta_0}{\beta_1}\right) = \frac{1}{2} \, .$$

$$f\left(-\frac{\beta_0}{\beta_1} - \Delta\right) = \frac{1}{1 + \mathrm{e}^{\beta_1 \Delta}} = 1 - f\left(-\frac{\beta_0}{\beta_1} + \Delta\right)$$



# Estimation of parameters in GLMs

The estimation of parameters $\beta_j$ in Generalised Linear Models is based on the Maximum Likelihood Method.

The fact that estimation is based on the likelihood function implies that, unlike what happens with the Linear Model, in GLMs distributional assumptions are crucial for parameter estimation.

Since the distributions considered in GLMs belong to the exponential family of distributions, there are some specificities in the estimation.

# Likelihood in the exponential family

The likelihood function for $n$ independent observations, $y_1, y_2, ..., y_n$, in a distribution of the exponential family is:

$$\mathbf{L}(\vec{\theta}, \vec{\phi}\ ;\ y_1, y_2, ..., y_n) = \prod_{i=1}^{n} f(y_i; \theta_i, \phi_i) = e^{\sum_{i=1}^{n}\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)} + c(y_i, \phi_i)\right]}$$

Maximising likelihood is equivalent to maximising log-likelihood:

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

Where are the model parameters, $\beta_j$?

# Maximum Likelihood in GLMs

In a GLM, the systematic component and the expected value of the response variable are related by $g(E[Y]) = \vec{\mathbf{x}}^t\vec{\boldsymbol{\beta}}$.

When using a canonical link function, we have $\theta = \vec{\mathbf{x}}^t\vec{\boldsymbol{\beta}}$.

In general, we can write the log-likelihood as a function of the model parameters: $\mathscr{L}(\vec{\boldsymbol{\beta}})$.

Estimating the parameters by maximum likelihood means choosing the vector $\vec{\boldsymbol{\beta}}$ that maximises the log-likelihood function $\mathscr{L}(\vec{\boldsymbol{\beta}})$.

# Maximum Likelihood in GLMs (cont.)

A necessary condition to maximise the function of $p+1$ variables $\mathscr{L}(\vec{\boldsymbol{\beta}})$ is:

$$\frac{\partial \mathscr{L}(\vec{\boldsymbol{\beta}})}{\partial \beta_j} = 0, \qquad \forall\, j = 0 : p\ ,$$

assuming that the functions $a(\cdot)$, $b(\cdot)$ e $c(\cdot)$ are sufficiently regular so that all the partial derivatives exist.

In the case of a generic Generalised Linear Model, there is no guarantee that the log-likelihood function has a maximum (with admissible values of the parameters $\vec{\boldsymbol{\beta}}$), nor that, if maxima exist, they are unique.

For the specific GLMs covered in this course, this is not a problem.

# Example: the Logistic Regression case

In the Logistic Regression Model, with $n$ independent observations from Bernoulli distributions, we have the following likelihood function:

$$\mathbf{L}(\vec{\mathbf{p}}\ ;\ \vec{\mathbf{y}}) = \prod_{i=1}^{n} \mathrm{e}^{\ln(1-p_i)+y_i \ln\left(\frac{p_i}{1-p_i}\right)}$$

and the log-likelihood is:

$$\mathscr{L}(\vec{\mathbf{p}}\ ;\ \vec{\mathbf{y}}) = \sum_{i=1}^{n} \left[\ln(1-p_i) + y_i \ln\left(\frac{p_i}{1-p_i}\right)\right]$$

Since the link function is $g(p) = \ln\left(\frac{p}{1-p}\right) = \vec{\mathbf{x}}^t\vec{\boldsymbol{\beta}}$, the log-likelihood is the following function of the parameters $\vec{\boldsymbol{\beta}}$:

$$\mathscr{L}(\vec{\boldsymbol{\beta}}\ ;\ \vec{\mathbf{y}}) \ = \ \sum_{i=1}^{n} \left[-\ln\left(1+\mathrm{e}^{\vec{\mathbf{x}}_i^t\vec{\boldsymbol{\beta}}}\right) + y_i\,\vec{\mathbf{x}}_i^t\vec{\boldsymbol{\beta}}\right]$$

# Estimation in a Logistic Regression (cont.)

We have:

$$\mathscr{L}(\vec{\boldsymbol{\beta}}) = \sum_{i=1}^{n}\left(\beta_0 y_i + \sum_{k=1}^{p} y_i\, x_{k(i)}\beta_k\right) - \sum_{i=1}^{n} \ln\left(1+\mathrm{e}^{\beta_0+\Sigma_{k=1}^{p} x_{k(i)}\beta_k}\right)$$

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

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

These $p+1$ normal equations form a non-linear system of equations in $p+1$ unknowns $\hat{\beta}_j$ ($j = 0:p$).

# Estimation algorithms

In general, the system of $p+1$ normal equations associated with the maximization of the log-likelihood function in a Generalised Linear Model is a non-linear system:

$$\frac{\partial \mathscr{L}(\vec{\boldsymbol{\beta}})}{\partial \beta_j} = 0 \qquad , \qquad \forall\, j = 0 : p.$$

Numerical algorithms used to solve these equations for GLMs are adaptations of the Newton-Raphson algorithm, which are known by various names: Iterative Weighted Least Squares, IWLS (or Reweighted, IRLS), or Fisher Scoring Method.

The Newton-Raphson method works with a second order approximation (using Taylor's formula) of the log-likelihood function, applied at a point that is an initial estimate of the vector $\vec{\boldsymbol{\beta}}$.

# Estimation algorithms (cont.)

## Newton-Raphson Method

Let:

- $\vec{\boldsymbol{\beta}}^{[0]}$ be an initial estimate of $\vec{\boldsymbol{\beta}}$;
- $\vec{\nabla}\mathscr{L}_{\vec{\boldsymbol{\beta}}}$ be the gradient vector of $\mathscr{L}$ (vector of partial derivatives $\frac{\partial \mathscr{L}}{\partial \beta_j}$) in $\vec{\boldsymbol{\beta}}$;
- $\mathscr{H}_{\vec{\boldsymbol{\beta}}}$ be the Hessian matrix of partial derivatives of second order of function $\mathscr{L}(\cdot)$

The second order approximation given by Taylor's formula is:

$$\mathscr{L}(\vec{\boldsymbol{\beta}}) \approx \mathscr{L}_*(\vec{\boldsymbol{\beta}}) = \mathscr{L}(\vec{\boldsymbol{\beta}}^{[0]}) + (\vec{\nabla}\mathscr{L}_{\vec{\boldsymbol{\beta}}^{[0]}})^t(\vec{\boldsymbol{\beta}} - \vec{\boldsymbol{\beta}}^{[0]}) + \frac{1}{2}(\vec{\boldsymbol{\beta}} - \vec{\boldsymbol{\beta}}^{[0]})^t \mathscr{H}_{\vec{\boldsymbol{\beta}}^{[0]}}(\vec{\boldsymbol{\beta}} - \vec{\boldsymbol{\beta}}^{[0]})$$

Instead of maximising $\mathscr{L}(\vec{\boldsymbol{\beta}})$, we maximise the approximation $\mathscr{L}_*(\vec{\boldsymbol{\beta}})$.

# Estimation algorithms (cont.)

Computing gradient vectors of inner products or quadratic forms is simple:

If $h(\vec{\mathbf{x}}) = \vec{\mathbf{a}}^t\vec{\mathbf{x}}$ , we have $\frac{\partial h(\vec{\mathbf{x}})}{\partial \vec{\mathbf{x}}} = \frac{\partial(\vec{\mathbf{a}}^t\vec{\mathbf{x}})}{\partial \vec{\mathbf{x}}} = \vec{\mathbf{a}}$.

If $h(\vec{\mathbf{x}}) = \vec{\mathbf{x}}^t\mathbf{A}\vec{\mathbf{x}}$ , we have $\frac{\partial h(\vec{\mathbf{x}})}{\partial \vec{\mathbf{x}}} = \frac{\partial(\vec{\mathbf{x}}^t\mathbf{A}\vec{\mathbf{x}})}{\partial \vec{\mathbf{x}}} = 2\mathbf{A}\vec{\mathbf{x}}$.

Thus,

$$\nabla\vec{\mathscr{L}}_{*_{\vec{\boldsymbol{\beta}}}} = \nabla\vec{\mathscr{L}}_{\vec{\boldsymbol{\beta}}^{[0]}} + \mathscr{H}_{\vec{\boldsymbol{\beta}}^{[0]}}(\vec{\boldsymbol{\beta}} - \vec{\boldsymbol{\beta}}^{[0]}).$$

Assuming that $\mathscr{H}_{\vec{\boldsymbol{\beta}}^{[0]}}$ has an inverse, we have:

$$\nabla\vec{\mathscr{L}}_{*_{\vec{\boldsymbol{\beta}}}} = \vec{\mathbf{0}} \quad \Leftrightarrow \quad \vec{\boldsymbol{\beta}} = \vec{\boldsymbol{\beta}}^{[0]} - \mathscr{H}^{-1}_{\vec{\boldsymbol{\beta}}^{[0]}} \cdot \nabla\vec{\mathscr{L}}_{\vec{\boldsymbol{\beta}}^{[0]}}.$$

The Newton-Raphson algorithm iterates this relation.

# Estimation algorithms (cont.)

Take:

$$\vec{\boldsymbol{\beta}}^{[i+1]} = \vec{\boldsymbol{\beta}}^{[i]} - \mathscr{H}^{-1}_{\vec{\boldsymbol{\beta}}^{[i]}} \cdot \vec{\nabla}\mathscr{L}_{\vec{\boldsymbol{\beta}}^{[i]}}$$

Notes:

- Successfully applying this algorithm requires the existence and invertibility of the Hessian matrices of $\mathscr{L}$ at sucessive points $\vec{\boldsymbol{\beta}}^{[i]}$;
- The algorithm's convergence is not guaranteed, even if the log-likelihood function has a single maximum;
- If there is a single maximum, convergence tends to improve when the initial estimate $\vec{\boldsymbol{\beta}}^{[0]}$ is close to the maximum.
- There may be several local maxima and a bad choice of $\vec{\boldsymbol{\beta}}^{[0]}$ can then lead to convergence to a suboptimal solution.

# Estimation algorithms (cont.)

## Fisher Scoring Method

Computing the Hessian matrices of the log-likelihood function at points $\vec{\boldsymbol{\beta}}^{[i]}$ is computationally demanding.

The Fisher Scoring Method modifies the Newton-Raphson algorithm, by replacing the Hessian matrix with Fisher's information matrix, defined as minus the expected Hessian matrix:

$$\mathbf{I}_{\vec{\boldsymbol{\beta}}^{[i]}} = -E\left[\mathscr{H}_{\vec{\boldsymbol{\beta}}^{[i]}}\right]$$

Thus, the iteration associated with the Fisher Scoring Method is:

$$\vec{\boldsymbol{\beta}}^{[i+1]} = \vec{\boldsymbol{\beta}}^{[i]} + \mathbf{I}^{-1}_{\vec{\boldsymbol{\beta}}^{[i]}} \cdot \vec{\nabla}\mathscr{L}_{\vec{\boldsymbol{\beta}}^{[i]}}$$

# Estimation algorithms (cont.)

For a GLM with canonical link function, the Hessian matrix of the log-likelihood does not depend on the response variable $Y$, so the Hessian and its expected matrix coincide.

Therefore, in this case the Fisher and Newton-Raphson methods coincide.

This is one of the reasons why canonical links are important.

# Probit Regression

Another example of GLM is the probit model of Bliss (1935), very frequent in Toxicology studies.

## *Probit* Model

As in a Logistic Regression, we have:

- a binary response variable (with a Bernoulli distribution).
- a systematic component, given by a linear combination of predictors.

Different from a Logistic Regression is the link function.

# Probit Regression (cont.)

In a Logistic Regression, the link function gives $p = E[Y]$ as a logistic function of the systematic component $\eta = \vec{\mathbf{x}}^t \vec{\boldsymbol{\beta}}$.

In a *Probit* Model, the probability of success and expected value of $Y$, $p = E[Y]$, is given by a different sigmoid curve: the cumulative distribution function (c.d.f.), $\Phi$, of a Standardized Normal:

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

where $\Phi$ is the c.d.f. of a $\mathcal{N}(0,1)$.

This choice means that the link function is the inverse of a c.d.f. of a Standard Normal, that is, $g = \Phi^{-1}$:

$$\vec{\mathbf{x}}^t \vec{\boldsymbol{\beta}} \;=\; \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \;=\; g(E[Y]) \;=\; g\left(p(\vec{\mathbf{x}})\right) \;=\; \Phi^{-1}\left(p(\vec{\mathbf{x}})\right) \;.$$

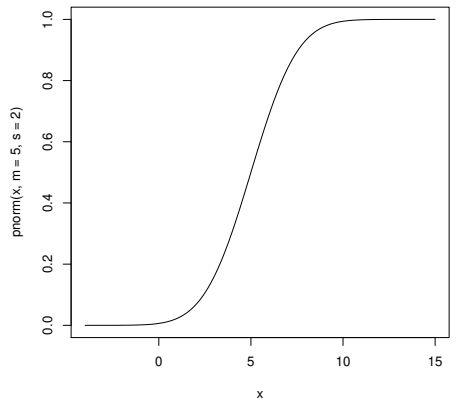
# Probit Regression (cont.)

If there is a single numerical predictor, we have:

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

with $\beta_0 = -\frac{\mu}{\sigma}$ and $\beta_1 = \frac{1}{\sigma}$.

Thus, the probability of success $p$ is related with the predictor variable $X$ by the c.d.f. of a $\mathcal{N}(\mu,\sigma^2)$, with $\sigma = \frac{1}{\beta_1}$ and $\mu = -\frac{\beta_0}{\beta_1}$.



# Probit Regression in toxicology

In a toxicological context, often:

- a predictor $X$ gives the dose (or log-dose) of a given toxic product;
- for each individual there is a tolerance level t: the dose of the toxic above which the individual dies;
- this tolerance level varies between individuals and may be represented by a random variable $T$.

Define a binary random variable $Y$:

$$Y = \begin{cases} 1 & , \quad \text{if individual dies} \\ 0 & , \quad \text{if individual survives} \end{cases}$$

Then:

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

# Probit Regression in toxicology (cont.)

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

Assuming that the tolerance $T$ follows a $\mathcal{N}(\mu, \sigma^2)$ distribution,

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

We have the Probit Model with $X$ as its single predictor.

Since $\beta_0 = -\frac{\mu}{\sigma}$ and $\beta_1 = \frac{1}{\sigma}$ (slide 62), the parameters of the distribution of tolerance $T$ are given by:

$$\mu = -\frac{\beta_0}{\beta_1} \qquad \text{e} \qquad \sigma = \frac{1}{\beta_1} .$$

# Probit Regression in R

We illustrate a Probit Regression in R, with the DAC data, considered before.

## Hosmer & Lemeshow dataset

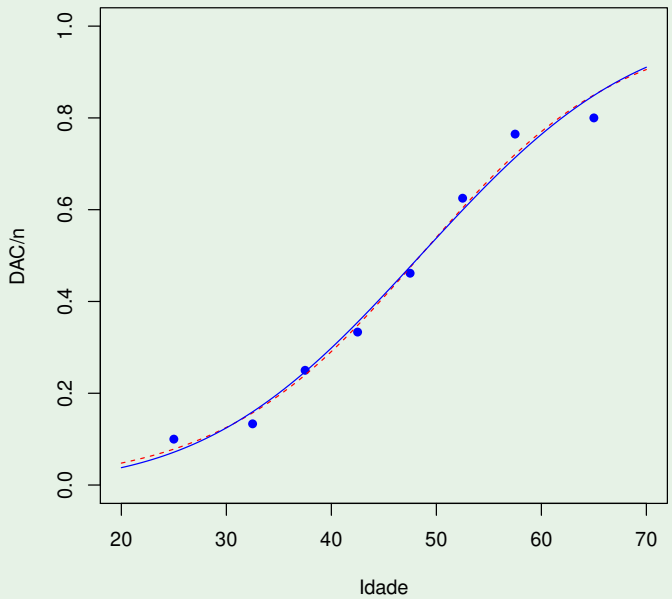
In a Probit regression, it is necessary to specify the link function, using the argument `family`, as follows:

```
> glm(cbind(DAC,nobs-DAC)~idade,family=binomial(link=probit),data=HL.tab)

Call:  glm(formula = cbind(DAC, nobs - DAC) ~ idade,
                family = binomial(link = probit), data = HL.tab)
Coefficients:
(Intercept)         idade
    -3.0245        0.0624
[...]
```

As in a Logistic Regression, the response variable may be a two-column matrix, indicating the number of "successes" and the number of "failures" (as above) or, alternatively, as a vector of zeros and ones.

# Probit Regression in R (cont.)

The equation of the fitted curve for probability of DAC over age ($x$) is: $p(x) = \Phi(-3.0245 + 0.0624\,x)$. Here is the curve, superimposed on the scatterplot (the dashed curve is the fitted logistic):




The curve was drawn with the following command:

```
> curve(pnorm(-3.0245+0.0624*x), add=TRUE, col="blue")
```

# The command update

The *probit* model that was now fitted only differs, in relation to the logistic regression model (HLtab.glm) in the argument family, which specifies the link function.

The R command update is useful in these cases, since it enables us to re-fit a model, changing some arguments.

```
> update(HLtab.glm, family=binomial(link=probit))

Call: glm(formula=cbind(DAC,nobs-DAC)~idade, family=binomial(link=probit), data=HL.tab)
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
```

Note: The object HLtab.glm was not changed. And the new model was not saved.

# Probit Regression (cont.)

For any number of predictors, the Probit Model probability of success $p = P[Y = 1]$ is given by a function with behaviour similar to that of the Logit Model:

- strictly increasing function,
- with a single point of inflection when the linear predictor $\mathbf{x}^t\vec{\boldsymbol{\beta}} = 0$,
- that corresponds to a probability of success $p(0) = 0.5$,
- with symmetry around the point of inflection, that is, $p(-\eta) = 1 - p(\eta)$, for all $\eta$.

Drawbacks:

- no easy interpretation of the meaning of the parameters $\beta_j$;
- the link function is non-canonical.

# Complementary log-log model

## Complementary log-log model

Within the same context of a binary response variable $Y$, a different frequent choice for link function, whose history goes back to 1922, in the study of infectious organisms, is to have the probability of success ($Y = 1$) given by:

$$p(\vec{\mathbf{x}}) \quad = \quad g^{-1}\left(\vec{\mathbf{x}}^t\vec{\boldsymbol{\beta}}\right) \quad = \quad 1 - \mathrm{e}^{-\mathrm{e}^{\vec{\mathbf{x}}^t\vec{\boldsymbol{\beta}}}}$$

The range of this function is the interval $]0,1[$.

The link function will then be:

$$\vec{\mathbf{x}}^t\vec{\boldsymbol{\beta}} \;=\; g\left(p(\vec{\mathbf{x}})\right) \;=\; \ln\left[-\ln(1-p(\vec{\mathbf{x}}))\right]$$

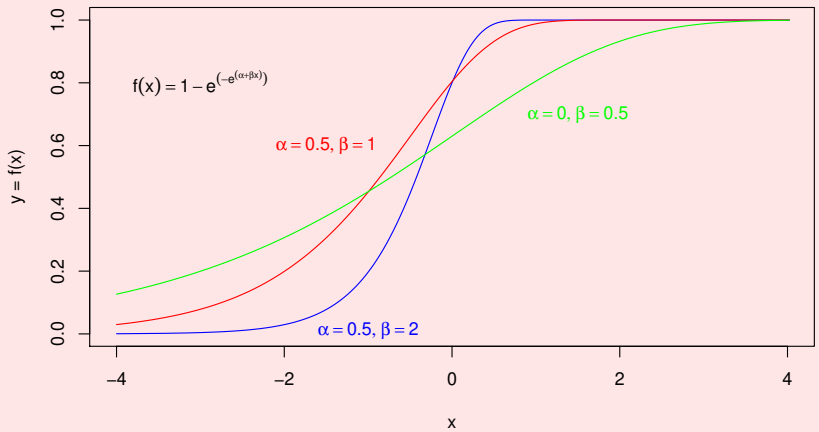
hence the name complementary log-log.

# Complementary log-log model (cont.)

With a single numerical predictor, the function $p$ is 1 minus a Gompertz curve ($y = \alpha e^{-\beta e^{-\gamma x}}$) with asymptote at $\alpha = 1$. That is,

$$p(x) = 1 - e^{-\beta e^{-\gamma x}} .$$

Function $p(x)$ is the cumulative distribution function of a Gumbel distribution:



# The complementary log-log model (cont.)

This function for $p$ has similarities and differences in behaviour, when compared with those for the Logit and Probit Models:

- it too is strictly monotonous;
- it too has a single point of inflection, when $\eta = 0$;
- but the corresponding probability is no longer half-way up the probability scale, but rather at $p(0) = 1 - \frac{1}{\mathrm{e}}$;
- this means that the "initial phase" of the probability curve will occur until a higher probability ($1 - \frac{1}{\mathrm{e}} \approx 0.632$) than in *Logit* and *Probit* Regressions.

As in the Probit Model, the coefficients $\beta_j$ in the systematic component can not be easily interpreted, unlike in a Logistic Regression.

# Complementary log-log model in R

To fit the model with the complementary log-log link function, we give argument `link` the value `cloglog`.

```
> update(HLtab.glm, family=binomial(link=cloglog))


Call: glm(formula=cbind(DAC, nobs-DAC)~idade,
                family=binomial(link=cloglog), data=HL.tab)

Coefficients:
(Intercept)         idade
   -4.00470       0.07311
[...]
```

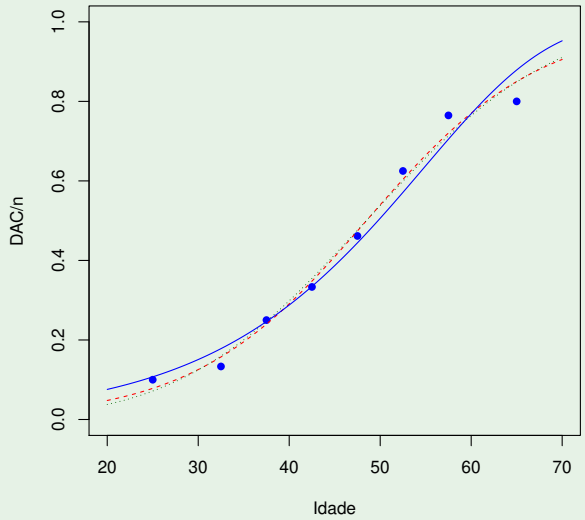
The fitted curve is:

$$p(x) = 1 - \mathrm{e}^{-\mathrm{e}^{-4.00470+0.07311\,x}}.$$

# Complementary log-log model in R (cont.)

The fitted curve, superimposed on the scatterplot for the DAC example, is:

$$p(x) = 1 - \mathrm{e}^{-\mathrm{e}^{-4.00470+0.07311\,x}}$$




The dashed curve is from the `probit` model and the dotted curve is for the logistic regression. The new curve was drawn with the following commands:

```
> cloglog <- function( b0,b1,x ){1-exp(-exp(b0+b1*x))}
> curve(cloglog(b0=-4.0047,b1=0.07311,x), add=TRUE, col="blue")
```

# Other link functions for binary responses

We considered three link functions for models with Bernoulli (or Binomial/n) responses, whose inverses are sigmoidal. In two of those cases, inverses of cumulative distribution functions were used:

- c.d.f. of a Standard Normal, in the Probit Model;
- c.d.f. of a Gumbel, in the Complementary log-log Model.

An obvious generalisation is to use other c.d.f.s of continuous random variables, generating new GLMs for binary responses.

In `R`, besides the two choices discussed before, we can use the c.d.f. of a Cauchy distribution (`link=cauchit`).

# Inference: properties of ML estimators

## Properties of Maximum Likelihood estimators

Maximum Likelihood estimators $\vec{\hat{\beta}}$ are:

- asymptotically Multinormal
- asymptotically unbiased ($E[\vec{\hat{\beta}}] \to \vec{\beta}$).
- asymptotically have variance-covariance matrix $\mathbb{I}_{\vec{\beta}}^{-1}$, where

$$\mathbb{I}_{\vec{\beta}} = -E[\mathscr{H}_{\vec{\beta}}]$$

is the Fisher Information matrix and $\mathscr{H}_{\vec{\beta}}$ is the Hessian matrix of the log-likelihood $\mathscr{L}$, at point $\vec{\beta}$, whose $(j,m)$-th element is:

$$\left(\mathscr{H}_{\vec{\beta}}\right)_{(j,m)} = \frac{\partial^2 \mathscr{L}}{\partial \beta_j \, \partial \beta_m}$$

Conclusion: (asymtpotic) inference can be carried out in GLMs!

# Inference in GLMs

## Asymptotic distribution of $\vec{\hat{\beta}}$ in a GLM

In a Generalised Linear Model, the vector of Maximum Likelihood estimators, $\vec{\hat{\beta}}$, has, asymptotically:

$$\vec{\hat{\beta}} \sim \mathcal{N}_{p+1}\left(\vec{\beta}, \mathbb{I}_{\vec{\beta}}^{-1}\right)$$

where $\mathbb{I}_{\vec{\beta}}$ is the Fisher information matrix for the sample log-likelihood, computed at point $\vec{\beta}$.

The sample size should be large to ensure a good asymptotic approximation.

Notice the similarity with the distributional result that underlies inference in a linear model. The properties of a Multinormal can be used to obtain similar results.

# Inference in GLMs (cont.)

## Distribution of a linear combination of the parameters

Given a GLM (with certain regularity conditions) and a non-random vector $\vec{\mathbf{a}} \in \mathbb{R}^{p+1}$, the Maximum Likelihood estimators $\vec{\hat{\boldsymbol{\beta}}}$ are, asymptotically:

$$\frac{\vec{\mathbf{a}}^t \vec{\hat{\boldsymbol{\beta}}} - \vec{\mathbf{a}}^t \vec{\boldsymbol{\beta}}}{\sqrt{\vec{\mathbf{a}}^t \mathbb{I}_{\beta}^{-1} \vec{\mathbf{a}}}} \sim \mathcal{N}(0,1) \ .$$

This result enables (approximate) confidence intervals and hypotheses tests for linear combinations of the parameters $\vec{\boldsymbol{\beta}}$.

# Inference in GLMs (cont.)

To build CIs and Hypothesis Tests, we must overcome the issue that the inverse of the information matrix is to be calculated at an unknown point $\vec{\boldsymbol{\beta}}$.

This unknown matrix is replaced by a known matrix: the information matrix computed at the estimated $\hat{\vec{\boldsymbol{\beta}}}$.

For distributions with an unknown dispersion parameter $\phi$ (a situation not yet considered), there is an additional problem: estimating $\phi$.

All this strengthens the advisability of working with large samples, so that we may have confidence in the results.

# Inference in GLMs (cont.)

## Confidence Intervals (asymptotic)

An asymptotic $(1-\alpha)\times 100\%$ confidence interval for the linear combination $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}$ is:

$$\left]\ \vec{\mathbf{a}}^t\vec{\mathbf{b}} - z_{\frac{\alpha}{2}}\cdot\sqrt{\vec{\mathbf{a}}^t\mathbb{I}_{\hat{\beta}}^{-1}\vec{\mathbf{a}}}\quad,\quad \vec{\mathbf{a}}^t\vec{\mathbf{b}} + z_{\frac{\alpha}{2}}\cdot\sqrt{\vec{\mathbf{a}}^t\mathbb{I}_{\hat{\beta}}^{-1}\vec{\mathbf{a}}}\ \right[$$

where $\mathbb{I}_{\hat{\beta}}^{-1}$ is the inverse of the Fisher information matrix for the log-likelihood, computed at point $\vec{\hat{\boldsymbol{\beta}}}$.

# Inference in GLMs (cont.)

## Hypothesis Tests (asymptotic)

In a GLM, an (asymptotic) two-sided Hypothesis Test for a linear combination of the $\beta_j$ is:

- Hypotheses:

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

- Test Statistic:

$$Z = \frac{\vec{\mathbf{a}}^t \hat{\vec{\boldsymbol{\beta}}} - \vec{\mathbf{a}}^t \vec{\boldsymbol{\beta}}_{|H_0}}{\sqrt{\vec{\mathbf{a}}^t \mathbb{I}^{-1}_{\hat{\vec{\boldsymbol{\beta}}}} \vec{\mathbf{a}}}} \quad \sim \quad \mathcal{N}(0,1) \ ,$$

- Critical Region: Two-sided. Reject $H_0$ if $|Z_{calc}| > z_{\frac{\alpha}{2}}$ .

One-sided tests can be defined, with hypotheses and Critical Regions similar to those for the Linear Model.

# The `summary` command

The command `summary` has a method for GLMs, producing results analogous to those for Linear Models.

The table `Coefficients` has similar columns:

- `Estimate` – estimated values of the parameters $\beta_j$;
- `Std.Error` – the corresponding standard errors, $\hat{\sigma}_{\hat{\beta}_j}$, i.e., the square roots of the diagonal elements of matrix $\mathscr{I}^{-1}_{\vec{\hat{\boldsymbol{\beta}}}}$;
- `z value` – computed value of the statistic $Z = \frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}}$, to test the hypotheses $H_0 : \beta_j = 0$ *vs.* $H_1 : \beta_j \neq 0$;
- `Pr(>|z|)` – the (two-sided) *p-value* of the statistic in the previous column (computed on a $\mathscr{N}(0,1)$).

This test may help in deciding whether some predictor can be dropped from the model.

# Inference in R

The `summary` command output has the basic information for Confidence Intervals or Hypothesis Tests on parameters values in a GLM.

```
> summary(HLtab.glm)

Call:
glm(formula=cbind(DAC, n-DAC) ~ idade, family=binomial, data=HL.tab)
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 ***
[...]
Number of Fisher Scoring iterations: 4      <-- no. steps of algorithm
```

# The covariance matrix of the estimators in R

The command `vcov` produces the (co-)variance matrix of the estimators $\vec{\hat{\beta}}$, that is, the inverse of Fisher's Information matrix, $\mathbb{I}_{\hat{\beta}}^{-1}$.

```
> vcov(HLtab.glm)
             (Intercept)          idade
(Intercept)  1.20457613 -0.0247424726
idade       -0.02474247  0.0005325726
```

This matrix is used in an asymptotic $(1-\alpha)\times 100\%$ confidence interval for the linear combination $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}$:

$$\left]\ \vec{\mathbf{a}}^t\vec{\mathbf{b}} - z_{\frac{\alpha}{2}}\cdot\sqrt{\vec{\mathbf{a}}^t\mathbb{I}_{\hat{\beta}}^{-1}\vec{\mathbf{a}}}\quad,\quad \vec{\mathbf{a}}^t\vec{\mathbf{b}} + z_{\frac{\alpha}{2}}\cdot\sqrt{\vec{\mathbf{a}}^t\mathbb{I}_{\hat{\beta}}^{-1}\vec{\mathbf{a}}}\ \right[$$

# Confidence intervals for $\beta_j$ in 

Confidence intervals for individual parameters $\beta_j$ are given by the function `confint.default`.

```
> confint.default(HLtab.glm)
                  2.5 %     97.5 %
(Intercept) -7.24185609 -2.9396103
idade        0.05978799  0.1502503
```

Venables & Ripley, in package `MASS`, provide an alternative (computationally more demanding) method of building confidence intervals in GLMs, called *profiling*. It is invoked by default, when command `confint` is invoked:

```
> confint(HLtab.glm)
Waiting for profiling to be done...
                  2.5 %     97.5 %
(Intercept) -7.42548805 -3.0887956
idade        0.06276942  0.1539715
```

# GLMs for Poisson response variables

Consider models in which the random component $Y$ has a Poisson distribution.

Poisson distributions are frequently used when counting random events (which can be assumed non-simultaneous).

If $Y$ has a Poisson distribution, its values are in $\mathbb{N}_0$, with probability $P[Y = y] = \frac{\lambda^y}{y!}\mathrm{e}^{-\lambda}$, for some $\lambda > 0$.

This distribution is not appropriate when a maximum number of possible counts has been previously specified, as happens with a Binomial.

# Poisson

Recall that a discrete random variable has a Poisson distribution if its probability mass function is: $P[Y = y] = \frac{\lambda^y}{y!}\mathrm{e}^{-\lambda}$, for $y \in \mathbb{N}_0$.

We can re-write the Poisson probability mass function as:

$$f(y|\lambda) = \mathrm{e}^{-\lambda}\frac{\lambda^y}{y!} = \mathrm{e}^{-\lambda}\cdot\mathrm{e}^{\ln(\lambda^y/y!)} = \mathrm{e}^{-\lambda+\ln(\lambda^y)-\ln(y!)} = \mathrm{e}^{-\lambda+y\ln(\lambda)-\ln(y!)}$$

This belongs to the exponential family $f(y\,|\,\theta,\phi) = \mathrm{e}^{\frac{y\theta-b(\theta)}{a(\phi)}+c(y,\phi)}$, with:

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

# Link functions and canonical link

The expected value of $Y \frown Po(\lambda)$ is the parameter $\lambda = E[Y]$.

A link function is a function $g(\cdot)$ such that: $g(\lambda) = g(E[Y]) = \vec{\mathbf{x}}^t\vec{\boldsymbol{\beta}}$, where $\vec{\mathbf{x}}^t\vec{\boldsymbol{\beta}}$ is the Model's systematic component.

The natural parameter in a Poisson distribution is $\theta = \ln(\lambda)$.

Thus, the canonical link function for a random component with Poisson distribution is the (natural) logarithmic function:

$$g(\lambda) = \ln(\lambda) = \vec{\mathbf{x}}^t\vec{\boldsymbol{\beta}} \quad \Leftrightarrow \quad \lambda = g^{-1}\left(\vec{\mathbf{x}}^t\vec{\boldsymbol{\beta}}\right) = e^{\vec{\mathbf{x}}^t\vec{\boldsymbol{\beta}}}$$

A Model thus defined is called a Log-Linear Model.

# Log-linear Models

## Log-linear Models

Are models with:

- Poisson random component;
- logarithmic link function, which is canonical link for Poissons.

In these models, the expected value of the Poisson response variable is given by:

$$\lambda = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_p x_p}$$

Note: the link function only gives positive values of the parameter $\lambda$, which is in structural agreement with the characteristics of parameter $\lambda$ in a Poisson distribution.

# Interpretation of the parameters $\beta_j$

With a single predictor $X$, the relation between parameter $\lambda = E[Y]$ of a Poisson distribution and the predictor is:

$$\lambda(x) = e^{\beta_0} \cdot e^{\beta_1 x}$$

An increase of one unit in the predictor multiplies the expected value of the response variable by $e^{\beta_1}$:

$$\lambda(x+1) = e^{\beta_0} \cdot e^{\beta_1 (x+1)} = e^{\beta_0} \cdot e^{\beta_1 x} \cdot e^{\beta_1} = \lambda(x) \cdot e^{\beta_1} \ .$$

This interpretation extends to more than one predictor. With $p$ predictors we have:

$$\lambda(x) = e^{\beta_0} \, e^{\beta_1 x_1} \, e^{\beta_2 x_2} \, \ldots \, e^{\beta_p x_p}.$$

An increase of one unit in the value of predictor $x_j$, keeping the other predictors fixed, multiplies the expected value of $Y$ by $e^{\beta_j}$.

# Factors as predictors and contingency tables

With an indicator variable $X_j$, an observation belonging to the category associated with indicator $X_j$ sees its Poisson parameter $\lambda$ multiplied by $e^{\beta_j}$.

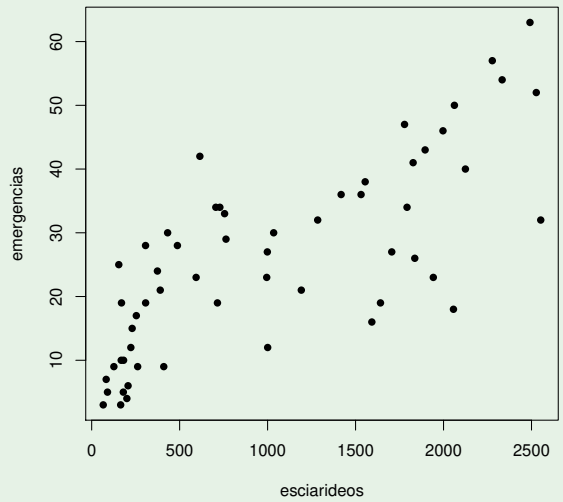
Log-linear models have great importance in the study of contingency tables, whose margins correspond to different factors and with counts of the number of observations for each combination of factor levels.

As in previous cases, other link functions are conceivable for Poisson response variables.

# Example: Exercise 5 (Log-linear Model)

## Elisa1 data (emergence of a predator)

We wish to model the number of a given species of predators that reach the adult stage (emergence) `emergencias` ($Y$), as a function of the number of gnats (`esciarideos`, $x$) in the substrate which feeds the predator's larvae.
Dados: *data frame* `Elisa1`.




There is curvilinear growth in the mean number of emergences.
If the growth is exponential, we have $E[Y] = \gamma e^{\beta_1 x} = e^{\beta_0 + \beta_1 x}$.

# Example: Exercise 5 (cont.)

## Elisa1 data (emergence)

Assuming $Y$ has a Poisson distribution, $E[Y] = \lambda$. Exponential growth of $E[Y]$ as a function of $x$ produces a log-linear model (the canonical model for a Poisson):

$$\lambda = E[Y] = e^{\beta_0 + \beta_1 x} \quad \Leftrightarrow \quad \ln(\lambda) = \beta_0 + \beta_1 x.$$

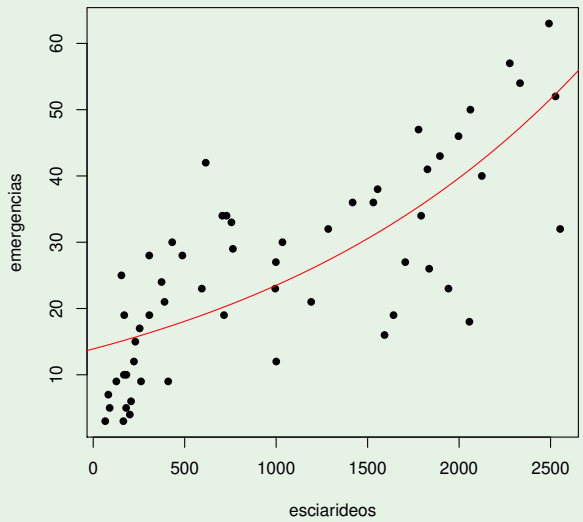
```
> Elisa1.glm <- glm(emergencias~esciarideos,family=poisson,data=Elisa1)

> summary(Elisa1.glm)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 2.632e+00  5.076e-02   51.85   <2e-16 ***
esciarideos 5.248e-04  3.209e-05   16.36   <2e-16 ***
--
(Dispersion parameter for poisson family taken to be 1)  <---
[...]
```

The fitted curve is: $\lambda = E[Y] = e^{b_0 + b_1 x} = e^{2.632 + 0.0005248 x}$

# Example: Exercise 5 (cont.)

## Elisa1 data (emergence of predator)

This is the fitted curve $\lambda = E[Y] = e^{b_0 + b_1 x} = e^{2.632 + 0.0005248\, x}$:




For each additional 100 gnats, the mean emergence number is given by:

$$\lambda_{x+100} = e^{100\, b1} \times \lambda_x = e^{0.05248} \times \lambda_x = 1.053881 \times \lambda_x \, .$$

For each additional 100 gnats, there is an increase in the mean number of emergences of $\approx 5,4\%$.

# Assessing goodness of fit in a GLM

An important concept when assessing the goodness of fit of a GLM is that of model Deviance (*desvio* in Portuguese).

The deviance plays, in GLMs, a role similar to that of the Residual Mean Squares in Linear Models.

In the study of a Linear Model we introduced the Null Model: a Model in which the linear combination of predictors is just a constant and all the variability in observed values of the response is residual variability, which the Model cannot explain.

In the study of Generalised Linear Models, also useful is a concept that is on the opposite extreme in the range of possible models: a Saturated Model, which has as many parameters as there are available observations of $Y$.

# Null Model and Saturated Model (cont.)

If in a Null Model all variability is residual (unexplained) variability, in a Saturated model all variability is "explained" by the model, with no residual variability.

In a Saturated Model, the fit is "perfect", but artificially so: the estimate of each expected value of $Y$ fully coincides with the corresponding observed value of $Y$, that is, $\hat{\mu}_i = \widehat{E[Y_i]} = Y_i$.

This "perfect" fit of the model to the data is an illusion created by the saturation (insufficient information). But it is a useful benchmark when measuring the goodness of fit of a GLM to a dataset, since we can measure the deviation from this "ideal" fit.

This is at the root of the notion of Deviance.

# Deviance

## Deviance

Consider a Generalised Linear Model fitted with $n$ independent observations of the response variable $Y$. Let:

- $\mathscr{L}_M$ be the log-likelihood of the estimated vector $\vec{\hat{\boldsymbol{\beta}}}_M$ of the model parameters (which is maximum for the dataset);
- $\mathscr{L}_T$ be the log-likelihood of the corresponding saturated model, that is, the log-likelihood calculated replacing each expected value $\mu_i$ with the corresponding observation $y_i$.

We define the deviance as:

$$D^* = -2\left(\mathscr{L}_M - \mathscr{L}_T\right)$$

# Deviance (cont.)

In a distribution of the exponential family, we have:

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

Denote by the letters M and T the estimators associated with the $\theta$ parameter in the fitted and saturated model, respectively, and assume that the dispersion parameter $\phi$ is known. The deviance is given by:

$$D^* = -2[\mathscr{L}(\hat{\theta}^M) - \mathscr{L}(\hat{\theta}^T)] = 2 \sum_{i=1}^{n} \left[ \frac{y_i(\hat{\theta}_i^T - \hat{\theta}_i^M) - [b(\hat{\theta}_i^T) - b(\hat{\theta}_i^M)]}{a(\phi_i)} \right]$$

# Deviance (cont.)

The expression for the deviance is simpler when the dispersion parameter is a known constant, that does not have to be estimated. This happens with the Poisson and Bernoulli or Binomial/n distributions:

- $\phi = 1$ for a Poisson;
- $\phi = 1$ for a Bernoulli or Binomial/n.

For distributions in the exponential family with unkown parameter $\phi$, $\phi$ must be estimated from the data in order to compute the deviance. This complicates the issue.

# Deviance in a Poisson model

We saw (slide 86) that when $Y$ has a Poisson distribution:

$$\theta = \ln(\lambda) \quad ; \quad b(\theta) = e^{\theta} = \lambda \quad ; \quad \phi = a(\phi) = 1 \ .$$

In the saturated model: $\hat{\lambda}_i^T = y_i$. In the fitted model: $\hat{\lambda}_i^M = \hat{\lambda}_i$.

The general expression for the deviance (slide 97) becomes:

$$\begin{aligned} D^* &= -2\left[\mathscr{L}(\hat{\theta}^M) - \mathscr{L}(\hat{\theta}^T)\right] = 2\sum_{i=1}^{n}\left[y_i(\hat{\theta}_i^T - \hat{\theta}_i^M) - \left[b(\hat{\theta}_i^T) - b(\hat{\theta}_i^M)\right]\right] \\ &= 2\sum_{i=1}^{n}\left\{y_i\left[\ln(y_i) - \ln(\hat{\lambda}_i)\right] - (y_i - \hat{\lambda}_i)\right\} \\ \Leftrightarrow \quad D^* &= 2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{y_i}{\hat{\lambda}_i}\right) - (y_i - \hat{\lambda}_i)\right] \end{aligned}$$

Note: The expression for $\hat{\lambda}$ (and therefore, the Deviance) also depends on the link function that is used.

# Comparing models: Likelihood ratio

To test the admissibility of a Submodel we can resort to a very general result: the Likelihood Ratio Test (LRT) (Teste da Razão de Verosimilhanças).

Let $\vec{\mathbf{Y}} = (Y_1, Y_2, ..., Y_n)$ be a random sample. Let $L(\vec{\boldsymbol{\theta}}|\vec{\mathbf{Y}})$ be its likelihood, where $\vec{\boldsymbol{\theta}}$ denotes a vector of parameters. Let $\Theta_0$ and $\Theta_1$ be two sets of alternative conditions on the values of the parameters $\boldsymbol{\theta}$.

For a Generalised Linear Model with known $\phi$, the parameters $\boldsymbol{\theta}$ are the $p+1$ coefficients $\beta_j$ in the linear combination defining the model's systematic component.

Let:

- $\Theta_0$ be the vector with the constraints for the submodel: $H_0 : \vec{\boldsymbol{\beta}}_{\overline{S}} = \vec{\mathbf{0}}$.
- $\Theta_1$ be the complementary condition: $H_1 : \vec{\boldsymbol{\beta}}_{\overline{S}} \neq \vec{\mathbf{0}}$.
- $\Theta_0 \cup \Theta_1$ represents any vector $\vec{\boldsymbol{\beta}}_{\overline{S}}$, without constraints.

# Teorema de Wilks

The Likelihood Ratio is:

$$R_n(\mathbf{x}) = \frac{\max\limits_{\boldsymbol{\theta} \in \Theta_0} L(\boldsymbol{\theta}|\mathbf{x})}{\max\limits_{\boldsymbol{\theta} \in (\Theta_0 \cup \Theta_1)} L(\boldsymbol{\theta}|\mathbf{x})}$$

Wilks' Theorem guarantees that, under $H_0$ (and given some regularity conditions for the likelihood function) $\boldsymbol{\Lambda} = -2\ln(R_n)$ has an asymptotic $\chi^2_q$ distribution, where $q$ is the number of constraints imposed on the parameters by $H_0$:

$$\boldsymbol{\Lambda} \quad = \quad -2\left(\max_{\boldsymbol{\theta} \in \Theta_0} \mathscr{L}(\boldsymbol{\theta};\mathbf{x}) - \max_{\boldsymbol{\theta} \in (\Theta_0 \cup \Theta_1)} \mathscr{L}(\boldsymbol{\theta};\mathbf{x})\right) \quad \sim \quad \chi^2_q \, .$$

Thus, $\boldsymbol{\Lambda}$ can be used as a test statistic for the Hypotheses:

$$H_0 : \boldsymbol{\theta} \in \Theta_0 \qquad \text{vs.} \qquad H_1 : \boldsymbol{\theta} \in \Theta_1 \, .$$

with a one-sided, right-tail, critical region.

# Wilks' Test for Submodels

When comparing models and submodels in GLMs,

- $q$ is the difference between the number of parameters in the full model ($\Theta_0 \cup \Theta_1$) and the submodel ($\Theta_0$): $q = p - k$;
- the Deviance of the full model, $D^*_M = -2(\mathscr{L}(\hat{\theta}^M) - \mathscr{L}(\hat{\theta}^T))$ is calculated from the log-likelihood $\mathscr{L}(\hat{\theta}^M)$ of the ML estimates of the Full Model;
- the Deviance of the submodel, $D^*_S = -2(\mathscr{L}(\hat{\theta}^S) - \mathscr{L}(\hat{\theta}^T))$ is calculated from the log-likelihood $\mathscr{L}(\hat{\theta}^S)$ of the ML estimates of the Submodel;
- The log-likelihood $\mathscr{L}(\hat{\theta}^T)$ of the saturated model is equal in both cases (the expected values of $Y$ are always estimated by the observed $y$-values);
- The Test statistic is merely the difference of the deviances:

$$\Lambda = D^*_S - D^*_M$$

# Wilks' Test on Submodels

The Wilks' Test statistic for nested models is just the difference in the Submodel and Model Deviances.

## Wilks' Test for Nested Models

Hypotheses:

$$\begin{array}{rlll} & H_0 : \beta_j = 0, \quad \forall j \notin S & \textit{vs.} & H_1 : \exists j \notin S, \text{ t.q. } \beta_j \neq 0 \\ \Leftrightarrow & H_0 : \vec{\boldsymbol{\beta}}_{\overline{S}} = \vec{\mathbf{0}} & \textit{vs.} & H_1 : \vec{\boldsymbol{\beta}}_{\overline{S}} \neq \vec{\mathbf{0}} \\ & [\text{ Submodel OK}] & \textit{vs.} & [\text{ Model better}] \end{array}$$

Test Statistic: $\boldsymbol{\Lambda} = D^*_S - D^*_M \sim \chi^2_{p-k}$ ,

Critical Region: One-sided, right-tail Reject $H_0$ if $\boldsymbol{\Lambda}_{calc} > \chi^2_{\alpha;(p-k)}$ .

Note: When the dispersion parameter $\phi$ is unknown, calculating $D^*$ (which involves $\phi$) is an issue. Alternative approaches are needed, or else the results are approximated, using an estimate of $\phi$. This problem does not exist for Binomial or Poisson distributions.

# Wilks' Test for Goodness-of-fit

For GLMs whose systematic component has an additive constant, the concept of model goodness-of-fit is similar to that used in Linear Models: compare the fit of the Model and the Null (Sub)model, without any predictors (the systematic component is just the constant $\beta_0$).

For the Null Submodel:

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

Thus, $E[Y]$ is a constant.

If the Model being studied does not have a significantly better fit than the Null Submodel, the Model cannot be considered useful.

# Wilks' Test for Goodness-of-fit (cont.)

For models in which it is not necessary to estimate the dispersion parameter $\phi$, we have:

## Wilks' Test for Goodness-of-fit of a GLM

Hypotheses:

$$H_0 : \beta_j = 0, \quad \forall j = 1 : p \quad \textit{vs.} \quad H_1 : \exists j = 1 : p, \text{ s.t. } \beta_j \neq 0$$
$$[\text{Model useless}] \quad \textit{vs.} \quad [\text{Better than Null Model}]$$

Test Statistic: $\Lambda = D_N^* - D_M^* \sim \chi^2_p,$

Critical Region: One-sided, right-tail Reject $H_0$ if $\Lambda_{calc} > \chi^2_{\alpha;p}$.

$D_N^*$ denotes the Deviance of the Null Model.

# Example: Exercise 5 (cont.)

## Elisa1 data (emergence)

To test the significance of the reduction in Deviance (*vis a vis* the Null Model), we use Wilks' Test:

```
> anova(Elisa1.glm, test="Chisq")

Analysis of Deviance Table
Model: poisson, link: log
Response: emergencias

            Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                           56     513.00
esciarideos  1   268.78         55     244.22 < 2.2e-16 ***
```

The model has a significantly better fit than the Null Model.

# Desviance for Bernoulli

We saw (slide 16) that if $Y$ has a Bernoulli distribution:

$$\theta = \ln\left(\frac{p}{1-p}\right) \quad ; \quad b(\theta) = \ln(1+e^{\theta}) = -\ln(1-p) \quad ; \quad \phi = a(\phi) = 1 \ .$$

In the saturated model: $\hat{p}_i^T = y_i$. In the fitted model: $\hat{p}_i^M = \hat{p}_i$.

Substituting in the general expression for the Deviance (slide 97), we get:

$$\begin{aligned} D^* &= -2[\mathscr{L}(\hat{\theta}^M) - \mathscr{L}(\hat{\theta}^T)] = 2\sum_{i=1}^{n}\left[y_i(\hat{\theta}_i^T - \hat{\theta}_i^M) - [b(\hat{\theta}_i^T) - b(\hat{\theta}_i^M)]\right] \\ &= 2\sum_{i=1}^{n}\left\{y_i\left[\ln\left(\frac{y_i}{1-y_i}\right) - \ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right)\right] + \left[\ln(1-y_i) - \ln(1-\hat{p}_i)\right]\right\} \\ \Leftrightarrow \quad D^* &= 2\sum_{i=1}^{n}\left\{y_i \ln\left(\frac{y_i}{\hat{p}_i}\right) + (1-y_i)\ln\left(\frac{1-y_i}{1-\hat{p}_i}\right)\right\} \end{aligned}$$

# Deviance for Binomial/n

We saw (slide 30) that if $Y$ has a 'Binomial/n' distribution:

$$\theta = \ln\left(\frac{p}{1-p}\right) \quad ; \quad b(\theta) = -\ln(1-p) \quad ; \quad \phi = 1 \quad ; \quad a(\phi) = \frac{1}{n} \, .$$

For the saturated model: $\hat{p}_i^T = y_i$. For the fitted model: $\hat{p}_i^M = \hat{p}_i$.

Substituting in the general expression for the Deviance (slide 97):

$$D^* = -2\left[\mathscr{L}(\hat{\theta}^M) - \mathscr{L}(\hat{\theta}^T)\right] = 2\sum_{i=1}^{n} \frac{\left[y_i(\hat{\theta}_i^T - \hat{\theta}_i^M) - [b(\hat{\theta}_i^T) - b(\hat{\theta}_i^M)]\right]}{a(\phi_i)}$$

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

The expression for $\hat{p}$ (hence, for the Deviance) also depends on the link function.

The Deviance is different for Bernoulli and Binomial/n.

# Example: Exercise 5 (cont.)

## Elisa1 data (emergence data)

```
> Elisa1.glm <- glm(emergencias~esciarideos,family=poisson,data=Elisa1)
> summary(Elisa1.glm)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 2.632e+00  5.076e-02   51.85   <2e-16 ***
esciarideos 5.248e-04  3.209e-05   16.36   <2e-16 ***
--
(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 513.00  on 56  degrees of freedom   <---
Residual deviance: 244.22  on 55  degrees of freedom   <---
AIC: 526.32
```

The model Deviance is the `Residual Deviance` (244.22, in our case).

The Deviance of the Null Model is the `Null Deviance` (513.00, is our case).

# Example: Exercises 1 and 10

Consider GLMs with both numeric and factor predictors.

## Example: tobacco larvae (Venables & Ripley)

A study in the resistence of the tobacco budworm (*heliothis virescens*) to doses of a given toxic substance.

Lots of 20 moths of each sex were exposed to different doses of the substance (in $\mu$g). After 3 days, the number of dead individuals in each lot was registered. The results are given in the table below.

| | Dose | | | | | |
|---|---|---|---|---|---|---|
| Sexo (Sex) | 1 | 2 | 4 | 8 | 16 | 32 |
| Machos (Males) | 1 | 4 | 9 | 13 | 18 | 20 |
| Fêmeas (Females) | 0 | 2 | 6 | 10 | 12 | 16 |

The data can be considered observations of a Binomial response variable (number of deaths in each experimental situation, with $n_i = 20$ larvae exposed to the toxic).

# Tobacco budworm example (cont.)

## Example: tobacco budworm

Creating the *data frame* `tabaco` with the data:

```
> 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
```

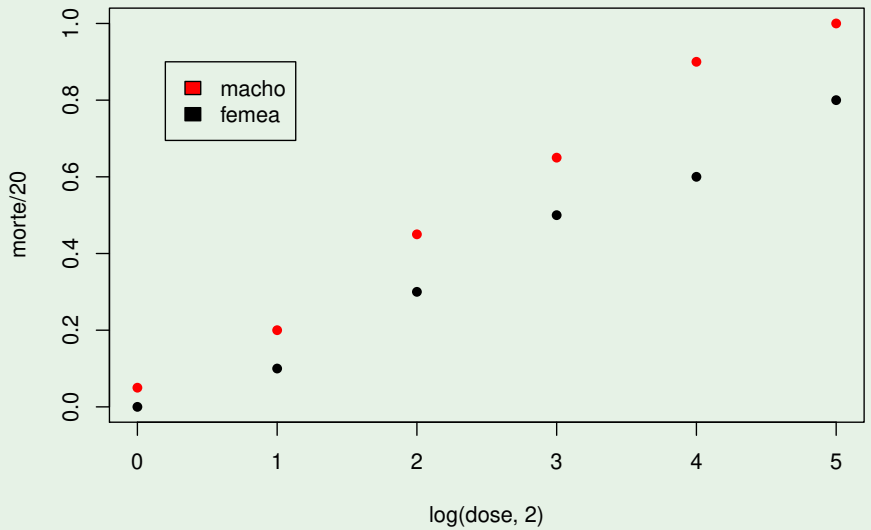
```
   morte  sexo dose
1      1 macho    1
2      4 macho    2
3      9 macho    4
4     13 macho    8
5     18 macho   16
6     20 macho   32
7      0 femea    1
8      2 femea    2
9      6 femea    4
10    10 femea    8
11    12 femea   16
12    16 femea   32
```

It is usual in toxicology to successively double the doses and so a logarithmic transformation with base 2 of the doses makes sense.

# Exercise 1

## Example: tobacco budworm

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




Although a linear relation seems adequate, a sigmoidal relation is structurally more appropriate, forcing the fitted values of $\hat{p}$ to be in the interval $]0,1[$.

# Exercise 1 in R (cont.)

To fit a Probit Regression, use the option `link=probit` when specifying argument `family`.

## Example: tobacco budworm

```
> glm(cbind(morte,20-morte) ~ log(dose,2),
+       family=binomial(link=probit), data=tabaco)

Call:  glm(formula = cbind(morte, 20 - morte) ~ log(dose, 2),
+         family = binomial(link = probit), 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
```

The fitted relation is: $p(x) = \Phi(-1.6431 + 0.5966 \log_2(x))$, where $x$ is the dose.

The model Deviance is given as `Residual Deviance` (16.41).

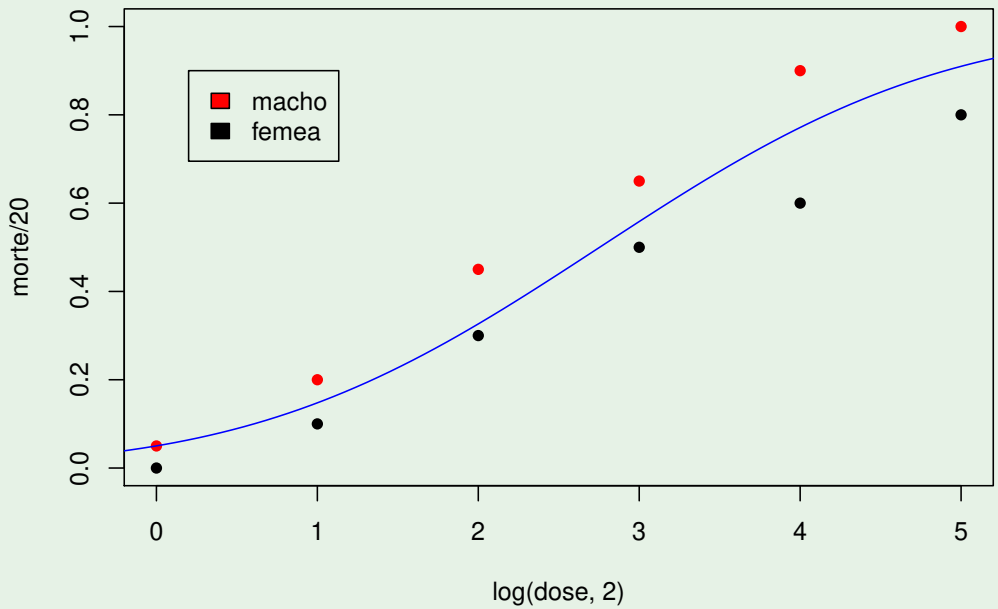
The Deviance of the Null Model is the `Null Deviance` (124.9).

# Exercise 1 in R (cont.)

## Example: tobacco budworm

We draw the fitted curve on the scatterplot with the command:

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



# Toxicological interpretation

## Tobacco budworm

In base 2 logarithms, the doses used were $0, 1, 2, 3, 4, 5$.

In a toxicological context, we can state (slide 64) that the tolerance to $\log_2$-dosage follows the distribution $T \frown \mathcal{N}\left(\mu = -\frac{\beta_0}{\beta_1}, \sigma^2 = \frac{1}{\beta_1^2}\right)$.

With the estimated values, we have:
$\hat{\mu} = \frac{1.6431}{0.5966} = 2.754107$ e $\hat{\sigma}^2 = \frac{1}{0.5966^2} = 2.8096$.

Hence, tolerance to dosage (in base 2 logarithms) has the distribution:

$$T \frown \mathcal{N}(2.7541, \underbrace{2.8096}_{=\hat{\sigma}^2}).$$

# Exercise 1: Goodness-of-fit test in R

In R, a Wilks' test comparing a GLM model with the corresponding Null Model can be carried out with the command `anova` and argument `test="Chisq"`.

## Example: tobacco budworm

```
> 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 ***
```

As would be expected, the model has a significantly better fit than the Null model (which estimates a constant $p$ for all dosages).

# Exercise 10

## Example: an ANCOVA-type model for the tobacco budworm

We can also mix numeric and factor predictors, as in an ANCOVA.

```
> tabaco.glmSx <- glm(cbind(morte,20-morte) ~ log(dose,2) * sexo ,
+         family=binomial(link=probit), data=tabaco)
> summary(tabaco.glmSx)

(...)
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  <-- the deviance has decreased from 16.41 to 3.768
AIC: 41.878
```

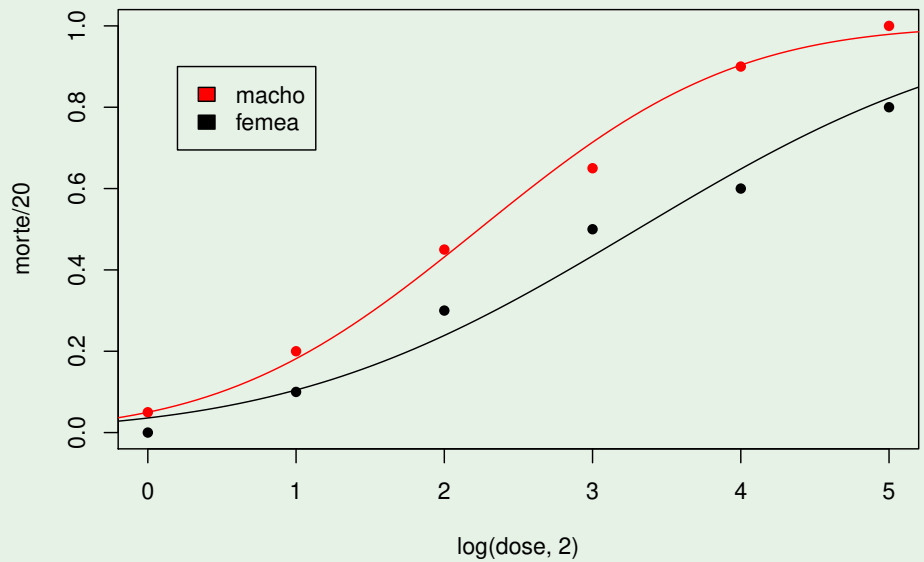
The reference level is females (alphabetical order). The estimated relations are:

- $p(x) = \Phi(-1.80072 + 0.54532 \log_2(x))$ for females; and
- $p(x) = \Phi((-1.80072 + 0.15479) + (0.54532 + 0.19165) \log_2(x))$ for males.

# Exercise 10 in R (cont.)

## Example: tobacco budworm

```
> plot(morte/20 ~ log(dose,2), col=sexo, data=tabaco, pch=16)
> curve(pnorm(-1.80072+0.54523*x), from=-1, to=6, 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"))
```

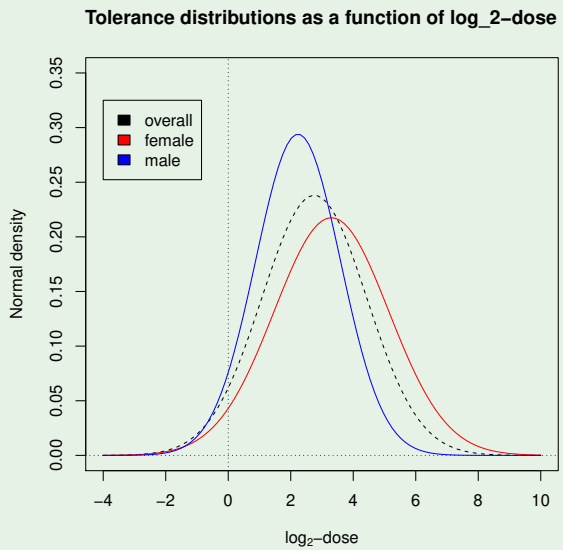


# Toxicological interpretation

## Tobacco budworm

The distribution of the tolerance to $\log_2$-dosage (slide 64), can now be calculated separatly for females and males:

$$T_F \frown \mathcal{N}(3.30268, 3.36388) \quad ; \quad T_M \frown \mathcal{N}(2.233647, 1.84164).$$




The log-dosage at which half the females die kills almost 80% of the males:

```
> pnorm(3.30268 , m=2.233647 , sd=sqrt(1.84164))
[1] 0.7845787
```

# Exercise 10: Wilks' test in 

To assess whether different models for each sex are preferable, we compare the models (question 10b), using Wilks' test.

## Example: tobacco budworm

```
> anova(tabaco.glm, tabaco.glmSx, 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 P(>|Chi|)
1        10     16.414
2         8      3.768  2   12.646  0.001795 **
```

The model with different equations for each sex is significantly better (predictable, given the scatterplot).

# AIC in GLMs

## Akaike Information Criterion (AIC)

Define, in a GLM with $m$ parameters, estimated by $\vec{\hat{\boldsymbol{\beta}}}$:

$$AIC = -2 \cdot \mathscr{L}(\vec{\hat{\boldsymbol{\beta}}}; \vec{\mathbf{Y}}) + 2m\,.$$

where $\mathscr{L}(\vec{\hat{\boldsymbol{\beta}}}; \vec{\mathbf{Y}})$ is the sample log-likelihood.

- The smaller the AIC (for comparable models), the better the model fit.
- The AIC can be used as a criterion to compare models with the same random component and fitted with the same data.
- Note that in a GLM, both the deviance $D^* = -2(\mathscr{L}_M - \mathscr{L}_T)$ and the AIC are defined using log-likelihoods.

# Selection of Submodels

As with Linear Models, The choice of an adequate submodel may result from considerations of a different type.

Of no specific submodel is proposed, a full search of all $2^p - 2$ possible submodels presents the same computational difficulties that were discussed for the Linear Model.

The `eleaps` function in the `R` package `subselect` can do complete searches for optimal GLM submodels of a given cardinality (as long as there are not too many predictors).

Alternatively, it is possible to use stepwise search algorithms, similar to those used fot Linear Models, but using as criteria for adding/dropping predictors their effect on the Deviance or AIC.

The `step` function automatically runs stepwise algorithm based on AIC.

# Algorithms to select predictors in R

## Backward elimination in R: Exercise 10

```
> step(tabaco.glmSx)
Start:  AIC=41.88
cbind(morte, 20 - morte) ~ log(dose, 2) * sexo
                    Df Deviance    AIC
- log(dose, 2):sexo  1    5.566 41.676       <- Notice the hierarchy in effect types. The
<none>                    3.768 41.878        first step only considers interaction effects.


Step:  AIC=41.68
cbind(morte, 20 - morte) ~ log(dose, 2) + sexo
               Df Deviance     AIC
<none>              5.566  41.676     <- Having excluded interaction, other effects are assessed.
- sexo           1   16.414  50.524
- log(dose, 2)   1  118.799 152.909


Call:  glm(formula = cbind(morte, 20 - morte) ~ log(dose, 2) + sexo,
    family = binomial(link = "probit"), data = tabaco)
Coefficients:
 (Intercept)  log(dose, 2)     sexomacho
     -2.0603        0.6324        0.6536
Degrees of Freedom: 11 Total (i.e. Null);  9 Residual
Null Deviance:      124.9
Residual Deviance: 5.566  AIC: 41.68
```

Final choice: model with $\beta_1$ equal on both sexes, but $\beta_0$ different.

# Models with Gamma response variables

## The Gamma distribution

The Gamma is a distribution for continuous random variables, with values in $\mathbb{R}^+$. A standard parametrization in GLMs (see Exercise GLM 7) is:

$$f(y \mid \mu, \nu) = \frac{\nu^\nu}{\mu^\nu \Gamma(\nu)} y^{\nu-1} \, \mathrm{e}^{-\frac{\nu y}{\mu}}$$

Specific cases: the Chi-squared distribution ($\chi^2_n$ if $\nu = \frac{n}{2}$ e $\mu = n$) and the Exponential distribution ($\nu = 1$).

If $Y \frown G(\mu, \nu)$, the variance is proportional to the square of the mean:

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

GLMs with Gamma random component may be useful when the variance of the data around the trend is not constant, but rather proportional to the square of the mean.

# The Gamma distribution in the exponential family

A random variable $Y$ has a Gamma distribution with parameters $\mu$ and $\nu$ if its values are in $\mathbb{R}^+$, with density function:

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

Gammas belong to the exponential family, $f(y|\theta,\phi) = \mathrm{e}^{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)}$, with:

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

# The canonical link function for Gammas

Since if $Y \frown G(\mu, \nu)$ then $E[Y] = \mu$, the link functions $g$ in a GLM with Gamma response relate the mean $\mu$ to the linear combinations of the predictors:

$$g(\mu) = \vec{\mathbf{x}}^t \vec{\boldsymbol{\beta}} = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$$

The canonical link function for models with Gamma distributions transform the expected value of $Y$ into the natural parameter $\theta = -\frac{1}{\mu}$.

Since the minus sign only affects the sign of the $\beta_j$s, the canonical link function for Gamma GLMs is usually defined as only the reciprocal function:

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

# Gamma model with a canonical link

The model equation can be written as:

$$g(\mu) = \frac{1}{\mu} = \vec{\mathbf{x}}^t\vec{\boldsymbol{\beta}} \quad \Leftrightarrow \quad \mu_{\vec{\mathbf{x}}} = g^{-1}\left(\vec{\mathbf{x}}^t\vec{\boldsymbol{\beta}}\right) = \frac{1}{\vec{\mathbf{x}}^t\vec{\boldsymbol{\beta}}}$$

Note: although the expected value of the response variable $Y$ must be positive (a variable $Y$ with Gamma distribution only takes positive values), in the above relation the expected value $\mu$ may be negative.

Thus, and unlike previous GLM models, there is no "structural guarantee" that the fitted values of $\mu$ are adequate.

# Familiar curves in a GLM context

When there is a single numeric predictor, the above relation states that the mean value of $Y$ follows a hyperbolic-type curve,

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

This function has been used in Agronomy to model curves for yield per plant ($Y$), as a function of crop density ($X$).

If we use the reciprocals of a single predictor, in other words the transformation $X^* = \frac{1}{X}$, the expected value becomes

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

so that the expected value of $Y$ follows a Michaelis-Menten curve (with the Shinozaki-Kira parametrisation).

# Summary table for the exponential family

| Distribution | $E[Y]$ | $V[Y]$ | $\theta$ | $b(\theta)$ | $\phi$ | $a(\phi)$ |
|---|---|---|---|---|---|---|
| Normal | $\mu$ | $\sigma^2$ | $\mu$ | $\frac{\theta^2}{2} = \frac{\mu^2}{2}$ | $\sigma^2$ | $\sigma^2$ |
| Poisson | $\lambda$ | $\lambda$ | $ln(\lambda)$ | $e^{\theta} = \lambda$ | 1 | 1 |
| Bernoulli | $p$ | $p(1-p)$ | $\ln\left(\frac{p}{1-p}\right)$ | $ln(1+e^{\theta}) = -ln(1-p)$ | 1 | 1 |
| Binomial/n | $p$ | $\frac{p(1-p)}{n}$ | $\ln\left(\frac{p}{1-p}\right)$ | $e^{\theta} = \lambda$ | 1 | $\frac{1}{n}$(*) |
| Gamma | $\mu$ | $\frac{\mu^2}{\nu}$ | $-\frac{1}{\mu}$ | $-ln(-\theta) = -ln\left(\frac{1}{\mu}\right)$ | $\frac{1}{\nu}$ | $\frac{1}{\nu}$ |

(*) Except in this case, we always have $a(\phi) = \phi$.

# Unknown dispersion parameter $\phi$

In GLMs with Poisson or Bernoulli/Binomial response, the dispersion parameter is known: $\phi = 1$.

But in GLMs with a Gamma or Normal random component, the dispersion parameter is, in general, unknown:

- In a Normal, $\phi = \sigma^2$ (the variance);
- In a Gamma with parametrization $Y \frown G(\mu, \nu)$, $\phi = \frac{1}{\nu}$ (with variance $V[Y] = \mu^2 \phi$).

The unknown $\phi$ must be estimated and this raises problems.

It is often assumed that $\phi$ is common to all observations, or that it varies along observations only because of known constants. Assuming that each observation was free to have its own $\phi_i$ would render their estimation impossible.

# Deviance and Scaled deviance

When it is necessary to estimate $\phi$, we define the Scaled Deviance (Desvio Reduzido).

## Deviance and Scaled Deviance

Assuming $a(\phi_i) = \frac{\phi}{w_i}$, for some $\phi$ that is equal for all observations and with known weights $w_i$, the Deviance becomes:

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

$D^*$ is called the Scaled Deviance and the Deviance is $D$, defined as:

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

Note: For Poisson and Bernoulli-Binomial/n, Deviance and Scaled Deviance coincide, since $\phi = 1$.

# Deviance and Scaled Deviance for a Normal

We saw (slide 15) that for $Y$ with a Normal distribution:

$$\theta = \mu \quad ; \quad b(\theta) = \frac{\theta^2}{2} = \frac{\mu^2}{2} \quad ; \quad \phi = \sigma^2 \quad ; \quad a(\phi) = \phi = \sigma^2 \ .$$

In a saturated model: $\hat{\mu}_i^T = y_i$. In the fitted model: $\hat{\mu}_i^M = \hat{\mu}_i$.

Substituting in the general expression for the Deviance (slide 97):

$$\begin{aligned} D^* &= 2 \sum_{i=1}^{n} \frac{\left[ y_i(\hat{\theta}_i^T - \hat{\theta}_i^M) - [b(\hat{\theta}_i^T) - b(\hat{\theta}_i^M)] \right]}{a(\phi_i)} = 2 \sum_{i=1}^{n} \frac{\left[ y_i(y_i - \hat{\mu}_i) - \frac{y_i^2}{2} + \frac{\hat{\mu}_i^2}{2} \right]}{\sigma_i^2} \\ &= 2 \sum_{i=1}^{n} \frac{\left[ \frac{y_i^2}{2} - y_i \hat{\mu}_i + \frac{\hat{\mu}_i^2}{2} \right]}{\sigma_i^2} = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\sigma_i^2} \ . \end{aligned}$$

With the homogeneous variance assumption of Linear Models, $\sigma_i^2 = \sigma^2 = \phi$ for all $i$, and the Deviance of the Normal will be the Residual Sum of Squares:

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

# Deviance and Scaled Deviance in a Gamma

From slide 125 we have:

$$\theta = -\frac{1}{\mu} \quad ; \quad b(\theta) = -\ln(-\theta) = \ln(\mu) \quad ; \quad \phi = \frac{1}{\nu} \quad ; \quad a(\phi) = \phi = \frac{1}{\nu} \, .$$

Hence, (slide 97) the Scaled Deviance $D^*$ is:

$$\begin{aligned} D^* &= 2 \sum_{i=1}^{n} \frac{\left[ y_i(\hat{\theta}_i^T - \hat{\theta}_i^M) - [b(\hat{\theta}_i^T) - b(\hat{\theta}_i^M)] \right]}{a(\phi_i)} \\ &= 2 \sum_{i=1}^{n} \frac{\left[ y_i(-\frac{1}{y_i} + \frac{1}{\hat{\mu}_i}) - [\ln(y_i) - \ln(\hat{\mu}_i)] \right]}{\frac{1}{\nu}} = 2 \sum_{i=1}^{n} \nu_i \left[ \left( \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} \right) - \ln\left( \frac{y_i}{\hat{\mu}_i} \right) \right] \end{aligned}$$

Assuming that $a(\phi_i) = \phi = \frac{1}{\nu}$ (constant for all observations), the Deviance $D$ (slide 131) will be:

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

# Estimation of the dispersion parameter $\phi$

Estimating the dispersion parameter $\phi$ requires additional assumptions, such as that there is a common value of $\phi$ for all observations, or that a common value varies between observations only due to known contants.

One way of estimating $\phi$ is to use a maximum likelihood estimator.
But the most frequent way of estimating $\phi$ involves one type of residuals, called Pearson residuals, defined later.

Even in models with Binomial or Poisson responses, for which $\phi = 1$, such estimates of dispersion may be useful: a value of $\hat{\phi}$ very much larger than 1 indicates overdispersion, which suggests that the model be modified. In `R`, this estimate may be obtained using the option `quasi` in the definition of argument `family`, when specifying the GLM.

# Residuals and Model Validation

The usual concept of residuals in a Linear Model, $e_i = y_i - \hat{y}_i = y_i - \hat{\mu}_i$, has different adaptations in GLMs, where, unlike for Linear Models, additive random errors are not contemplated.

In Generalised Linear Models different definitions of residuals are used. The two main ones are

- Pearson residuals; and
- Deviance residuals.

In Pearson residuals, the difference between observed values $Y_i$ and corresponding fits of their expected values, $\widehat{E[Y_i]} = \hat{\mu}_i$, is divided by the square root of the model's so-called variance function.

# Variance function

## Variance function

Consider a GLM with random component $Y$ with mean $E[Y]$, variance $V[Y]$ and dispersion parameter $\phi$. The function $f_V(E[Y]) = \frac{V[Y]}{\phi}$ is called the model's variance function. Thus: $V[Y] = \phi \cdot f_V(E[Y])$.

The variance function is different for each distribution of $Y$:

- Normal: $f_V(\mu) = \frac{V[Y]}{\phi} = \frac{\sigma^2}{\sigma^2} = 1$.
- Bernoulli: $f_V(p) = \frac{V[Y]}{1} = V[Y] = p(1-p)$.
- Binomial/n: $f_V(p) = \frac{V[Y]}{1} = V[Y] = \frac{p(1-p)}{n}$.
- Poisson: $f_V(\lambda) = \frac{V[Y]}{1} = V[Y] = \lambda$.
- Gamma: $f_V(\mu) = \frac{V[Y]}{\phi} = \frac{\frac{\mu^2}{\nu}}{\frac{1}{\nu}} = \mu^2$.

# Pearson Residuals

## Pearson Residuals

Let $Y_1, Y_2, ..., Y_n$ be a random sample of a random component in a Generalised Linear Model. We define the Pearson Residual of each observation as:

$$r_i^P = \frac{Y_i - \hat{\mu}_i}{\sqrt{f_v(\hat{\mu}_i)}} .$$

- Normal: We have $f_v(\mu_i) = \frac{V[Y_i]}{\sigma_i^2} = 1$. The Pearson residual is the usual residual in Linear Models:

$$r_i^P = Y_i - \hat{\mu}_i$$

- Poisson: We have $f_v(\lambda_i) = \frac{V[Y_i]}{1} = \lambda_i$. The Pearson residual is:

$$r_i^P = \frac{Y_i - \hat{\lambda}_i}{\sqrt{\hat{\lambda}_i}}$$

# Pearson Residuals (cont.)

- Bernoulli: $f_v(p_i) = \frac{V[Y_i]}{1} = p_i(1-p_i)$. The Pearson residual is:

$$r_i^P = \frac{Y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}} \tag{1}$$

- Binomial/n: $f_v(p_i) = \frac{V[Y_i]}{1} = \frac{p_i(1-p_i)}{n_i}$. The Pearson residual is:

$$r_i^P = \frac{Y_i - \hat{p}_i}{\sqrt{\frac{\hat{p}_i(1-\hat{p}_i)}{n_i}}} \tag{2}$$

- Gamma: We have $f_v(\mu_i) = \frac{V[Y_i]}{\phi_i} = \frac{\frac{\mu_i^2}{v_i}}{\frac{1}{v_i}} = \mu_i^2$. The Pearson residual is:

$$r_i^P = \frac{Y_i - \hat{\mu}_i}{\hat{\mu}_i}$$

# Pearson Residuals (cont.)

The expressions of the Pearson residuals also depend on the link functions. For example, in binary response models, formulas (1) and (2) on slide 138 become,

- in a Logistic Regression:

$$\hat{p}_i = \frac{1}{1 + \mathrm{e}^{-(\hat{\beta}_0 + \hat{\beta}_1 x_{1(i)} + \ldots + \hat{\beta}_p x_{p(i)})}}$$

- In a Probit Regression:

$$\hat{p}_i = \Phi(\hat{\beta}_0 + \hat{\beta}_1 x_{1(i)} + \ldots + \hat{\beta}_p x_{p(i)})$$

- In a Complementary log-log model:

$$\hat{p}_i = 1 - \mathrm{e}^{-\mathrm{e}^{\hat{\beta}_0 + \hat{\beta}_1 x_{1(i)} + \ldots + \hat{\beta}_p x_{p(i)}}}$$

# The generalised Pearson statistic

On slide 137 we saw how, for GLMs with Poisson random component, the Pearson residual is given by: $r_i^P = \frac{Y_i - \hat{\lambda}_i}{\sqrt{\hat{\lambda}_i}}$. Since for those models $\hat{\lambda}_i = E[Y_i]$, the sum of squares of those residuals is the usual Pearson statistic in Chi-square tests, $\sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$. This gives rise to the following concept.

## Generalised Pearson statistic, $X^2$

Given a GLM with Pearson residuals $r_i^P$ we define the Generalised Pearson statistic as the sum of squares of those residuals:

$$X^2 = \sum_{i=1}^{n} (r_i^P)^2 .$$

Sometimes, this quantity is used instead of the Deviance to indicate the goodness-of-fit of a model.

# Estimation of the dispersion parameter $\phi$

In GLMs with unknown dispersion parameter $\phi$, one way of estimating $\phi$ uses Pearson residuals.

## Estimator of $\phi$

In a GLM with $m$ parameters in the systematic component and assuming a common dispersion parameter $\phi$ for all observations $Y_i$, an estimator of $\phi$ is:

$$\hat{\phi} = \frac{X^2}{n-m} = \frac{\sum_{i=1}^{n} (r_i^P)^2}{n-m} .$$

Note: For the Linear Model this estimator gives *QMRE*.

# Deviance Residuals

An alternative notion of residual is based on the individual terms of the definition of Deviance of a GLM (by analogy with what happens with the Residual Sum of Squares in a Linear Model).

## Deviance Residuals

Let $Y_1, Y_2, ..., Y_n$ be a random sample of a random component in a Generalised Linear Model. Let

$$D = \sum_{i=1}^{n} d_i$$

be its Deviance. We call Deviance residual of observation $i$ to:

$$r_i^D = sign(y_i - \hat{\mu}_i) \cdot \sqrt{d_i}$$

# Deviance Residuals (cont.)

Specifically:

- Normal: We have $d_i = (y_i - \hat{\mu}_i)^2$. The Deviance residual is:

$$r_i^D = y_i - \hat{\mu}_i$$

In a Linear Model, Deviance residuals are the usual residuals.

- Bernoulli: we have

$$d_i = -2 \cdot [y_i \ln(\hat{p}_i) + (1 - y_i) \ln(1 - \hat{p}_i)] = \begin{cases} -2 \ln(1 - \hat{p}_i) & \text{se} \quad y_i = 0 \\ -2 \ln(\hat{p}_i) & \text{se} \quad y_i = 1 \end{cases}$$

The Deviance residuals for a Bernoulli $Y$ are:

$$r_i^D = sign(y_i - \hat{p}_i) \cdot \sqrt{d_i} = \begin{cases} -\sqrt{-2 \ln(1 - \hat{p}_i)} & \text{se} \quad y_i = 0 \\ \sqrt{-2 \ln(\hat{p}_i)} & \text{se} \quad y_i = 1 \end{cases}$$

# Deviance Residuals (cont.)

- Binomial/n: we have

$$d_i = \begin{cases} -2n_i \left[ y_i \ln\left(\frac{y_i}{\hat{p}_i}\right) + (1-y_i)\ln\left(\frac{1-y_i}{1-\hat{p}_i}\right)\right] & \text{se} \quad y_i \neq 0,1 \\ -2n_i \left[ -y_i \ln(\hat{p}_i) - (1-y_i)\ln(1-\hat{p}_i)\right] & \text{se} \quad y_i \in \{0,1\}\ . \end{cases}$$

The Deviance residuals for Binomial/n $Y$ are:

$$r_i^D = \begin{cases} \sqrt{-2n_i \left[ y_i \ln\left(\frac{y_i}{\hat{p}_i}\right) + (1-y_i)\ln\left(\frac{1-y_i}{1-\hat{p}_i}\right)\right]} & \text{se} \quad y_i \neq 0,1 \\ \sqrt{2n_i \left[ y_i \ln(\hat{p}_i) + (1-y_i)\ln(1-\hat{p}_i)\right]} & \text{se} \quad y_i \in \{0,1\}\ . \end{cases}$$

- Poisson: In this case, $d_i = 2 \cdot \left[ y_i \ln\left(\frac{y_i}{\hat{\lambda}_i}\right) - (y_i - \hat{\lambda}_i)\right]$. The Deviance residuals for Poisson $Y$ are:

$$r_i^D = sign(y_i - \hat{\lambda}_i) \cdot \sqrt{2\left[ y_i \ln\left(\frac{y_i}{\hat{\lambda}_i}\right) - (y_i - \hat{\lambda}_i)\right]}$$

# Deviance Residuals (cont.)

- Gamma: in this case,

$$d_i = 2 \cdot \left[ \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} - \ln\left( \frac{y_i}{\hat{\mu}_i} \right) \right]$$

The Deviance residuals for Gamma $Y$ are:

$$r_i^D = sign(y_i - \hat{\mu}_i) \cdot \sqrt{2 \cdot \left[ \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} - \ln\left( \frac{y_i}{\hat{\mu}_i} \right) \right]}$$

As before, to each different link function $g$ corresponds a different expression for the fitted means $\hat{\mu}$, hence a different specific expression for the Deviance residuals.

# Residuals in R

As for the Linear Model, R provides commands to compute residuals.

- `residuals` calculates the, by default, deviance residuals. .

```
> residuals(tabaco.glm)
          1           2           3           4           5           6
-0.003720807  0.631866326  1.149311810  0.834445925  1.498556259  1.944020824
          7           8           9          10          11          12
-1.435052677 -0.632286846 -0.253437946 -0.523178851 -1.708321246 -1.502372783
```

- Pearson residuals can be calculated using the argument `type=''pearson''`.

```
> residuals(tabaco.glm, type="pearson")
         1          2          3          4          5          6
-0.00371868  0.65977673  1.17907319  0.82587899  1.37017323  1.40774747
          7          8          9         10         11         12
-1.02793742 -0.60081763 -0.25159052 -0.52497400 -1.82437610 -1.71522073
```

# Residuals and the Validation of a GLM

Residuals can be used to:

- assess the validity of the distributional assumption regarding the random component;
- assess the suitability of the systematic component;
- assess the suitability of the link function that is used;
- as diagnostic tools to identify special observations.

But the use of residuals for model validation has many specificities for each individual GLM, and an overall discussion becomes difficult.

For a more detailed discussion, check McCullagh & Nelder (1989) or other bibliographical references.

# GLMs in the study of contingency tables

GLMs can be defined with numeric, categorical or both kinds of predictors.

Log-linear Models are particularly important for the study of contingency tables.

In such models, the random component corresponds to counts (discrete variable), that we seek to explain with the levels of one or more factors.

It can often be assumed that the response variable has a Poisson distribution (or a binomial or its generalization, the multinomial).

# (No) Tabelas de contingência para 2 factores

Consideremos o caso frequente de tabelas de contingência com dois factores de classificação.

Exemplo: uma tabela de contagens de observações de espécies (primeiro factor) em vários locais (segundo factor).

| Níveis do Factor A | Níveis do Factor B | | | | | Marginal de A |
|---|---|---|---|---|---|---|
| | 1 | 2 | $\cdots$ | $b-1$ | $b$ | |
| 1 | $n_{11}$ | $n_{12}$ | $\cdots$ | $n_{1,(b-1)}$ | $n_{1,b}$ | $n_{1\cdot}$ |
| 2 | $n_{21}$ | $n_{22}$ | $\cdots$ | $n_{2,(b-1)}$ | $n_{2,b}$ | $n_{2\cdot}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $a-1$ | $n_{(a-1),1}$ | $n_{(a-1),2}$ | $\cdots$ | $n_{(a-1),(b-1)}$ | $n_{(a-1),b}$ | $n_{(a-1)\cdot}$ |
| $a$ | $n_{a1}$ | $n_{a2}$ | $\cdots$ | $n_{a,(b-1)}$ | $n_{a,b}$ | $n_{a\cdot}$ |
| Marginal de B | $n_{\cdot 1}$ | $n_{\cdot 2}$ | $\cdots$ | $n_{\cdot(b-1)}$ | $n_{\cdot b}$ | $n = n_{\cdot\cdot}$ |

# (No) Tabelas de contingência para 2 factores (cont.)

Quando não há restrições sobre o número total de observações, ou sobre qualquer das margens (como será o caso nas tabelas de locais $\times$ espécies), as contagens podem ser consideradas como observações independentes de distribuições de Poisson.

Numa situação dessas, será de considerar um modelo com algumas semelhanças aos modelos ANOVA, mas em que a variável resposta $Y_{ij} = n_{ij}$, tenha distribuição Poisson.

Neste contexto, um modelo tipo ANOVA factorial em que, além de efeitos principais de cada factor, se prevejam efeitos de interacção entre os dois factores, é um modelo saturado, uma vez que:

- há apenas 1 observação em cada uma das $ab$ células (a contagem $n_{ij}$);
- há $ab$ parâmetros num modelo factorial com interacção.

# (No) A hipótese de independência

Mais útil serão modelos associados a hipóteses mais restritivas sobre a natureza da relação entre os factores associados à tabela. Em particular a hipótese de independência entre os factores pode ser interessante.

Existindo independência entre os factores, os valores esperados de $Y_{ij} = n_{ij}$ serão dados (para qualquer $i$ e $j$) por:

$$E[Y_{ij}] = \lambda_{ij} = n\, p_{ij} = n\, p_{i.}\, p_{.j}$$

onde:

- $n$ é o número total de observações;
- $p_{ij}$ é a probabilidade duma observação recair na célula (i,j);
- $p_{i.}$ é a probabilidade marginal associada ao nível $i$ do Factor A;
- $p_{.j}$ é a probabilidade marginal associada ao nível $j$ do Factor B.

# (No) A hipótese de independência (cont.)

Uma vez que a distribuição Poisson é adequada à variável resposta, surge de forma natural a ideia de usar a função de ligação canónica para essa distribuição, ou seja, de logaritmizar $E[Y_{ij}]$:

$$\ln\left(E[Y_{ij}]\right) \quad = \quad \ln(n\, p_{i.}\, p_{.j}) \quad = \quad \ln(n) + \ln(p_{i.}) + \ln(p_{.j})$$

Trata-se duma relação do tipo ANOVA a dois factores, sem interacção:

$$\ln\left(E[Y_{ij}]\right) \quad = \quad \mu + \alpha_i + \beta_j$$

onde se pode considerar (embora mais tarde se modifique):

- $\mu = \ln(n)$ é uma constante comum a todas as observações;
- $\alpha_i = \ln(p_{i.})$ é um efeito associado ao nível $i$ do factor A;
- $\beta_j = \ln(p_{.j})$ é um efeito associado ao nível $j$ do factor B.

# (No) A hipótese de independência (cont.)

Estamos perante um Modelo Log-linear com:

- componente aleatória Poisson;
- função de ligação logarítmica (ligação canónica da Poisson);
- componente sistemática dada por variáveis indicatrizes de níveis de cada factor.

Tal como nas ANOVAs clássicas, há que impor restrições aos parâmetros e considerar a célula associada ao primeiro nível de cada factor como uma célula de referência, sendo a situação nas restantes células comparada com essa célula de referência.

# (No) As restrições aos parâmetros

Consideramos

$$\begin{aligned} \lambda_{11} &= E[Y_{11}] &&= n \cdot p_{1.} \cdot p_{.1} \\ \lambda_{ij} &= E[Y_{ij}] &&= n \cdot p_{i.} \cdot p_{.j} = \lambda_{11} \cdot \frac{p_{i.}}{p_{1.}} \cdot \frac{p_{.j}}{p_{.1}} \quad , \forall\, i = 1:a \,,\, j = 1:b \end{aligned}$$

Logaritmizando, temos as relações:

$$\ln(\lambda_{ij}) = \ln(E[Y_{ij}]) = \underbrace{\ln(\lambda_{11})}_{=\mu} + \underbrace{\ln\left(\frac{p_{i.}}{p_{1.}}\right)}_{=\alpha_i} + \underbrace{\ln\left(\frac{p_{.j}}{p_{.1}}\right)}_{=\beta_j} \quad , \forall\, i \,,\, j$$

Assim surgem de forma natural as restrições $\alpha_1 = 0$ e $\beta_1 = 0$.

# (No) Um modelo log-linear a dois factores

O valor de $n$, o número total de observações, é conhecido.

Os estimadores de máxima verosimilhança dos parâmetros $\mu$, $\alpha_i$ e $\beta_j$ são dados pelas frequências relativas marginais:

$$\hat{p}_{i.} = \frac{n_{i.}}{n} \quad \text{e} \quad \hat{p}_{.j} = \frac{n_{.j}}{n} \,,$$

pelo que

$$\begin{aligned}
\hat{\mu} &= \ln\left(n \cdot \hat{p}_{i.} \cdot \hat{p}_{.j}\right) = \ln\left(n \cdot \frac{n_{1.}}{n} \cdot \frac{n_{.1}}{n}\right) = \ln\left(\frac{n_{1.} \cdot n_{.1}}{n}\right) \\
\hat{\alpha}_i &= \ln\left(\frac{\hat{p}_{i.}}{\hat{p}_{1.}}\right) = \ln\left(\frac{n_{i.}}{n_{1.}}\right) \\
\hat{\beta}_j &= \ln\left(\frac{\hat{p}_{.j}}{\hat{p}_{.1}}\right) = \ln\left(\frac{n_{.j}}{n_{.1}}\right)
\end{aligned}$$

# (No) O Desvio mede afastamento da independência

Já se viu que saturar este modelo log-linear a dois factores corresponde a prever efeitos de interacção. Nesse modelo, cada célula é livre de ter o seu valor, sem qualquer estrutura especial associada à tabela.

O Desvio do modelo sem interacção

$$D^* = -2\left(\mathscr{L}_M(\vec{\hat{\boldsymbol{\beta}}}_M) - \mathscr{L}_T(\vec{\hat{\boldsymbol{\beta}}}_T)\right)$$

corresponde ao valor da estatística de Wilks para uma comparação do submodelo (M) sem interacção (isto é, a hipótese de independência) face ao modelo saturado (T), com interacção (sem qualquer relação especial). Quanto menor $D^*$, mais os dados se comportam de acordo com a hipótese de independência. Pelo contrário, quanto maior $D^*$, menos plausível a hipótese de independência.

# (No) Exemplo: modelo para tabela de contingência

## Dados `HairEyeColor` (para ambos os sexos)

Na *data frame* `cabelo.olho` há $n = 16$ contagens numa tabela cruzando 4 côres de cabelo e 4 côres de olhos, num grupo de $N = 592$ estudantes.

```
> cabelo.olho
   contagens    cabelo     olhos
1         68     preto castanhos
2        119 castanho castanhos
3         26    ruivo castanhos
4          7    louro castanhos
5         20     preto     azuis
6         84 castanho     azuis
7         17    ruivo     azuis
8         94    louro     azuis
9         15     preto cinzentos
10        54 castanho cinzentos
11        14    ruivo cinzentos
12        10    louro cinzentos
13         5     preto    verdes
14        29 castanho    verdes
15        14    ruivo    verdes
16        16    louro    verdes
```

```
> cabeloOlho
               Cabelo
Olhos         preto  castanho  ruivo  louro
   castanhos     68       119     26      7
   azuis         20        84     17     94
   cinzentos     15        54     14     10
   verdes         5        29     14     16
```

Nota: Estes dados encontram-se na *data frame* `HairEyeColor` da distribuição base do `R`, e resultam de somar os valores relativos a ambos os sexos.

# (No) Exemplo (cont.)

```
> cabelo.glm <- glm(contagens ~ cabelo + olhos, family=poisson, data=cabelo.olho)
> summary(cabelo.glm)

Call: glm(formula = contagens ~ cabelo + olhos, family = poisson, data = cabelo.olho)
(...)
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     4.64312    0.08036  57.776  < 2e-16 ***
cabelolouro    -0.81180    0.10663  -7.613 2.68e-14 ***
cabelopreto    -0.97386    0.11294  -8.623  < 2e-16 ***
cabeloruivo    -1.39331    0.13259 -10.508  < 2e-16 ***
olhoscastanhos  0.02299    0.09590   0.240    0.811
olhoscinzentos -0.83804    0.12411  -6.752 1.46e-11 ***
olhosverdes    -1.21175    0.14239  -8.510  < 2e-16 ***
(...)
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 453.31  on 15  degrees of freedom
Residual deviance: 146.44  on  9  degrees of freedom
AIC: 241.04

Number of Fisher Scoring iterations: 5
```

Nota: Neste contexto, o modelo ajustado corresponde à hipótese de independência.

O modelo Nulo corresponde a admitir que as contagens esperadas de todas as células são iguais, sendo estimadas por $\frac{N}{n} = \frac{592}{16} = 37$.

# (No) Exemplo (cont.)

O modelo log-linear de tipo ANOVA a 2 factores, mas com efeitos de interacção corresponde, como se viu, a um modelo saturado:

```
> cabelo.glmT <- glm(contagens ~ cabelo * olhos, family=poisson, data=cabelo.olho)
> summary(cabelo.glmT)
[...]
Deviance Residuals:
[1]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
[...]
Null deviance: 4.5331e+02  on 15  degrees of freedom
Residual deviance: 5.9952e-15  on  0  degrees of freedom
AIC: 112.6
```

O teste de Wilks comparando o modelo saturado e o modelo de independência avalia (e rejeita) a hipótese de independência:

```
> anova(cabelo.glm, cabelo.glmT, test="Chisq")
Analysis of Deviance Table

Model 1: contagens ~ cabelo + olhos
Model 2: contagens ~ cabelo * olhos
      Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1         9    146.44
2         0      0.00       9   146.44 < 2.2e-16 ***
```

# (No) Exemplo (cont.)

Por definição, o desvio é a soma dos quadrados dos resíduos do desvio.

```
> sum(residuals(cabelo.glm)^2)
[1] 146.4436
```

A soma dos quadrados dos resíduos de Pearson tem um valor próximo.

```
> sum(residuals(cabelo.glm, type="pearson")^2)
[1] 138.2898
```

Esta última soma de quadrados é também o valor da usual estatística do teste $\chi^2$ de independência:

```
> chisq.test(cabeloOlho)
Pearson's Chi-squared test
data:  cabeloOlho
X-squared = 138.29, df = 9, p-value < 2.2e-16
```

# (No) Tabelas de contingência (cont.)

O exemplo de uma tabela de dupla entrada foi sobretudo ilustrativo. O interesse maior de modelos log-lineares corresponde ao estudo de tabelas definidas por três ou mais factores.

A diferentes conceitos de independência envolvendo três ou mais factores (independência, independência mútua, independência conjunta, independência condicional, etc.) correspondem diferentes modelos log-lineares.

A validade de um ou outro conceito de independência pode ser estudada através da qualidade do ajustamento do correspondente modelo.

# (No) Tabela de independências

A tabela indica as designações mnemónicas para os vários tipos de modelos considerados até aqui.

| Notação | Tipo de Modelo | Equação do Modelo para $\ln(\lambda_{ijk})$ | Relação-base |
|---|---|---|---|
| (A,B,C) | Independência Mútua | $\mu+\alpha_i+\beta_j+\gamma_k$ | $p_{ijk}=p_{i..}\cdot p_{.j.}\cdot p_{..k}$ |
| (B:C) | Ind. conjunta (B,C) com A | $\mu+\alpha_i+\beta_j+\gamma_k+(\beta\gamma)_{jk}$ | $p_{ijk}=p_{i..}\cdot p_{.jk}$ |
| (A:B) | Ind. conjunta (A,B) com C | $\mu+\alpha_i+\beta_j+\gamma_k+(\alpha\beta)_{ij}$ | $p_{ijk}=p_{ij.}\cdot p_{..k}$ |
| (A:C) | Ind. conjunta (A,C) com B | $\mu+\alpha_i+\beta_j+\gamma_k+(\alpha\gamma)_{ik}$ | $p_{ijk}=p_{i.k}\cdot p_{.j.}$ |
| (A:C,B:C) | Ind. (A,B) condicional a C | $\mu+\alpha_i+\beta_j+\gamma_k+(\alpha\gamma)_{ik}+(\beta\gamma)_{jk}$ | $p_{ijk}=\frac{p_{i.k}\cdot p_{.jk}}{p_{..k}}$ |
| (A:B,B:C) | Ind. (A,C) condicional a B | $\mu+\alpha_i+\beta_j+\gamma_k+(\alpha\beta)_{ij}+(\beta\gamma)_{jk}$ | $p_{ijk}=\frac{p_{ij.}\cdot p_{.jk}}{p_{.j.}}$ |
| (A:B,A:C) | Ind. (B,C) condicional a A | $\mu+\alpha_i+\beta_j+\gamma_k+(\alpha\beta)_{ij}+(\alpha\gamma)_{ik}$ | $p_{ijk}=\frac{p_{ij.}\cdot p_{i.k}}{p_{i..}}$ |
| (A:B:C) | Modelo Saturado | $\mu+\alpha_i+\beta_j+\gamma_k+(\alpha\beta)_{ij}+(\alpha\gamma)_{ik}+(\beta\gamma)_{jk}+(\alpha\beta\gamma)_{ijk}$ | |