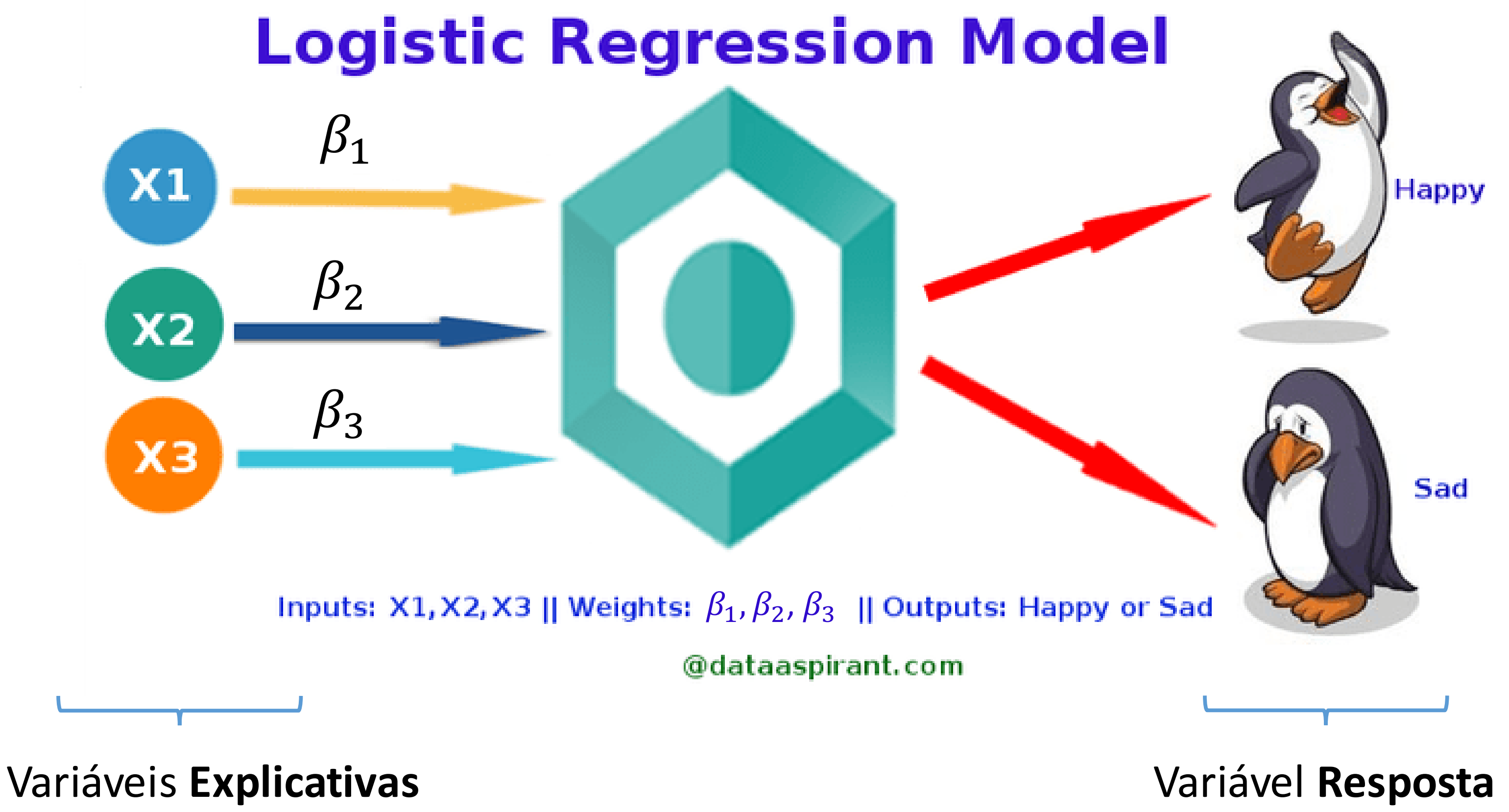
What is Logistic Regression?

Probabilistic model that aims to **explain** and/or **predict** a binary variable from a set of explanatory variables of any type, given a set of observations. (Berkson, 1944; Cox, 1960's)

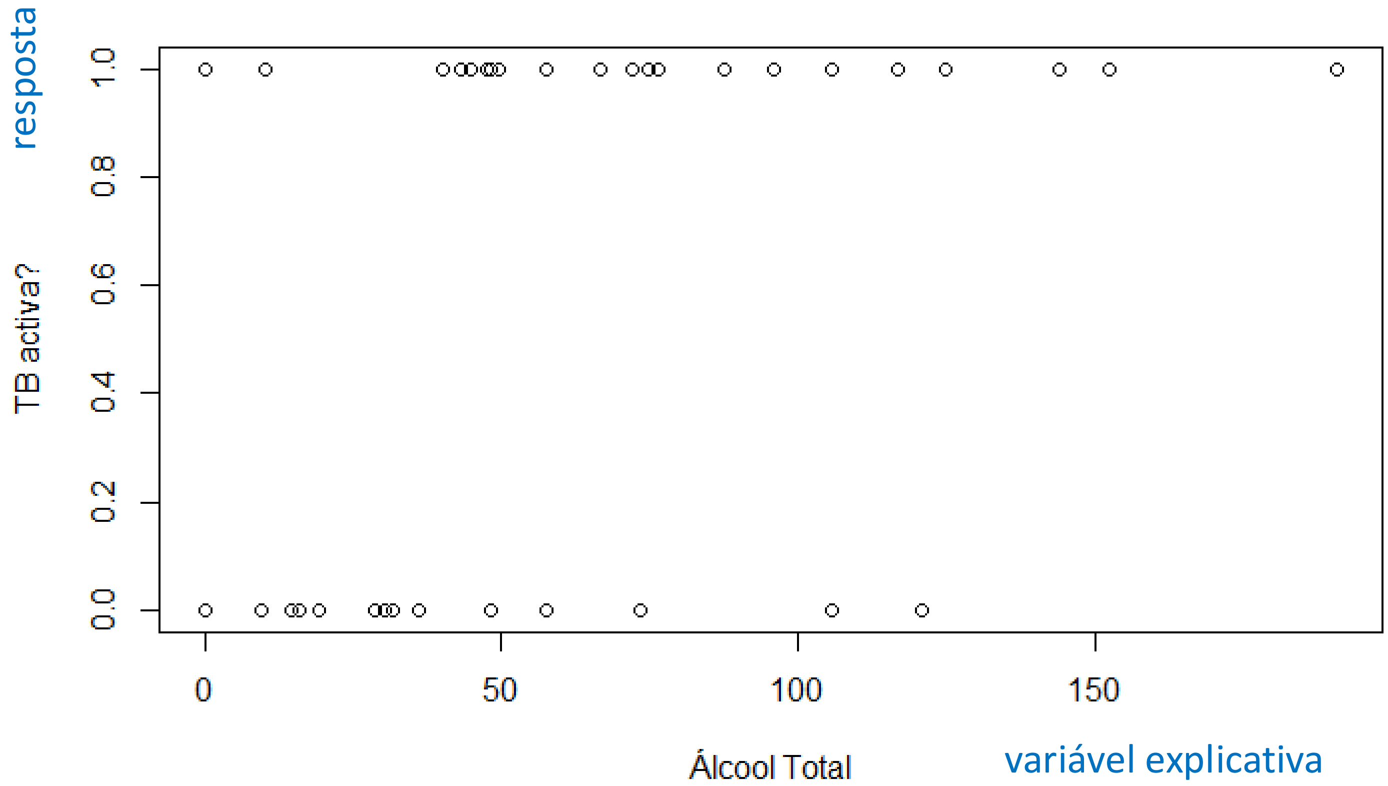


Examples:

- to study the effect of alcohol consumption on the existence (yes/no) of active pulmonary tuberculosis (TB)
- to evaluate the association between the existence (yes/no) of defects in a part and the material and temperature used in its production
- to evaluate the germination of a seed (yes/no) as a function of several experimental conditions
- to predict an individuals's voting behaviour (against or in favor of a political candidate) as a function of his/her's education level, ideologies, race and gender.

How should the response variable be described as a function of the explanatory variables?

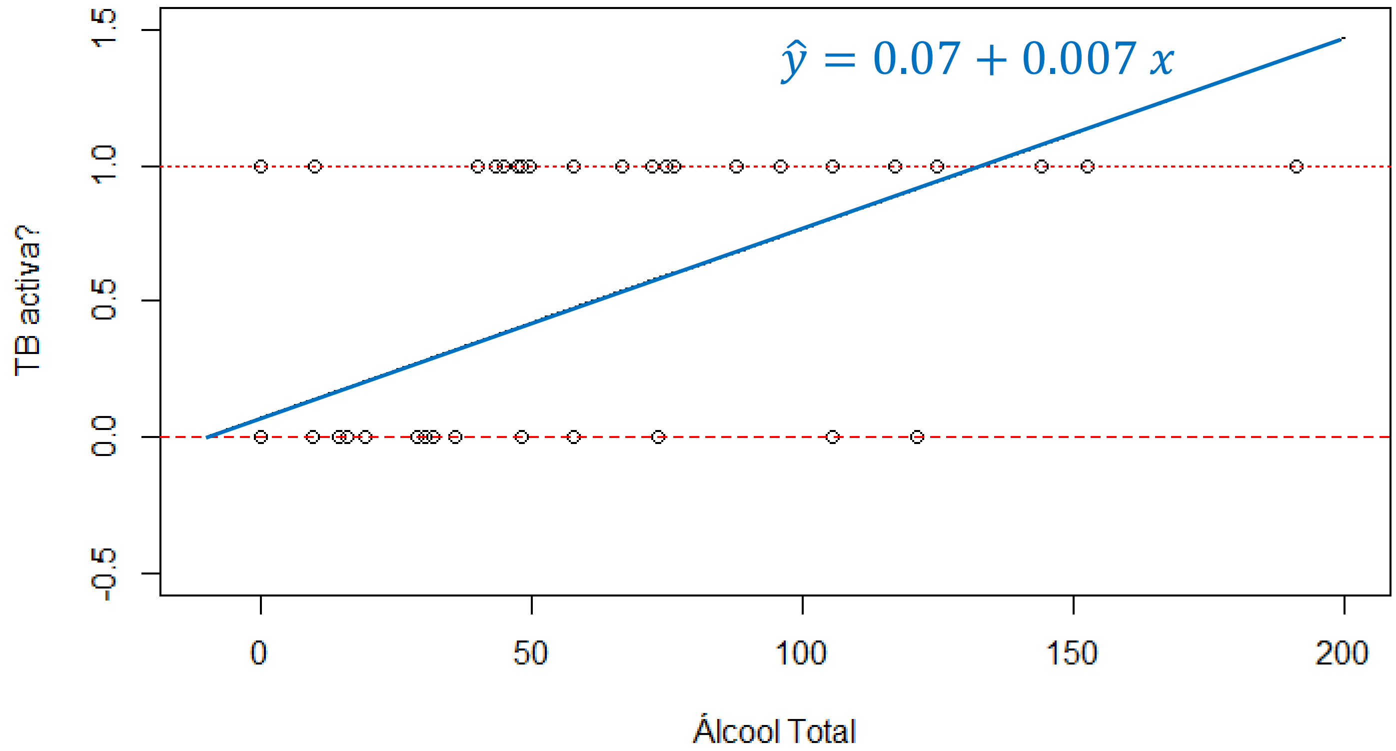
Example: effect of the total alcohol consumption on the existence (yes/no) of active pulmonary TB



Y_i: success occurence (1:yes/0:no) on individual i (random variable)

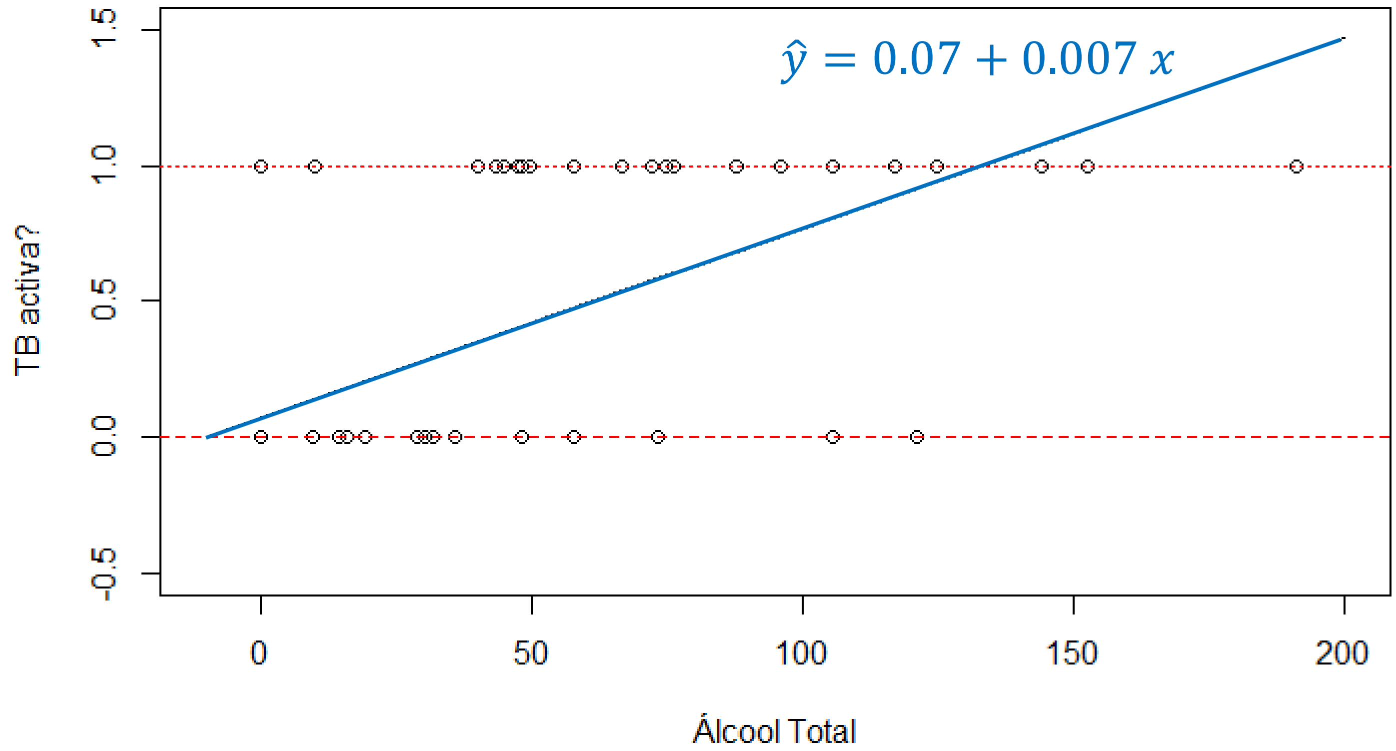
Response as a function of the explanatory variable?

Linear regression...



Response as a function of the explanatory variable?

Linear regression...



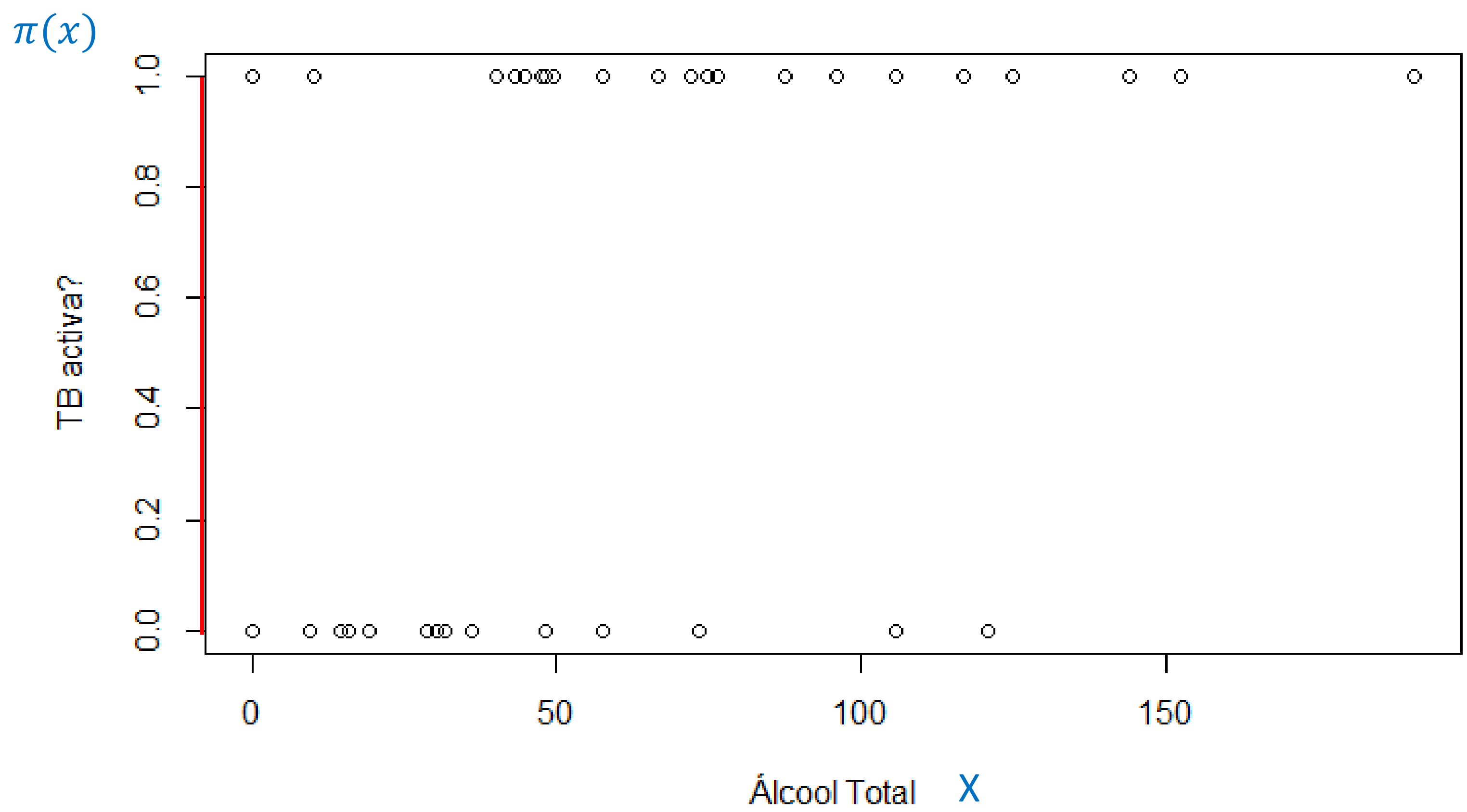
To model the **conditional probability of the occurence of the success**:

$$\boxed{\pi(x) = P(Y = 1 \mid X = x)} \qquad \text{(unknown)}$$

Response as a function of the explanatory variable?

Function $\pi(x) = P(Y = 1 \,|\, X = x)$:

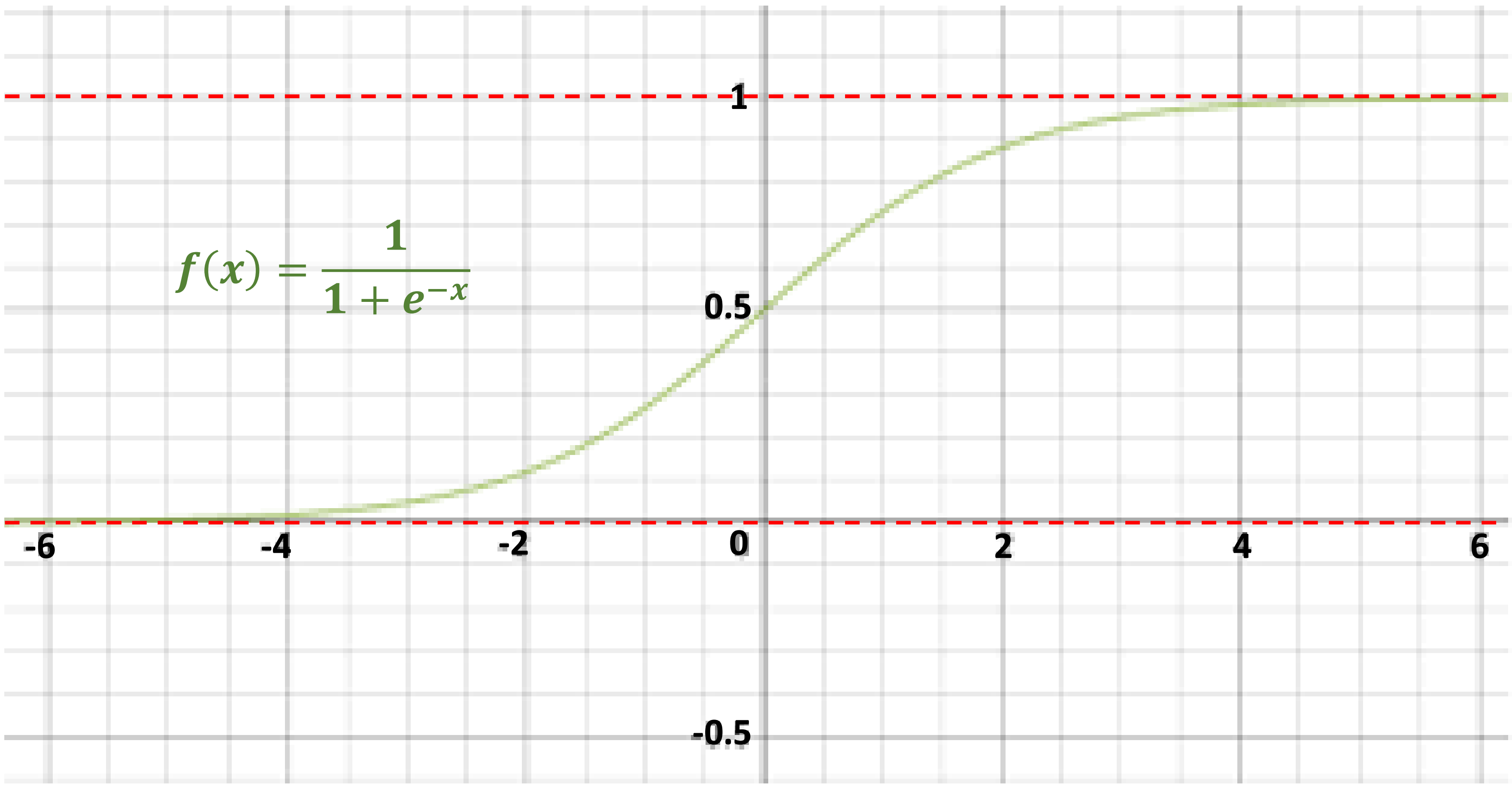
- takes values between 0 and 1
- graph with an S-shape (epidemiological interpretation...)



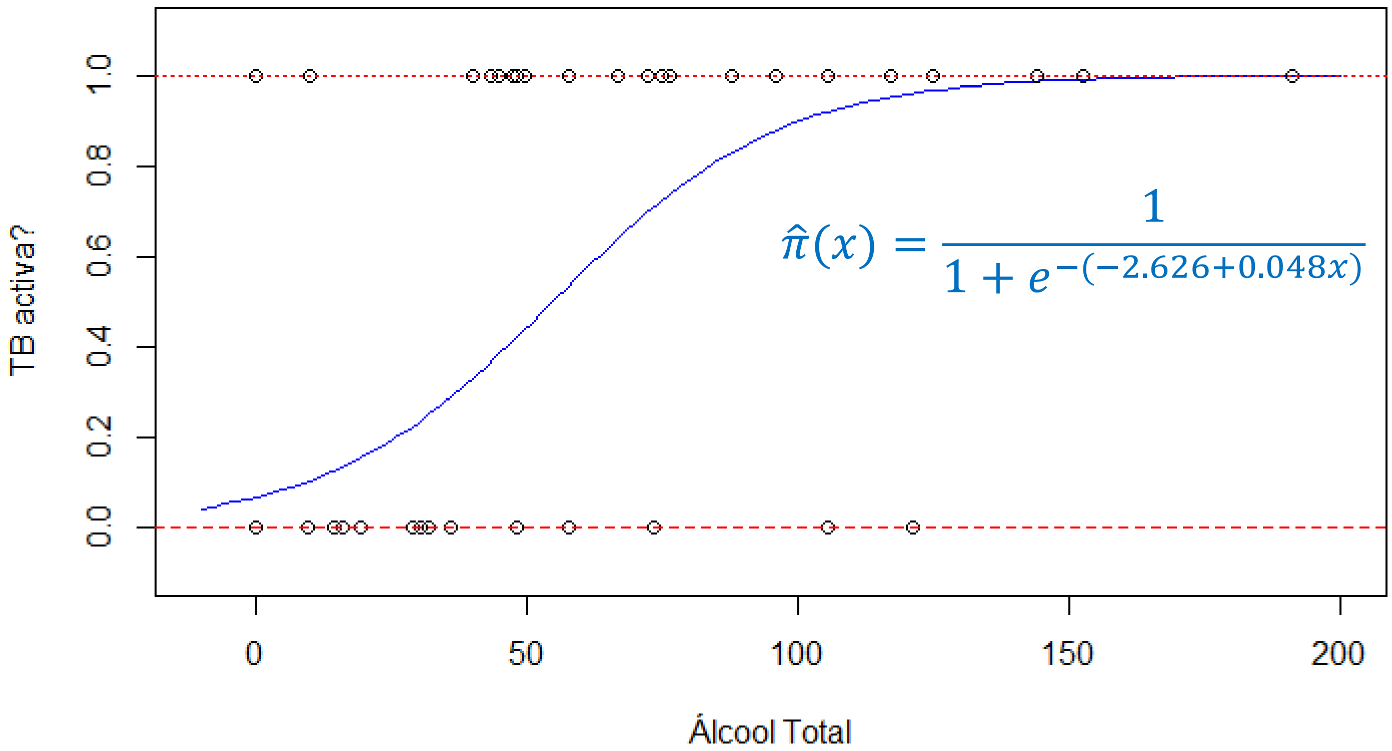
Response as a function of the explanatory variable

Logistic model for populacional growth (Verhulst, 1838)

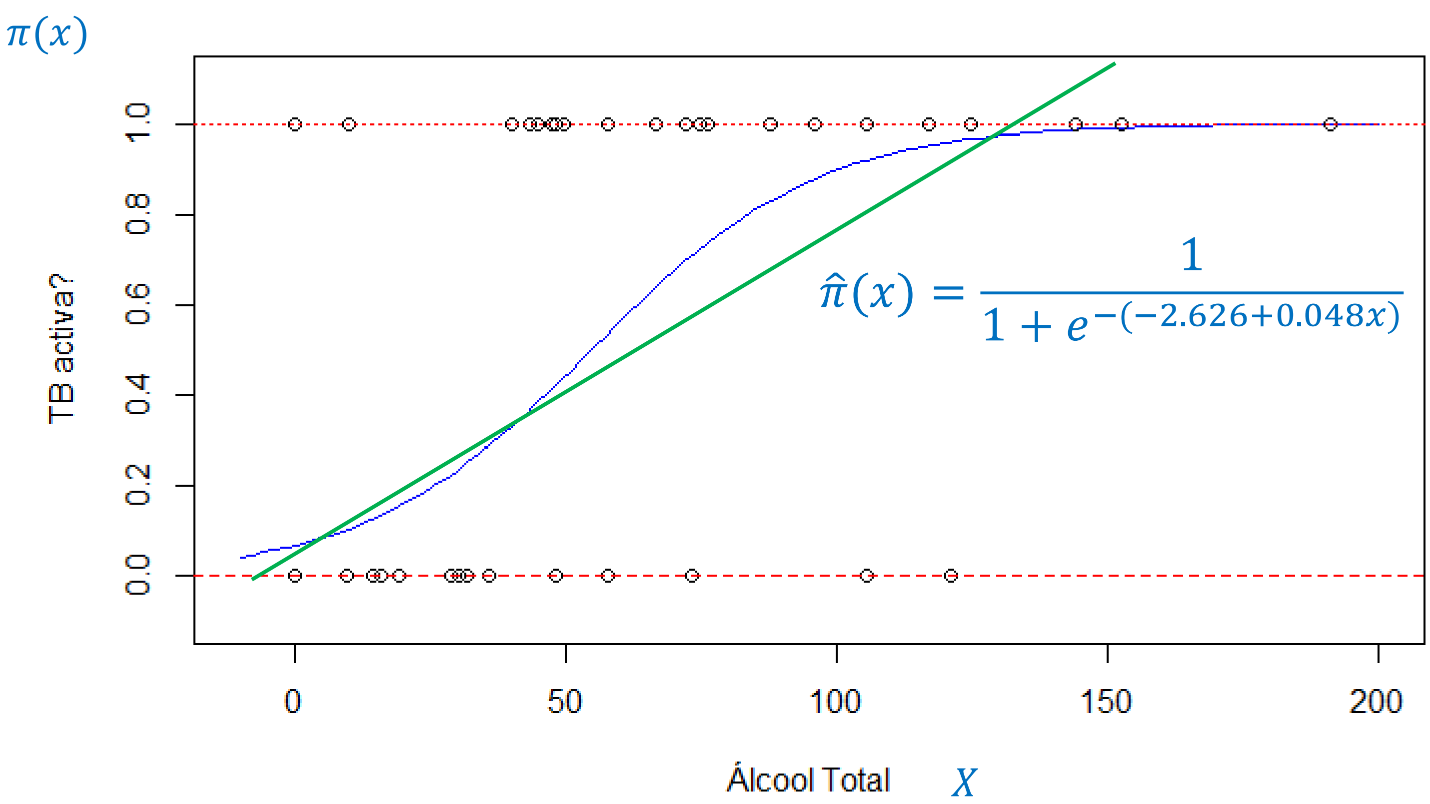
$$f(x) = \frac{1}{1 + be^{-rx}}$$



Response as a function of the explanatory variable

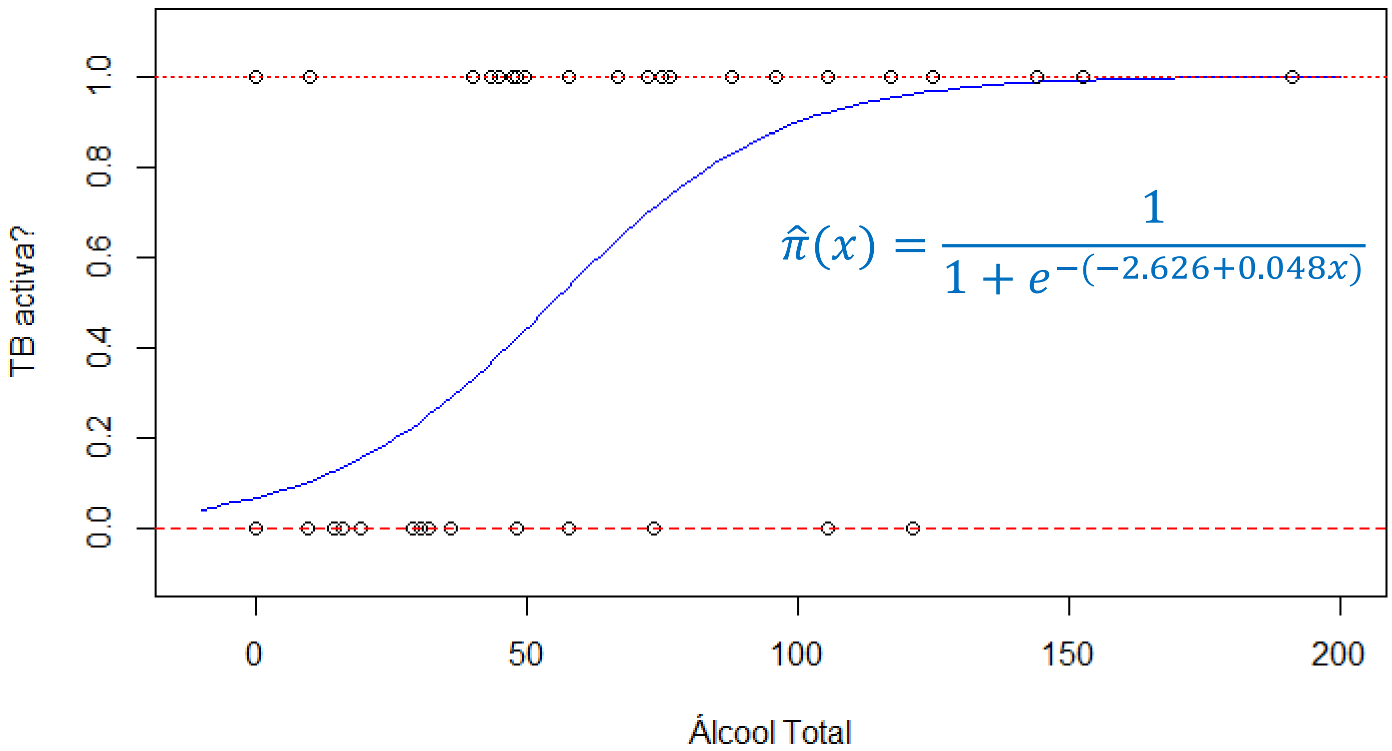


Response as a function of the explanatory variable



Simple (*univariate...*) logistic regression model

$$\pi(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \Leftrightarrow \boxed{\log\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \beta_0 + \beta_1 x}$$



Multiple (*multivariate...*) logistic regression model

- $X_1, X_2, \ldots, X_p$ explanatory variables
- $x = (x_1, x_2, \ldots, x_p)$ vector of observations from an individual
- $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ vector of parameters

$$\boxed{\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p}$$

$$\pi(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}}$$

Odds and Odds-Ratio

Definition: The ***odds* of an event** is $\frac{P(event)}{1-P(event)}$ $\left(= \frac{P(event)}{P(non\,event)}\right)$

Let E be an exposure variable (1: exposed; 0: nonexposed)

For the model $\boxed{\log\left(\frac{\pi(E)}{1-\pi(E)}\right) = \beta_0 + \beta_1 E}$ it can be shown that

$$e^{\beta_1} = \frac{\text{odds}(Y=1|E)}{\text{odds}(Y=1|\overline{E})} = OR(Y=1|E \text{ vs } \overline{E}) \qquad \text{Odds Ratio}$$

Odds and Odds-Ratio

Suppose $Y = 1$ denotes the presence of a disease and $OR(Y = 1|E \; vs \; \overline{E}) = 0.2$. Then:

- the odds for the disease among the exposed individuals is 20% of the odds for the disease among the nonexposed individuals
- the odds for the disease among the nonexposed individuals is 5 times the the odds for the disease among the exposed individuals
- the odds for the non-existence of the disease among the exposed individuals is 5 times the odds for the non-existence of the disease among the nonexposed individuals

Fact: $OR(Y = 1|E \; vs \; \overline{E}) = OR(E = 1|Y \; vs \; \overline{Y})$
Odds-ratio is invariant under study design (cohort or case-control)

Remark: Odds-Ratio and Relative Risk

The relative risk of the exposure variable E on the response Y is

$$RR = \frac{P(Y=1|E)}{P(Y=1|\overline{E})}$$

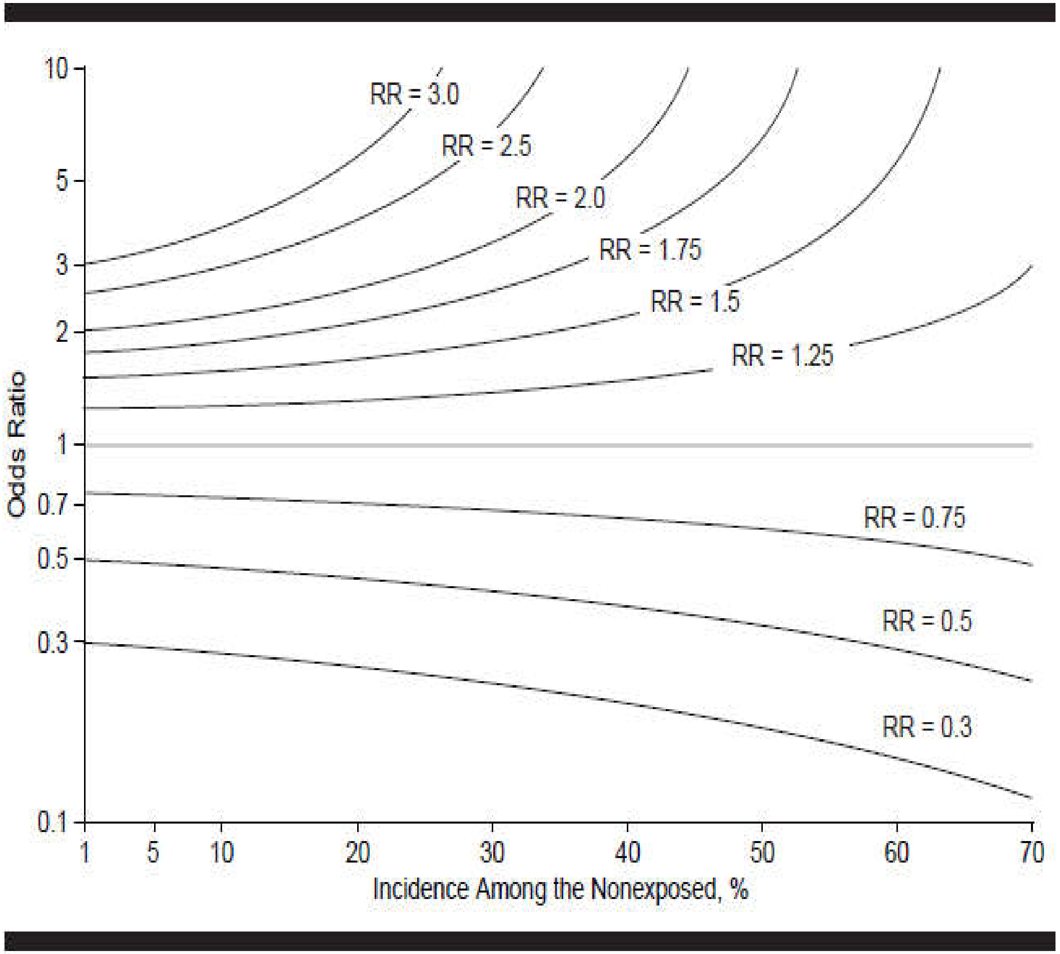
Rmk: If Y is binary, then P(Y=1| E)=E(Y|E)

It follows from the definitions of OR and RR that

$$\frac{OR}{RR} = \frac{1-P(Y=1|\overline{E})}{1-P(Y=1|E)}.$$

In particular, if $Y=1$ denotes the presence of a disease, **OR and RR are close** whenever the **disease is rare**

Remark: Odds-Ratio and Relative Risk



Relationship between OR and RR depending on the incidence of the outcome among the nonexposed(JAMA- Journal of the American Medical Association, 1998).

The 3 most used models

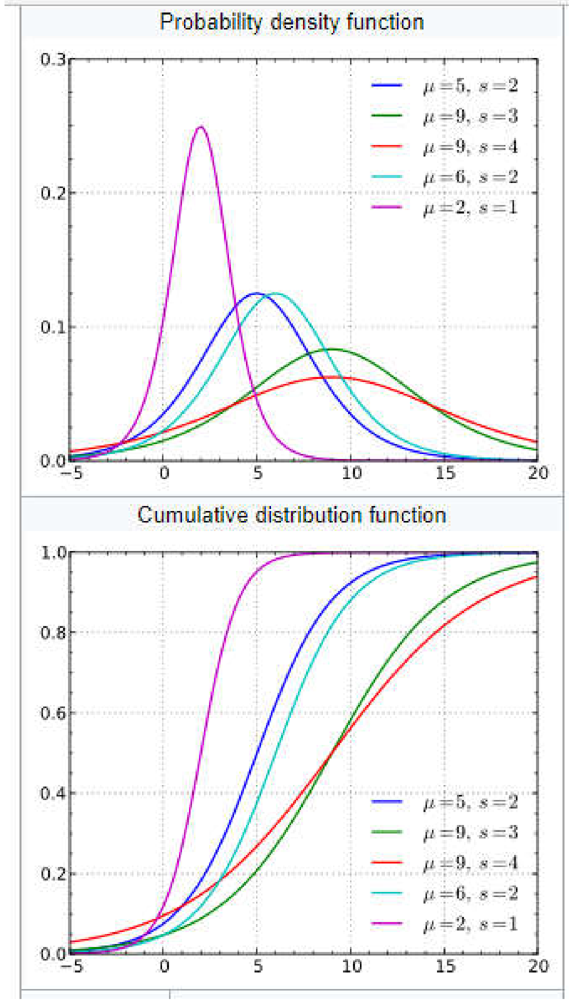
$$\boxed{\mathit{logit}(\pi(x)) = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p}$$

$$\boxed{\mathit{probit}(\pi(x)) = \Phi^{-1}(\pi(x)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p}$$

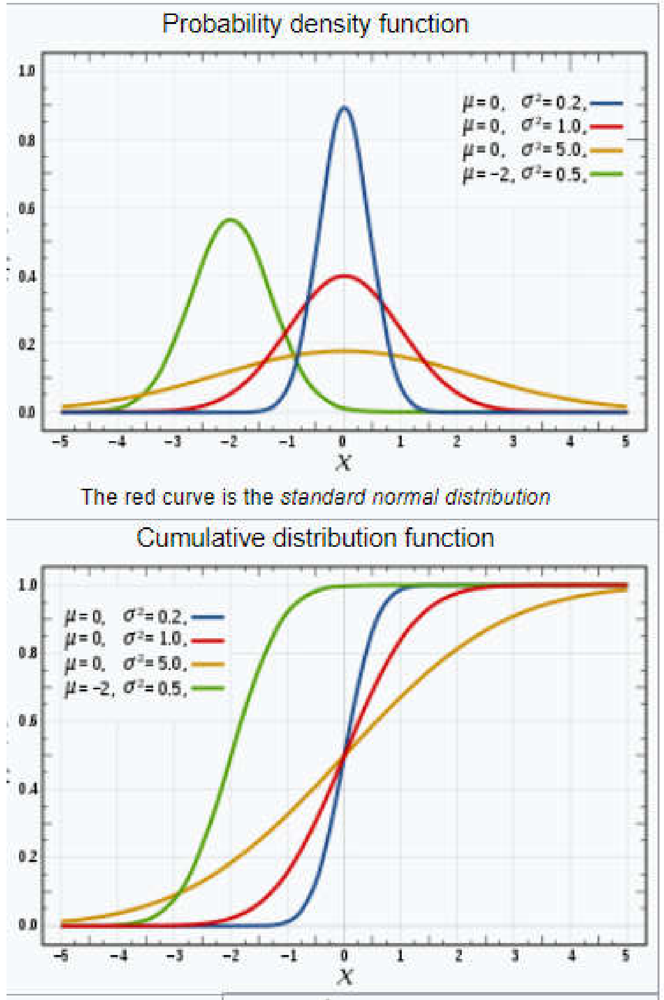
$$\boxed{\log(-\log(1-\pi(x))) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p}$$

Response as a function of the explanatory variable

Logistic Distribution



Normal (Gaussian) Distribution



The logistic model is the most popular

- all models can be applied to prospective data
- Breslow & Day, 1981; Prentice & Pike, 1979: the logistic model can be applied to either retrospective or cross-sectional data
  - **Prospective studies (cohort)**:

$$X_1, X_2, \ldots, X_p \quad \Longrightarrow \quad Y$$

    Ex: select a sample of newborns and register their sex, age, breastfeed or formula feed, ... Follow the sample for a year and register the occurence (or not) of respiratory infections
  - **Retrospective Studies (case-control)**:

$$X_1, X_2, \ldots, X_p \quad \Longleftarrow \quad Y$$

    Ex: several newborns go to a hospital with respiratory infections; their sex, age and method of feeding is registered.

The logistic model is the most popular

- easy (epidemiological) interpretation of the results:
  - direct modelling of the logarithm of the odds for success
  - $\exp(\beta_i)$: **odds-ratio**
  - $\pi(x)$: **risk for the disease** under conditions x (only for prospective data)
- the logit is the canonical link function for the binomial distribution hence there exists a sufficient and minimal statistics for β
- is implemented in the most common softwares of statistical analyses (SPSS, R, STATA, SAS, ...)

How are categorical explanatory variables included?

- categorical variables (gender, age group, severity of a disease, ...) are represented by a set of auxiliary variables, denoted by **dummy variables**, or simply **dummies**.
- a categorical variable X with $k+1$ categories, $\{1, 2, ..., k, k+1\}$ requires k binary dummies Z_1, Z_2, ..., Z_k
- for $i \in \{1, 2, ..., k\}$ each dummy Z_i is the indicator variable for category i:

$$
\begin{aligned}
Z_i(x) &= 1 \qquad \text{if} \quad x = i \\
Z_i(x) &= 0 \qquad \text{if} \quad x \neq i
\end{aligned}
$$

How are categorical explanatory variables included?

- there is no dummy for the last category, $k+1$; this is sad to be the **reference category**.
- the reference category can be any category of X; without loss of generality, it was chosen to be the last one, above.
- in Epidemiology, it is common to choose the reference category as the one that is the *healthiest* and usually not associated with the outcome.

How are categorical explanatory variables included? - example

Let X denote the age class of an individual, with the following possible values

$$1 :< 40\,\text{anos}, \qquad 2 : 40 \leq \text{anos} \leq 65, \qquad 3 :> 65\,\text{anos}.$$

For example, let hte first class be the reference class. The variable X will be represented by **two dummies**, Z_2 and Z_3, associated with classes 2 and 3, respectively.

- a 44 years-old individual is represented by $(z_2, z_3) = (1, 0)$
- a 32 years-old individual is represented by $(z_2, z_3) = (0, 0)$
- a 68 years-old individual is represented by $(z_2, z_3) = (0, 1)$

Remark: Due to the use of dummies, it is recommended to code all binary variables using 0's and 1's. The reference category will be that associated with 0.

Interpretation of the parameters β

- Model 1:

$$\log\left(\frac{\pi(E)}{1-\pi(E)}\right) = \beta_0 + \beta_1 E, \qquad E \text{ exposure variable}$$

It was already seen that

$$e^{\beta_1} = OR(Y = 1 | E \, vs \, \overline{E}).$$

- Model 2:

$$\log\left(\frac{\pi(X)}{1-\pi(X)}\right) = \beta_0 + \beta_1 X, \qquad X \text{ continuous variable}$$

It can be seen that (maths...)

$$e^{\beta_1} = OR(Y = 1 | X + 1 \, vs \, X).$$

Question: what happens to OR when X increases 3 units?

Interpretation of the parameters β

- Model 3:

$$\log\left(\frac{\pi(X)}{1-\pi(X)}\right) = \beta_0 + \beta_1 X, \qquad X \text{ categorical with } k \text{ categories}$$

The model has to include the $k-1$ dummies of X, say X_1, ..., X_{k-1}:

$$\log\left(\frac{\pi(X)}{1-\pi(X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{k-1} X_{k-1}$$

Assuming that the k^{th} category is the reference category, it can be shown that (maths...)

$$\begin{aligned} e^{\beta_1} &= OR(Y=1|X_1 \; vs \; X_k) \\ e^{\beta_2} &= OR(Y=1|X_2 \; vs \; X_k) \\ &\ldots \\ e^{\beta_{k-1}} &= OR(Y=1|X_{k-1} \; vs \; X_k) \end{aligned}$$

Interpretation of the parameters β

- Model 4: (generic model)

$$\log\left(\frac{\pi(X)}{1-\pi(X)}\right) = \beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_p X_p, \quad X_1,\ldots,X_p \text{ of any type}$$

Each parameter β_i can only be interpreted whenever all the remaining variables X_1, ..., X_{i-1}, X_{i+1}, ..., X_p are fixed

Interpretation of the parameters - example

A study conducted by Payne, 1987,[1] comprised 2074 children less than 1 year-old and its goal was to relate the incidence of pulmonary infections with the type of milk being administered and the sex of the child.

	Only Formula Milk	Breast Feeding with Supplement	Only Breast Feeding
Boys	77/458	19/147	47/494
Girls	48/384	16/127	31/464

We say we have 6 **covariate patterns**.[2]

- binary response: each studied children either has or not a pulmonary infection
- explanatory variables: sex (2 categories) and type of milk (3 categories - 2 dummies)

[1] Payne, C. (Ed.), *The GLIM System Release 3.77 Manual*. Oxford: Numerical Algorithms Group

[2] *padrões de covariáveis*

Interpretation of the parameters - example

Results from the fitting of the logistic regression model:

	Estimate	Std. Error	z value	p-value
Intercept	-1.613	0.112	-14.35	<0.001
sexGirl	-0.313	0.141	-2.22	0.027
foodBreast	-0.669	0.153	-4.37	< 0.001
foodSuppl	-0.173	0.206	-0.84	0.401

Start by noting the reference categories:

- the reference category for sex is 'being a boy'
- the reference category for type of feeding is 'only formula milk'

Interpretation:

- $\widehat{\beta}_0 = -1.613 \rightarrow$ the odds for a pulmonary infection in boys receiving only adapted milk is $e^{-1.613} = 0.20$. The probability ~~for~~ of not having an infection is 5 times greater than the probability ~~for~~ of having an infection.

Interpretation of the parameters - example

	Estimate	Std. Error	z value	p-value
Intercept	-1.613	0.112	-14.35	$<$0.001
sexGirl	-0.313	0.141	-2.22	0.027
foodBreast	-0.669	0.153	-4.37	$<$ 0.001
foodSuppl	-0.173	0.206	-0.84	0.401

Interpretation (cont'n):

- $\widehat{\beta}_{sexGirl} = -0.313 \rightarrow OR(infection|Girl\ vs\ Boy) = exp(-0.313) = 0.73$
  The odds ratio for the infection among girls is 0.73 times the odds ratio among boys, for the same type of feeding.
  Equivalently, the odds for the infection among girls is 100% - 73% = 27% lower than the odds among boys.
  Or else, the odds for the infection among boys is $1/0.73 = 1.37$ times the odds among girls (therefore 37% higher).

  Being a boy is positively associated with the infection (while being a girl is negatively associated)

Interpretation of the parameters - example

	Estimate	Std. Error	z value	p-value
Intercept	-1.613	0.112	-14.35	<0.001
sexGirl	-0.313	0.141	-2.22	0.027
foodBreast	-0.669	0.153	-4.37	< 0.001
foodSuppl	-0.173	0.206	-0.84	0.401

Interpretation (cont'n):

- $\beta_{foodBreast} = -0.669 \rightarrow$ the odds for an infection among children being breastfed is $exp(-0.669) = 0.51$ times the odds for an infection among children being fed only with formula milk.
  Equivalently, in comparison with formula milk, breast-feeding reduces the odds for an infection by approximately half.

Estimation of the parameters β: maximum likelihood

- $\log\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}$
- $Y_1, \ldots, Y_n$ random sample (r.s.)
  $Y_i \sim B(1, \pi(x_i))$; $\quad P(Y_i = y_i) = \pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i}$
- $y = (y_1, \ldots, y_n)$ realization of the random sample

likelihood function $\qquad L(\beta \mid y) = \prod_{i=1}^{n} \pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i}$

Problem 1: To find β maximizing L.

$$\begin{aligned}
\ell(\beta \mid y) &= \log(L)(\beta|y) \\
&= \sum_{i=1}^{n} \left(y_i \ln(\pi(x_i)) + (1-y_i)\ln(1-\pi(x_i))\right) \\
&= \sum_{i=1}^{n} \left(y_i x_i^t \beta - \log(1 + e^{x_i^t \beta})\right) \qquad x_i^t = (1, x_{1i}, \ldots, x_{pi})
\end{aligned}$$

Problem 2: To find β maximizing ℓ.

Estimation of the parameters β: maximum likelihood

In *several* models, $\ell(\beta)$ is strictly concave and upper bounded hence it has a unique (global) maximum.

Maximum Likelihood Estimator **MLE**:

$$\widehat{\beta} \quad \text{such that} \quad \underbrace{\frac{\partial \ell}{\partial \beta_j}(\widehat{\beta}) = 0, \quad j = 0, 1, \ldots, p}_{\text{likelihood equations } (p+1)}$$

The likelihood equations are nonlinear on β
$\Rightarrow$ estimation of $\widehat{\beta}$ requires iterative numerical algorithms

Estimation of β: properties of the MLE

$\widehat{\theta}_{MV}^{(n)}$ MLE of θ associated with a r.s. $Y_1, Y_2, \ldots, Y_n$

$\theta^{\#}$ real value of θ

(a) assymptotic existence and uniqueness

(b) $\widehat{\theta}_{MV}^{(n)}$ assimptotically unbiased

$$E(\widehat{\theta}_{MV}^{(n)}) \overset{n \to +\infty}{\longrightarrow} \theta^{\#}$$

(c) $\widehat{\theta}_{MV}^{(n)} \overset{a}{\sim} N(\theta^{\#}, I^{(-1)}(\theta^{\#}))$

(d) $\ldots$ (consistency, efficiency, sufficiency, invariance)

★ One can use hypothesis tests (and confidence intervals) to test $H_0 : \beta = \beta_0$ and to evaluate the goodness of fit of the model

Problems in the estimation process

- ★ the Fisher algorithm does not converge
- ★ MLE with infinite values
- ★ finite LME values but with large standard deviations
- ★ compromised assymptotic convergence (invalid inferences)

Separability

- the success has a *very* low (or high) prevalence
  (assymptotic properties of the MLE are not applicable; inference from hypothesis tests is not valid)

  Eg: Identification of factors associated with low-weight newborns. In 320 newborns, 14 (4.4%) were low-weighted.

- for a particular (combination of) explanatory variables, *almost all* observations correspond to successes (or failures)

  Eg: next file

- there exists a continuous explanatory variable that *essencially* predicts the success

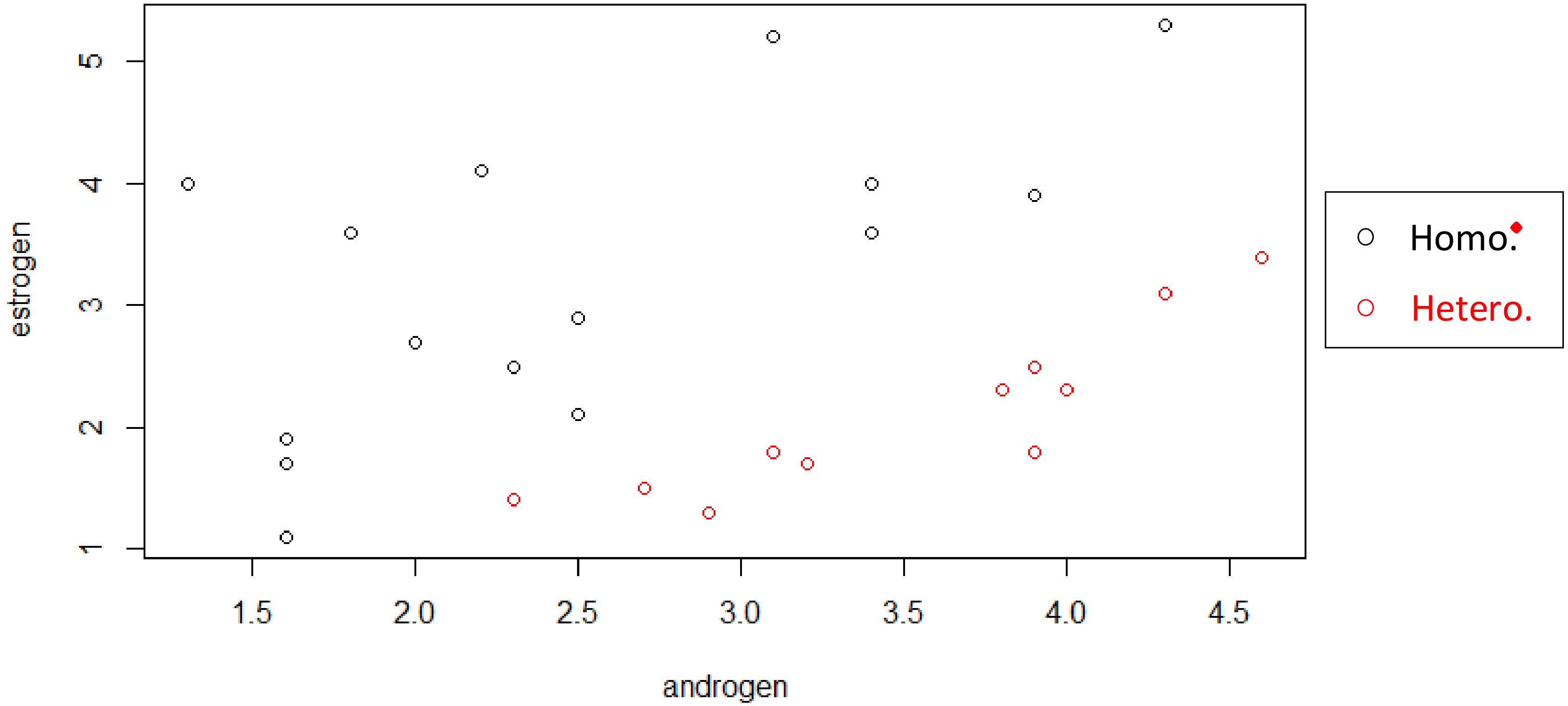
  Eg: next file

Separability

Variables	Total (n=48)	GOS 1-3 (n=9)	GOS 4-5 (n=39)	Simple Analysis OR(95%CI)	p-value
Initial BRS	11.7 (1.5-38.7)	4.3 (1.5-13.9)	12.6 (3.2-38.7)	**1.359 (1.088, 1.698)**	**0.007**
WFNS (2 cat)					
I (1+2)	35 (72.9)	2 (22.2)	33 (84.6)	**1.0**	
II (3+4+5)	13 (27.1)	7 (77.8)	6 (15.4)	**0.052 (0.009, 0.313)**	**0.001**
WFNS (2 cat)					
I				19.231 (3.195, 111.111)	
II				1.0	
Fisher (2 cat)					
I (1+2+3)	25 (52.1)	1 (11.1)	24 (61.5)	**1.0**	
II (4)	23 (47.9)	8 (88.9)	15 (38.5)	**0.078(0.009, 0.688)**	**0.022**
Fisher					
I				12.821 (1.453, 111.111)	
II				1.0	
Sedation					
No	41 (85.4)	4 (44.4)	37 (94.9)	**1.0**	
Yes	7 (14.6)	5 (55.6)	2 (5.1)	**0.043 (0.006, 0.300)**	**0.001**
Sedation					
No				23.256 (3.333, 166.667)	
Yes				1.0	
GCS	15 (3-15)	5 (3-13)	15 (8-15)	**2.151 (1.324, 3.494)**	**0.002**
HR_Avg (n=47)	68.0 (44.3-105.3)	72.5 (53.9-105.3)	65.97 (44.3-85.1)	**0.928(0.866, 0.996)**	**0.037**

Separability

- 26 adult men (Margolese, 1970)
- **Question:** Is it possible to predict the sexual orientation (only classified as straight/gay) only from the androgen and estrogen values?



Separability

```
➢mod1 <- glm(orientation ~ estrogen +androgen,
             data = hormone, family = binomial)

  Warning messages:
  1: glm.fit: algorithm did not converge
  2: glm.fit: fitted probabilities numerically 0 or 1 occurred

➢summary(mod1)

 Coefficients:  Estimate   Std. Error   z value   Pr(>|z|)
 (Intercept)      -84.49    136095.03    -0.001      1.000
  estrogen        -90.22     75910.98    -0.001      0.999
  androgen        100.91     92755.62     0.001      0.999

Null deviance: 3.5426e+01 on 25 degrees of freedom
Residual deviance: 2.3229e-09 on 23 degrees of freedom
```

A very good fitting but no significant explanatory variables!

Separability: solutions

- exact logistic regression (1970's) (generalization of the Fisher's test)
- Firth logistic regression (1993)
- logistic regression with bias correction (King & Zheng, 2001)
- conditional logistic regression (small number of cases)
- Bayesian methods
- ...

Referências bibliográficas

- **Hosmer, Lemeshow; Applied Logistic Regression, 2nd edition, Wiley Series in Probability and Statistics, 2000**
- Kleinbaum, Klein; Logistic Regression, Springer Verlag
- Faraway; Extending the Linear Model with R, Chapman and Hall/CRC, 2006. (available online).
- Chatterjee, Hadi; Regression Analysis by Example
- McCullag, Nelder; Generalized Linear Models
- Fahrmeir; Multivariate Statistical Modelling Based on Generalized Linear Models, Springer Verlag
- German Rodriguez; Princeton University, online lecture notes, http://data.princeton.edu/wws509/
- Agresti; An introduction to categorical data analysis, 2nd edition, Wiley Series in Probability and Statistics, 1996.