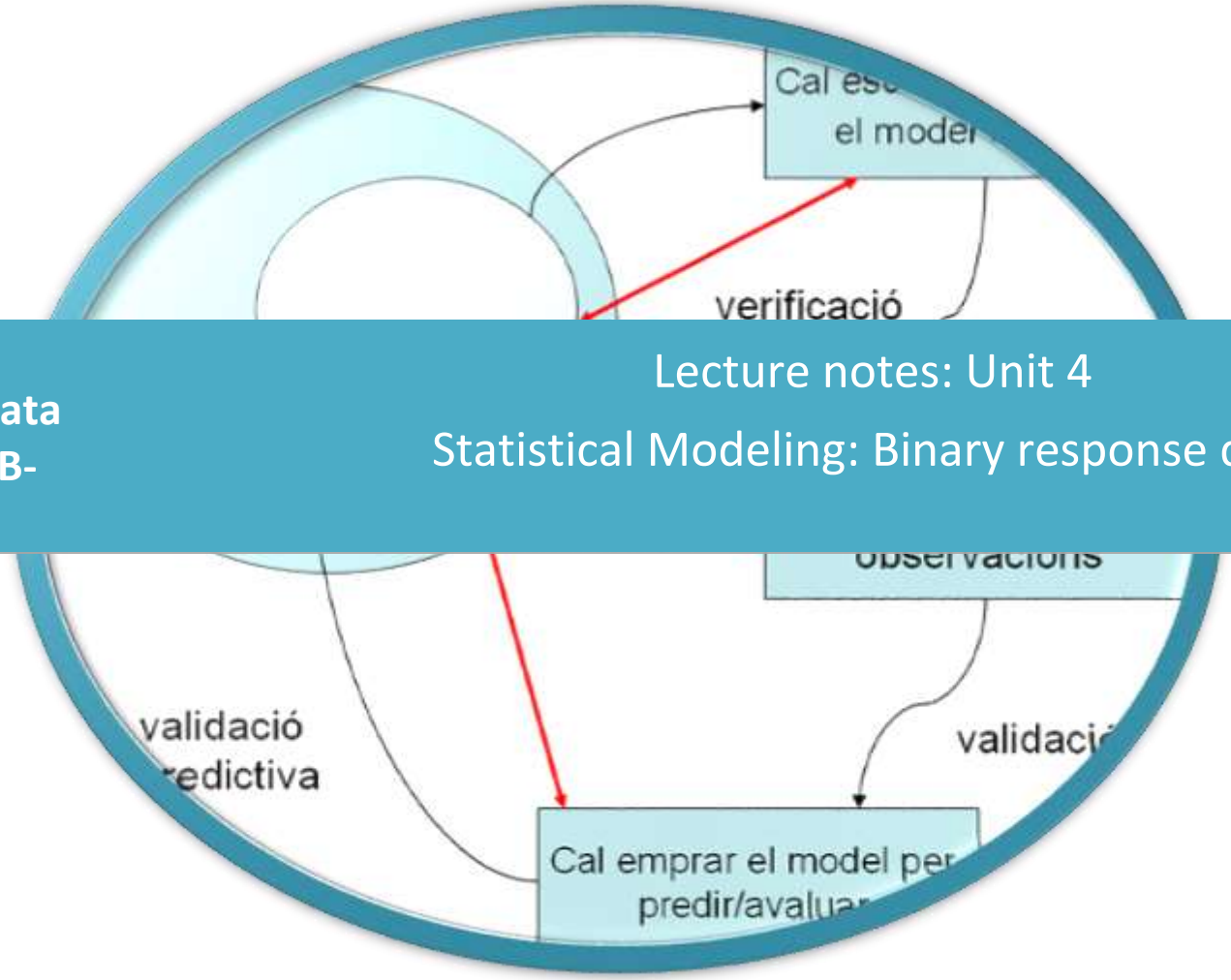
UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Departament d'Estadística
i Investigació Operativa

2022-2023 Academic Year



SIM course.
Master in Data Science – FIB-UPC

Lecture notes: Unit 4

Statistical Modeling: Binary response data

TABLE OF CONTENTS

4-1. BINARY RESPONSE DATA. BINOMIAL MODELS

4-1.1 Components of generalized linear models

Generalized linear models are extensions of classic multiple regression models.

Let $\mathbf{y}^T = (y_1, \ldots, y_n)$ be a vector of n components randomly drawn from vector $\mathbf{Y}^T = (Y_1, \ldots, Y_n)$, whose variables are statistically independent and distributed with expectation $\boldsymbol{\mu}^T = (\mu_1, \ldots, \mu_n)$:

The random component assumes that mutual independence holds and each random variable in $\mathbf{Y}^T = (Y_1, \ldots, Y_n)$ belongs to the exponential family with one parameter distribution $Y_i|X_i \sim F(.; \mu_i, \phi) = Bern(\pi_i), \phi = 1 \; or \; m_i$ and expected values $E(Y_i|X_i) = \mu_i = m_i\pi_i$ and $V(Y_i|X_i) = \phi\pi_i(1-\pi_i)$.

➡ **At the disaggregated level for each individual observation, the response is dichotomous and we are dealing with the Bernoulli distribution. For grouped data, binomial distributions are suitable.**

- **The systematic component in the model** specifies a vector $\boldsymbol{\eta}$. The linear predictor vector is a linear combination from a limited number of explanatory variables $\mathbf{X} = (X_1, \ldots, X_p)$ or regressors and parameters $\boldsymbol{\beta}^T = (\beta_1, \ldots, \beta_p)$ to be estimated. In matrix notation, $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$ where $\boldsymbol{\eta}$ is *nx1*, $\mathbf{X}$ is *nxp* and $\boldsymbol{\beta}$ is *px1*.

BINARY RESPONSE DATA. BINOMIAL MODELS

For each observation, the expected value $\boldsymbol{\mu}$ is related to the linear predictor $\boldsymbol{\eta}$ through the scalar ***link function*, denoted *g(.)*** (logit, probit, clog-log) and thus $g(\pi_i) = \mathbf{x_i^T}\boldsymbol{\beta} = \eta_i$.

The response function is $\mu_i = m_i\pi_i = m_i g^{-1}(\mathbf{x_i^T}\boldsymbol{\beta}) = m_i g^{-1}(\eta_i)$

In ordinary least squares models for normal data, the identity link used is $\boldsymbol{\eta = \mu}$.

For binary data, several link functions are commonly used and will be presented in a later section.

Since ML estimates: $\widehat{\boldsymbol{\beta}}\ \forall i \rightarrow \hat{\eta}_i = \mathbf{x_i^T}\widehat{\boldsymbol{\beta}} \rightarrow \hat{\pi}_i = g^{-1}(\hat{\eta}_i) \rightarrow \hat{\mu}_i = m_i\hat{\pi}_i$

- For disaggregated data: $\hat{\mu}_i = m_i\hat{\pi}_i = 1 \cdot \hat{\pi}_i$ is the probability of positive response for observation i.
- For grouped data: $\hat{\mu}_i = m_i\hat{\pi}_i$ is the probability $\hat{\pi}_i$ of positive response for observations in group i **by** group size, that is, the expected number of positive outcomes in group i.

BINARY RESPONSE DATA. BINOMIAL MODELS

4-1.2 Classification of statistical linear models

Explanatory variables	Response variable				
	Dichotomous or binary	**Polytomous**	***Counts (discrete)***	***Continuous***	
				Normal	***Time between events***
Dichotomous	Contingency tables Logistic regression Log-linear models	Contingency tables Log-linear models	Log-linear models	Tests for 2 subpopulation means: t.test	Survival analysis
Polytomous	Contingency tables Logistic regression Log-linear models	Contingency tables Log-linear models	Log-linear models	ONEWAY, ANOVA	Survival analysis
Continuous (covariates)	Logistic regression	*	Log-linear models	Multiple regression	Survival analysis
Factors and covariates	Logistic regression	*	Log-linear models	Analysis of covariance	Survival analysis
Random effects	Mixed models	Mixed models	Mixed models	Mixed models	Mixed models

4-2. INTRODUCTION TO BINARY RESPONSE DATA. BINOMIAL MODELS

This variable appears when, given a sample, each individual has or does not have a target characteristic being studied, which is codified as (*Y=1*) or not (*Y=0*).

For example, regarding mode selection in transportation models, one might be interested in the modal choice between public (metro, bus, light rail, etc.) or private (car, motorcycle, etc.) modes in the study area of home to work trips. In such models, the response variable can be defined for a commuter as *Y=1 (positive response or success, for example, public modes)* or *Y=0 (negative response or failure, in this case, private modes)*.

- It is possible to have more than two levels or categories in the response variable.
- The probability of success is denoted π , such that

$P(Y_k = 1) = \pi_k$: Probability of positive response (success) for k*th* individual in sample.

$P(Y_k = 0) = 1 - \pi_k$: Probability of negative response (failure) for k*th* individual in sample.

Each individual in the sample is characterized by a set of variables (some of which are covariates such as income and age and some of which are factors such as gender, grades, etc.) that defines $\mathbf{x_k^T} = \begin{pmatrix} x_{k1} & \dots & x_{kp} \end{pmatrix}$.

INTRODUCTION TO BINARY RESPONSE DATA. BINOMIAL MODELS

➡ Explanatory variables that will form the linear predictor $\mathbf{x_k^T} = \begin{pmatrix} x_{k1} & \cdots & x_{kp} \end{pmatrix}$ might be:

- Quantitative variables or covariates.
- Transformation of original variables.
- Polynomial regressors built from covariates.
- Dummy variables to represent factors.
- Dummy variables to represent interaction between factors and covariates.

For example, in the public-private binary modal choice model, for each commuter variable, such as income, gender, car availability, distance to local public transport, value of time, etc.

➡ In this subject, the goal relies on studying the relationship between response variable y and the explanatory variables in order to model the probability of a positive response: $\pi = \pi(\mathbf{x})$.

➡ In designing the experiment, groups of individual units are defined and each group receives a combination of experimental conditions that are shared by all the units in the group. In general, factors are considered explanatory variables, and a kth experimental *condition is modeled by a common set of values for all the explanatory variables of individual units in the group* and thus apply to m_k individual units.

INTRODUCTION TO BINARY RESPONSE DATA. BINOMIAL MODELS

➡ The total number of units in the sample is the sum of the size of the groups and, thus, the number of experimental conditions or groups is defined by n and the total number of units is $N = m_1 + \ldots + m_n$.

Each group or combination of experimental conditions defines a covariate class in which all individual units belonging to the covariate class share the same values for explanatory variables.

The difference between individual and covariate class is critical when specifying data to statistical packages. In general, both representations are fully disaggregated (each individual outcome is broken down) or aggregation at some level is allowed according to covariate classes:

1. Some analysis methods are well suited to aggregated data and perform badly when applied to disaggregated data, such as asymptotic approximations of normality.
2. Asymptotic approximations for aggregated data are based either on the asymptotic evolution of the number of covariate classes (or groups) ($\mathbf{m} \to \infty$) or on the total number of individual units ($N \to \infty$).
3. Disaggregated data is suitable for asymptotic approximations based on the total size.

INTRODUCTION TO BINARY RESPONSE DATA. BINOMIAL MODELS

➡ Let's use a simple example to see differences in the representation. The table shows an experiment consisting of dichotomous factors *A* and *C* and thus n=4=2x2 is the number of covariate classes, but the total number of individuals is N=7. In our example, factor *A* is gender (two levels, coded as male and female) and factor *C* is car availability (1 car or more than 1).

Disaggregated data			***Grouped data***		
Individual unit	***Variables***	***Response***	***Covariate class***	***Class Size***	***Positive***
1	(male,1)	0	(1,1)	2	1
2	(male,2)	1	(1,2)	3	2
3	(male,2)	0	(2,1)	1	0
4	(female,1)	0	(2,2)	1	1
5	(female,2)	1			
6	(male,2)	1			
7	(male,1)	1			

INTRODUCTION TO BINARY RESPONSE DATA. BINOMIAL MODELS

```
> dfgrup
   Region   m ypos yneg
1 Rgn.Nrd 440  182  258
2 Rgn.Lvn 210   56  154
3 Regn.Sr 144   17  127
4 Rgn.Cnt 138   26  112
5 Rgn.NrO 119   47   72
6 Rgn.NrC 117   39   78
7 R.BCNAM 203   71  132
8 Rgn.MAM  76   19   57
```

Example: What would be the grouped data format for the Cordorniu data and the model *glm(Codorniu~Region)?*

Codorniu is a binary variable built from P24REC, indicating a positive outcome when Codorniu was the first trademark to come to mind when thinking of cava.

INTRODUCTION TO BINARY RESPONSE DATA. BINOMIAL MODELS

Aggregated/grouped versus disaggregated data

- Considering aggregated data is more efficient and consumes less memory. It makes observing the effect significantly simpler.
- Aggregated data implies that the serial order is lost. If additional variables are present, only average values can be considered, possibly leading to *ecological fallacy situations*.
- Aggregated data implies a response variable model of the binomial type, since observed positive responses are $y_1/m_1, \ldots, y_n/m_n$, where $0 \leq y_k \leq m_k$ is the number of positive responses in kth covariate class, the size of which is m_k.

- The size of covariate classes in vector form is called the ***binomial index vector*** and it is denoted $\mathbf{m} = (m_1 \quad \ldots \quad m_n)$. For disaggregated data, each individual unit defines a binomial response for a group of size 1 and thus, $\mathbf{m} = (1 \quad \ldots \quad 1)$.

4-3. BINOMIAL MODELS FOR BINARY DATA

$$p_Y(y) = P([Y = y]) = \binom{m}{y} \pi^y (1-\pi)^{m-y}$$

$$F_Y(y) = \begin{cases} 0 & y < 0 \\ \sum_{i=0}^{\lfloor y \rfloor} \binom{m}{i} \pi^i (1-\pi)^{m-i} & 0 \le y \le m \\ 1 & y > m \end{cases}$$

$$\mathrm{E}[Y] = m \cdot \pi$$

$$\mathrm{V}[Y] = m \cdot \pi \cdot (1-\pi)$$

➡ Usually considered and defined in introductory courses to statistical analysis or theory of probability:

Let $Y \approx B(m, \pi)$ be a binomial variable that models the number of positive responses in m independent trials in a Bernoulli process and, thus, each one with a common probability π.

BINOMIAL MODELS FOR BINARY DATA

4-3.1 Link functions

➡ The goal consists in establishing a functional relationship between the probability of a positive result π and the vector of explanatory variables (predictors, in general; covariates if they are continuous) $\mathbf{x}_\mathbf{k}^\mathbf{T} = \begin{pmatrix} x_{k1} & \dots & x_{kp} \end{pmatrix}$: $\boxed{\pi = \pi(\mathbf{x})}$.

- In generalized linear models, the link function relates the linear predictor scale to the expected value of the probabilistic variable selected to model the random response. In the case of a binomial model concerned with the probability of positive response for a dichotomous individual response, the linear predictor η might be any value in the real axis for a given observation, but the probability of a positive answer belongs to the open interval (0, 1).

➡ Vector π is related to the linear predictor η through *the link function*, denoted ***g(.)***, $\eta = \mathbf{g}(\pi)$, π is *nx1*.

➡ The canonical link for binomial data is the *logit function* $\eta = \theta = \mathbf{logit}(\pi)$.

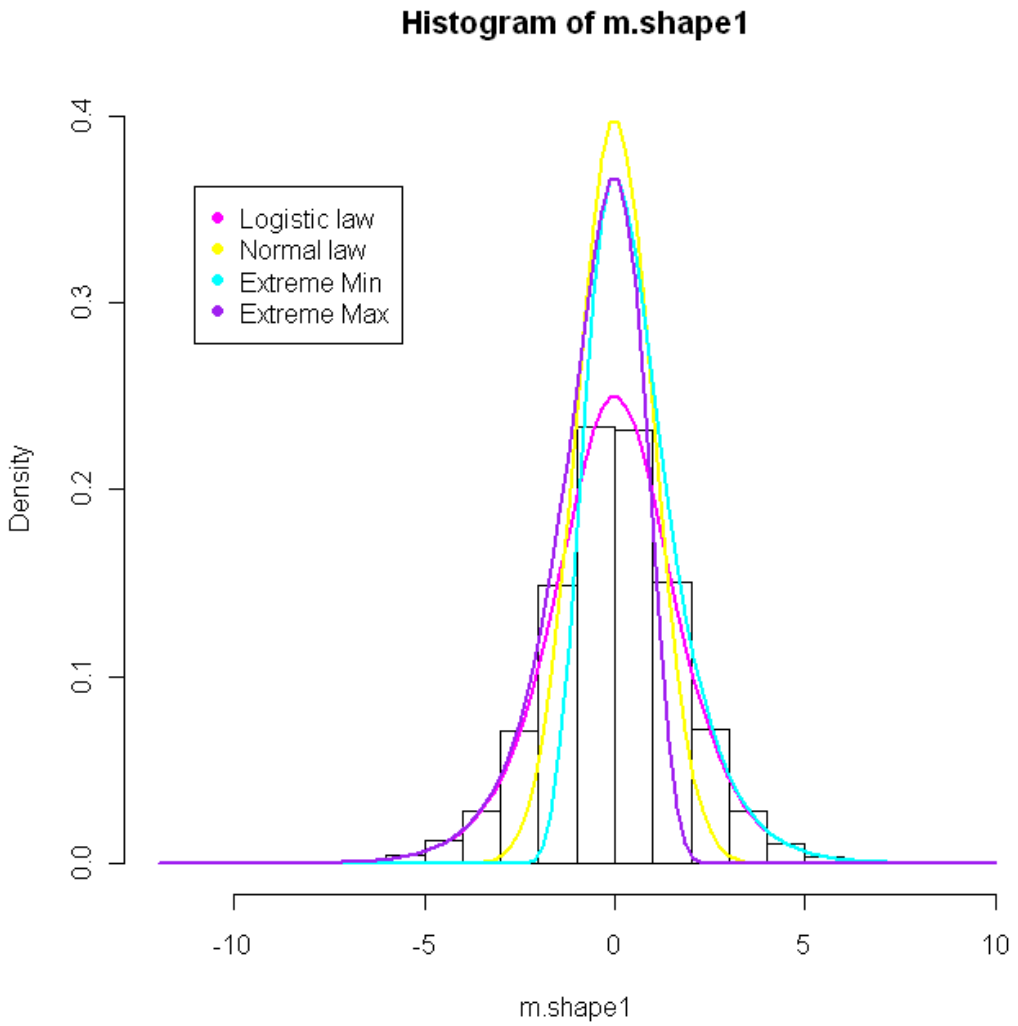
BINOMIAL MODELS FOR BINARY DATA

- Logit link function is the most frequently used link, but it's important to understand the role of link functions and not act automatically.
- Some common *link* functions for binary response data are:

1. **The logit link (sometimes erroneously called the logistic link)**

$$\eta = g_1(\pi) = \mathbf{logit}(\pi) = \mathbf{log}\left(\frac{\pi}{1-\pi}\right).$$

and $\pi_1(\eta) = g_1^{-1}(\eta) = \dfrac{\exp(\eta)}{1+\exp(\eta)}$; this is the distribution function for a standard logistic variable, whose density function is $\left(g_1^{-1}\right)'(\eta) = \dfrac{\exp(\eta)}{(1+\exp(\eta))^2}$ with a mean of 0 (position parameter) and variance $\pi^2/3$ (scale parameter 1). This is a continuous and symmetric variable, quite similar to a normal distribution.



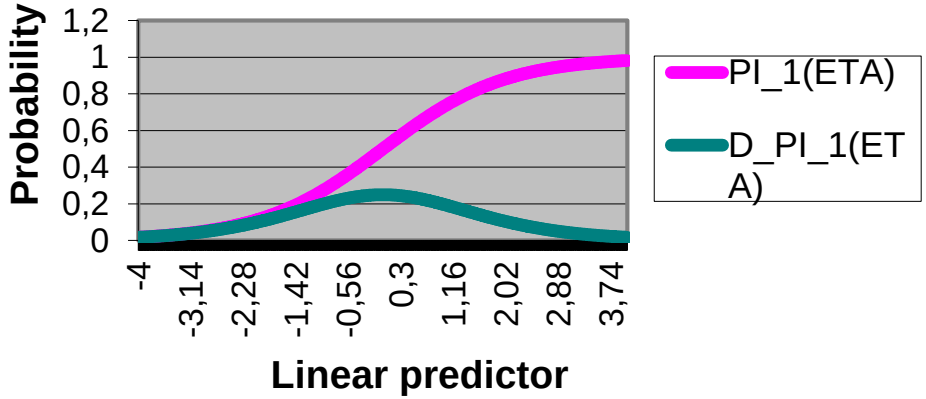
BINOMIAL MODELS FOR BINOMIAL DATA: LINK FUNCTIONS

➡ *Link* function for binary response data

2. **Probit link or standard normal** inverse: $\eta = g_2(\pi) = \Phi^{-1}(\pi)$ and $\pi_2(\eta) = g_2^{-1}(\eta) = \Phi(\eta)$.
   Standard normal (mean 0 and variance 1).

➡ Logit link:

$$\pi_1(\eta) = g_1^{-1}(\eta) = \frac{\exp(\eta)}{1+\exp(\eta)} \quad \text{and} \quad (g_1^{-1})'(\eta) = \frac{\exp(\eta)}{(1+\exp(\eta))^2} = \pi_1(\eta)(1-\pi_1(\eta))$$



In general,

$$\pi_i = \mathrm{P}(\eta_i) = \mathrm{P}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}),$$

where **P(.)** indicates a distribution function for continuous variables and transforms real values for the linear predictor into the $[0,1]$ interval.

Transformations should depend on data characteristics and not be selected on a straightforward basis.

BINOMIAL MODELS FOR BINARY DATA: LINK FUNCTIONS

➡ *Logit* and *probit* links are related to changes in scales:

Probability π	Odds $\frac{\pi}{1-\pi}$	Log-odds $\log\left(\frac{\pi}{1-\pi}\right)=\mathbf{x}\beta$	Probit $\Phi^{-1}(\pi)=\mathbf{x}\beta$
0,01	0,0101	-4,5951	-2,3263
0,05	0,0526	-2,9444	-1,6449
0,10	0,1111	-2,1972	-1,2816
0,15	0,1765	-1,7346	-1,0364
0,20	0,2500	-1,3863	-0,8416
0,25	0,3333	-1,0986	-0,6745
0,30	0,4286	-0,8473	-0,5244
0,50	1,0000	0,0000	0,0000
0,70	2,3333	0,8473	0,5244
0,75	3,0000	1,0986	0,6745
0,80	4,0000	1,3863	0,8416
0,85	5,6667	1,7346	1,0364
0,90	9,0000	2,1972	1,2816
0,95	19,0000	2,9444	1,6449
0,99	99,0000	4,5951	2,3263

BINOMIAL MODELS FOR BINARY DATA: LINK FUNCTIONS

4-3.1.1 Comparison of common link functions

➡ **logit** link is symmetric,

$$\eta = g_1(\pi) = \mathbf{log}\left(\frac{\pi}{1-\pi}\right) = \mathbf{log}(\pi) - \mathbf{log}(1-\pi) = -(-\mathbf{log}(\pi) + \mathbf{log}(1-\pi)) = -\mathbf{log}\left(\frac{1-\pi}{\pi}\right) = -g_1(1-\pi)$$

➡ **probit** link is symmetric, if $\eta = g_2(\pi) = \Phi^{-1}(\pi)$ then,

$$\pi = \Phi(\eta) = 1-\Phi(-\eta) \rightarrow 1-\pi = \Phi(-\eta) \rightarrow \Phi^{-1}(1-\pi) = -\eta \rightarrow \eta = -\Phi^{-1}(1-\pi) = -g_2(1-\pi)$$

➡ **logit** link has a straightforward interpretation:

Linear model in log-odd scale: $g_1(\pi_i) = \mathbf{log}\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$.

Multiplicative model in odd scale:

$\frac{\pi_i}{1-\pi_i} = \mathbf{exp}(\beta_1 x_{i1} + \ldots + \beta_p x_{ip}) = \mathbf{exp}(\beta_1 x_{i1}) \cdots \mathbf{exp}(\beta_p x_{ip}) = \mathbf{exp}(\beta_1)^{x_{i1}} \cdots \mathbf{exp}(\beta_p)^{x_{ip}}$. A one unit increment in variable j, x_{ij} to x_{ij}+1, thus multiplies π_i odds by $\mathbf{exp}(\beta_j)$.

Probability scale approximated interpretation for variable j: $\bar{\pi}(1-\bar{\pi})\beta_j$ where $\bar{\pi}$ is the sample positive outcome probability.

BINOMIAL MODELS FOR BINARY DATA: LINK FUNCTIONS

➡ New table taking into account cloglog and loglog common links:

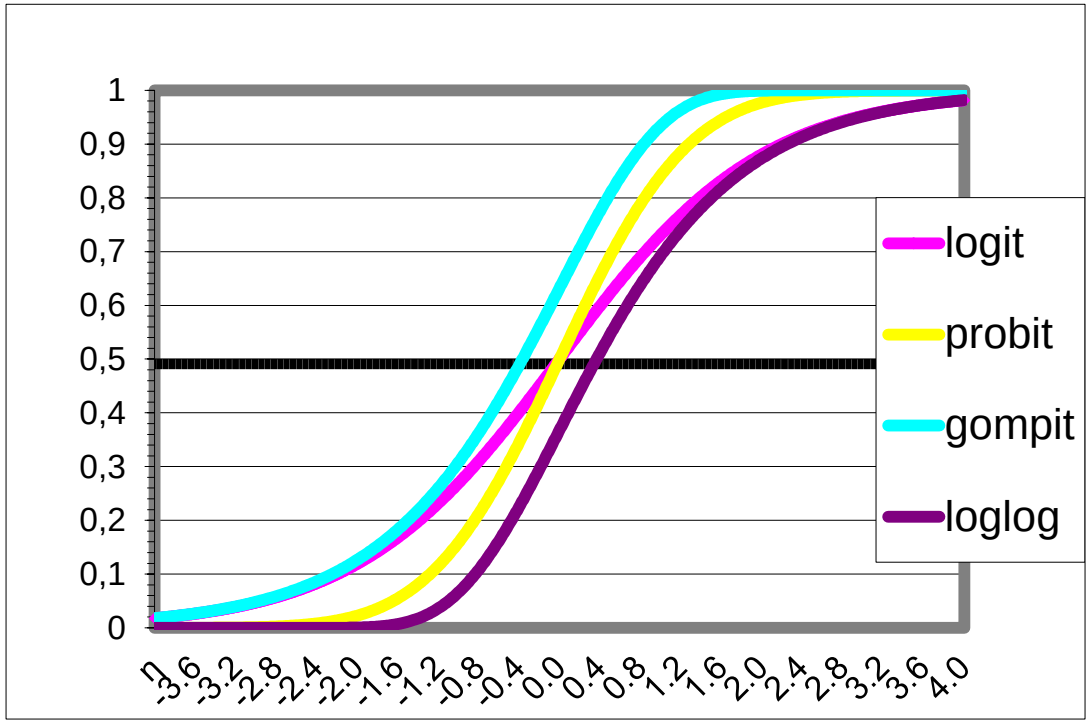
Probabilidad π	**Odds** $\frac{\pi}{1-\pi}$	**Log-odds** $\log\left(\frac{\pi}{1-\pi}\right)=\mathbf{x}\beta$	**Probit** $\Phi^{-1}(\pi)=\mathbf{x}\beta$	**C_log-log** $\log\left(\log\left(\frac{1}{1-\pi}\right)\right)=\mathbf{x}\beta$	**Log-log** $-\log\left(\log\left(\frac{1}{\pi}\right)\right)=\mathbf{x}\beta$
0,01	0,0101	-4,5951	-2,3263	**-4,60015**	**-1,52718**
0,05	0,0526	-2,9444	-1,6449	**-2,97020**	**-1,09719**
0,10	0,1111	-2,1972	-1,2816	**-2,25037**	**-0,83403**
0,15	0,1765	-1,7346	-1,0364	**-1,81696**	**-0,64034**
0,20	0,2500	-1,3863	-0,8416	**-1,49994**	**-0,47588**
0,25	0,3333	-1,0986	-0,6745	**-1,24590**	**-0,32663**
0,30	0,4286	-0,8473	-0,5244	**-1,03093**	**-0,18563**
0.3679	0.5820282	-0.5412364	-0.3374204	**-0.7793422**	**0,0000**
0.4296	0.7532291	-0.2833858	-0.1773318	**-0.5772**	**0,1685**
0,50	1,0000	0,0000	0,0000	**-0,36651**	**0,36651**
0.5704	1.327747	0.2834833	0.1773926	**-0.1685361**	**0.5772**
0.6321	1.71813	0.5412364	0.3374204	**0,0000**	**0.7793**
0,70	2,3333	0,8473	0,5244	**0,18563**	**1,03093**
0,75	3,0000	1,0986	0,6745	**0,32663**	**1,24590**
0,80	4,0000	1,3863	0,8416	**0,47588**	**1,49994**
0,85	5,6667	1,7346	1,0364	**0,64034**	**1,81696**
0,90	9,0000	2,1972	1,2816	**0,83403**	**2,25037**
0,95	19,0000	2,9444	1,6449	**1,09719**	**2,97020**
0,99	99,0000	4,5951	2,3263	**1,52718**	**4,60015**

BINOMIAL MODELS FOR BINARY DATA: LINK FUNCTIONS

- **log-log and c-log-log** link functions are also related by the following equation:

$$g_3(\pi) = \log\left(\log\left(\frac{1}{1-\pi}\right)\right) = -\left(\log\left(\log\left(\frac{1}{1-\pi}\right)\right)\right) = -g_4(1-\pi)$$

- All link function are continuous and increasing functions in the open interval *(0,1)*.
- **logit and probit** link functions show an almost linear relationship between 0.1 and 0.9.
- For small probabilities, **logit and complementary log-log links are rather similar**.
- For large probabilities near 1, **complementary log-log** has a less steepy increment than logit function..
- For large probabilities near 1, **logit and log-log** link function are **rather similar**.



BINOMIAL MODELS FOR BINARY DATA: THEORY

- Bernoulli in particular and Binomial laws belong to exponential family of one parameter. Parameters for binomial laws are group size, m and positive outcome probability, π, where total counts in the group is my, meaning $Y \approx B(m,\pi)/m$

$$f_Y(y,\theta,\phi)=\binom{m}{my}\pi^{my}(1-\pi)^{m-my}=\exp\left(my\log(\pi)-(my-m)\log(1-\pi)+\log\left(\binom{m}{my}\right)\right)=$$

$$=\exp\left(\frac{y\log\left(\frac{\pi}{1-\pi}\right)-\log\left(\frac{1}{1-\pi}\right)}{1/m}+\log\left(\binom{m}{my}\right)\right)=\exp\left(\frac{y\theta-\log(1+e^{\theta})}{1/m}+\log\left(\binom{m}{my}\right)\right)$$

$a(\phi)=1/m$, $b(\theta)=\log(1+e^{\theta})$, $\theta=\text{logit}\,\pi=\log\frac{\pi}{1-\pi}$, $c(y,\phi)=\log\left(\binom{m}{my}\right)$ and $\pi(\theta)=\frac{e^{\theta}}{1+e^{\theta}}$

$$\ell(\theta,\phi,y)=\log f_Y(y,\theta,\phi)=\frac{y\theta-b(\theta)}{a(\phi)}+c(y,\phi)=\frac{y\theta-\log(1+e^{\theta})}{1/m}+\log\left(\binom{m}{my}\right)$$

BINOMIAL MODELS FOR BINARY DATA: THEORY

➡ First and second order properties for scaled binomial distribution $B(m,\pi)/m$ are,

$$\mathrm{E}[Y]=\pi=\mu \text{ , } \mu(\theta)=b'(\theta)=\frac{d}{d\theta}\left(\log\left(1+\exp(\theta)\right)\right)=\frac{\exp(\theta)}{1+\exp(\theta)}. \text{ y } \theta(\mu)=\operatorname{logit}\mu=\log\left(\frac{\mu}{1-\mu}\right)$$

$$V[Y]=a(\phi)b''(\theta)=\frac{1}{m}\frac{\exp(\theta)}{\left(1+\exp(\theta)\right)^2} \quad \text{and } V[\mu]=b''(\theta)=\frac{\exp(\theta)}{\left(1+\exp(\theta)\right)^2}=\mu\cdot(1-\mu).$$

BINOMIAL MODELS FOR BINARY DATA: THEORY

➡ Directly for binomial law, $B(m,\pi)$, where **y** is the total number of positive outcomes (not observed probabilities) (thus, y/m is a probability), implies some minor modifications to the former expressions:

$$B(m,\pi) \rightarrow \text{y}_\text{i}, \quad \theta_i = \mathbf{log}\frac{\pi_i}{1-\pi_i} = \mathbf{log}\frac{\mu_i/m_i}{1-\mu_i/m_i} \quad \text{and} \quad \ell(\theta,\phi,y) = \frac{y\theta - m\,\mathbf{log}\left(1+e^{\theta}\right)}{1} + \mathbf{log}\left(\binom{m}{y}\right),$$

$$\mathrm{E}[Y_i] = m_i\pi_i = \mu_i \;,\; b(\theta_i) = m_i\,\mathbf{log}\left(1+e^{\theta}\right), \quad \mu(\theta_i) = b'(\theta_i) = m_i\frac{\mathbf{exp}(\theta_i)}{1+\mathbf{exp}(\theta_i)} \quad \text{and} \quad a(\phi) = \phi = 1$$

$$V[Y] = a(\varphi)b''(\theta) = b''(\theta) = m\frac{\exp(\theta)}{(1+\exp(\theta))^2} \quad \text{and} \quad V[\mu] = \frac{\mu}{m}\cdot(m-\mu).$$

...

BINOMIAL MODELS FOR BINARY DATA: THEORY

➡ … Scaled deviance (measure of discrepancy) for a binomial law $B(m,\pi)$, where **y** are ***observations not probabilities,***

$$D'(\mathbf{y},\hat{\boldsymbol{\mu}}) = 2\,\ell(\mathbf{y},\phi,\mathbf{y}) - 2\,\ell(\hat{\boldsymbol{\mu}},\phi,\mathbf{y}) =$$

$$= \sum_{i=1}^{n}\left\{2\left(y_i\ \mathbf{log}\left(\frac{y_i/m_i}{1-y_i/m_i}\right) - m_i\,\mathbf{log}\left(1+\frac{y_i/m_i}{1-y_i/m_i}\right) + \mathbf{log}\left(\binom{m_i}{y_i}\right)\right)\right\} -$$

$$- \sum_{i=1}^{n}\left\{2\left(y_i\,\mathbf{log}\left(\frac{\hat{\mu}_i/m_i}{1-\hat{\mu}_i/m}\right) - m_i\,\mathbf{log}\left(1+\frac{\hat{\mu}_i/m_i}{1-\hat{\mu}_i/m_i}\right) + \mathbf{log}\left(\binom{m_i}{y_i}\right)\right)\right\} =$$

$$= 2\sum_{i=1}^{n}\left\{ y_i\,\mathbf{log}\left(\frac{y_i}{\hat{\mu}_i}\right) - y_i\,\mathbf{log}\left(\frac{m_i-y_i}{m_i-\hat{\mu}_i}\right) + m_i\left(\mathbf{log}\left(\frac{m_i-y_i}{m_i}\right) - \mathbf{log}\left(\frac{m_i-\hat{\mu}_i}{m_i}\right)\right)\right\} =$$

$$= 2\sum_{i=1}^{n}\left\{ y_i\,\mathbf{log}\left(\frac{y_i}{\hat{\mu}_i}\right) + (m_i-y_i)\mathbf{log}\left(\frac{m_i-y_i}{m_i-\hat{\mu}_i}\right)\right\}$$

4-4. ESTIMATION OF MODEL PARAMETERS

➡ The estimation process relies on unconstricted maximization of the log-likelihood function,

$$Max_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta},\mathbf{y})=\sum_{i=1}^{n}\log f\left(y_i,\beta_p\right),\ \boldsymbol{\beta}^{\mathbf{T}}=\left(\beta_1,\ldots,\beta_p\right) \text{ and } \mathbf{y}^{T}=\left(y_1,\ldots,y_n\right).$$

➡ The iterative process to compute the estimates is called the score method, a second-order Newton-type method specialized for the properties of the log likelihood function. The method converges fast but is not globally convergent.

➡ Existence and unicity for estimates under any of the aforementioned link functions, if $0<y_i<m_i$ for any covariate class/observation.

➡ The quality of the initialization is usually not very important since the algorithm has fast convergence properties. It is not globally convergent, however, so an extreme initial point might lead to divergence.

ESTIMATION OF MODEL PARAMETERS: CODORNIU DATA

```
> m0<-glm(Codorniu~1,family=binomial,data=df)
> m1<-glm(Codorniu~Region,family=binomial,data=df)
> summary(m1)

Call: glm(formula = Codorniu ~ Region, family = binomial, data = df)

…

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.34895    0.09680  -3.605 0.000312 ***
RegionRgn.Lvn -0.66265    0.18363  -3.609 0.000308 ***
RegionRegn.Sr -1.66202    0.27580  -6.026 1.68e-09 ***
RegionRgn.Cnt -1.11145    0.23824  -4.665 3.08e-06 ***
RegionRgn.NrO -0.07757    0.21104  -0.368 0.713210
RegionRgn.NrC -0.34419    0.21871  -1.574 0.115538
RegionR.BCNAM -0.27117    0.17616  -1.539 0.123713
RegionRgn.MAM -0.74966    0.28204  -2.658 0.007861 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1804.9  on 1446  degrees of freedom
Residual deviance: 1735.3  on 1439  degrees of freedom
AIC: 1751.3
```

4-5. GOODNESS OF FIT

- If $\hat{\boldsymbol{\beta}}$ is an estimated model parameter, then a linear predictor for each observation *i* might be computed as $\hat{\eta}_i = \mathbf{x}_i^T \cdot \hat{\boldsymbol{\beta}}$ and, thus, through the response function (inverse of the selected link function), fitted values can be computed: $\boxed{\hat{\pi}_i = g^{-1}(\hat{\eta}_i)}$.

- Scaled deviance can be calculated from the maximum likelihood function at convergence, $D'(\mathbf{y}, \hat{\boldsymbol{\mu}}) = 2\ell(\mathbf{y}, \mathbf{y}) - 2\ell(\hat{\boldsymbol{\mu}}, \mathbf{y})$.

- And so, deviance under the binomial distribution is identical, since $\varphi = 1$

$$D(\mathrm{y}, \hat{\mu}) = D'(\mathrm{y}, \hat{\mu})\varphi = D'(\mathrm{y}, \hat{\mu}) \text{ if } Y_i \approx B(m_i, \pi_i)$$

BINOMIAL MODELS FOR BINARY DATA: GOODNESS OF FIT

➡ Deviance is expressed as:

$$D(\mathbf{y},\hat{\boldsymbol{\mu}}) = D(\mathbf{y},\hat{\boldsymbol{\pi}}) = 2\sum_{i=1}^{n}\left\{ y_i \log\left(\frac{y_i}{m_i\hat{\pi}_i}\right) + (m_i - y_i)\log\left(\frac{(m_i - y_i)}{(m_i - m_i\hat{\pi}_i)}\right)\right\}$$

➡ Sometimes, the deviance statistic is written in an alternative way:

$$\boxed{D = 2\sum_{postive,negative}\sum_{i=1}^{n} o_i \log\frac{o_i}{e_i}}$$ where

1. Observed values for positive responses for observation i, $o_i = y_i$.
2. Observed values for negative responses for observation i, $o_i = m_i - y_i$.
3. Expected positive responses for observation i, $e_i = m_i\hat{\pi}_i$.
4. Expected negative responses for observation i, $e_i = m_i - m_i\hat{\pi}_i$.

BINOMIAL MODELS FOR BINARY DATA: GOODNESS OF FIT

- Asymptotic distribution for model (M) with p parameters $D_M = D(\mathbf{Y}, \hat{\pi})$ is χ^2_{n-p} (***not to be confused with*** χ^2_{N-p}). Asymptotic conditions are not met with individual data.

Thus, a goodness of fit test can be formulated as H0 "The current model properly fits the data" and the p value for the test is $P\left(\chi^2_{n-p} > D_M\right) = p\_value$:

- If pvalue <<0.05 then there is evidence to reject H0 and therefore the model (M) does not properly fit the data. There is statistical evidence of discrepancy between observations and fitted values provided by model (M).
- If pvalue >> 0.05 then there is no evidence to reject H0 and therefore H0 is accepted, leading to the conclusion that model (M) does properly fit the data, since discrepancy between observed and fitted values is not statistically significant.

- AIC (Akaike Information Criteria, 1974) is defined as a trade-off between the goodness of fit provided by a model (M) and the number of parameters p in the model (as an indicator for model complexity). Let **M be a model with p parameters** $AIC_M = 2\left(-\ell\left(\hat{\pi}_M, \mathrm{y}\right) + p\right)$. Models with minimum AIC are preferred.

BINOMIAL MODELS FOR BINARY DATA: GOODNESS OF FIT

➡ In order to consider sample size, another statistic known as BIC (*Bayesian Information Criteria*) (in SAS©) or Schwartz criteria is proposed $BIC_{\mathbf{B}} = -2\,\ell(\hat{\pi}_B, \mathrm{y}) + p \log n$. Minimum BIC models are preferred (`AIC(model,k=log(n))`).

➡ **AIC and BIC may be used to compare unnested models.**

➡ **Following McCullagh, the test for Generalized LM equivalent to F.Test in classical linear regression consists in comparing differences in scaled deviance in two hierarchical models (nested models):**

Let M_A be a model with q parameters nested in model M_B with p > q parameters; let $\hat{\pi}_A$ and $\hat{\pi}_B$ be fitted probabilities for both models, **such that, the set of parameters for M_B are those common and specific to M_A; i.e.,** $\boldsymbol{\beta}_B^T = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)$ and $\boldsymbol{\beta}_A^T = (\boldsymbol{\beta}_1^T)$ with dim($\boldsymbol{\beta}_A$)=q<p, then

$$\Delta D_{AB} = D(\mathbf{y}, \hat{\pi}_A) - D(\mathbf{y}, \hat{\pi}_B) = 2\,\ell(\hat{\pi}_B, \mathrm{y}) - 2\,\ell(\hat{\pi}_A, \mathrm{y})$$

is asymptotically distributed χ^2_{p-q}.

And for testing $\mathrm{H}_0 : \beta_2 = \mathbf{0}$

$$P(\chi^2_{p-q} > \Delta D_{AB}) \rightarrow \begin{cases} << \alpha & \mathrm{H}_0 \text{ Rejected} \\ >> \alpha & \mathrm{H}_0 \text{ Accepted} \end{cases}$$

This is a contrast for multiple coefficients! Large values indicate non-equivalence of models.

BINOMIAL MODELS FOR BINARY DATA: GOODNESS OF FIT

In R software: Deviance test is as Fisher test for normal models

```
anova(modelA, modelB, ...,     test = c("F", "Chisq")) # Deviance Test
waldtest(modelA, modelB, ...,     test = c("F", "Chisq")) # Wald Test lmtest library
glm.scoretest(modelA, x2='additional columns') # Scoretest statmod library,Z~N(0,1)
```

- Rao's score test is a type of asymptotic test that is an alternative to Wald tests or likelihood ratio tests (LRTs) (Dunn and Smyth, 2018). All three types of tests (Wald, score and LRT) are asymptotically equivalent under ideal circumstances, but the score and LRT tests are invariant under-reparametrization whereas Wald tests are not.
  - Wald tests are computed by dividing parameter estimates by their standard errors.
  - LRTs are computed from differences in the log-likihoods between the null and alternative hypotheses.
  - Score tests are computed from log-likelihood derivatives.

- Deviance for a GLM plays a role similar to the residual sum of squares in classical regression. Thus, we can define a generalized R^2, or pseudo R^2

$$R^2 = 1 - \frac{D(\mathbf{y},\pi_A)}{D(\mathbf{y},\pi_0)} = \frac{G(\mathbf{y},\pi_A)}{G(\mathbf{y},\pi_A)+D(\mathbf{y},\pi_A)} \quad where\ G(\mathbf{y},\pi_A) = D(\mathbf{y},\pi_0) - D(\mathbf{y},\pi_A), \quad 0 \le R^2 \le 1$$

```
> PseudoR2( model, which='all' ) # library(DescTools)
```

BINOMIAL MODELS FOR BINARY DATA: GOODNESS OF FIT

- Goodness of fit using the generalized Pearson **X^2** statistic, asymptotically distributed as χ^2_{n-p} :

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - m_i\hat{\pi}_i)^2}{m_i\hat{\pi}_i(1-\hat{\pi}_i)} \qquad \left(= \sum_{i=1}^{n} \frac{m_i(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i(m_i - \hat{\mu}_i)} \right) \left(= \sum_{+,-} \sum_{i=1}^{n} \frac{(o_i - e_i)^2}{e_i} \right)$$

- Hosmer-Lemeshow Goodness of Fit (GOF) Test is based on dividing the sample up according to their predicted probabilities. The observations in the sample are then split into g groups (usually 10) according to their predicted probabilities. Then the first group consists of the observations with the lowest 10% predicted probabilities. The second group consists of the 10% of the sample whose predicted probabilities are next smallest, etc etc.
- Then, how many $Y = 1$ observations we would expect is calculated, by taking the average of the predicted probabilities in the group, and multiplies this by the number of observations in the group. The test also performs the same calculation for $Y = 0$, and then calculates a Pearson goodness of fit statistic distributed with g - 2 degrees of freedom.

```
hoslem.test(y, fitted(model), g=10) # ResourceSelection library
```

GOODNESS OF FIT AND MODEL SELECTION: CODORNIU DATA

- Region has global significance: gross effect!
- Step(model): Selection of the best model according to AIC criteria is available for glm()

```
> anova(m0,m1,test="Chisq")
Analysis of Deviance Table

Model 1: Codorniu ~ 1
Model 2: Codorniu ~ Region
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      1446     1804.9
2      1439     1735.3  7   69.589 1.789e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> m3aic<-step(m4gran)
Start:  AIC=1751.8
Codorniu ~ Region * Brandist + Age_Resp * Brandist + HH_nb

                    Df Deviance    AIC
- Region:Brandist   21   1705.8 1737.8
- HH_nb              1   1679.2 1751.2
<none>                   1677.8 1751.8
- Brandist:Age_Resp  3   1687.1 1755.1

Step:  AIC=1737.75
Codorniu ~ Region + Brandist + Age_Resp + HH_nb + Brandist:Age_Resp
```

GOODNESS OF FIT AND MODEL SELECTION: CODORNIU DATA

```
                     Df Deviance    AIC
- HH_nb               1   1707.1 1737.1
<none>                    1705.8 1737.8
- Brandist:Age_Resp   3   1715.4 1741.4
- Region              7   1763.2 1781.2

Step:  AIC=1737.12
Codorniu ~ Region + Brandist + Age_Resp + Brandist:Age_Resp

                     Df Deviance    AIC
<none>                    1707.1 1737.1
- Brandist:Age_Resp   3   1716.6 1740.6
- Region              7   1764.9 1780.9
>
```

MODELS FOR BINARY DATA: GOODNESS OF FIT AND MODEL SELECTION

4-5.1 ROC curve and confusion matrices

ROC (Receiver Operating Characteristic) curve analysis has been widely accepted as the standard for describing and comparing the accuracy of predictions.

If the ROC curve rises rapidly towards the upper right-hand corner of the graph or if the area under the curve is large, we can say the model performs well. If the area is close to 1.0, the model is good. If the area is close to 0.5, the model is bad.

Confusion matrix for a binary model (M) shows predicted response versus observed response (positive/negative outcomes).

Let the predicted response be $\hat{y}_i = 1$ if $\hat{\pi}_i > s$, where s is a threshold between 0 and 1. For each s, a confusion matrix can be built for model (M):

s	Y=1	Y=0	Total
$\hat{y}_i = 1$	a/TP	b/FP	a+b
$\hat{y}_i = 0$	c/FN	d/TN	c+d
	a+c	b+d	n

- **Sensibility** *is the proportion of observed positive outcomes (Y=1) predicted to be positive* ($\hat{y}_i = 1$): Sn =a/(a+c). Recall or True Positive Rate (TPR).
- **Specificity** *is the proportion of observed negative outcomes (Y=0) predicted to be negative* ($\hat{y}_i = 0$): Sp = d/(b+d). True Negative Rate (TNR)

MODELS FOR BINARY DATA: GOODNESS OF FIT AND MODEL SELECTION

- Accuracy: The correct number of predictions made by the model over all the observed values. The accuracy is calculated by Equation (3), where TP refers to true positive, TN refers to true negative, FP refers to false positive and FN refers to false negative:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (3)$$

- Precision: Precision gives how many of the correctly predicted cases actually turned out to be positive. The proportion is calculated with the formula shown in Equation (4):

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

- Recall: Recall gives how many of the actual positive cases the model is able to predict correctly. It is calculated as follows:

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

- F1-score: F1-score is a metric which takes into account both precision and recall and is defined as follows:

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (6)$$

Harmonic mean of precision and recall

MODELS FOR BINARY DATA: GOODNESS OF FIT AND MODEL SELECTION

➡ **ROC curves** show, for each s, **1-Sp (true negative rate)** on the x-axis and **Sensibility - Sn (true positive rate)** on the y-axis.

- The point (0,1) is the perfect classifier: it classifies all positive cases and negative cases correctly. This is (0,1) because the false positive rate is 0 (none), and the true positive rate is 1 (all).
- The point (0,0) is a classifier that predicts all cases to be negative.
- The point (1,1) is a classifier that predicts every case to be positive.
- A good online source to understand ROC curves is http://gim.unmc.edu/dxtests/ROC1.htm.

Guidelines for interpreting ROC curves

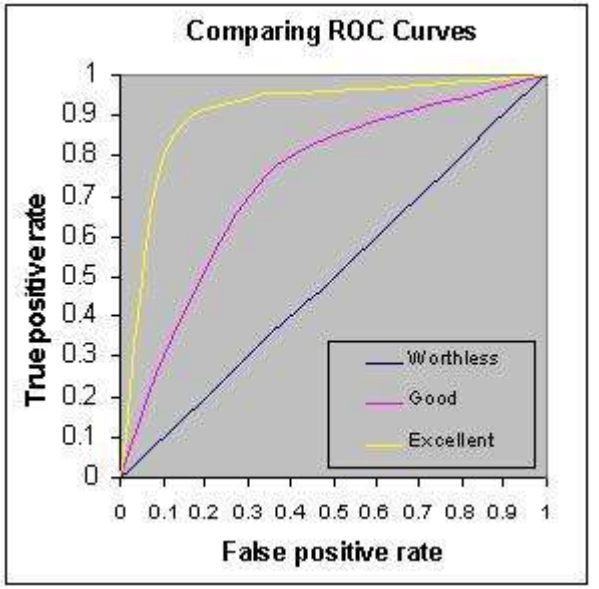
.90-1 = excellent(A)

.80-.90 = very good (B)

.70-.80 = good (C)

.60-.70 = bad (D)

.50-.60 = very bad (F)



MODELS FOR BINARY DATA: GOODNESS OF FIT AND MODEL SELECTION

➡ Goodness of fit statistics commonly used:

Kendall Tau = (C-D)/H	Gamma = (C-D)/(C+D)
Sommer D = (C-D)/(C+D+T) – Gini Coefficient	C=0.5(1+ Sommer D)

Good properties show statistics near 1 and c corresponds to the area under ROC curve. Calculated by MINITAB.

BINOMIAL MODELS FOR BINARY DATA: GOODNESS OF FIT – ROC CURVE

How to compute in R, Pearson's X^2 statistic - Pearson's residual sum of squares:

```
sum( resid( model, 'pearson') ^2 )
```

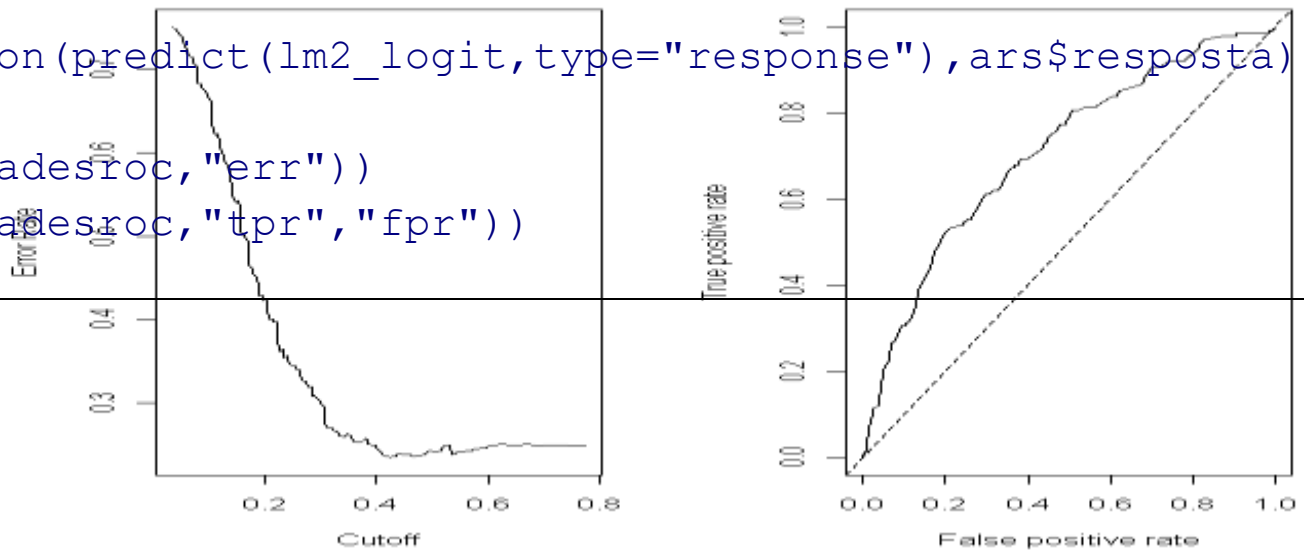
As in the case for deviance residuals:

```
sum( resid( model, 'deviance') ^2 )   ==   model$deviance
```

Package rms contains the specific method lrm(.) for logistic regression with additional diagnostics (c, Naglekerke R^2, and so on). `NagelkerkeR2` is also in the **fmsb** package.

To compute ROC curves: Install package ROCR; specific performance plots are available.

```
> library("ROCR")
> dadesroc<-prediction(predict(lm2_logit,type="response"),ars$resposta)
> par(mfrow=c(1,2))
> plot(performance(dadesroc,"err"))
> plot(performance(dadesroc,"tpr","fpr"))
> abline(0,1,lty=2)
```



4-6. MODEL DIAGNOSTICS

4-6.1 Residuals in GLMz

Normal regression methods extended to generalized linear model:

➡ Pearson residuals are casewise components of the Pearson (standardized) goodness of fit statistic for the model

$$e_i^P = \frac{(y_i - \hat{\mu}_i)}{\sqrt{V[\hat{\mu}_i]/\hat{\phi}}} \text{ and } e_i^{PS} = \frac{(y_i - \hat{\mu}_i)}{\sqrt{(1-h_{ii})V[\hat{\mu}_i]/\hat{\phi}}} \text{ leading to } X^2 = \sum_{i=1}^{n} \left(e_i^P\right)^2$$

This is a basic set of residuals for use with a GLM because of their direct analogy to linear models. For a model named M, the R command `residuals(M, type="pearson")` returns the Pearson residuals.

➡ Deviance residuals, $r_i^D = sign(y_i - \hat{\mu}_i)\sqrt{d_i}$ are the square roots of the casewise components of the residual deviance $D(\mathrm{y}, \hat{\mu}) = \sum_{i=1,\ldots,n} r_{D_i}^2$, attaching the sign of $y_i - \hat{\mu}_i$.

- In the linear model, the deviance residuals reduce to the Pearson residuals.
- The deviance residuals are often the preferred form of residual for GLMs and are returned by the command `residuals(M, type="deviance")`.

➡ Studentized Residuals have been defined by combining Deviance and Pearson residuals: use rstudent(model) from car library.

MODEL DIAGNOSTICS: RESIDUALS

➡ The following functions (some in standard R and some in the car package) have methods for GLMs: rstudent, hatvalues, cooks.distance, dfbetas, outlierTest, avPlots, residualPlots, marginalModelPlots, crPlots, etc.

➡ **Hat matrix for generalized linear models can be defined, although it depends on Y (through W) and x's values,**

$$\mathbf{H} = \mathbf{W}^{1/2}\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}^{1/2}$$

H matrix is symmetric with diagonal values between 0 and 1, h_{ii}, named leverages and average value ***p/n***. It corresponds to the last iteration (convergence) of the IWLS for estimating model parameters.

The h_{ii} are taken from the final iteration of the iterative weighted least squares procedure for fitting the model and have the usual interpretation, except that, unlike in a linear model, the hat-values in a GLM depend on y as well as on the configuration of the xs.

MODEL DIAGNOSTICS: RESIDUALS

4-6.2 Influence data in GLMz

➡ **Influence data are detected by and adapted to Cook's statistic** derived from the Wald statistic for multiple hypothesis testing: H_0: $\boldsymbol{\beta} = \boldsymbol{\beta}_0$,

$$Z_0^2 = \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)^T \hat{V}\left[\hat{\boldsymbol{\beta}}\right]^{-1}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) = \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)^T \mathbf{X}^{\mathbf{T}}\mathbf{W}\mathbf{X}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)$$

Let the Wald statistic be, for observation I, $Z^2_{(-i)}$ for testing H_0: $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_{(-i)}$, the distance between $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_{(-i)}$ $\left(\mathbf{d_i} = \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)}\right)$.

And thus, $Z^2_{(-i)} = \left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)}\right)^T \mathbf{X}^{\mathbf{T}}\mathbf{W}\mathbf{X}\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)}\right) = \dfrac{\left(e_i^{PS}\right)^2 h_{ii}}{p\left(1 - h_{ii}\right)}$

MODEL DIAGNOSTICS: RESIDUALS

4-6.3 Diagnostic plots

- plot(model) default diagnostic tool for normal response: DOES NOT WORK!!!
- A scatterplot showing Pearson residuals (Y-axis) (3-5 cut-off) and leverage (h_{ii}, diagonal of **H**). Cut-offs can be included at 2p/n or 3p/n.
- A scatterplot showing Pearson residuals versus each of the predictors in turn.
- A scatterplot showing Pearson residuals against fitted values. However, residualPlots shows residuals against the estimated linear predictor, η(x).
- Examine leverage for observations.
- Examine Cook's distance for observations.
  - In binary regression **for disaggregated data**, the plots of Pearson residuals or deviance residuals are strongly patterned, especially the plot against the linear predictor, where the residuals can take on only two values, depending on whether the response is equal to 0 or 1.
  - A suitable model requires that the conditional mean function in any residual plot be constant as we move across the plot; smoothers help in this purpose.
- In R, residualPlots(M) in each panel in the graph by default includes a smooth fit; a lack-of-fit test is provided only for the numeric predictor

```
residualPlots(mod.working, layout=c(1, 3))
influenceIndexPlot(mod.working, vars=c("Cook", "hat"), id.n=3)
```

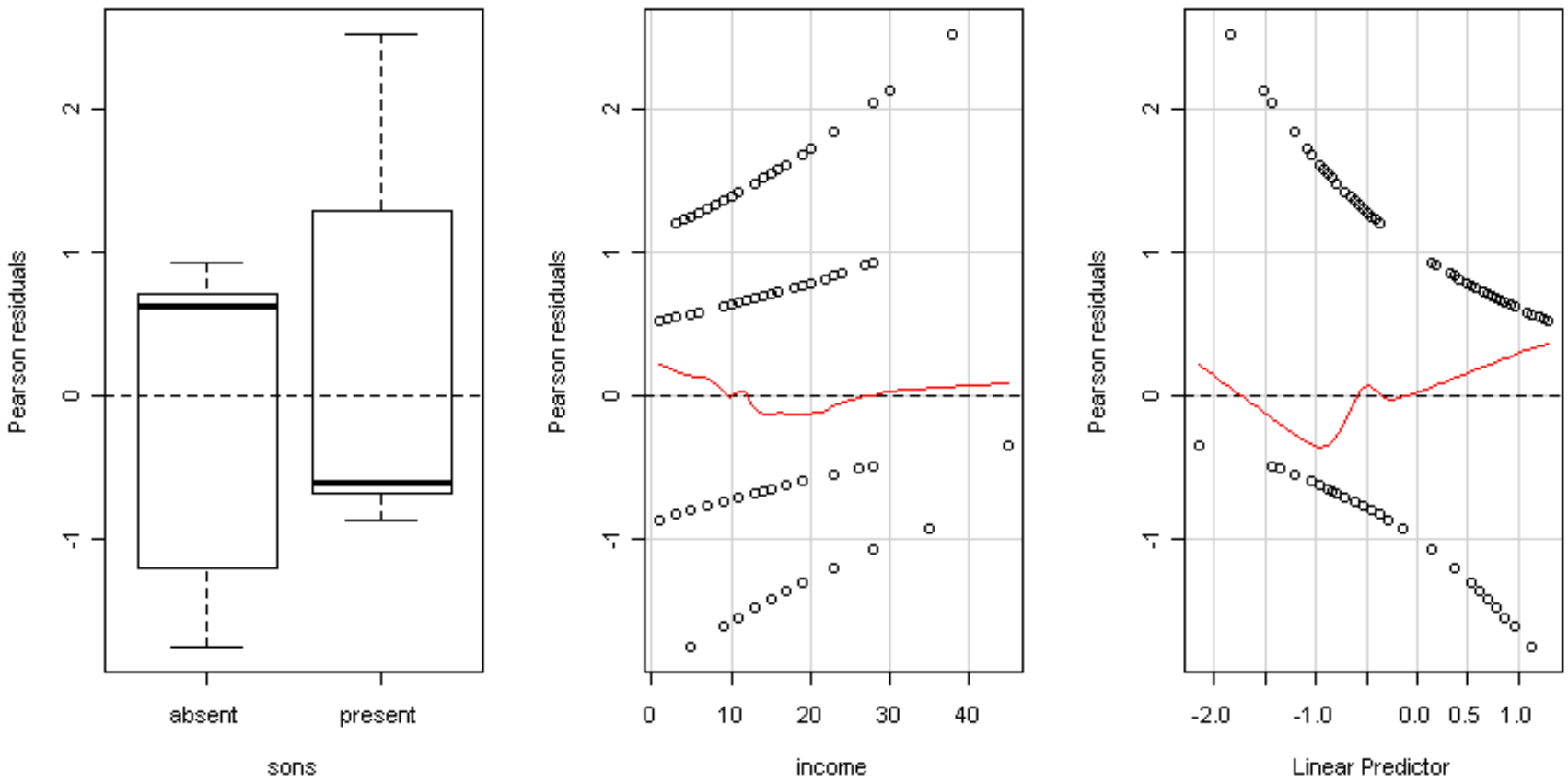
MODEL DIAGNOSTICS: RESIDUALS

Example of diagnostic plots for binary outcomes:

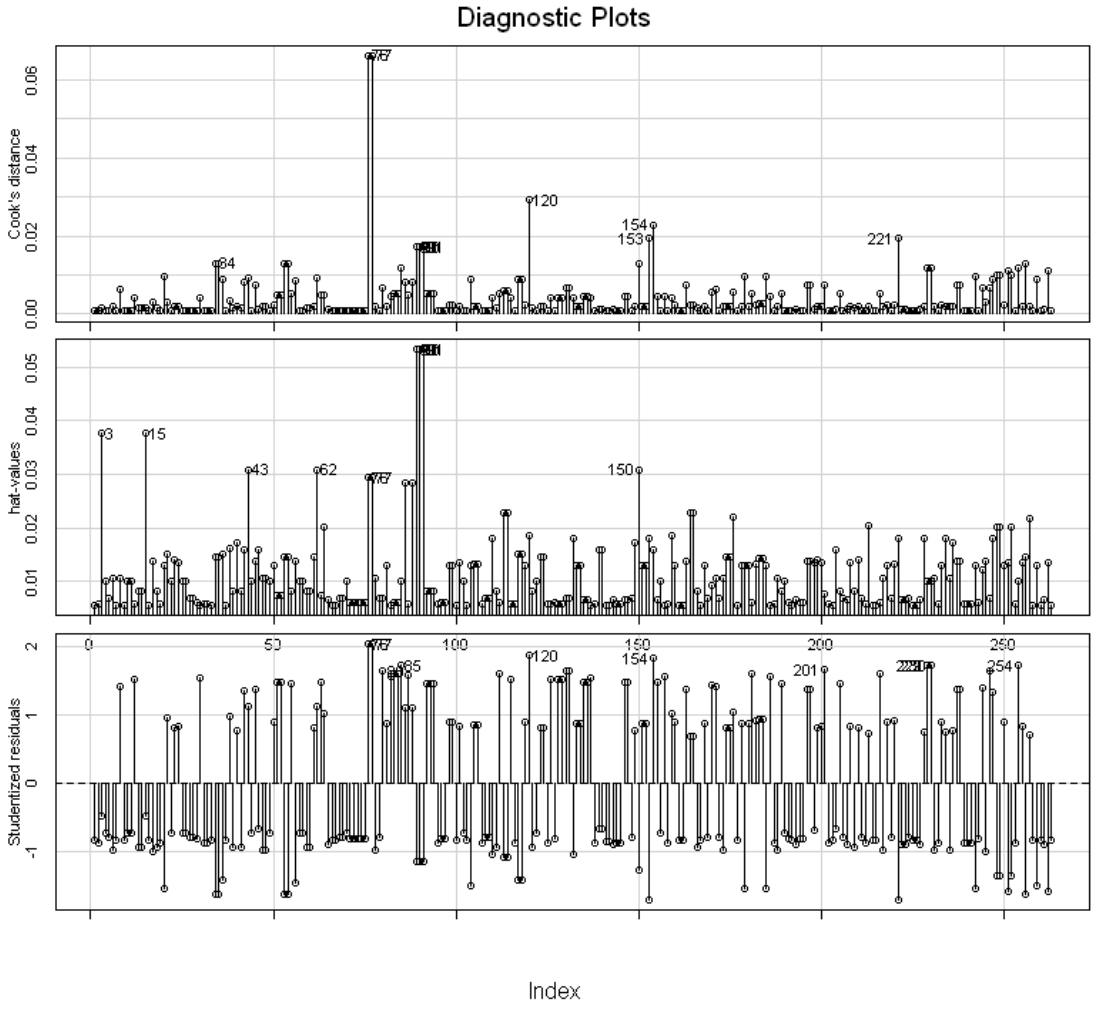
```
> options(contrasts=c("contr.treatment","contr.treatment"))
> bm3 <-glm( bwork~sons+income,family=binomial, data=womenlf )
> bm6 <-glm( bwork~sons*income, family=binomial, data=womenlf )
> anova(bm3,bm6,test='Chisq')
Analysis of Deviance Table

Model 1: bwork ~ sons + income
Model 2: bwork ~ sons * income
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1       260     319.73
2       259     319.12  1  0.60831    0.4354
>
> ## Diagnosi
> library(car)
> residualPlots(bm3, layout=c(1, 3))
       Test stat Pr(>|t|)
sons           NA       NA
income      1.226    0.268
> influenceIndexPlot(bm3, id.n=10)
> matplot(dfbetas(bm3),type='l')
> abline(h=sqrt(2/(dim(womenlf)[1])),lty=3,col=6)
> abline(h=-sqrt(2/(dim(womenlf)[1])),lty=3,col=6)
> lines(sqrt(cooks.distance(bm3)),lwd=3,col=1)
>legend(locator(n=1),legend=c(names(as.data.frame(dfbetas(bm3))),"Cook      D"),      col=c(1:3,1),
lty=c(3,3,3,1) )
```

MODEL DIAGNOSTICS: RESIDUALS

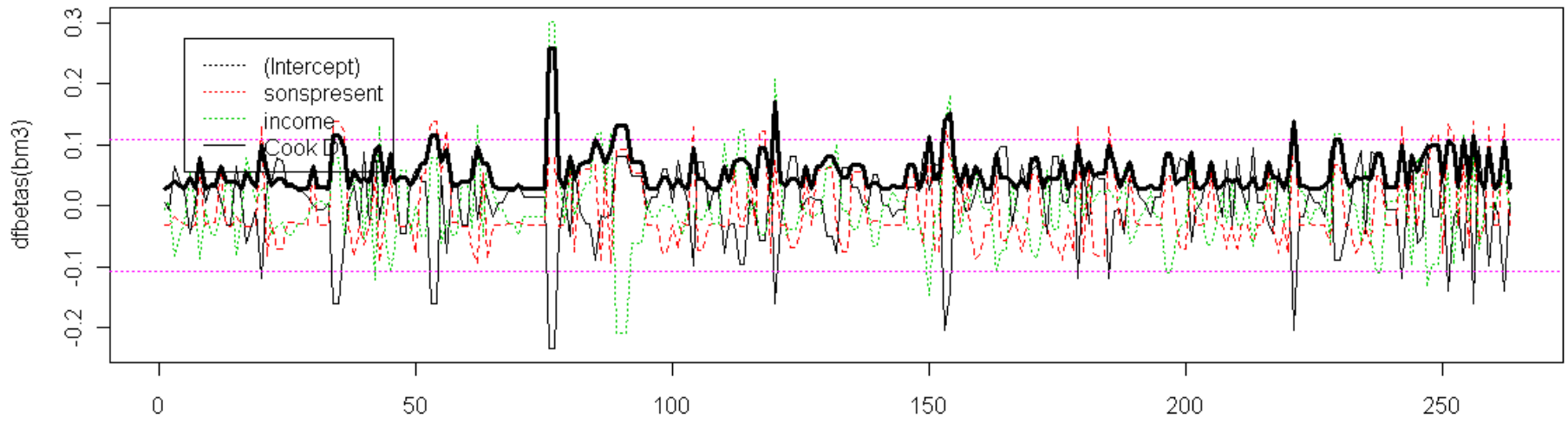


MODEL DIAGNOSTICS: RESIDUALS



MODEL DIAGNOSTICS: RESIDUALS

```
>
```



4-7. EXAMPLE 1: ACCIDENTS WITH INJURED PEOPLE ACCORDING TO SEAT-BELT USE – AGRESTI (2002)

Data about 68,694 accidents at Main. Accident severity and gender, environment and seat-belt use are available. The presence of injured people (No, Yes) will be studied as the target. (ref. NoInjured)

gender	environment	belt	severity	y	gender	environment	belt	severity	y
Female	Urban	No	None	7287	Male	Urban	No	Hospital	566
Female	Urban	Yes	None	11587	Male	Urban	Yes	Hospital	259
Female	Rural	No	None	3246	Male	Rural	No	Hospital	710
Female	Rural	Yes	None	6134	Male	Rural	Yes	Hospital	353
Male	Urban	No	None	10381	Female	Urban	No	StayHospital	91
Male	Urban	Yes	None	10969	Female	Urban	Yes	StayHospital	48
Male	Rural	No	None	6123	Female	Rural	No	StayHospital	159
Male	Rural	Yes	None	6693	Female	Rural	Yes	StayHospital	82
Female	Urban	No	NoHospital	175	Male	Urban	No	StayHospital	96
Female	Urban	Yes	NoHospital	126	Male	Urban	Yes	StayHospital	37
Female	Rural	No	NoHospital	73	Male	Rural	No	StayHospital	188
Female	Rural	Yes	NoHospital	94	Male	Rural	Yes	StayHospital	74
Male	Urban	No	NoHospital	136	Female	Urban	No	Mortal	10
Male	Urban	Yes	NoHospital	83	Female	Urban	Yes	Mortal	8
Male	Rural	No	NoHospital	141	Female	Rural	No	Mortal	31
Male	Rural	Yes	NoHospital	74	Female	Rural	Yes	Mortal	17
Female	Urban	No	Hospital	720	Male	Urban	No	Mortal	14
Female	Urban	Yes	Hospital	577	Male	Urban	Yes	Mortal	1
Female	Rural	No	Hospital	710	Male	Rural	No	Mortal	45
Female	Rural	Yes	Hospital	564	Male	Rural	Yes	Mortal	12

EXAMPLE 1

Models	logit(π_{ijk})	Deviance	n-p	AIC
1	η	1912.5	7	1981.2
SeatBelt – A (Cinturón)	$\eta+ \alpha_i$	1144.4	6	1215.1
Environment – C (Entorno)	$\eta+ \beta_j$	1192.8	6	1263.5
Gender –D (Genero)	$\eta+ \gamma_k$	1670.7	6	1741.4
A + D	$\eta+ \alpha_i+ \beta_j$	795.82	5	868.52
A + C	$\eta+ \alpha_i+ \gamma_k$	411.02	5	483.73
D + C	$\eta+ \beta_j+ \gamma_k$	911.01	5	983.71
A D	$\eta+ \alpha_i+ \beta_j+ (\alpha\beta)_{ij}$	795.32	4	870.03
A C	$\eta+ \alpha_i+ \gamma_k+ (\alpha\gamma)_{ik}$	408.31	4	483.01
A + D + C	$\eta+ \alpha_i+ \beta_j+ \gamma_k$	7.4645	4	82.167
A D + C	$\eta+ \alpha_i+ \beta_j+ \gamma_k+ (\alpha\beta)_{ij}$	7.3826	3	84.085
A C + D	$\eta+ \alpha_i+ \beta_j+ \gamma_k+ (\alpha\gamma)_{ik}$	3.5914	3	80.294
A + D C	$\eta+ \alpha_i+ \beta_j+ \gamma_k+ (\beta\gamma)_{jk}$	4.4909	3	81.193
A D + A C	$\eta+ \alpha_i+ \beta_j+ \gamma_k+ (\alpha\beta)_{ij}+ (\alpha\gamma)_{ik}$	3.5624	2	82.265
A D + D C	$\eta+ \alpha_i+ \beta_j+ \gamma_k+ (\alpha\beta)_{ij}+ (\beta\gamma)_{jk}$	4.372	2	83.074
A C + D C	$\eta+ \alpha_i+ \beta_j+ \gamma_k+ (\alpha\gamma)_{ik}+ (\beta\gamma)_{jk}$	1.3670	2	80.07
A D + A C + D C	$\eta+ \alpha_i+ \beta_j+ \gamma_k+ (\alpha\beta)_{ij}+ (\alpha\gamma)_{ik}+ (\beta\gamma)_{jk}$	1.3253	1	82.028

EXAMPLE 1

```
> summary(acc3)
    gender    environment seatbelt            severity        y                  y.bin          ones        target
 Male  :20   Rural:20    Yes:20   Hospitalizaci¢n:8   Min.   :    1.00   NoWounded: 8   Min.   :1   Min.   :  0.0
 Female:20   Urban:20    No :20   Hospital:8   1st Qu.:   66.75   Wounded  :32   1st Qu.:1   1st Qu.:  9.5
                                  NoHospital:8   Median :  138.50                  Median :1   Median : 74.0
                                  Mortal     :8   Mean   : 1717.35                  Mean   :1   Mean   :156.8
                                  None       :8   3rd Qu.:  710.00                  3rd Qu.:1   3rd Qu.:163.0
                                                  Max.   :11587.00                  Max.   :1   Max.   :720.0
> tapply(acc3$y,acc3$y.bin,sum);sum(acc3$y)
NoWounded   Wounded
    62420      6274
[1] 68694
```

- Taking as a response variable the presence of wounded people (f.heridos), globally there are 6274 accidents out of a total of 68694, with a probability of injured people of 0.0913. Odd is 6274/62420 or 0.1005 to 1 and the log-odds is log (0.1005) = -2.297472.
- It is proposed to initially compare the presence of injured people (response) according to Seat-Belt Use Factor (2 levels, base-line Yes).

Seat-Belt	y.bin - Wounded (positive outcome)	y.bin - NonWounded	m
Yes (ref)	2409	35383	37792
No	3865	27037	30902
	6274	***62420***	***68694***

P('Accident with Injured')=0.0913=6274/68694

EXAMPLE 1

There are only 2 possible models: the null model that assumes homogeneity in the Use in the two groups defined by the Factor (M1) and the complete model (M2) that proposes different proportions in the Use between the two groups:

(M1) $\log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta$

(M2) $\log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta + \alpha_i \quad i = 1,2 \quad \alpha_1 = 0$

```
> dfbelt
  seatbelt     m ypos  yneg
1      Yes 37792 2409 35383
2       No 30902 3865 27037
> prob <- sum(dfbelt$ypos)/sum(dfbelt$m);prob
[1] 0.09133258
> dfbelt$ypos0<-dfbelt$m*prob
> dfbelt$yneg0<-dfbelt$m*(1-prob)
> dfbelt
  seatbelt     m ypos  yneg    ypos0    yneg0
1      Yes 37792 2409 35383 3451.641 34340.36
2       No 30902 3865 27037 2822.359 28079.64
> m0<-glm(cbind(ypos,yneg)~1, family=binomial, data=dfbelt)
> summary(m0)
Call:
glm(formula = cbind(ypos, yneg) ~ 1, family = binomial, data = dfbelt)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.29747    0.01324  -173.5   <2e-16 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 768.03  on 1  degrees of freedom
Residual deviance: 768.03  on 1  degrees of freedom
AIC: 789.55

> m0$deviance
[1] 768.0317
> m1 <- glm(cbind(ypos,yneg)~seatbelt, family=binomial, data=dfbelt)
> summary(m1)
Call:
glm(formula = cbind(ypos, yneg) ~ seatbelt, family = binomial,
    data = dfbelt)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.68702    0.02106 -127.61   <2e-16 ***
seatbeltNo   0.74178    0.02719   27.29   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
    Null deviance:  7.6803e+02  on 1  degrees of freedom
Residual deviance: -5.7800e-12  on 0  degrees of freedom
AIC: 23.523

> m1$deviance
[1] -5.780043e-12
> residuals(m0,'pearson')
       Si        No
-18.61742  20.58856
> xpea<-sum(residuals(m0,'pearson')^2);xpea
[1] 770.4972
```

EXAMPLE 1

Pearson Statistic for (M2) is 0 and for (M1): $X_P^2 = \sum_{i=1,2} \frac{m_i\left(y_i - \hat{\mu}_i\right)^2}{\hat{\mu}_i\left(m_i - \hat{\mu}_i\right)} = 770.4972 \approx \chi^2_{n-p=2-1=1}$

(M2) Deviance is 0 and (M1) de: $D = 2\sum_{i=1,2}\left\{ y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) + \left(m_i - y_i\right)\log\left(\frac{m_i - y_i}{m_i - \hat{\mu}_i}\right)\right\} = 768.3 \approx \chi^2_{n-p=2-1=1}$.

Both statistics are highly significant, implying that the model does not fit the data well.

In (M1) the estimator $\hat{\eta} = -2.29747$, the logit of the sample proportion.

In (M2), the estimator $\hat{\eta}$, is the logit of the reference level (Yes) (logit of the proportion of wounded in group that Uses belt, logit (2409/37792) = - 2,687) and the effect of the No level on the logit of the proportion of injured (difference of logits between the No level and the reference level Yes: logit (3865/30902) -logit (2409/37792) = 0.742.

```
> dfbelt
  seatbelt     m ypos  yneg
1      Yes 37792 2409 35383
2       No 30902 3865 27037
```

$$\frac{\pi_i}{1-\pi_i} = \begin{cases} e^{\eta} & \iota = 1\, Yes \\ e^{\eta} e^{\alpha_2} & \iota = 2\ No \end{cases}$$

$$odds - ratio\, NovsYes = e^{\alpha_2} = 2.1$$

The odds of having injuries among accidents that do not use seat-belt are more than twice the odds of having injuries among those who wear seat-belt.

EXAMPLE 1

Models with 2 Predictors: Seat-Belt and Environment

There are 4 groups, the number of accidents with injuries in the i-th Seat-Belt group and the j-th Environment group, where the reference levels are 'Yes' for Seat-Belt (Factor A) and' Urban'for Factor C.

```
> df2
  seatbelt environment     m ypos  yneg
1      Yes       Urban 23695 1139 22556
2       No       Urban 19476 1808 17668
3      Yes       Rural 14097 1270 12827
4       No       Rural 11426 2057  9369
```

There are 5 models of interest applicable to the systematic structure of the previous data (M1) to (M5), whose returns and details of the estimation are detailed below.

Model		*n-p*	*Deviance*	ΔD	*Contrast*	*g.l.*	**Modeo**
1	**1**	3	**1504.1**		All Significant		Constane: η
2	**A**	2	**736.11**	**767.99**	(M2) vs (M1)	1	Seat-belt: $\eta+\alpha_i$
3	**C**	2	**784.53**	**719.57**	(M3) vs (M1)	1	Environment: $\eta+\beta_j$
4	**A+C**	1	2.7116	**733.4**	(M4) vs (M2)	1	Additive: $\eta+\alpha_i+\beta_j$
				781.8	(M4) vs (M3)	1	
5	**A*C**	0	0	2.7116	(M5) vs (M4)	1	Interacción Factores: $\eta+\alpha_i+\beta_j+\alpha\beta_{ij}$

EXAMPLE 1

```
> sum(df2[,3]);sum(df2[,4]);sum(df2[,5])
[1] 68694
[1] 6274
[1] 62420
> m1<-glm(cbind(ypos,yneg)~1, family=binomial, data=df2)
> summary(m1)
Call:
glm(formula = cbind(ypos, yneg) ~ 1, family = binomial, data = df2)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.29747    0.01324  -173.5   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1504.1  on 3  degrees of freedom
Residual deviance: 1504.1  on 3  degrees of freedom
AIC: 1542.4

Number of Fisher Scoring iterations: 4

> m1$deviance
[1] 1504.141
> m2<-glm(cbind(ypos,yneg)~seatbelt, family=binomial, data=df2)
> summary(m2)

Call:
glm(formula = cbind(ypos, yneg) ~ seatbelt, family = binomial, data = df2)

Coefficients:
```

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.68702    0.02106 -127.61   <2e-16 ***
seatbeltNo   0.74178    0.02719   27.29   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1504.14  on 3  degrees of freedom
Residual deviance:  736.11  on 2  degrees of freedom
AIC: 776.34

Number of Fisher Scoring iterations: 4

> m2$deviance
[1] 736.109
> m3<-glm(cbind(ypos,yneg)~environment, family=binomial, data=df2)
> summary(m3)
Call:
glm(formula = cbind(ypos, yneg) ~ environment, family = binomial,
    data = df2)
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -2.61368    0.01908 -136.96   <2e-16 ***
environmentRural  0.71584    0.02664   26.87   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1504.14  on 3  degrees of freedom
Residual deviance:  784.53  on 2  degrees of freedom
AIC: 824.76

Number of Fisher Scoring iterations: 4
```

```
> m3$deviance
[1] 784.5302
> m4<-glm(cbind(ypos,yneg)~seatbelt+environment, family=binomial, data=df2)
> summary(m4)

Call:
glm(formula = cbind(ypos, yneg) ~ seatbelt + environment, family = binomial,
    data = df2)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -3.01397    0.02544 -118.48   <2e-16 ***
seatbeltNo        0.75265    0.02734   27.53   <2e-16 ***
environmentRural  0.72721    0.02682   27.12   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1504.1407  on 3  degrees of freedom
Residual deviance:    2.7116  on 1  degrees of freedom
AIC: 44.938

Number of Fisher Scoring iterations: 3

> m4$deviance
[1] 2.711593
> m5<-glm(cbind(ypos,yneg)~seatbelt*environment, family=binomial, data=df2)
> summary(m5)

Call:
glm(formula = cbind(ypos, yneg) ~ seatbelt * environment, family = binomial,
    data = df2)

Deviance Residuals:
[1]  0  0  0  0
```

```
Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                -2.98585    0.03037 -98.318   <2e-16 ***
seatbeltNo                  0.70632    0.03914  18.046   <2e-16 ***
environmentRural            0.67331    0.04228  15.925   <2e-16 ***
seatbeltNo:environmentRural 0.09006    0.05468   1.647   0.0996 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1.5041e+03  on 3  degrees of freedom
Residual deviance: 6.5707e-12  on 0  degrees of freedom
AIC: 44.226

Number of Fisher Scoring iterations: 2

> m5$deviance
[1] 6.570744e-12
```

EXAMPLE 1

➡ The additive model fits the data well, but there is still some deviance to explain:

```
> summary(acc3)
    gender    environment seatbelt       damage         y                 y.bin        target
 Female:20   Urban:20     Yes:20   NoWounded:8   Min.   :    1.00   NoWounded: 8   Min.   :  0.0
 Male  :20   Rural:20     No :20   Mild      :8   1st Qu.:   66.75   Wounded  :32   1st Qu.:  9.5
                                   Severe    :8   Median :  138.50                  Median : 74.0
                                   Hospital  :8   Mean   : 1717.35                  Mean   :156.8
                                   Death     :8   3rd Qu.:  710.00                  3rd Qu.:163.0
                                                  Max.   :11587.00                  Max.   :720.0
> df3
  seatbelt environment gender     m ypos  yneg
1      Yes       Urban Female 12346  759 11587
2       No       Urban Female  8283  996  7287
3      Yes       Rural Female  6891  757  6134
4       No       Rural Female  4219  973  3246
5      Yes       Urban   Male 11349  380 10969
6       No       Urban   Male 11193  812 10381
7      Yes       Rural   Male  7206  513  6693
8       No       Rural   Male  7207 1084  6123
```

EXAMPLE 1

```
> summary(m1)
Call: glm(formula = cbind(ypos, yneg) ~ 1, family = binomial, data = df3)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.29747    0.01324  -173.5   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 1912.5  on 7  degrees of freedom
Residual deviance: 1912.5  on 7  degrees of freedom
AIC: 1981.2

> summary(m2)
Call: glm(formula = cbind(ypos, yneg) ~ seatbelt + environment + gender,
    family = binomial, data = df3)


Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -2.03350    0.02714  -74.94   <2e-16 ***
seatbeltNo       0.81710    0.02765   29.55   <2e-16 ***
environmentUrban -0.75806    0.02697  -28.11   <2e-16 ***
genderFemale     -0.54483    0.02727  -19.98   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 1912.4532  on 7  degrees of freedom
Residual deviance:    7.4645  on 4  degrees of freedom
AIC: 82.167
```

EXAMPLE 1

➡ The next step could be to add an interaction between 2 of the factors: A * C or A * D or C * D.

Model		n-p	Deviance	ΔD	Contrast	g.l.	Model
1	A+C+D	4	7.4645				Additive: $\eta+\alpha_i+\beta_j+\gamma_k$
2	A*C+D	3	**3.5914**	**3.8730**	(M2) vs (M1)	1	Interaction Seat.Belt-Environ. : $\eta+\alpha_i+\beta_j+\gamma_k+\alpha\beta_{ij}$
3	A*D+B	3	7.3826	0.0818	(M3) vs (M1)	1	Interaction Seat.Belt-Gender: $\eta+\alpha_i+\beta_j+\gamma_k+\alpha\gamma_{ik}$
4	C*D+A	3	**4.4909**	**2.9736**	(M4) vs (M1)	1	Interaction Environ. - Gender: $\eta+\alpha_i+\beta_j+\gamma_k+\beta\gamma_{jk}$

Strictly only the interaction between Seat.Belt and Environment is statistically significant, although the interaction between Environment and Gender has a value of 8% according to the deviance contrast with the additive model. The best model so far seems to have all 3 factors and 2 double interactions: one Belt Use – Environment and the second, Belt Use –Environment.

```
> m3<-glm(cbind(ypos,yneg)~(seatbelt+environment+gender)^2, family=binomial, data=df3)
> summary(m3)

Call:
glm(formula = cbind(ypos, yneg) ~ (seatbelt + environment + gender)^2,
    family = binomial, data = df3)

Deviance Residuals:
```

```
      1         2         3         4         5         6         7         8
 0.3861   -0.3457   -0.3938    0.3760   -0.5309    0.3754    0.4731   -0.3373

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -2.07710    0.03599 -57.706   <2e-16 ***
seatbeltNo                     0.85855    0.04673  18.373   <2e-16 ***
environmentUrban              -0.66304    0.04717 -14.056   <2e-16 ***
genderFemale                  -0.51318    0.05071 -10.119   <2e-16 ***
seatbeltNo:environmentUrban   -0.09685    0.05547  -1.746   0.0808 .
seatbeltNo:genderFemale        0.01144    0.05603   0.204   0.8382
environmentUrban:genderFemale -0.08176    0.05469  -1.495   0.1349
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1912.4532  on 7  degrees of freedom
Residual deviance:    1.3253  on 1  degrees of freedom
AIC: 82.028

Number of Fisher Scoring iterations: 3

> Anova(m3)
Analysis of Deviance Table (Type II tests)

Response: cbind(ypos, yneg)
                    LR Chisq Df Pr(>Chisq)
seatbelt              901.71  1     <2e-16 ***
environment           787.94  1     <2e-16 ***
gender                404.72  1     <2e-16 ***
seatbelt:environment    3.05  1     0.0809 .
seatbelt:gender         0.04  1     0.8382
environment:gender      2.24  1     0.1347
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> m4 <- step(m3)
Start:  AIC=82.03
cbind(ypos, yneg) ~ (seatbelt + environment + gender)^2

                       Df Deviance    AIC
- seatbelt:gender       1   1.3670 80.069
<none>                     1.3253 82.028
- environment:gender    1   3.5624 82.265
- seatbelt:environment  1   4.3720 83.074

Step:  AIC=80.07
cbind(ypos, yneg) ~ seatbelt + environment + gender + seatbelt:environment +
    environment:gender

                       Df Deviance    AIC
<none>                     1.3670 80.069
- environment:gender    1   3.5914 80.294
- seatbelt:environment  1   4.4909 81.193
> summary(m4)

Call:
glm(formula = cbind(ypos, yneg) ~ seatbelt + environment + gender +
    seatbelt:environment + environment:gender, family = binomial,
    data = df3)

Deviance Residuals:
      1        2        3        4        5        6        7        8
 0.4522  -0.4043  -0.3212   0.3063  -0.6204   0.4396   0.3851  -0.2750

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                -2.07989    0.03333 -62.395   <2e-16 ***
seatbeltNo                  0.86389    0.03872  22.310   <2e-16 ***
environmentUrban           -0.66274    0.04718 -14.048   <2e-16 ***
genderFemale               -0.50634    0.03806 -13.305   <2e-16 ***
seatbeltNo:environmentUrban -0.09773    0.05529  -1.768   0.0771 .
```

```
environmentUrban:genderFemale -0.08148    0.05466  -1.491   0.1360
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1912.453  on 7  degrees of freedom
Residual deviance:    1.367  on 2  degrees of freedom
AIC: 80.069

Number of Fisher Scoring iterations: 3
> xpea2<-sum(residuals(m2,'pearson')^2);xpea2
[1] 7.487384
> 1-pchisq( xpea2, m2$df.residual )
[1] 0.1122669
> xpea3<-sum(residuals(m3,'pearson')^2);xpea3
[1] 1.324618
> 1-pchisq( xpea3, m3$df.residual )
[1] 0.249765
> xpea4<-sum(residuals(m4,'pearson')^2);xpea4
[1] 1.365019
> 1-pchisq( xpea4, m4$df.residual )
[1] 0.5053472
> anova(m2,m4,test="Chisq")
Analysis of Deviance Table

Model 1: cbind(ypos, yneg) ~ seatbelt + environment + gender
Model 2: cbind(ypos, yneg) ~ seatbelt + environment + gender + seatbelt:environment +
    environment:gender
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         4     7.4645
2         2     1.3670  2   6.0975  0.04742 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

EXAMPLE 1

The next step would be to analyze the models with 2 interactions between the factors, since the A * C + D model fits the data well, but still leaves a 3.5914 return for explaining in 3 degrees of freedom.

Modelo		*n-p*	*Devianza*	ΔD	*Contraste*	*g.l.*	**Modelo**
1	**A*C+A*D**	2	3.562410	2.2371	(M1) vs (M4)	1	Interactions Seatbelt-Environment and Seatbelt-Gender : $\eta + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{jk}$
2	**A*D+C*D**	2	4.371979	**3.0467**	(M2) vs (M4)	1	Interactions Seatbelt-Gender and Environment-Gender: $\eta + \alpha_i + \beta_j + \gamma_k + \alpha\gamma_{ik} + \beta\gamma_{jk}$
3	**A*C+C*D**	2	1.367022	0.04171	(M3) vs (M4)	1	Interactions Seatbelt-Environment and Environment-Gender : $\eta + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \beta\gamma_{jk}$
4	**A*C+C*D+ A*D**	1	1.325317				$\eta + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk}$

➡ The model does not require further analysis, there are no significant differences between the model with the 3 double interactions and any of the models with 2 pairs of interactions.

EXAMPLE 1

The next step would be to analyze the models with 2 interactions between the factors and compare them with the additive model, to see if 2 double interactions are simultaneously significant.

Model		*n-p*	*Deviance*	ΔD	*Contrast*	*g.l.*	***Model***
1	**A*C+A*D**	2	3.562410	3.9021	(M1) vs (M4)	1	Interactions Seatbelt-Environment and Seatbelt-Gender : $\eta+\alpha_i+\beta_j+\gamma_k+\alpha\beta_{ij}+\alpha\gamma_{jk}$
2	**A*D+C*D**	2	4.371979	3.0925	(M2) vs (M4)	1	Interactions Seatbelt-Gender and Environment-Gender: $\eta+\alpha_i+\beta_j+\gamma_k+\alpha\gamma_{ik}+\beta\gamma_{jk}$
3	**A*C+C*D**	2	1.367022	**6.0975**	(M3) vs (M4)	1	Interactions Seatbelt-Environment and Environment-Gender: $\eta+\alpha_i+\beta_j+\gamma_k+\alpha\beta_{ij}+\beta\gamma_{jk}$
4	**A+C+D**	4	7.4645				$\eta+\alpha_i+\beta_j+\gamma_k$

➡ The model does not require further analysis, since 2 interactions are simultaneously significant Belt-Environment and Environment-Gender.

EXAMPLE 1

Comparing the best model with 1 double interaction (Belt-Environment) with the model that has 2 double interactions (Belt-Environment and Environment-Gender) the p value of the contrast of the Environment-Gender interaction is 0.14, therefore, not significant once Belt-Environment is in the model, but with an uncomfortable value.

```
> m5 <- glm(cbind(ypos,yneg)~seatbelt + environment + gender + seatbelt:environment,
family=binomial, data=df3)
> anova( m5, m4, test="Chisq")
Analysis of Deviance Table

Model 1: cbind(ypos, yneg) ~ seatbelt + environment + gender + seatbelt:environment
Model 2: cbind(ypos, yneg) ~ seatbelt + environment + gender + seatbelt:environment +
    environment:gender
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         3     3.5914
2         2     1.3670  1   2.2244   0.1358
```

It is proposed to finalize the analysis evaluating the model with 2 double interactions and the best model with 1 double interaction according to the information criterion of Akaike and the step () method in R.

It is preferred to keep the 2 double interactions.

At the beginning, a summary table is given with the residual liability and the AIC for all the models that have been calculated.