UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
School of Professional & Executive Development

Industria 4.0, Modelización, Simulación y Materialización.

P.01 Industria 4.0 y Gestión de Datos

Statistical Modelling – Binary Outcome

Lídia Montero Mercadé
lidia.montero@upc.edu

Barcelona, 15, 21 Noviembre y 5 Diciembre 2025

CONTENTS

READINGS

Basic References:

- 🕮 Fox, J. Applied Regression Analysis and Generalized Linear Models. Sage Publications, Edition 2015.
- 🕮 Fox and Weisberg ***An R Companion to Applied Regression****.* Sage Publications, Edition 2011.
- 🕮 Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. Springer New York, 2009.

1.2-1 BINARY RESPONSE DATA. A GENERALIZED LINEAR MODEL CASE

1.2-1.1 Components of generalized linear models

Generalized linear models are extensions of the classical multiple regression models.

Let $\mathbf{y}^T = (y_1, \ldots, y_n)$ be a vector of n components, sample of a random vector $\mathbf{Y}^T = (Y_1, \ldots, Y_n)$, with statistical independent components, identically distributed with means $\boldsymbol{\mu}^T = (\mu_1, \ldots, \mu_n)$:

- **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 distributions with jointly expected values $\mathrm{E}[\mathbf{Y}] = \boldsymbol{\mu}$.**

➡ **At desaggreted leved for each individual observation, the response is dycothomic and we are concerned with Bernoulli distribution. For grouped data, binomial distributions are suitable.**

- **The systematic component in the model** model specifies a vector $\boldsymbol{\eta}$, the predictor lineal vector is a linear combination from a limited number of explicative 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*, **X** is *nxp* and $\boldsymbol{\beta}$ is *px1*.

BINARY RESPONSE DATA. A GENERALIZED LINEAR MODEL CASE

- For each observation, the expected value μ is related to the linear predictor η, through the ***link function*, notated as *g(.)***, and thus $\eta = \mathrm{g}(\mu)$.

Response function is: $\mu = \mathrm{g}^{-1}(\eta)$

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

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

Since ML estimates $\boxed{\hat{\beta} \rightarrow \forall i \quad \hat{\eta}_i = x_i^T \hat{\beta} \rightarrow \hat{\mu}_i = g^{-1}(\hat{\eta}_i)}$

BINARY RESPONSE DATA. A GENERALIZED LINEAR MODEL CASE

1.2-1.2 Classification of statistical linear models

Explicative Variables	**Response Variable**				
	Dicothomic or Binary	***Polythomic***	***Counts (discrete)***	***Continuous***	
				Normal	***Time between events***
Dicothomic	Contingency tables Logistic regression Log-linear models	Contingency tables Log-linear models	Log-linear models	Tests for 2 subpopulation means: t.test	Survival Analysis
Polythomic	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	Covariance Analysis	Survival Analysis
Random Effects	Mixed models	Mixed models	Mixed models	Mixed models	Mixed models

1.2-2. INTRODUCTION TO BINARY RESPONSE DATA. BINOMIAL MODELS

This variable appears when given a sample each individual holds or does not hold a target characteristic of the study and this is codified as (*Y=1*) or not (*Y=0*).

For example, on mode selection in transportation models, one might be interested in the modal choice between public (metro, bus, light train, etc) or private modes (car, motorcycle, etc) in the study area for home to work trips. In those 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 fail, in our case private modes)*.

➡ It is possible the extension to more than 2 levels or categories in the response variable.

➡ The probability of success is notated by π , in such a way that,

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

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

Each individual in the sample is characterized by a set of variables some covariates (income, age) some of them factors (gender, grades, etc) that defines: $\mathbf{x}_\mathbf{k}^\mathbf{T} = \begin{pmatrix} x_1 & \dots & x_p \end{pmatrix}$.

➡ Explicative variables that will form the linear predictor $\mathbf{x}_{\mathbf{k}}^{\mathbf{T}} = \left(x_1 \quad \dots \quad x_p\right)$ might be:

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

Por example, in the public-private binary modal choice model, for each commuter variables 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 the response variable y and the explicative variables in order to model the probability of positive response: $\pi = \pi(\mathbf{x})$.

➡ In design of experiment groups of individual units are defined, and each group receives a combination of experimental conditions in such a way that are shared by all the unit in the group, in general, factors are considered as explicative variables, and k*th* experimental *condition is modeled by a common set of values for all the explicative variables of individual units in the group* $\mathbf{x}_{\mathbf{k}}^{\mathbf{T}} = \left(x_1 \quad \dots \quad x_p\right)$ and thus apply to $\boxed{m_k}$ individual units.

1.2-2. 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 $\boxed{n}$ and the total number of units is $N = m_1 + \ldots + m_n$.

Each group or combination of experimental conditions defines a *covariate class* where all individual units belonging to the covariate class share the same values for explicative variables.

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

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

1.2-2. INTRODUCTION TO BINARY RESPONSE DATA. BINOMIAL MODELS

➡ … Let us work with a simple example to see differences in the representation…

Disaggregated data			*Aggregated data*		
Individual unit	***Variables***	***Response***	***Covariate class***	***Size of the class***	***Positive responses***
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			

Former table shows an experiment consisting on *dicothomic* 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 can be gener (two levels, coded as *male* and *female*) and factor C is the car availability (1 car or more than 1).

1.2-2. INTRODUCTION TO BINARY RESPONSE DATA. BINOMIAL MODELS

... **Aggregated versus disaggregated data ...**

- More efficient and less memory consumption when aggregated data is considered. It simplifies significant effect detected at a glance.

- Aggregated data implies 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 sample observed positive responses are $\boxed{y_1/m_1, \ldots, y_n/m_n}$, where $\boxed{0 \le y_k \le m_k}$ being the number of positive responses in kth covariate class which size is $\boxed{m_k}$.

- Size of covariate classes in a vector form are called ***binomial index vector*** and notated by $\mathbf{m} = \begin{pmatrix} m_1 & \ldots & m_n \end{pmatrix}$. For disaggregated data, each individual unit defines a binomial response for a group of size 1 and thus, $\mathbf{m} = \begin{pmatrix} 1 & \ldots & 1 \end{pmatrix}$.

1.2-2. INTRODUCTION TO BINARY RESPONSE DATA. BINOMIAL MODELS

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

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

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

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

1.2-2.1 **Binomial Distribution**

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

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

1.2-2. BINOMIAL MODELS FOR BINARY DATA

1.2-2.2 Link functions

➡ The goal consists on stablishing a functional relationship between the probability of a positive result π and the vector of explicative variables (predictors in general, covariates if they are continuous) $\mathbf{x}^{\mathrm{T}} = \begin{pmatrix} x_1 & \dots & x_p \end{pmatrix}$: $\boxed{\pi = \pi(\mathbf{x})}$.

- In Generalized Linear Models, the link function relates the linear predictor scale with 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 dicothomic individual response, the linear predictor η might for a given observation any value in the real axis, but the probability of positive answer belongs to the open interval (0, 1).

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

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

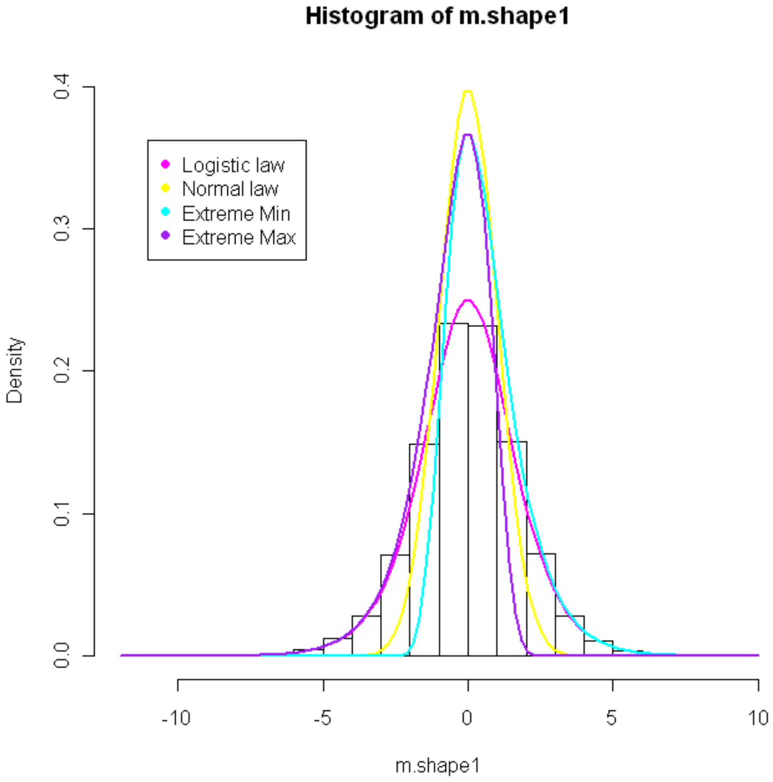
1.2-2. BINOMIAL MODELS FOR BINARY DATA

➡ Logit link function is the most frequently used link, but one should understand the role of link functions and do not act authomatically.Some common *link* function for binary response data are:

1. **The logit link (sometimes bad-called logistic link)**:

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

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



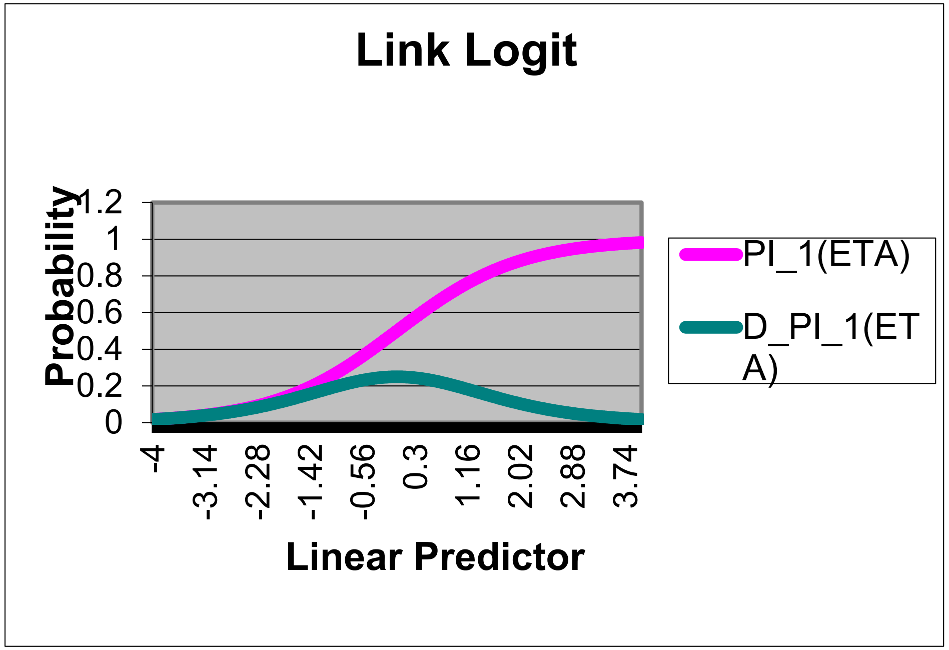
1.2-2. BINOMIAL MODELS FOR BINOMIAL DATA : LINK FUNCTIONS

- … *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:

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



In general,

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

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

Transformations should depend on characteristics of data and not to be selected in an straight forward way.

1.2-2. BINOMIAL MODELS FOR BINOMIAL DATA : LINK FUNCTIONS

➡ *logit* and *probit* links can be seen 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

1.2-3. ESTIMATION OF MODEL PARAMETERS

- ➡ The estimation process relies on an 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 method of the scores, a second order method Newton-type, specialized to the properties of the log likelihood function. The method converges fast, but it is not globally convergent.
- ➡ Existence and unicity for estimates under any of the formerly presented link functions, if $0<y_i<m_i$ for any covariate class/observations.

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

1.2-4. GOODNESS OF FIT

- Let $\hat{\beta}$ the estimates of the model parameters, thus a linear predictor for each observation *i* might be computed $\hat{\eta}_i = \mathrm{x}_\mathrm{i}^\mathrm{T}\beta$ and thus through the response function (inverse of the selected link function) fitted values can be computed: $\boxed{\hat{\pi}_i = g^{-1}(\hat{\eta}_i)}$.
- The scaled deviance can be computed, $D'(\mathbf{y},\hat{\boldsymbol{\mu}}) = 2\,\ell(\mathbf{y},\mathbf{y}) - 2\,\ell(\hat{\boldsymbol{\mu}},\mathbf{y})$.
- And so, the deviance that under the binomial distribution is identical, since $\varphi = 1$

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

- The saturated model $\ell(\mathbf{y},\mathbf{y})$ implies fitted probabilities as observed probabilities, notated in the following as $\tilde{\pi}_i = \dfrac{y_i}{m_i}$, for all i = 1,…n.

BINOMIAL MODELS FOR BINARY DATA: GOODNESS OF FIT

- Expression for the deviance:

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

- Sometimes, the deviance statistic is ,

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

where,

1. Observed values for positive response for observation i, $o_i = y_i$.
2. Observed values for negative response 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

➡ Assimptotic 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}). Assymptotic conditions are

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

➡ If $\boxed{\text{pvalue <<0.05}}$ then there is evidence to reject H0 and thus, the model (M) does not fit properly data. There is an statistical evidence of discrepancy between observations and fitted values provided by model (M).

➡ If $\boxed{\text{pvalue >> 0.05}}$ then there is not evidence to reject H0 and thus, H0 is accepted leading to the conclusion that model (M) does fit properly data, since discrepance between observed and fitted values is not significative in statistical terms.

➡ AIC (Akaike Information Criteria, 1974) is defined as a trade-off between a goodness of fit provided by a model (M) and the number of parameters *p* in the model (as an indicator for model complexity) kaike. 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 named BIC (*Bayesian Information Criteria*) is proposed (in SAS©), Schwartz criteria $BIC_{\mathbf{B}} = -2\,\ell\left(\hat{\pi}_B, \mathrm{y}\right) + p \log n$. Minimum BIC models are preferred (`AIC(model,k=log(n)`).

- **AIC and BIC might be used to compare unnested models.**

- **Following McCullagh, the test for Generalized LM equivalent to F.Test in classical linear regression, consists on comparing differences on scaled deviance on 2 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 $\hat{\pi}_B$ fitted probabilities for both models, Such that, the set of parameters for M$_B$ are those common to M$_A$ and those specific; i.e., $\boldsymbol{\beta}_B^T = \left(\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T\right)$ and $\boldsymbol{\beta}_A^T = \left(\boldsymbol{\beta}_1^T\right)$ with dim($\boldsymbol{\beta}_A$)=q<p, then

$\boxed{\Delta D_{AB} = D\left(\mathbf{y}, \hat{\pi}_A\right) - D\left(\mathbf{y}, \hat{\pi}_B\right) = 2\,\ell\left(\hat{\pi}_B, \mathrm{y}\right) - 2\,\ell\left(\hat{\pi}_A, \mathrm{y}\right)}$ is asymptotically distributed χ^2_{p-q} .

And for testing $\boxed{\mathrm{H}_0 : \beta_2 = \mathbf{0}}$ $P\left(\chi^2_{p-q} > \Delta D_{AB}\right) \rightarrow \begin{cases} << \alpha & \mathrm{H}_0 \text{ Rejected} \\ >> \alpha & \mathrm{H}_0 \text{ Accepted} \end{cases}$

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

BINOMIAL MODELS FOR BINARY DATA: GOODNESS OF FIT

In R software:

```
anova(modelA, modelB, ...,    test = c("F", "Chisq")) # Deviance Test
```

- Deviance for a GLM plays a role similar to Residual Sum of Squares in classical regression, thus it allows to define a Generalized ***R*²**, or pseudo ***R*²**

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

$$0 \le R^2 \le 1$$

- Goodness of fit by generalized Pearson **X²** statistic, asymptotically distributed as:

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

BINOMIAL MODELS FOR BINARY DATA: GOODNESS OF FIT

1.2-4.1 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 value of area under the curve is large, we can say the model performs well. If the area is close to 1.0, it indicates that the model is good. While if the area is to 0.5, it shows that the model is bad.

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

Let the prediction in response be $\hat{y}_i = 1$ if $\hat{\pi}_i > s$, where is a threshold $\boxed{s}$ 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	b	**a+b**
$\hat{y}_i = 0$	c	d	**c+d**
	a+c	b+d	**n**

- **Sensibility** *is the proportion of observed positive outcomes (Y=1) predicted as positive (* $\hat{y}_i = 1$ *):* Sn =a/(a+c).
- **Specificity** *is the proportion of observed negative outcomes (Y=0) predicted as negative (* $\hat{y}_i = 0$ *):* Sp = d/(b+d).

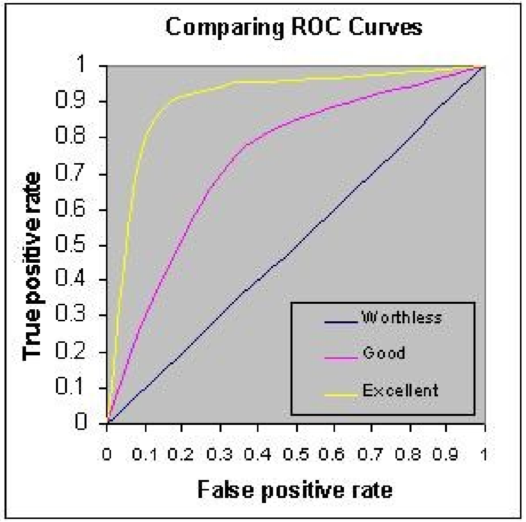
BINOMIAL MODELS FOR BINARY DATA: GOODNESS OF FIT

➡ **ROC Curve** shows in X axis for each s, **1-Sp (false negative rate)** and **Sensibility** - Sn (true positive rate) in Y-Axis.

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

A guideline for interpreting ROC curve is:

.90-1 = excellent(A)
.80-.90 = very good (B)
.70-.80 = good (C)
.60-.70 = bad (D)
.50-.60 = very bad (F)



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

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

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

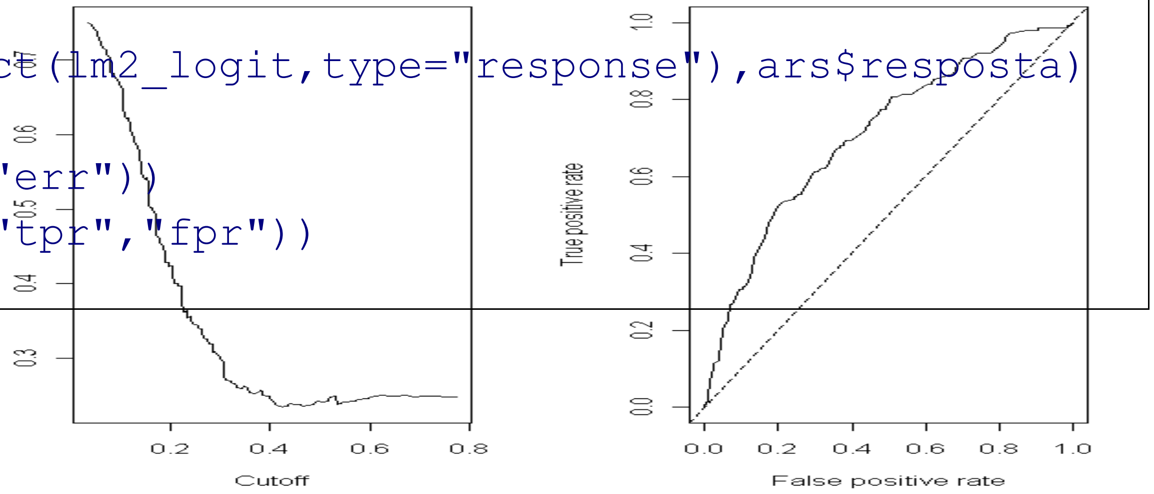
As in the case, for deviance residual:

```
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)
```



1.2-5. MODEL DIAGNOSTICS

1.2-5.1 Residuals in GLMz

Extension to Generalized linear Models of Normal regression methods:

- Pearson residuals are casewise components of the Pearson goodness of fit statistic for the model,

$$e_i^P = \frac{(y_i - \hat{\mu}_i)}{\sqrt{V[\hat{\mu}_i]/\hat{\phi}}} \quad \text{and} \quad X^2 = \sum_{i=1}^{n} r_{P_i}{}^2$$

These are 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, $e_i^D = sign\,(y_i - \hat{\mu}_i)\sqrt{d_i}$ are the square roots of the casewise components of the residual deviance $D(\mathbf{y}, \hat{\mu}) = \sum_{i=1,\ldots,n} r_{D_i}^2$, 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")`.

1.2-5. 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, hii, 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 x*s*.

1.2-5. MODEL DIAGNOSTICS: RESIDUALS

1.2-5.2 Influence data in GLMz

➡ **Influence data are detected by and adapted Cook's statitstic** derived from Wald statistic for multiple hypothesis testing: H$_0$: $\boxed{\beta = \beta_0}$,

$$Z_0^2 = \left(\hat{\beta} - \beta_0\right)^T \hat{V}\left[\hat{\beta}\right]^{-1}\left(\hat{\beta} - \beta_0\right) = \left(\hat{\beta} - \beta_0\right)^T \mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}\left(\hat{\beta} - \beta_0\right)$$

Let Wald statistic be for observation I, $Z_{(-i)}^2$ for testing H$_0$: $\boxed{\beta = \hat{\beta}_{(-i)}}$, distance between $\hat{\beta}$ and $\hat{\beta}_{(-i)}$ ($\mathbf{d}_{\mathbf{i}} = \hat{\beta} - \hat{\beta}_{(-i)}$).

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

1.2-5. MODEL DIAGNOSTICS: RESIDUALS

1.2-5.3 **Graphics for diagnostics**

- A scatterplot showing Standardized Pearson residuals (Y-axis) and *leverage* (h_{ii}, diagonal of **H**). Cut-offs can be included at 2p/n.
- A scatterplot showing Pearson residuals versus each of the predictors in turn.
- A scatterplot showing Pearson residuals against fitted values, however, residualPlots plots 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—particularly 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 correct model requires that the conditional mean function in any residual plot be constant as we move across the plot, and 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)
```

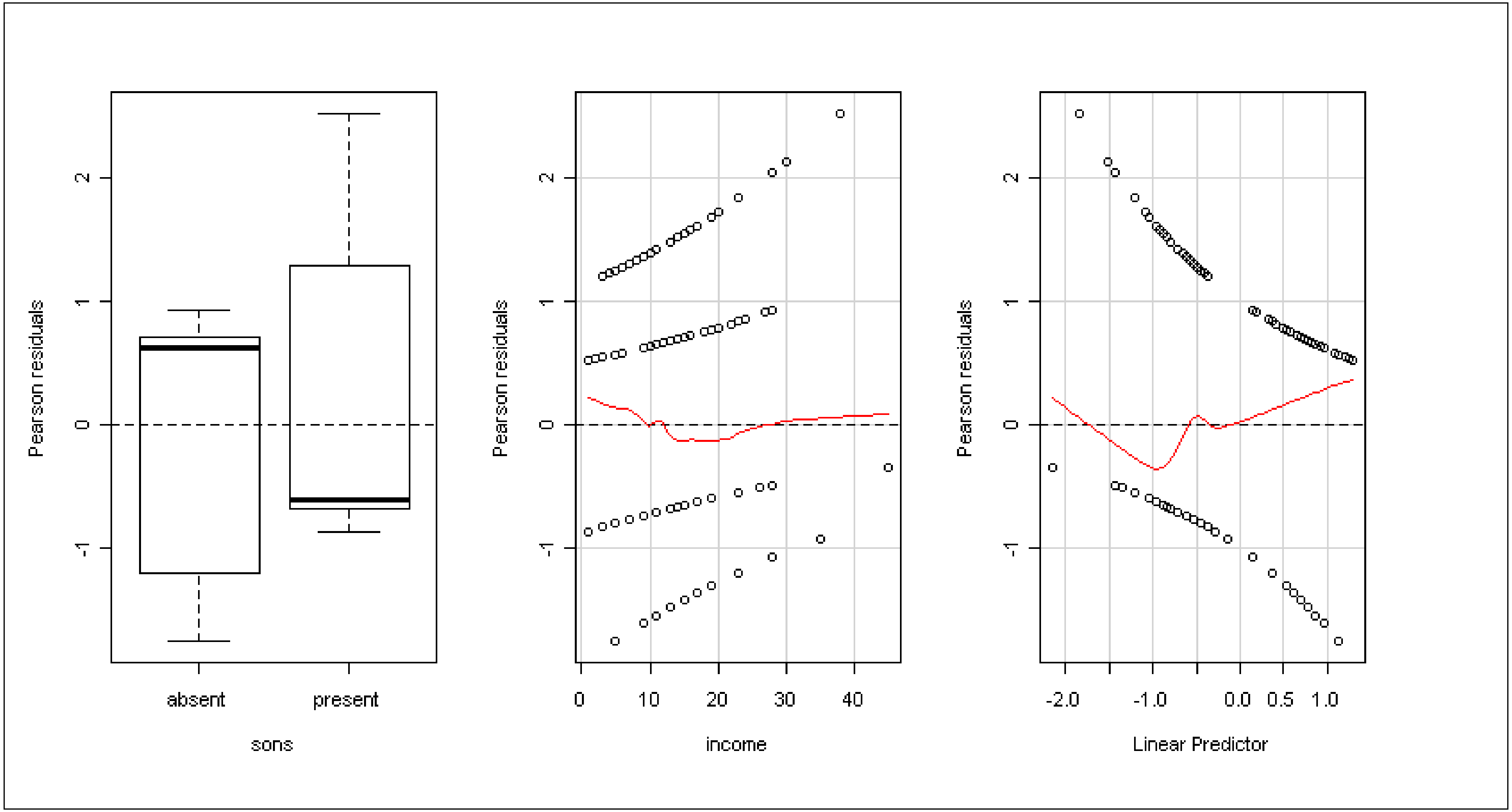
1.2-5. MODEL DIAGNOSTICS: RESIDUALS

Example of diagnostic plots for binary outcomes:

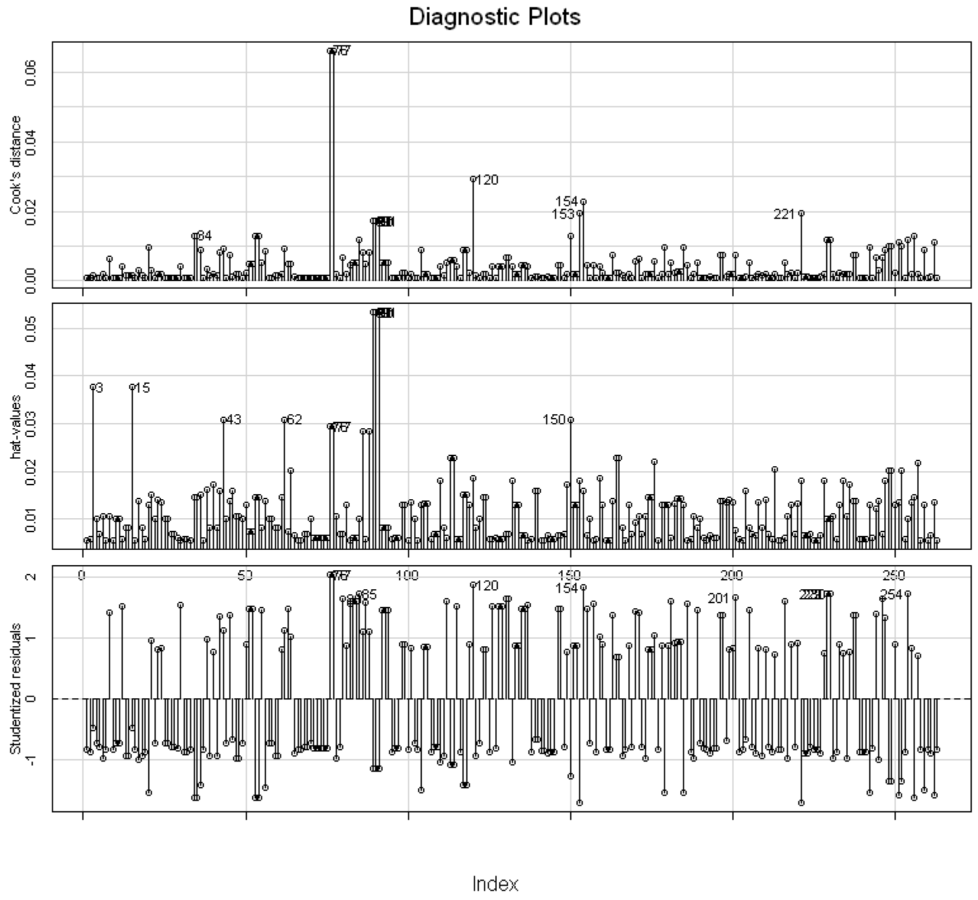
```
> options(contrasts=c("contr.treatment","contr.treatment"))
> bm3 <-glm( bwork~sons+income,family=binomial, data=womenlf )
> bm6 <-glm( bwork~sons*income, family=binomial, data=womenlf )
> anova(bm3,bm6,test='Chisq')
Analysis of Deviance Table

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) )
```

1.2-5. MODEL DIAGNOSTICS: RESIDUALS

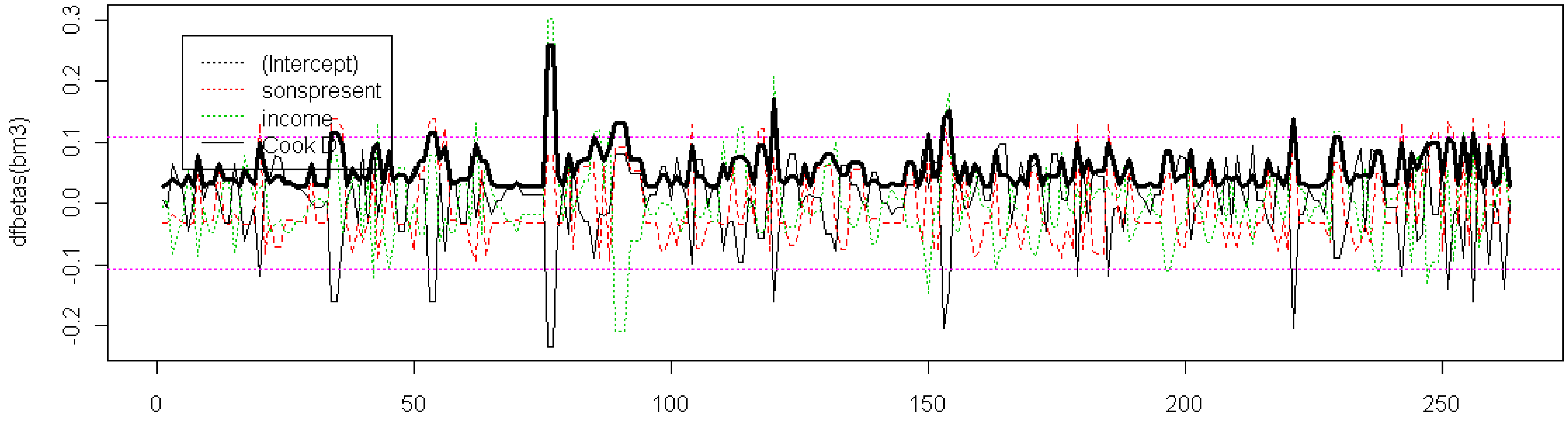


1.2-5. MODEL DIAGNOSTICS: RESIDUALS



1.2-5. MODEL DIAGNOSTICS: RESIDUALS

```
>
```



1.2-6. EXEMPLE 1: ACCIDENTS WITH INJURED PEOPLE DEPENDING ON 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)

genero	entorno	cinturon	gravedad	y
Mujer	Urbano	No	SinHeridos	7287
Mujer	Urbano	Si	SinHeridos	11587
Mujer	NoUrbano	No	SinHeridos	3246
Mujer	NoUrbano	Si	SinHeridos	6134
Hombre	Urbano	No	SinHeridos	10381
Hombre	Urbano	Si	SinHeridos	10969
Hombre	NoUrbano	No	SinHeridos	6123
Hombre	NoUrbano	Si	SinHeridos	6693
Mujer	Urbano	No	LeveSinHospital	175
Mujer	Urbano	Si	LeveSinHospital	126
Mujer	NoUrbano	No	LeveSinHospital	73
Mujer	NoUrbano	Si	LeveSinHospital	94
Hombre	Urbano	No	LeveSinHospital	136
Hombre	Urbano	Si	LeveSinHospital	83
Hombre	NoUrbano	No	LeveSinHospital	141
Hombre	NoUrbano	Si	LeveSinHospital	74
Mujer	Urbano	No	LeveConHospital	720
Mujer	Urbano	Si	LeveConHospital	577
Mujer	NoUrbano	No	LeveConHospital	710
Mujer	NoUrbano	Si	LeveConHospital	564
Hombre	Urbano	No	LeveConHospital	566
Hombre	Urbano	Si	LeveConHospital	259
Hombre	NoUrbano	No	LeveConHospital	710
Hombre	NoUrbano	Si	LeveConHospital	353
Mujer	Urbano	No	Hospitalización	91
Mujer	Urbano	Si	Hospitalización	48
Mujer	NoUrbano	No	Hospitalización	159
Mujer	NoUrbano	Si	Hospitalización	82
Hombre	Urbano	No	Hospitalización	96
Hombre	Urbano	Si	Hospitalización	37
Hombre	NoUrbano	No	Hospitalización	188
Hombre	NoUrbano	Si	Hospitalización	74
Mujer	Urbano	No	Mortal	10
Mujer	Urbano	Si	Mortal	8
Mujer	NoUrbano	No	Mortal	31
Mujer	NoUrbano	Si	Mortal	17
Hombre	Urbano	No	Mortal	14
Hombre	Urbano	Si	Mortal	1
Hombre	NoUrbano	No	Mortal	45
Hombre	NoUrbano	Si	Mortal	12

EXAMPLE 1

Models	logit(π_{ijk})	Deviance	n-p	AIC
1	η	1912.5	7	1981.2
SeatBelt - A	$\eta+\alpha_i$	1144.4	6	1215.1
Environment - C	$\eta+\beta_j$	1192.8	6	1263.5
Gender -D	$\eta+\gamma_k$	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(acc)
    genero        entorno     cinturon                 gravedad          y                f.heridos
 Hombre:20    NoUrbano:20   Si:20     Hospitalización:8   Min.   :    1.00   Sin: 8
 Mujer :20    Urbano   :20   No:20     LeveConHospital:8   1st Qu.:   66.75   Con:32
                                        LeveSinHospital:8   Median :  138.50
                                        Mortal         :8   Mean   : 1717.35
                                        SinHeridos     :8   3rd Qu.:  710.00
                                                            Max.   :11587.00
> tapply(acc$y,acc$f.heridos,sum);sum(acc$y)
  Sin   Con
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. The odds 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	**With Injured (positive outcome)**	**NoInjured**	**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:

$$\textbf{(M1)}\quad \log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta \qquad\qquad \textbf{(M2)}\quad \log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta + \alpha_\iota \quad i = 1,2 \quad \alpha_1 = 0$$

```
> dfc
    Seat-belt    m ypos  yneg
Yes        Si 37792 2409 35383
No         No 30902 3865 27037
>
> acc.m1 <-glm(cbind(ypos,yneg)~1, family=binomial(link=logit), data=dfc)
> summary(acc.m1)

Call:
glm(formula = cbind(ypos, yneg) ~ 1, family = binomial(link = logit),
    data = dfc)

Deviance Residuals:
    Si       No
-19.59   19.60

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
>
> acc.m2 <-glm(cbind(ypos,yneg)~seatbelt, family=binomial(link=logit), data=dfc)
> summary(acc.m2)

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

Deviance Residuals:
[1]  0  0

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.68702    0.02106 -127.61   <2e-16 ***
Seatbelt.No  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: -4.3099e-13  on 0  degrees of freedom
AIC: 23.523
> residuals(acc.m1,'pearson')
       Si        No
-18.61742  20.58856
> xpea<-sum(residuals(acc.m1,'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 Si: logit (3865/30902) -logit (2409/37792) = 0.742.

$$\boxed{\frac{\pi_i}{1-\pi_i} = \begin{cases} e^{\eta} & \iota = 1\ Yes \\ e^{\eta}e^{\alpha_2} & \iota = 2\ No \end{cases}} \quad odds-ratio\ No vs Yes = e^{\alpha_2} = 2.1$$

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' NonUrbanno 'for Factor C.

```
> df2
  cinturon  entorno     m ypos  yneg
1       Si NoUrbano 14097 1270 12827
2       No NoUrbano 11426 2057  9369
3       Si   Urbano 23695 1139 22556
4       No   Urbano 19476 1808 17668
```

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
> acc.m20 <-glm(cbind(ypos,yneg)~1, family=binomial(link=logit), data=df2)
> summary(acc.m20)

Call:
glm(formula = cbind(ypos, yneg) ~ 1, family = binomial(link = logit),
    data = df2)

Deviance Residuals:
       1          2          3          4
 -0.5131   29.4486  -25.2217    0.7247

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

> acc.m21 <-glm(cbind(ypos,yneg)~environment, family=binomial(link=logit), data=df2)
> summary(acc.m21)
```

EXAMPLE 1

```
Call:
glm(formula = cbind(ypos, yneg) ~ environment, family = binomial(link = logit),
    data = df2)

Deviance Residuals:
     1        2        3        4
-14.92   15.04  -12.97   12.94

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         -1.89784    0.01859 -102.08   <2e-16 ***
environment Urban    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

> acc.m22 <-glm(cbind(ypos,yneg)~seat-belt, family=binomial(link=logit), data=df2)
> summary(acc.m22)

Call:
glm(formula = cbind(ypos, yneg) ~ seat-belt, family = binomial(link = logit),
    data = df2)
```

EXAMPLE 1

```
Deviance Residuals:
     1        2        3        4
 12.10    16.82   -10.30   -14.17

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.68702    0.02106 -127.61   <2e-16 ***
seat-belt No   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

> acc.m23 <-glm(cbind(ypos,yneg)~cinturon+entorno, family=binomial(link=logit), data=df2)
> summary(acc.m23)

Call:
glm(formula = cbind(ypos, yneg) ~ seat-belt + environment, family = binomial(link = logit),
    data = df2)

Deviance Residuals:
      1         2         3         4
-0.8793   0.7358   0.9220  -0.7396
```

EXAMPLE 1

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         -2.28676    0.02465  -92.78   <2e-16 ***
Seat-belt No         0.75265    0.02734   27.53   <2e-16 ***
Environment Urban   -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

> xpea<-sum(residuals(acc.m21,'pearson')^2);xpea
[1] 787.0698
> xpea<-sum(residuals(acc.m22,'pearson')^2);xpea
[1] 761.8445
> xpea<-sum(residuals(acc.m20,'pearson')^2);xpea
[1] 1618.284
> xpea<-sum(residuals(acc.m23,'pearson')^2);xpea
[1] 2.712893
> 1-pchisq(xpea,1)
[1] 0.09954032
>
```

EXAMPLE 1

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

```
> summary(acc)
    genero        entorno    cinturon             gravedad           y              f.heridos    heridos     
 Hombre:20    Urbano  :20    Si:20    Hospitalización:8   Min.   :    1.00   Sin: 8    Min.   :  0.0  
 Mujer :20    NoUrbano:20    No:20    LeveConHospital:8   1st Qu.:   66.75   Con:32    1st Qu.:  9.5  
                                      LeveSinHospital:8   Median :  138.50             Median : 74.0  
                                      Mortal         :8   Mean   : 1717.35             Mean   :156.8  
                                      SinHeridos     :8   3rd Qu.:  710.00             3rd Qu.:163.0  
                                                          Max.   :11587.00             Max.   :720.0  
>
> df3
  cinturon  entorno genero     m ypos  yneg
1       Si   Urbano Hombre 11349  380 10969
2       No   Urbano Hombre 11193  812 10381
3       Si NoUrbano Hombre  7206  513  6693
4       No NoUrbano Hombre  7207 1084  6123
5       Si   Urbano  Mujer 12346  759 11587
6       No   Urbano  Mujer  8283  996  7287
7       Si NoUrbano  Mujer  6891  757  6134
8       No NoUrbano  Mujer  4219  973  3246
```

EXAMPLE 1

```
> summary(acc.m331)

Call:
glm(formula = cbind(ypos, yneg) ~ cinturon + entorno + genero,
    family = binomial(link = logit), data = df3)

Deviance Residuals:
      1        2        3        4        5        6        7        8
-0.5055  -0.7976   0.2133   0.9023   1.7426  -0.4639  -1.5365   0.3172

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -3.33639    0.03114 -107.14   <2e-16 ***
cinturonNo       0.81710    0.02765   29.55   <2e-16 ***
entornoNoUrbano  0.75806    0.02697   28.11   <2e-16 ***
generoMujer      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

Number of Fisher Scoring iterations: 3
```

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.

```
glm(formula = cbind(ypos, yneg) ~ cinturon * entorno + genero,     family = binomial, data = df3)
Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
 (Intercept)               -3.30342    0.03509 -94.149   <2e-16 ***
cinturonNo                  0.76173    0.03933  19.366   <2e-16 ***
entornoNoUrbano             0.69360    0.04239  16.362   <2e-16 ***
generoMujer                 0.54594    0.02729  20.007   <2e-16 ***
cinturonNo:entornoNoUrbano  0.10800    0.05486   1.968    0.049 *
```

EXAMPLE 1

```
> summary(acc.m331)

Call:
glm(formula = cbind(ypos, yneg) ~ cinturon + entorno + genero,
    family = binomial(link = logit), data = df3)

Deviance Residuals:
      1        2        3        4        5        6        7        8
-0.5055  -0.7976   0.2133   0.9023   1.7426  -0.4639  -1.5365   0.3172

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -3.33639    0.03114 -107.14   <2e-16 ***
cinturonNo       0.81710    0.02765   29.55   <2e-16 ***
entornoNoUrbano  0.75806    0.02697   28.11   <2e-16 ***
generoMujer      0.54483    0.02727   19.98   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    Null deviance: 1912.4532  on 7  degrees of freedom
Residual deviance:    7.4645  on 4  degrees of freedom
AIC: 82.167

Number of Fisher Scoring iterations: 3

> summary(acc.m332)

Call:
glm(formula = cbind(ypos, yneg) ~ cinturon + entorno * genero,
    family = binomial(link = logit), data = df3)
```

EXAMPLE 1

```
…
Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -3.36383    0.03519 -95.592   <2e-16 ***
cinturonNo                   0.81618    0.02765  29.521   <2e-16 ***
entornoNoUrbano              0.80907    0.04010  20.177   <2e-16 ***
generoMujer                  0.59306    0.03914  15.152   <2e-16 ***
entornoNoUrbano:generoMujer -0.09345    0.05422  -1.724   0.0848 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    Null deviance: 1912.4532  on 7  degrees of freedom
Residual deviance:    4.4909  on 3  degrees of freedom
AIC: 81.193

Number of Fisher Scoring iterations: 3

> summary(acc.m333)

Call:
glm(formula = cbind(ypos, yneg) ~ cinturon * entorno + genero,
    family = binomial(link = logit), data = df3)

…
Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                -3.30342    0.03509 -94.149   <2e-16 ***
cinturonNo                  0.76173    0.03933  19.366   <2e-16 ***
entornoNoUrbano             0.69360    0.04239  16.362   <2e-16 ***
generoMujer                 0.54594    0.02729  20.007   <2e-16 ***
cinturonNo:entornoNoUrbano  0.10800    0.05486   1.968    0.049 *
```

EXAMPLE 1

```
    Null deviance: 1912.4532  on 7  degrees of freedom
Residual deviance:    3.5914  on 3  degrees of freedom
AIC: 80.294

Number of Fisher Scoring iterations: 3

> summary(acc.m334)

Call:
glm(formula = cbind(ypos, yneg) ~ cinturon * genero + entorno, 
    family = binomial(link = logit), data = df3)

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)    
(Intercept)            -3.34236    0.03755 -89.014   <2e-16 ***
cinturonNo              0.82621    0.04220  19.579   <2e-16 ***
generoMujer             0.55459    0.04370  12.691   <2e-16 ***
entornoNoUrbano         0.75792    0.02698  28.096   <2e-16 ***
cinturonNo:generoMujer -0.01598    0.05586  -0.286    0.775    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    Null deviance: 1912.4532  on 7  degrees of freedom
Residual deviance:    7.3826  on 3  degrees of freedom
AIC: 84.085

Number of Fisher Scoring iterations: 3
```

EXAMPLE 1

```
> anova(acc.m331,acc.m332,test="Chisq")
Analysis of Deviance Table
Model 1: cbind(ypos, yneg) ~ cinturon + entorno + genero
Model 2: cbind(ypos, yneg) ~ cinturon + entorno * genero
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1         4     7.4645
2         3     4.4909  1   2.9736    0.0846
> anova(acc.m331,acc.m333,test="Chisq")
Analysis of Deviance Table

Model 1: cbind(ypos, yneg) ~ cinturon + entorno + genero
Model 2: cbind(ypos, yneg) ~ cinturon * entorno + genero
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1         4     7.4645
2         3     3.5914  1   3.8730    0.0491
> anova(acc.m331,acc.m334,test="Chisq")
Analysis of Deviance Table

Model 1: cbind(ypos, yneg) ~ cinturon + entorno + genero
Model 2: cbind(ypos, yneg) ~ cinturon * genero + entorno
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1         4     7.4645
2         3     7.3826  1   0.0818    0.7748
> xpea<-sum(residuals(acc.m332,'pearson')^2);xpea
[1] 4.496567
> 1-pchisq(xpea,3)
[1] 0.2125967
> xpea<-sum(residuals(acc.m333,'pearson')^2);xpea
[1] 3.580126
> 1-pchisq(xpea,3)
[1] 0.3105178
```

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	Interacción Cinturón-Entorno Y Cinturón-Género : $\eta+\alpha_i+\beta_j+\gamma_k+\alpha\beta_{ij}+\alpha\gamma_{jk}$
2	**A*D+C*D**	2	4.371979	**3.0467**	(M2) vs (M4)	1	Interacción Cinturón-Género Y Entorno-Género : $\eta+\alpha_i+\beta_j+\gamma_k+\alpha\gamma_{ik}+\beta\gamma_{jk}$
3	**A*C+C*D**	2	1.367022	0.04171	(M3) vs (M4)	1	Interacción Cinturón-Entorno Y Entorno-Género : $\eta+\alpha_i+\beta_j+\gamma_k+\alpha\beta_{ij}+\beta\gamma_{jk}$
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	Interacción Cinturón-Entorno Y Cinturón-Género : $\eta+\alpha_i+\beta_j+\gamma_k+\alpha\beta_{ij}+\alpha\gamma_{jk}$
2	**A*D+C*D**	2	4.371979	3.0925	(M2) vs (M4)	1	Interacción Cinturón-Género Y Entorno-Género : $\eta+\alpha_i+\beta_j+\gamma_k+\alpha\gamma_{ik}+\beta\gamma_{jk}$
3	**A*C+C*D**	2	1.367022	**6.0975**	(M3) vs (M4)	1	Interacción Cinturón-Entorno Y Entorno-Género : $\eta+\alpha_i+\beta_j+\gamma_k+\alpha\beta_{ij}+\beta\gamma_{jk}$
4	**A+C+D**	4	7.4645				$\boldsymbol{\eta+\alpha_i+\beta_j+\gamma_k}$

➡ 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.

```
> anova(acc.m333,acc.m43,test="Chisq")
Analysis of Deviance Table

Model 1: cbind(ypos, yneg) ~ cinturon * entorno + genero
Model 2: cbind(ypos, yneg) ~ cinturon * entorno + entorno * genero
  Resid. Df Resid. Dev Df Deviance P(>|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 end a summary table is given with the residual liability and the AIC for all the models that have been calculated.

EXAMPLE 1

```
> acc.res<-step(acc.m34)
Start:  AIC=82.7
cbind(ypos, yneg) ~ cinturon * genero * entorno

                          Df  Deviance    AIC
- cinturon:genero:entorno  1     1.325 82.028
<none>                       2.411e-12 82.702

Step:  AIC=82.03
cbind(ypos, yneg) ~ cinturon + genero + entorno + cinturon:genero + 
    cinturon:entorno + genero:entorno

                   Df Deviance    AIC
- cinturon:genero   1    1.367 80.069
<none>                   1.325 82.028
- genero:entorno    1    3.562 82.265
- cinturon:entorno  1    4.372 83.074

Step:  AIC=80.07
cbind(ypos, yneg) ~ cinturon + genero + entorno + cinturon:entorno + 
    genero:entorno

                   Df Deviance    AIC
<none>                   1.367 80.069
- genero:entorno    1    3.591 80.294
- cinturon:entorno  1    4.491 81.193
>
```