Binary response

Generalized Linear Models

Grau d'Estadística

09/11/2018

Outline

1. Generalized linear models. Introduction
2. Binomial models. Response and Explicative variables
3. Binomial models. Link functions
4. Binomial models. Interpretation of model parameters
5. Binomial models. Estimation of model parameters
6. Binomial models. Goodness of fit
    - Deviance
    - Wald
    - Pearson
    - Hosmer-Lemeshow
    - AIC/BIC
    - Residuals
7. Binomial models. Prediction
    - ROC curve
    - Confussion matrix
8. Example: Agresti (2002)

Generalized linear models. Components

- Generalized linear models are extensions of the classical multiple regression models.
- Let $y^T = (y_1, \ldots, y_n)$ be a vector of n components, sample of a random vector $Y^T = (Y_1, \ldots, Y_n)$, with:
  - Statistical independent components
  - Identically distributed components
  - Expected values equal to $\mu^T = (\mu_1, \ldots, \mu_n)$.
- Then, we can distinguish two components of the model:
  - **The random component** (Y) that belongs to the exponential family with one parameter distribution and jointly expected values $E[Y] = \mu$.
  - **The systematic component** (η) specifies a vector (linear predictor) that is a linear combination from a limited number of explicative variables $X = (X_1, \ldots, X_p)$. The vector of parameters $\beta^T = (\beta_1, \ldots, \beta_p)$ has to be estimated. In matrix notation:

$$\eta = X\beta$$

where η is *nx1*; **X** is *nxp*; and β is *px1*.

Generalized linear models. Link functions

- For each observation, the expected value μ is related to the linear predictor η, through the **link function**, notated as $g(.)$:

$$\eta = g(\mu) \to \mu = g^{-1}(\eta).$$

- Examples:
  - In ordinary least squares models (normal response), the *identity* link is used:

$$\eta = \mu$$

  - For binary data, several link functions are commonly used. We will see them in detail:
    - *logit*
    - *probit*
    - *complementary log-log*
    - *log-log*

Generalized linear models. Classification

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

Binomial Models. Response

- A Binomial Random Variable (RV) appears when each observation holds or does not hold a target characteristic. This RV can take values $Y = 1$ (*Yes*) or $Y = 0$ (*No*).
- The probability of success is notated by π:
  - $P(Y_k = 1) = \pi_k$: probability of positive response (*success*) for kth observation in the sample
  - $P(Y_k = 0) = 1 - \pi_k$: probability of negative response (*failure*) for kth observation in the sample
- Examples:
  - One might be interested in the choice between public (metro, bus, etc) or private modes (car, motorcycle, etc) for home to work trips. The response variable can be defined as $Y = 1$ (positive response or success, public modes), or $Y = 0$ (negative response or fail, private modes).
  - In a clinical study, it is common to consider the presence/absence of a disease as the primary endpoint

Binomial Models. Explicative variables

- Each individual in the sample is characterized by a set of covariates (income, age, etc) and factors (gender, grades, etc) that defines:

$$X_k^T = (x_1, \ldots, x_p)$$

- Explicative variables that will form the linear predictor $x_k^T = (x_1 \ldots x_p)$ 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.
- Example: in the public-private binary modal choice model, each commuter has several variables as income, gender, car availability, distance to local public transport, value of time, etc. . . .
- The *goal* is to study the relationship between the response $\boldsymbol{Y}$ and the explicative variables by modelling the **probability** of positive response: $\pi = \pi(x)$.

Binomial Models. Aggregated/Disaggregated data (I)

- In experimental designs, groups of observations are defined and each group receives a combination of experimental conditions (*factors*) shared by all the units in the group:
  - Factors are explicative variables
  - The *kth* experimental condition is modeled by a common set of values of all the explicative variables and thus apply to m_k individual units.
- The total number of units (N) in the sample is the sum of the size of the groups (n groups):

$$N = m_1 + \cdots + m_n.$$

- Each group or combination of experimental conditions defines a *covariate class*: individual units sharing the same values for explicative variables.

Binomial Models. Aggregated/Disaggregated data (II)

- Therea are two possible representations of the data:
  - **Disaggregated data**. The unit is each **observation**. Individual outcome (0 or 1) is detailed for each observation. **Bernoulli response**
  - **Aggregated data**. The unit is each **covariate class**. Each class is fully defined by the number of individuals (m_k) and the number of positive responses (y_k). **Binomial response**.
- It is only possible having aggregated data if all explicative variables are categorical.
- Some analysis methods are well suited for aggregated data, and perform badly when applied to disaggregated data, for example asymptotic approximations to normality:
  - Asymptotic approximations for *aggregated data* are based either on the asymptotic evolution of the **number of covariate classes** ($m \to \infty$) or **on the total number of individual units** ($N \to \infty$).
  - *Disaggregated data* is only suitable for asymptotic approximations based on the **total size**.

Binomial Models. Aggregated/Disaggregated data (III)

- **Disaggregated** data: the response is **dycothomic** for each observation (Bernoulli distribution).
- **Grouped** or **Aggregated** data: the response is the number of positive outcomes (Binomial distribution).

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

- Table shows an experiment with *dicothomic* factors:
  - *gender* (two levels, 1:*male* and 2:*female*)
  - *car availability* (1:*1* car or 2:$>$*1*)
- There are $N = 7$ individuals
- There are $n = 4 = 2x2$ covariate classes

Binomial Models. Aggregated/Disaggregated data (IV)

- **Aggregated data** implies more efficient and less memory consumption.
  - It simplifies significant effect detected at a glance.
- **Aggregated data** implies to lose the serial order.
  - If additional variables are present, only average values can be considered possibly leading to *ecological fallacy situations*.
- **Aggregated data** implies a binomial response variable, since sample observed positive responses are $y_1/m_1, \ldots, y_n/m_n$, being $0 \leq y_k \leq m_k$ the number of positive responses in *kth* covariate class which size is $m_k \rightarrow m = (m_1 \ldots m_n)$.
- For **disaggregated data**, each individual unit defines a binomial response for a group of size 1 and thus, $m = (1 \ldots 1)$.

Binomial Models. Aggregated/Disaggregated data (V)

- When factor levels define covariate classes, as in our example, a **contingency table** is a good representation for aggregated data
- Our convention is to place the response Y in columns:

	FACTOR C						
	$\mathbf{C}_1 = 1$			$\mathbf{C}_2 = 2$			
FACTOR A	**FACTOR B: Y**			**FACTOR B: Y**			**TOTAL**
	B_1 (Y=0)	B_2 (Y=1)	SUBTOTAL	B_1 (Y=0)	B_2 (Y=1)	SUBTOTAL	
$\mathbf{A}_1$ **= male**	1	1	2	1	2	3	**5**
$\mathbf{A}_2$ **= female**	1	0	1	0	1	1	**2**
SUBTOTAL	2	1		1	3		
TOTAL	**3**			**4**			**7**

Binomial Models. Probability and Distribution functions

- Let $Y \sim B(m, \pi)$ a **Binomial** RV for the number of positive responses in $\boldsymbol{m}$ independent trials of a Bernoulli process with a common probability π.
- **Probability function**:

$$p_Y(y) = P(Y = y) = \left(\begin{array}{c} m \\ y \end{array} \right) \pi^y (1 - \pi)^{m-y}$$

- **Distribution function**:
  - $F_Y(y) = 0$ $\qquad y < 0$
  - $F_Y(y) = \sum_{i=0}^{\lfloor y \rfloor} \left(\begin{array}{c} m \\ i \end{array} \right) \pi^i (1 - \pi)^{m-i}$ $\qquad 0 \leq y \leq m$
  - $F_Y(y) = 1$ $\qquad y > m$

- Indicators:

$$E[Y] = m \cdot \pi$$

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

Binomial Models. Link functions

- Remember that the goal consists on stablishing a functional relationship between the probability of a positive result π and the vector of explicative variables (factors or covariates):

$$x^T = (x_1 \dots x_p) \leftrightarrow \pi = \pi(x).$$

- 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:
- Problem: the linear predictor η might be any value in the real axis, but the probability of positive answer belongs to the open interval (0, 1).
- We need a **link function g(.) to relate the vector π with the linear predictor η:

$$\eta = g(\pi), \text{ where } \pi \text{ is a vector } (nx1)$$

- **Canonic link** for binomial data is the ***logit*** function:

$$\eta = \theta = \mathrm{logit}(\pi)$$

Binomial Models. Logit link

- The **Logit link** is the most frequently used for its esay interpretation. The **Logit link** (sometimes bad-called logistic link) is:

$$\eta = g_1(\pi) = \text{logit}(\pi) = \log(\tfrac{\pi}{1-\pi}).$$

- The distribution function for a standard logistic variable is:

$$\pi_1(\eta) = g_1^{-1}(\eta) = \tfrac{\exp(\eta)}{1+\exp(\eta)} = \tfrac{1}{1+\exp(-\eta)}$$

- The density function is $(g_1^{-1})'(\eta) = \frac{\exp(\eta)}{(1+\exp(\eta))^2}$
  - It has 0 mean (position parameter) and variance $\pi^2/3$ (scale parameter 1)
  - This is a continuous and symmetric variable, quite similar to Standard Normal distribution.

Binomial Models. Other link functions

- **Probit link**. This link is the inverse of the Standard Normal distribution, with position and scale parameters taken values 0 and 1, respectively:

$$\eta = g_2(\pi) = \Phi^{-1}(\pi) \rightarrow \pi_2(\eta) = g_2^{-1}(\eta) = \Phi(\eta).$$

- **Complementary Log-log**. This link is the inverse of the distribution function for the **minimum extreme value** or **Gompertz**, with position and scale parameters taken values 0 and 1, respectively:

$$\eta = g_3(\pi) = \log(\log(1/(1-\pi))) \rightarrow \pi_3(\eta) = g_3^{-1}(\eta) = 1 - \exp(-\exp(\eta)).$$

- **Log-log**. This link is the inverse of distribution of probability function for the **maximum extreme value** or **Gumbel law**, with position and scale parameters taken values 0 and 1, respectively:

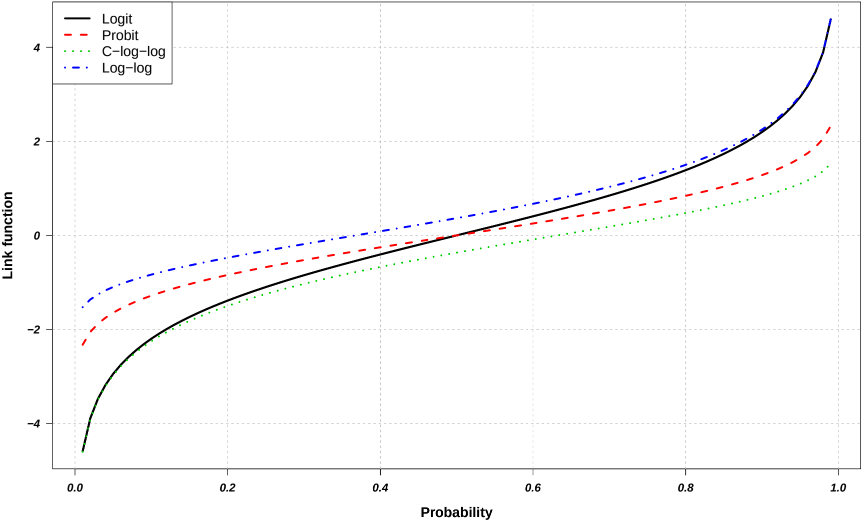
$$\eta = g_4(\pi) = -\log(\log(1/\pi)) \rightarrow \pi_4(\eta) = g_4^{-1}(\eta) = 1 - \exp(-\exp(-\eta))$$

- All link functions are related to well-known inverse of distribution functions for continuous RV.

Binomial Models. Link functions characteristics (I)

Probability	Odds	Log-odds	Probit	C-log-log	Log-log
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
0.50	1.0000	0.0000	0.0000	-0.36651	0.36651
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. Link functions characteristics (II)



Binomial Models. Link functions characteristics (III)

- All link functions are continuous and monotone increasing functions in the (0,1).
- **Logit** and **Probit** links can be seen related to changes in scales: They show an almost linear relationship in the 0.1 to 0.9 subinterval.
- Relantionship between **Log-log** and **C-log-log** link functions:

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

- For probabilities closed to 0, **Logit** and **C-log-log** functions are quite similar.
- For probabilities closed to 1, **C-log-log** trends slowly to 1 than **Logit** function does.
- For probabilities closed to 1, **Logit** and **Log-log** links are quite similar.

Binomial Models. Interpretation under logit link (I)

- The **odd** of an event represent the ratio of the *probability that the event will occur* and the *probability that the event will not occur* and the **log-odd** is its logarithm:

$$odd = \tfrac{\pi}{1-\pi} \rightarrow logodd = log\left(\tfrac{\pi}{1-\pi}\right).$$

- For instance, if the linear predictor has 2 covariates X_1 and X_2, then the **log-odd** of a *positive response* would be:

$$\eta = \log\left(\tfrac{\pi}{1-\pi}\right) = \begin{bmatrix} 1 & x_1 & x_2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = x^T\beta$$

- The **odd** for a *positive response* can be obtained as a function of η:

$$\tfrac{\pi}{1-\pi} = \exp(\eta) = \exp\left(x^T\beta\right) = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$$

- The **probability** of a *positive response* given a logit link is:

$$\pi = g_1^{-1}(\eta) = \tfrac{\exp(\eta)}{1+\exp(\eta)} = \tfrac{\exp\left(x^T\beta\right)}{1+\exp\left(x^T\beta\right)} = \tfrac{\exp(\beta_0+\beta_1 x_1+\beta_2 x_2)}{1+\exp(\beta_0+\beta_1 x_1+\beta_2 x_2)}$$

Binomial Models. Interpretation under logit link (II)

- The **probability** of a negative response is:

$$1-\pi = \frac{1}{1+\exp(\eta)} = \frac{1}{1+\exp(X\beta)} = \frac{1}{1+\exp(\beta_0+\beta_1 x_1+\beta_2 x_2)}$$

- Change in the reference level of a factor:
    - In the **log-odds** scale, a change of the reference level in a factor modeled through the dummy variable x_2 whose parameter is β_2 have an effect of a change in the sign of β_2.
    - In the **odd** scale, the effect would be multiplicative, a change of the reference level implies to multiply the odd obtained for positive response by $exp(\beta_2)$, leaving all the other variables in the same values.
    - In the **probability** scale, it is much more difficult since the effect depends not only of one factor, but on the rest of model variables and thus, only an approximation can be given:
        - We can use the partial derivative effect on the positive response π is $\frac{\partial \pi}{\partial x_2} = \pi\ (1-\pi)\,\beta_2$, that means a greater effect when the positive response probability π is closer to 0.5.

Binomial Models. Interpretation under logit link (III). Exercise

- Data
  - **bronch** chronical bronchial reaction, no = 0, yes = 1
  - **dust** dust concentration (mg/cm^3) at working place
  - **smoke** employee smoker?, no = 1, yes = 2

```
##   bronch dust smoke years
## 1      0 0.20     1     5
## 2      0 0.25     1     4
## 3      0 0.25     1     8
## 4      0 0.25     1     4
```

- Summary

```
##      bronch              dust           smoke        years      
##  Min.   :0.0000   Min.   : 0.2000   0:325   Min.   : 3.00  
##  1st Qu.:0.0000   1st Qu.: 0.4925   1:921   1st Qu.:16.00  
##  Median :0.0000   Median : 1.4050           Median :25.00  
##  Mean   :0.2343   Mean   : 2.8154           Mean   :25.06  
##  3rd Qu.:0.0000   3rd Qu.: 5.2475           3rd Qu.:33.00  
##  Max.   :1.0000   Max.   :24.0000           Max.   :66.00  
```

Binomial Models. Interpretation under logit link (IV). Exercise

```
summary(m <- glm(bronch ~ dust + smoke,dust, family=binomial))
```

```
##
## Call:
## glm(formula = bronch ~ dust + smoke, family = binomial, data = dust)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6601  -0.7281  -0.6939  -0.5144   2.0440
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.99330    0.17162 -11.614  < 2e-16 ***
## dust         0.10117    0.02295   4.408 1.04e-05 ***
## smoke1       0.65265    0.17111   3.814 0.000137 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1356.8  on 1245  degrees of freedom
## Residual deviance: 1321.8  on 1243  degrees of freedom
## AIC: 1327.8
##
## Number of Fisher Scoring iterations: 4
```

Binomial Models. Interpretation under logit link (V). Exercise

- What is the logodd of a smoker person with 2.5 mg/cm^3 of dust if his/her working place?
- And the odd?
- And the probability?
- And the Odds Ratio for the smoker factor?

Binomial Models. Parameter estimation (I)

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

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

- The iterative process to compute the estimates is based on a second order **Newton-type** method, specialized to the properties of the log likelihood function. The method converges fast, but it is not globally convergent.
- The quality of the **initialization point** 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.
- **Existence** and **unicity** is guaranteed for estimates under any of the presented link functions if $0 < y_i < m_i$ for any covariate class/observations.

Binomial Models. Parameter estimation (II)

- There are some examples where convergence of the estimates does not hold.
- When some values for variables perfectly separate positive from negative responses, then observations have a null probability of positive response equal to 0 or equal to 1: $y_i = 0$ or $y_i = m_i$. This is called the **separation** or the **quasi-separation** problem.
- Estimates for such parameters β do not converge, but fitted values $\hat{\pi}$ and deviance tend to a limiting value. For example:

$$\textit{if} \quad x_i = 1 \Rightarrow \pi_i = 1 \;\Rightarrow \beta_2 \to \infty \;\Rightarrow \pi_i = \frac{\exp(\beta_1+\beta_2 x_i)}{1+\exp(\beta_1+\beta_2 x_i)} \to 1$$

or

$$\textit{if} \quad x_i = 0 \Rightarrow \pi_i = 0 \;\Rightarrow \beta_2 \to -\infty \;\Rightarrow \pi_i = \frac{\exp(\beta_1+\beta_2 x_i)}{1+\exp(\beta_1+\beta_2 x_i)} \to 0$$

Binomial Models. Parameter estimation (III). Separation problem

Heinze et al. (2002) studied several solutions for the separation problem:

1. **Remove the variable** that cause the problem from the model. This solution is not recommended because, usually, these variables are strong related with the response.
2. Changing to a **diferent type of model**. Other models whose parameters have different interpretations not risk-related may be less appealing. This option is not recommended.
3. Use of an **ad hoc adjustment**. Clogg et al. (1991) proposed to artificially increase the observed frequenciesin such a way that the estimates were not biased.
4. **Exact logistic regression**. It permits replacement of the unsuitable maximum likelihood estimate by a median unbiased estimate. However, it has problems with a single continuous covariate factor or with multiple dichotomous factors.
5. Standard analysis with conflictive β **set to a 'high' value** (for example, the value of β of that iteration at which the log-likelihood changed by less than 10^{-6}). It implies extreme inflation of var(β) and leads to insignificant Wald test that may not be plausible for a very strong effect.
6. **Penalized logistic regression**. Proposed by Firth (1993), it is based on a modification of the score function of logistic regression with the aim of reducing the bias of maximum likelihood estimates since these they are biased away from zero and the infinite parameter estimates in situations of separation can be interpreted as an extreme consequence of this property. This would be the best solution.

Binomial Models. Goodness of fit.

- We will study the following methods to assess the goodness of fit of a model:
  - Statistics:
    - Deviance
    - Wald
    - Pearson
    - Hosmer-Lemeshow.
  - Graphical tools:
    - Calibration plot
    - Residual plots
- We will study the following methods to compare models:
  - Nested models:
    - Deviance
    - Wald
  - Any model:
    - AIC
    - BIC

Binomial Models. Goodness of fit. Deviance statistic (I)

- Let $\hat{\beta}$ the estimates of the model parameters, thus a linear predictor for each observation i might be computed $\hat{\eta}_i = x_i^T \hat{\beta}$ and thus through the response function (inverse of the selected link function) fitted values can be computed:

$$\hat{\pi}_i = g^{-1}\left(\hat{\eta}_i\right) = \frac{exp\left(x_i^T \hat{\beta}\right)}{1+exp\left(x_i^T \hat{\beta}\right)}$$

- The scaled deviance is:

$$D'\left(y, \hat{\mu}\right) = 2\ell(y, y) - 2\ell\left(\hat{\mu}, y\right)$$

- The deviance under the binomial distribution is identical to scaled deviance since $\phi = 1$:

$$D\left(y,\, \hat{\mu}\right) = D'\left(y,\, \hat{\mu}\right) \cdot \phi = D'\left(y,\, \hat{\mu}\right) \quad \text{if} \quad Y_i \sim B\left(m_i, \pi_i\right)$$

- With aggregated data, in the saturated model $\ell(y, y)$, fitted probabilities are equal to observed probabilities:

$$\hat{\pi}_i = \frac{y_i}{m_i} \quad i = 1, \ldots, n,$$

Binomial Models. Goodness of fit. Deviance statistic (II)

- In the binomial model, the specific expression for the deviance is:

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

- And it can be rewritten (it is the same formula) as:

$$D = 2\sum_{pos,neg}\sum_{i=1}^{n} o_i \cdot \log\frac{o_i}{e_i}$$

where:

$o_i = y_i \rightarrow$ **Observed** values for **positive** response for observation i.

$o_i = m_i - y_i \rightarrow$ **Observed** values for **negative** response for observation i.

$e_i = m_i\hat{\pi}_i \rightarrow$ **Expected positive** responses for observation i.

$e_i = m_i - m_i\hat{\pi}_i \rightarrow$ **Expected negative** responses for observation i.

Binomial Models. Goodness of fit. Deviance statistic (III)

- Assimptotic distribution of the deviance statistic for the model (M) under H_0 with p parameters is:

$$D_M = D\left(Y, \hat{\pi}\right) \sim \chi^2_{n-p} \qquad \textit{Remark: not to be confused with } \chi^2_{N-p}$$

- Thus, a goodness of fit test can be formulated as:

$$\begin{cases} H_0 : \textit{The current model fits properly the data} \\ H_1 : \textit{The current model does not fit properly the data} \end{cases}$$

- The **p value** for the test is $P\left(\chi^2_{n-p} > D_M\right)$
  - If *pvalue*$<<$0.05 then **there is evidence to reject** H_0 and thus, the model (M) does not fit properly data. There is an statistical evidence of discrepancy between observations and fitted values provided by model (M).
  - If *pvalue*$>>$0.05 then **there is not evidence to reject** H_0 and thus, leading to the conclusion that there is not evidence that the model (M) does not fit properly data, since discrepance between observed and fitted values is not significative in statistical terms.
- Deviance residual calculation in R can be performed with the next options:

```
sum(resid(model,'deviance')^2)  # D_m: option 1
model$deviance                  # D_m: option 2
```

Binomial Models. Deviance statistic (IV) for comparing nested models

- Let M_A be a model with q parameters nested in model M_B with $p > q$ parameters and $\hat{\pi}_A$ and $\hat{\pi}_B$ the fitted probabilities for both models. Then, the parameters for M_B are those common to M_A and those specific, i.e., $\beta_B^T = \left(\beta_1^T, \beta_2^T\right)$ and $\beta_A^T = \beta_1^T$.
- Following McCullagh (1989), the test for GLMZ equivalent to classical F Test in linear regression compares the scaled deviances between 2 hierarchical (nested) models through their difference:

$$\Delta D_{AB} = D\left(y, \hat{\pi}_A\right) - D\left(y, \hat{\pi}_B\right) = 2\ell\left(\hat{\pi}_B, y\right) - 2\,\ell\left(\hat{\pi}_A, y\right) \sim \chi^2_{p-q}.$$

- For testing $H_0 : \beta_2 = 0 \rightarrow M_B$ does not provide additional information:

$$P\left(\chi^2_{p-q} > \Delta D_{AB}\right) \rightarrow \begin{cases} << \alpha & H_0 \textit{ Rejected} \\ >> \alpha & H_0 \textit{ Not rejected} \end{cases}$$

- It is a contrast for multiple coefficients: **large values indicate non-equivalence of models**

Binomial Models. Goodness of fit. Wald test (I)

- *t-Student* test for a parameter in linear regression has the equivalence in ***Wald test*** for binomial models:

$$H_0 : \beta_j = \tilde{\beta}_j \to Z_0 = \tfrac{\tilde{\beta}_j - \hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim N(0,1) \text{ if } H_0 \text{ is true.}$$

- Wald Test is natural in a ML estimation framework since asymptotically:

$$\hat{\beta} - \beta \sim N_p \left(0,\, \Im^{-1}\right)$$

where $\Im$ is the Fisher Expected Information Matrix (score variance) approximated by the last iteration (convergence) of Score Method: $X^T W X$.

- Then, Wald (W) statistic is:

$$W = \left(\hat{\beta} - \beta\right)^T \Im \left(\hat{\beta} - \beta\right) \sim \chi^2_p$$

- Asymptotic bilateral confidence interval at level $100(1-\alpha)\%$ for a parameter is: $\hat{\beta}_j \pm z_{1-\alpha/2}\hat{\sigma}_{\hat{\beta}_j}$ where $z_{1-\alpha/2}$ is the quantile $(1-\alpha/2)$ from a $N(0,1)$ distribution.

Binomial Models. Goodness of fit. Wald test (II)

- Multiple hypothesis $H_0 : \beta = \beta_0$ testing is also possible with Wald Test

$$W = \left(\hat{\beta} - \beta_0\right)^T V\left[\hat{\beta}\right]^{-1}\left(\hat{\beta} - \beta_0\right) \sim \chi^2_p$$

- If $dim(\beta) = 1$ then we have the same expression of previous slide:

$$H_0 : \beta = 0 \rightarrow Z = \frac{\hat{\beta}}{\sqrt{V[\hat{\beta}]}} \sim N\left(0,\ 1\right).$$

- Deviance for a GLM plays a role similar to Residual Sum of Squares in classical regression. Therefore, it is possible to define a Generalized R^2, or *pseudo*-R^2:

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

$$0 \leq R^2 \leq 1$$

```
library(lmtest)
   anova(modelA, modelB, ..., test = "Chisq") # Deviance Test
waldtest(modelA, modelB, ..., test = "Chisq") # Wald Test
```

Binomial Models. Goodness of fit. Pearson statistic

- The Generalized Pearson statistic (X^2) is asymptotically distributed as:

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

- When disaggregated data is used ($m_i = 1$) asymptotic theory does not apply. As a rule of thumb, at a fully disaggregated level, a model is good if Deviance or Pearson X^2 statistics are similar to degrees of freedom of the model. Remember: **Large values indicate a lack of fit of the model**.
- How to compute in R Pearson X^2 statistic as sum of the squares of Pearson's residuals:

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

Binomial Models. Goodness of fit. Hosmer Lemeshow statistic

- The *Hosmer-Lemeshow Statistic* (1980,1989) is another measure of goodness of fit.
- Procedure to calculate the statistic:
  - Partitioning the observations into ***G*** (e.g, 10) **equal sized** groups according to their predicted probabilities
    - Remark: do **not** choose equispaced probabilities, i.e 0–0.1, 0.1–0.2, . . . , 0.9–1.
    - Remark: it is possible to perform the test with fewer than 10 groups with small sample size.
  - For each group, the observed number of responses (positive, negative) in the sample and the expected number of responses (positive, negative) are compared by calculating the Pearson Statistic from the 2xG table of observed and expected frequencies.
  - The statistic is defined as follows and distributed as a χ^2 with G-2 degrees of freedom:

$$X_{HL}^2 = \sum_{g=1}^{G} \frac{(o_g - m_g \hat{\pi}_g)^2}{m_g \hat{\pi}_g (1 - \hat{\pi}_g)} \sim \chi^2_{G-2}$$

- Hosmer & Lemeshow later realized that their proposed statistic was not suitable for assessing the goodness of fit.

Binomial Models. Compare unnested models. AIC and BIC.

- **AIC** (Akaike Information Criteria), proposed by Akaike (1974), is defined as a trade-off between a goodness of fit provided by a model (M) and the number of parameters p (as an indicator for model complexity). Let M be a model with p parameters:

$$AIC_M = 2 \cdot (p - \ell\,(\hat{\pi}_M, y))$$

- In order to take into account the sample size, **BIC** (*Bayesian Information Criteria*) is proposed by Schwartz (1978):

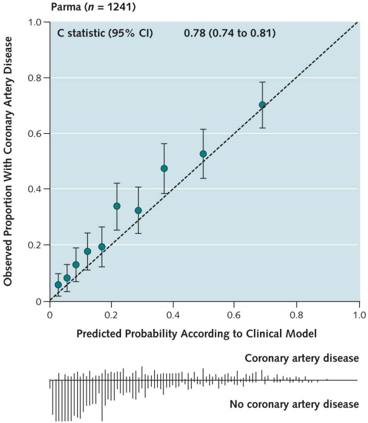
$$BIC_M = p \log(n) - 2\,\ell\,(\hat{\pi}_M, y)$$

- Models with minimum AIC and BIC are preferred
- AIC and BIC might be used to compare **nested** or **unnested** models.

```
AIC(model,k=2)                # AIC
AIC(model,k=log(nrow(dataset)))  # BIC
```

Binomial Models. Goodness of fit. Calibration plot.

- **Calibration plot** is a graphical tool to assess the model goodness of fit.
- It represents the observed probabilities (with their 95% CI) as function of the predicted probabilities.
- If most of the confidence intervals crossess the identity line, the model is validated.



Binomial Models. Model diagnostic. Residuals

- This section shows an extension to Generalized Linear Models of Normal regression methods for residual analysis.
- According to Pregibon (1981), Williams (1987) and Fox (2008) a residual can be defined for each covariate class *i* as:

$$e_i = y_i - \hat{y}_i = y_i - m_i \hat{\pi}_i$$

- However, the response residuals are not used in diagnostics of GzLM because they ignore the non-constant variance that is a paramount issue.
- **Pearson residuals** are casewise components of the Pearson statistic:

$$e_i^P = \frac{(y_i - \hat{\mu}_i)}{\sqrt{V[\hat{\mu}_i]\hat{\phi}}} \rightarrow X^2 = \sum_{i=1}^{n} \left(e_i^P\right)^2$$

- These set of residuals for GzLM have a direct analogy to linear models.
- For a model *M*, the next R command returns the Pearson residuals.

```
residuals(M, type="pearson")
```

- To compute the **standardized Pearson residuals**, we need to define the hat-values h_{ii} for GzLMs. They are corrected for conditional response variation and for the leverage are:

$$e_i^{PS} = \frac{(y_i - \hat{\mu}_i)}{\sqrt{V[\hat{\mu}_i](1 - h_{ii})}}$$

Binomial Models. Model diagnostic. Residuals

- **Deviance residuals** are the square roots of the casewise components of the residual deviance, attaching the sign of $y_i - \hat{\mu}_i$:

$$e_i^D = sign\ \left(y_i - \hat{\mu}_i\right)\sqrt{d_i} \rightarrow D\left(y,\, \hat{\mu}\right) = \sum_{i=1}^{n}\left(e_i^D\right)^2.$$

- In the linear model, the deviance residuals reduce to the Pearson residuals.
- The deviance residuals are often the preferred form of residual for GzLMs. The R command to obtain them from the model M is:

```r
residuals(M, type="deviance")
```

- **Standardized deviance residuals** are:

$$e_i^{DS} = \frac{e_i^D}{\sqrt{\hat{\phi}(1-h_{ii})}}$$

- **Studentized deviance residuals** are approximations to the scaled difference between the response and the fitted value computed without case i:

$$e_i^{Stu} = sign\left(y_i - \hat{\mu}_i\right)\sqrt{\left(1-h_{ii}\right)\left(e_i^{DS}\right)^2 + h_{ii}\left(e_i^{PS}\right)^2}.$$

Binomial Models. Model diagnostic. Residuals

- The following functions, some in base R and some in the *car* package, have methods for GzLMs:
  - ***rstudent***. Student residuals.
  - ***hatvalues***. Compute the h_{ii} values (leverage).
  - ***cooks.distance***. Measure of influence.
  - ***dfbetas***. Compute some of the regression (leave-one-out deletion) diagnostics for GzLMs.
  - ***outlierTest***. To detect outliers.
  - ***avPlots***. Construct added-variable, also called partial-regression, plots for GzLMs.
  - ***residualPlots***. Plots the residuals versus each term in a mean function and versus fitted values
  - ***marginalModelPlots***. Plot of the response (Y-axis) versus a linear combination *u* of regressors in the mean function on the horizontal axis
  - ***crPlots***. Construct partial-residual plots for GzLM.

Binomial Models. Model diagnostic. A priori influece data.

- **Hat matrix** (H) for GzLM can be defined, although **it depends on Y** (through W) and **X's values**:

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

- **H** is symmetric with diagonal values (h_{ii}) between 0 and 1 named leverages.
- The average value of h_{ii} is p/n.
- Cut-off for considering an observation with a priori influence is $2p/n$.
- This matrix is obtained through the last iteration (convergence) of the Iterative Weighted Least Squares (*IWLS*) for estimating model parameters.
- The h_{ii} have the usual interpretation, except that, unlike in a linear model, the hat-values in a GzLM depend on Y as well as on the configuration of the Xs.

Binomial Models. Model diagnostic. A posteriori influece data.

- **(Posterior) Influence data** are detected by an adapted **Cook's statitstic** derived from the Wald statistic for multiple hypothesis testing:

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

- Let Wald statistic be for observation i ($Z_{(-i)}^2$) for testing $H_0 : \beta = \hat{\beta}_{(-i)}$, the distance between $\hat{\beta}$ and $\hat{\beta}_{(-i)}$

$$d_i = \hat{\beta} - \hat{\beta}_{(-i)}$$

- Then, the **Cook's statistic** is:

$$Z_{(-i)}^2 = \left(\hat{\beta} - \hat{\beta}_{(-i)}\right)^T X^T W X \left(\hat{\beta} - \hat{\beta}_{(-i)}\right) = \frac{\left(e_i^{PS}\right)^2 h_{ii}}{p(1-h_{ii})}$$

Binomial Models. Model diagnostic. Plots

- Scatterplot: Standardized Pearson Residuals (Y-axis) versus *Leverage* (h_{ii}) (X-axis).
- Scatterplot: Pearson residuals (Y-axis) versus each of the predictors in turn (X-axis).
- Scatterplot: Pearson residuals (Y-axis) versus fitted values ($\eta(x)$)(X-axis) in the linear predictor scale. R function: *residualPlots*
- Examine leverage for each individual observation.
- Examine Cook's distance for each individual observation.
- Remarks
  - In binary regression for **disaggregated data**, the plots of Pearson or Deviance residuals are strongly patterned – particularly the plot against the linear predictor – where the residuals can take 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 to see this smoothers help in the purpose. R function *residualPlots* includes a smooth fit in each panel in the graph by default.
  - A lack-of-fit test is provided only for the numeric predictor.

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

Binomial Models. Prediction. Confusion matrix

- **Confusion matrix** for a binary model (M) shows predicted response versus observed response (positive/negative outcomes)
- To dichotomize the predicted probabilities, a threshold s $(0 < s < 1)$ have to be defined.
- Let the prediction in response be $\hat{y}_i = 1$ if $\hat{\pi}_i > s$ or 0, otherwise. 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**

Binomial Models. Prediction. Measures

- From the confusion matrix, several measures of predictive capability might be taked into account.
  - **Sensibility (Sens)**. Proportion of predicted as positive ($\hat{y}_i = 1$) within observed positive outcomes ($Y = 1$): $Sens = a/(a+c)$.
  - **Specificity (Sp)**. Proportion of predicted as negative ($\hat{y}_i = 0$) within observed negative outcomes ($Y = 0$): $Sp = d/(b+d)$.
  - **Positive predictive value (PPV)**. Proportion of observed positive outcomes ($Y = 1$) within those ones predicted as positive ($\hat{y}_i = 1$): $PPV = a/(a+b)$.
  - **Negative predictive value (NPV)**. Proportion of observed negative outcomes ($Y = 0$) within those ones predicted as negative ($\hat{y}_i = 0$): $NPV = d/(c+d)$.
  - **Positive Likelihood Ratio (PLR)**. It represents, in a observation predicted as positive, how much more likely a positive response is respect to a negative response: $PLR = PPV/(1-NPV)$
  - **Negative Likelihood Ratio (NLR)**. It represents, in a observation predicted as negative, how much more likely a negative response is respect to a positive response: $NLR = NPV/(1-PPV)$
- **Sens** and **Sp** depend on the model, but not on the data. On the other hand, **VPN** and **NPV** depend on the proportion of positive/negative responses in the population.
- To interpret the results is more useful the predictive values because they provide the probability of an specific response given the prediction of the model (that is the thing we know)

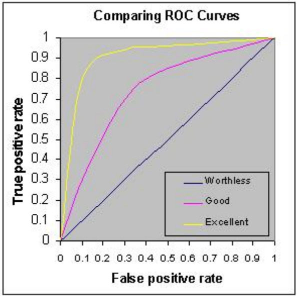
Binomial Models. Prediction. Measures. Exercise

- Calculate ***Sens, Sp, PPV, NPV, PLR, NLR*** based on the data of the table and considering a cutpoint of 0.4.

Predicted Probability	**Response**
0.08	0
0.27	0
0.34	1
0.36	0
0.44	0
0.52	0
0.53	1
0.54	0
0.80	1
0.96	1

Binomial Models. Prediction. ROC curve (I)

- ROC (Receiver Operating Characteristic) curve analysis has been widely accepted as the standard for describing and comparing the accuracy of predictions.
- For each threshold *s*, ROC Curve shows **1-Sp (false negative rate)** in the *X* axis and **Sens (true positive rate)** in *Y* axis.
- If the ROC curve rises rapidly towards the upper right-hand corner of the graph, or if the value of the Area Under the Curve (AUC) is large, the model performs well
  - If the AUC is close to 1.0, it indicates that the model is good
  - If the AUC is close to 0.5, it shows that the model is bad.



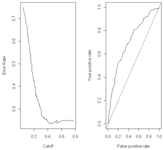
Binomial Models. Prediction. ROC curve (II)

- Interpretation:
  - The point (0,1) is the perfect classifier: it classifies all positive cases and negative cases correctly, i.e., 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.
  - ***Website*** to understand ROC curves.
  - ***Video*** to understand ROC curves.
- A guideline for interpreting ROC curve is:
  - 0.90 - 1 = excellent(A)
  - 0.80 - 0.90 = very good (B)
  - 0.70 - 0.80 = good (C)
  - 0.60 - 0.70 = bad (D)
  - 0.50 - 0.60 = very bad (F)
- Package *rms* contains the specific method *lrm(.)* for logistic regression with additional diagnostics:
  - C statistic (equivalent to AUC)
  - Naglekerke R^2. It is a $pseudo - R^2$. *NagelkerkeR2* is also in the *fmsb* package. Different expressions for $pseuro - R^2$ can be found ***here***.

Binomial Models. Prediction. ROC curve

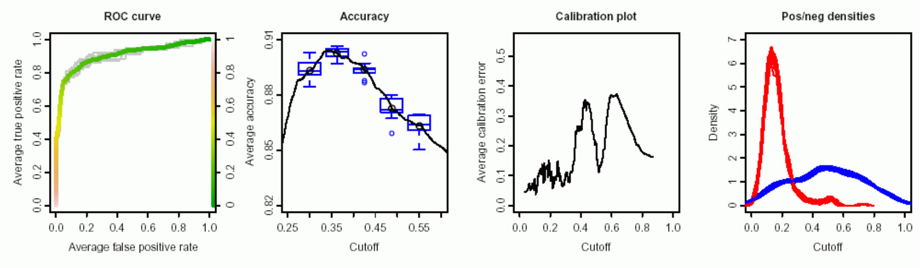
- AUC represents the proportion of couple of individuals (positive-negative response) with higher predicted probability in the individual with positive response.
- ROC curves and other performance plots are available in package *ROCR*

```
library("ROCR")
dadesroc <- prediction(predict(model,type="response"),dades$resposta)
par(mfrow=c(1,2))
plot(performance(dadesroc,"err"))          # Error rate
plot(performance(dadesroc,"tpr","fpr"))    # ROC curve
abline(0,1,lty=2)
```



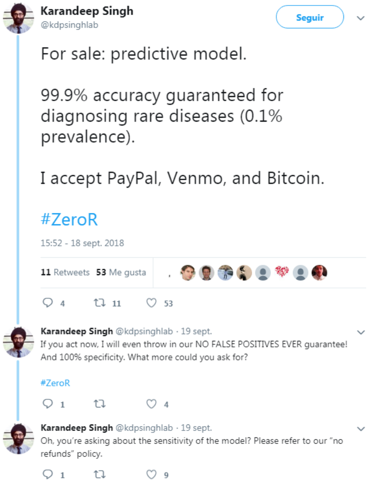
Binomial Models. Prediction. ROC curve (III)

- Here, there are several examples of performance plots provided by the *ROCR* package.
  - ROC curve
  - Accuracy
  - Calibration plot
  - Positive/Negative densities



Binomial models. Prediction.Exercise (I)

- Comment this tweet:



Karandeep Singh
@kdpsinghlab

For sale: predictive model.

99.9% accuracy guaranteed for diagnosing rare diseases (0.1% prevalence).

I accept PayPal, Venmo, and Bitcoin.

#ZeroR

15:52 - 18 sept. 2018

Karandeep Singh @kdpsinghlab · 19 sept.
If you act now, I will even throw in our NO FALSE POSITIVES EVER guarantee! And 100% specificity. What more could you ask for?

#ZeroR

Karandeep Singh @kdpsinghlab · 19 sept.
Oh, you're asking about the sensitivity of the model? Please refer to our "no refunds" policy.

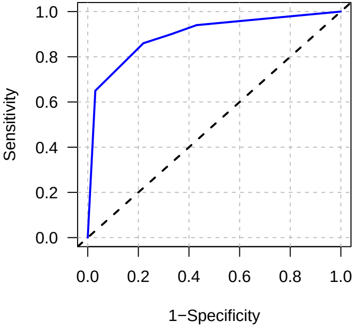
Binomial models. Prediction.Exercise (II)

- Build the ROC curve associated to this data:

	Non Disease		***Disease***	
cutpoint	***Test +***	***Test -***	***Test +***	***Test -***
1	25	33	48	3
2	19	39	46	5
3	23	45	44	7
4	2	56	33	18

Binomial models. Prediction. Exercise (II). Solution

cutpoint	*1-Specifity*	*Sensitivity*
-	1	1
1	0.43	0.94
2	0.33	0.90
3	0.22	0.86
4	0.03	0.65
-	0	0



Binomial models. Prediction.

- We can obtain the predictions in terms of the linear predictor or in terms of probabilities with the function *predict*.

```
predict(model, type='link')       # Linear predictor for observations to fit
predict(model, type='response')  # Probabilities for observations to fit th
predict(model, type='terms')      # terms (contribution of each term to link

predict(model,newdata, type='link')       # Linear predictor for new observa
predict(model,newdata, type='response')  # Probabilities for new observatio
predict(model,newdata, type='terms')      # terms (contribution of each term
```

Prediction. Samples.

There are several methods to assess the predictive capacity of a model

1. **Apparent performance**. A performance indicator is used that is obtained directly from the data used to estimate the model.
   - Advantage: very simple
   - Drawback: when we fit a statistical model to data, we find the parameters so that the fit is as good as possible and the errors are as small as possible and this gives an optimistic estimate (biased predictive capacity).
2. **Split the sample**. The sample is ¿randomly? divided into 2 sub-samples: training sample (construction of the model) and test sample (to obtain the performance measure)
   - Advantages:
     - Simple
     - Avoids bias
   - Drawback:
     - Inefficient since not all the data is used for fitting
     - It provides performance measures biased upwards if sample is splited at random

Prediction. Samples.

3. **Cross-validation**. The sample is randomly divided into k (e.g., k $=$ 10) groups of the same size and in each iteration are considered k-1 sub-samples as training samples and 1 as test sample. The performance in one iteration is obtained as the average of the k-1 yields and the overall yield is the average of all these.
    - Advantage: unlike the previous method, there are more chances of finding "particular" data partitions
    - Drawback: it requires programming. However, there are several functions in R implementing this method.
4. **Leave-out-one cross validation**. Only one single observation is used as test in each iteration.
    - Advantage: Useful with small datasets
    - Drawback: Probability too dependent on the response of the observation that is used as a test
5. **Bootstrap**. The original data is re-sampled R times taking N observations with replacement and a model is obtained in each iteration. The performance of this model is obtained from the re-sampled data and from the originals.
    - Advantage: provides a system for estimating overfit by comparing the two performance measures.
    - Drawback: it requires programming. However, there are several functions in R implementing this method.

Prediction. Samples. Split the data (observations)

- As in general linear model, one can attempt to *validate* a model built using one dataset by finding a second independent dataset and checking how well the second dataset outcomes are predicted from the model built using the first data set.
- If we randomly split our original data, little is gained because by definition, the two halves must have the same model. Any lack of fit is then just by chance, and any evidence for good fit brings no new information. Thus, it is better using all the data to build the best model.
- The best choice is to split the data **not at random**, but taken into account some relevant factor related with time or space in order to make the results more reproducible. Then, at least, one can check how well the first model predicts observations from the second model:
  - If it does fit, there is some assurance of generalisability of the first model to other contexts.
  - If the model does not fit, however, one cannot tell if the lack of fit is owing to the different contexts of the two datasets, or true “lack of fit” of the first model.

Overfitting. Many variables

- If there is no effect of any variable on the classification, it is still the case that the number of cases correctly classified increases in the sample that was used to derive the classifier as the number of variables increases
- But the statistical significance is usually not there
- If the variables used are selected from many, the apparent statistical significance and the apparent success in classification is greatly inflated, causing end-stage delusionary behavior in the investigator
- This problem can be improved using cross validation or other resampling methods

Example: Agresti (2002)

- Data of 68694 accidents occurred in the state of Main. The severity and explanatory variables of gender, environment and use of the belt are collected. The incidence in the presence of wounded of the factors will be studied, therefore a dichotomous factor is created: Without - With Wounds (Y)

Example: Agresti (2002)

```
summary(acc_des)
```

```
##     Gender       Location     SeatBelt          Y
##  Female:31739   Rural:25523   No :30902   Min.   :0.00000
##  Male  :36955   Urban:43171   Yes:37792   1st Qu.:0.00000
##                                           Median :0.00000
##                                           Mean   :0.09133
##                                           3rd Qu.:0.00000
##                                           Max.   :1.00000
```

```
table(acc_des$Y)
```

```
##
##     0     1
## 62420  6274
```

- Taking as a response variable the presence of injured, globally there are 6274 accidents out of a total of 68694, with a probability of 0.0913. The odds is $6274/62420 = 0.1005$ and the logodds is $log(0.1005) = -2.297472$
- It is proposed to initially compare the presence of wounded according to the Belt Use Factor (2 levels, baseline-Yes).

Example: Agresti (2002)

```r
addmargins(with(acc_des,table(SeatBelt,Y)))
```

```
##           Y
## SeatBelt      0     1   Sum
##       No  27037  3865 30902
##       Yes 35383  2409 37792
##       Sum 62420  6274 68694
```

$$P(Injured) = 6274/68694 = 0.0913$$

- 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:

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

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

Example: Agresti (2002). Null model

```
acc.m1 <- glm(cbind(Y_Yes,Y_No)~1, family=binomial(link=logit), data=acc_S)
summary(acc.m1)
```

```
##
## Call:
## glm(formula = cbind(Y_Yes, Y_No) ~ 1, family = binomial(link = logit),
##     data = acc_S)
##
## Deviance Residuals:
##      1       2
##  19.60  -19.59
##
## 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
##
## Number of Fisher Scoring iterations: 4
```

Example: Agresti (2002)

```r
xpea <- sum(residuals(acc.m1,'pearson')^2)  # pearson statistic
xdev <- sum(residuals(acc.m1,'deviance')^2) # deviance statistic
```

- The Pearson Statistic for $M1$ has the expression:

$$X_P^2 = \sum_{i=1,2} \frac{m_i(y_i-\hat{\mu}_i)}{\hat{\mu}_i(m_i-\hat{\mu}_i)}^{2} = 770.49 \approx \chi^2_{n-p=2-1=1}$$

- The deviance for $M1$ has the expression:

$$D = 2\sum_{i=1,2} \left\{ y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) + (m_i - y_i)\log\left(\frac{m_i-y_i}{m_i-\hat{\mu}_i}\right)\right\} = 768.3 \approx \chi^2_{n-p=2-1=1}$$

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

Example: Agresti (2002). Model with Seat Belt

```r
acc.m2 <- glm(cbind(Y_Yes,Y_No) ~ SeatBelt, family=binomial(link=logit), data=acc_S)
summary(acc.m2)
```

```
##
## Call:
## glm(formula = cbind(Y_Yes, Y_No) ~ SeatBelt, family = binomial(link = logit),
##     data = acc_S)
##
## Deviance Residuals:
## [1]  0  0
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.94524    0.01720 -113.12   <2e-16 ***
## SeatBeltYes -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: -8.5421e-13  on 0  degrees of freedom
## AIC: 23.523
##
## Number of Fisher Scoring iterations: 2
```

Example: Agresti (2002)

- In $M1$, the estimator $\hat{\eta} = 2.30$ is the logit of the sample proportion.
- In $M2$, the estimator $\hat{\eta} = -1.95$ is the logit of the reference level (logit of the proportion of wounded in group that does not uses belt), that is: $logit(3865/30902) = -1.95)$
- The effect of the level of Belt Use on the logit of the proportion of injured (difference of logits between the levels of Use and no use) is $logit(2409/37792) - logit(3865/30902) = -0.742$. Then, the Odds Ratio ($OR$) is:

$$OR = exp(-0.742) = 0.48$$

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

Example: Agresti (2002)

- Now, we proceed to analyze the incidence of accidents with injuries according to the gender of the driver (reference gender: female).

```
addmargins(with(acc_des,table(Gender,Y)))
```

```
##         Y
## Gender       0     1   Sum
##   Female 28254  3485 31739
##   Male   34166  2789 36955
##   Sum    62420  6274 68694
```

$$P(\mathit{Injured}) = 6274/68694 = 0.0913$$

- There are only 2 possible models: the null model that assumes homogeneity between both genders ($M1$) and the saturated model ($M2$) that proposes different proportions in men and women:

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

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

Example: Agresti (2002). Null model

```r
acc.m1g <- glm(cbind(Y_Yes,Y_No)~1, family=binomial(link=logit), data=acc_G)
summary(acc.m1g)
```

```
##
## Call:
## glm(formula = cbind(Y_Yes, Y_No) ~ 1, family = binomial(link = logit),
##     data = acc_G)
##
## Deviance Residuals:
##      1       2
##  11.10  -10.88
##
## 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: 241.72  on 1  degrees of freedom
## Residual deviance: 241.72  on 1  degrees of freedom
## AIC: 263.29
##
## Number of Fisher Scoring iterations: 4
```

Example: Agresti (2002)

```r
xpea <- sum(residuals(acc.m1g,'pearson')^2)
xdev <- sum(residuals(acc.m1g,'deviance')^2)
```

- The Pearson Statistic for $M1$ has the expression:

$$X_P^2 = \sum_{i=1,2} \frac{m_i(y_i-\hat{\mu}_i)}{\hat{\mu}_i(m_i-\hat{\mu}_i)}^2 = 242.497 \approx \chi^2_{n-p=2-1=1}$$

- The deviance for $M1$ has the expression:

$$D = 2\sum_{i=1,2} \left\{ y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) + (m_i - y_i) \log\left(\frac{m_i-y_i}{m_i-\hat{\mu}_i}\right) \right\} = 241.72 \approx \chi^2_{n-p=2-1=1}$$

Both statistics are highly significant.

Example: Agresti (2002). Model with Gender

```
acc.m2g <- glm(cbind(Y_Yes,Y_No)~Gender, family=binomial(link=logit), data=acc_G)
summary(acc.m2g)
```

```
##
## Call:
## glm(formula = cbind(Y_Yes, Y_No) ~ Gender, family = binomial(link = logit),
##     data = acc_G)
##
## Deviance Residuals:
## [1]  0  0
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.09277    0.01795 -116.56   <2e-16 ***
## GenderMale  -0.41278    0.02665  -15.49   <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: 2.4172e+02  on 1  degrees of freedom
## Residual deviance: 2.4771e-12  on 0  degrees of freedom
## AIC: 23.571
##
## Number of Fisher Scoring iterations: 2
```

Example: Agresti (2002)

- In $M1$, the estimator $\hat{\eta} = -2.29$ is the logit of the sample proportion.
- In $M2$, the estimator $\hat{\eta} = -2.09$ is the logit of the reference level (logit of the proportion of wounded in women), that is: $logit(3485/31739) = -2.09$) and the effect of the gender on the logit of the proportion of injured (difference of logits between genders) is $logit(2789/36955) - logit(3485/31739) = -0.41$.
- Then, the Odds Ratio (OR) is:

$$OR = exp(-0.41) = 0.66$$

- The odds of having accidents with injuries increase 51% in women (or decraese 34% in men).

Example: Agresti (2002). Model with Location.

- Finally, the latest univariate model has the location as explanatory factor: the odds of injuries decrease by $(1 - exp(-0.72)) \cdot 100\% = 51\%$ if it occurs in urban area, i.e., the odds of urban are $exp(-0.72) = 0.49$ times the odds of non-urban.

```
acc.m2e <- glm(cbind(Y_Yes,Y_No)~Location, family=binomial(link=logit), data=acc_L)
summary(acc.m2e)
```

```
##
## Call:
## glm(formula = cbind(Y_Yes, Y_No) ~ Location, family = binomial(link = logit),
##     data = acc_L)
##
## Deviance Residuals:
## [1]  0  0
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.89784    0.01859 -102.08   <2e-16 ***
## LocationUrban -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: 7.1961e+02  on 1  degrees of freedom
## Residual deviance: 8.3795e-12  on 0  degrees of freedom
## AIC: 23.564
##
## Number of Fisher Scoring iterations: 2
```

Example: Agresti (2002). Seat Belt and Location

```
##    Location SeatBelt Y_Yes  Y_No
## 1:    Urban       No  1808 17668
## 2:    Urban      Yes  1139 22556
## 3:    Rural       No  2057  9369
## 4:    Rural      Yes  1270 12827
```

- There are 5 models of interest applicable to the systematic structure of the previous data *M*1 to *M*5, whose deviances and details of the estimate with R are detailed below.

id	factors	df	deviance	dif_dev	contrast	df_contrast	pvalue
M1	1	3	1504.1	-	All significant	-	-
M2	A	2	784.5	719.6	M2 vs. M1	1	0
M3	C	2	736.1	48.4	M3 vs. M2	1	0
M4	A+C	1	2.7	733.4	M4 vs. M3	1	0
M5	A*C	0	0.0	2.7	M5 vs. M4	1	0.0996

Example: Agresti (2002). Model with Seat Belt and Location

```
##
## Call:
## glm(formula = cbind(Y_Yes, Y_No) ~ Location + SeatBelt, family = binomial(link =
##     data = acc_LS)
##
## Deviance Residuals:
##       1        2        3        4
## -0.7396   0.9220   0.7358  -0.8793
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.53411    0.02190  -70.05   <2e-16 ***
## LocationUrban -0.72721    0.02682  -27.12   <2e-16 ***
## SeatBeltYes   -0.75265    0.02734  -27.53   <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
```

Example: Agresti (2002)

- The additive model fits the data well. We will interpret its parameters:
    - $\eta = -1.53$ is the logit of the baseline probability: accidents when not using a belt in a rural environment.
    - $\alpha_2 = -0.72$ shows a decreasing effect of the incidence of accident victims when the accident occurs in urban surroundings.
    - $\beta_2 = -0.75$ shows a decreasing effect of the incidence of accident victims when the belt is used.
    - $exp(\alpha_2) = exp(-0.72) = 0.49$ is the OR for the location. The odds of suffering injuries in urban area is approximately half the odds in rural environment.
    - $exp(\beta_2) = exp(-0.75) = 0.47$ is the OR of suffering injuries depnding on belt use. The use of bet is more than half the odds of injuries when the belt is not used.
- The final attempt is to consider all available explanatory variables, that is, consider three factors A, C and D (location, Belt and Gender).

Example: Agresti (2002)

- Final table: Location (A); SeatBelt (B); Gender (D)

id	factors	df	deviance	AIC	BIC
M0		7	1912.5	1981.2	1981.2
M1	A	6	1192.8	1263.5	1263.7
M2	B	6	1144.4	1215.1	1215.3
M3	D	6	1670.7	1741.4	1741.6
M4	A+B	5	411.0	483.7	484.0
M5	A+D	5	911.0	983.7	984.0
M6	D+B	5	795.8	868.5	868.8
M7	A+B+A:B	4	408.3	483.0	483.3
M8	A+D+A:D	4	906.2	980.9	981.2
M9	D+B+D:B	4	795.3	870.0	870.3
M10	A+B+D	4	7.5	82.2	82.5
M11	A+B+D+A:B	3	3.6	80.3	80.7
M12	A+B+D+B:D	3	7.4	84.1	84.5
M13	A+D+B+D:B	3	7.4	84.1	84.5
M14	A+B+D+A:B+A:D	2	1.4	80.1	80.5
M15	B+A+D+B:A+B:D	2	3.6	82.3	82.7
M16	D+A+B+D:A+D:B	2	4.4	83.1	83.6
M17	A+B+D+A:B+A:D+B:D	1	1.3	82.0	82.6
M18	D+A+B+D:A+D:B+A:B+D:A:B	0	0.0	82.7	83.3

Binomial Models. References

1. Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control. 1974. AC-19:716-23
2. Clogg CC, Rubin DB, Schenker N, Schultz B, Weidman L. Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression. Journal of the American Statistical Association 1991; 86:68 -78.
3. Firth D. Bias reduction of maximum likelihood estimates. Biometrika 1993; 80:27-38.
4. Heinze, Georg, and Michael Schemper. 2002. A solution to the problem of separation in logistic regression. Statistics in Medicine 21:2409-19.
5. Hosmer D, Lemeshow S. Goodness-of-fit tests for the multiple logistic regression model. Commun Stat Part A Theor Meth. 1980;A10:1043-1069.
6. Hosmer, D. W., & Lemeshow, S. (1989). Applied logistic regression. New York: Wiley
7. P. McCullagh, John A. Nelder. (1989). Generalized Linear Models. Champman & Hall: CRC