

Generalized Linear Models. Binary response

Modelització Estadística

Grau en Intel·ligència Artificial

Curs 2024-25



Outline

- 1 Introduction
- 2 Data type
- 3 Link functions
- 4 Logit Link
- 5 Parameter estimation
- 6 Goodness of fit
- 7 Prediction
- 8 Example

Generalized linear models. Components. Recap

Random component

Let $y^T = (y_1, \dots, y_n)$ be a vector of length n , sample of a random vector $Y^T = (Y_1, \dots, Y_n)$ with:

- Statistical independent components
- Identically distributed components
- Expected values equal to $\mu^T = (\mu_1, \dots, \mu_n)$.

The random component (Y) that belongs to the **exponential family** with one parameter distribution and jointly expected values $E[Y] = \mu$

Systematic component

The systematic component (η) specifies a vector that is a linear combination from a limited number of explanatory variables $X = (X_1, \dots, X_p)$. The vector of parameters $\beta^T = (\beta_1, \dots, \beta_p)$ has to be estimated. In matrix notation:

$$\eta = X\beta$$

where η is $nx1$; \mathbf{X} is nxp ; and β is $px1$.

Generalized linear models. Link functions

Link function

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

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

- In ordinary least squares models, the *identity* link is used:

$$\eta = \mu$$

Links with binary response

For binary data, we will see several link functions commonly used:

- **Logit**
- **Probit**
- **Complementary log-log**
- **Log-log**

Binomial Family. Response variable (Y)

Binomial Random Variable

A Binomial Random Variable (RV) arises when each observation holds or does not hold a target characteristic.

- The response takes 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 k^{th} observation
 - $P(Y_k = 0) = 1 - \pi_k$. Probability of **negative** response (*failure*) for k^{th} observation

Example

- The presence/absence of a specific disease.
- Death during surgery.
- If a consumer purchases a product.
- Choice between public (metro, bus, etc) or private modes (car, motorcycle, etc) for home to work trips.

Binomial Family. Response variable (Y)

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

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

- Cumulative probability function:**

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

- Indicators:**

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

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

Binomial Family. Explanatory variables (X)

- Each individual in the sample is characterized by a set of **covariates** (eg, age) and **factors** (eg, gender) that defines the linear predictor:

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

- Explanatory variables might be:
 - Quantitative variables (covariates)
 - Transformed variables
 - Polynomial regressors
 - Dummy variables (factors or interactions)

Example

In the **public-private mode**, each commuter has several explanatory variables as income, gender, car availability, distance to local public transport, ...

Goal

To study the relationship between the response and the explanatory variables through the **probability of positive response**, which is the expected value:

$$\pi = \pi(x) = \mu_x$$

Aggregated/Disaggregated data

- Sometimes all the explanatory variables are **factors**:
 - In experimental designs, groups are defined by a combination of conditions (**covariate class**) shared by all the units in the group.
 - The common set of values of the **covariate class** applies to m_k individual units in the group.

- The total sample size (N) is the **sum of the sizes** of the n groups:

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

- With small numbers of factor levels, the number of distinct covariate vectors (n) is considerably **fewer than the number of individuals (N)**.

Aggregated/Disaggregated data

Disaggregated data			Aggregated data		
Individual unit	Variables	Response	Covariate class	Size of the class (m_k)	Positive resp. (y_k)
1	(male, 1)	0	(male,1)	2	1
2	(male,2)	1	(male,2)	3	2
3	(male,2)	0	(female,1)	1	0
4	(female,1)	0	(female,2)	1	1
5	(female,2)	1			
6	(male,2)	1			
7	(male,1)	1			

- Experiment with two *dichotomous* factors:
 - Gender: male or female
 - Car availability: “1 car” (1) or “more than 1 car” (2)
- There are $N = 7$ individuals
- There are $n = 4 = 2 \times 2$ covariate classes

Aggregated/Disaggregated data

Disaggregated data

The unit is each **observation** and each one has an individual outcome (0 or 1). **Bernoulli response**.

Aggregated data

The unit is each **covariate class** and each class is fully defined by the number of individuals (m_k) and the positive responses (y_k). **Binomial response**.

When is it possible to use aggregated data?

- All explanatory variables are factors or
- There are very few discrete variables and the sample size is large enough to form covariate classes with several individuals

Aggregated/Disaggregated data. Pros and cons

● Aggregated data

- It implies more efficient and **less memory consumption**.
- It simplifies significant effect **detected at a glance**.
- It implies to **lose the serial order** if new variables are added.
- It implies a **binomial response**. Sample observed positive responses are $y_1/m_1, \dots, y_n/m_n$, being $0 \leq y_k \leq m_k$ the number of positive responses in k^{th} covariate class which size is $m_k \rightarrow m = (m_1, \dots, m_n)$.
- The **use of Deviance and Pearson Statistic** to validate a single model is only suitable for aggregated data.
- The **use of Deviance Statistic** to compare models is suitable for aggregated data.

● Disaggregated data

- Each individual unit defines a **binomial response for a group of size 1** (bernoulli) and thus, $m = (1, \dots, 1)$.
- The **use of Deviance Statistic** to compare models is suitable for disaggregated data.

Link functions. Introduction

Goal in binary models

To establish a functional relationship between the probability of a positive result (π) and the vector of explanatory variables (factors or covariates):

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

- **Problem with binary response:** the linear predictor η might be any value in the real axis, but the probability of positive answer should belong to the open interval $(0, 1)$.
- In models with binomial response, the **link function** $g(\cdot)$ relates the vector π with the linear predictor η :

$$\eta = g(\pi), \text{ where } \pi \text{ is a vector } (n \times 1)$$

Link functions. Logit link

Logit link

- It is the **canonical link**.
- It is the inverse of **Logistic** distribution.
- It is the most frequent link function for its easy interpretation

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

- The **cumulative probability function** for a standard logistic variable is:

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

- The **density function** is:

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

- This is a **continuous and symmetric** random variable, similar to the Standard Normal distribution.

Link functions. Other link functions

Probit link

It 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 link

It is the inverse of the **Minimum extreme value** or **Gompertz** distribution, with position and scale parameters taken values 0 and 1, respectively:

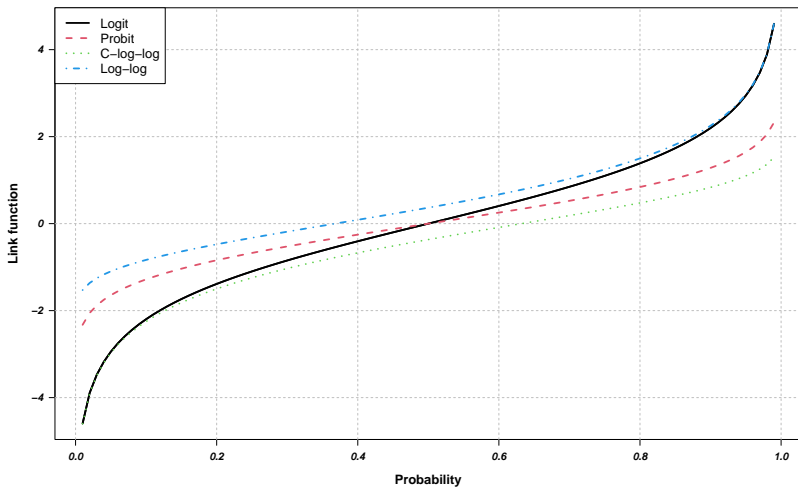
$$\eta = g_3(\pi) = \log(\log(1/(1 - \pi))) \rightarrow \pi_3(\eta) = g_3^{-1}(\eta) = 1 - \exp(-\exp(\eta)).$$

Log-log link

It is the inverse of the **Maximum extreme value** or **Gumbel** distribution, with position and scale parameters taken values 0 and 1, respectively:

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

Link functions. Link functions properties



Link functions. Link functions properties

- **All of them** are **continuous** and **monotone increasing** functions in $(0,1)$.
- **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. It is usually difficult to discriminate between these two functions.
- There is a **relationship** between **Log-log** and **C-log-log** link functions:

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

- **Logit** and **C-log-log** functions are quite similar for probabilities close to 0, but **C-log-log** trends slowly to 1 than **Logit** function does.
- The reverse is true for **Logit** and **Log-log** functions: they are quite similar for probabilities close to 1.

Link functions. Symmetry

Unlike **C-log-log** and **Log-log** links, **Logit** and **Probit** links are **symmetrical** with respect to zero.

```
##-- Generate data -----  
set.seed(12345)  
n <- 100  
Y <- sample(0:1,n, rep=TRUE)      # Response  
X <- ifelse(Y,rnorm(n,1),rnorm(n,0)) # Predictor  
Y2 <- 1-Y                        # Inverted response  
  
##-- Link logit (symmetrical) -----  
mod11 <- glm(Y ~X, family = binomial(link="logit"))  
mod12 <- glm(Y2~X, family = binomial(link="logit"))  
cbind( predict(mod11, data.frame(X=0), type='response'),  
        1-predict(mod12, data.frame(X=0), type='response'))  
##           [,1]      [,2]  
## 1 0.3836603 0.3836603  
  
##-- Link clog-log (not symmetrical) -----  
mod21 <- glm(Y ~X, family = binomial(link="cloglog"))  
mod22 <- glm(Y2~X, family = binomial(link="cloglog"))  
cbind( predict(mod21, data.frame(X=0), type='response'),  
        1-predict(mod22, data.frame(X=0), type='response'))  
##           [,1]      [,2]  
## 1 0.3694927 0.4150991
```

Link functions. When to choose probit link?

Probit link

It is appropriate when Y is obtained from dichotomizing a normally distributed variable Z .

$$Z \sim N(\mu, \sigma) \rightarrow Y = \begin{cases} 0 & \text{if } Z \leq k \\ 1 & \text{if } Z > k \end{cases}$$

Example

Dichotomize patients into **High blood** or **Normal** pressure

Link functions. When to choose Complementary log-log link?

Complementary log-log

If Z is a count having a Poisson distribution, this link arises when:

$$Z \sim P(\lambda) \rightarrow Y = \begin{cases} 1 & \text{if } Z > 0 \\ 0 & \text{if } Z = 0 \end{cases}$$

Example

Patients with **Some episode of asthma** or **No episodes of asthma**

Link functions. When to choose logit link?

Logit link

We use this link in the remaining cases or by default because:

- It is the **more interpretable**
- It is the **canonical link**
- The effects can be estimated regardless of whether the data are sampled **prospectively or retrospectively** (property not shared with other links)

Example

Patients with **Blindness or Not**

Final advice

In any case, we can see the **empirical fit** to the data (i.e, the likelihood) to choose the link function.

Interpretation under logit link (I)

- The **odd** of an event represents the ratio of the probabilities that the event occurs and its complementary. The **log-odd** is the logarithm of this ratio:

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

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

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

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

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

- The **probability of a positive response**:

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

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

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

Interpretation under logit link (II)

- Relationship between logits and odds

```
##-- Generate data -----
set.seed(12345)
n <- 100

# Response:
Y <- sample(0:1,n, rep=TRUE)
# Two factors with different referent levels:
X <- ifelse(Y,sample(0:1,n,prob=c(0.6,0.4),rep=TRUE),
            sample(0:1,n,prob=c(0.5,0.5),rep=TRUE))
X2 <- 1 - X

mod1 <- glm(Y~X, family = binomial(link="logit"))
mod2 <- glm(Y~X2,family = binomial(link="logit"))

##-- Estimated beta coefficient in the linear predictor -----
c(beta1 <- coef(mod1)[2],
  beta2 <- coef(mod2)[2])
##          X          X2
## 0.2354314 -0.2354314

##-- Odds -----
p1 <- predict(mod1,data.frame(X=0),type='resp')
p2 <- predict(mod2,data.frame(X2=0),type='resp')
c(p1/(1-p1) * exp(coef(mod1)[2]),
  p2/(1-p2))
##          1          1
## 1.318182 1.318182
```

Interpretation under logit link (III). Odds Ratio

- **Odd:** $odd(Y = 1|\mathbf{X}) = \exp(\mathbf{X}\beta)$
- Let's assume than we keep all factors constant except X_j :
 - **Individual 1:** $odd(Y = 1|\mathbf{X}_1) = \exp(\beta_1 X_1 + \dots + \beta_j X_j \dots + \beta_p X_p)$
 - **Individual 2:** $odd(Y = 1|\mathbf{X}_2) = \exp(\beta_1 X_1 + \dots + \beta_j (X_j + d) \dots + \beta_p X_p)$
- **Ratio of odds** (=Odds Ratio, OR):

$$OR_{j,21}(d) = \frac{odd(Y = 1|\mathbf{X}_2)}{odd(Y = 1|\mathbf{X}_1)} = \exp(\beta_j d)$$

Odds Ratio (Usually, d=1)

- **Covariate.** The OR from one unit increase ($d = 1$) in a covariate on the *positive response* is the exponential of its coefficient: $OR_j = \exp(\beta_j)$
- **Factor.** The OR from a factor level regarding the reference ($d = 1$, because of dummy variables) on the *positive response* is, also, the exponential of its coefficient: $OR_j = \exp(\beta_j)$

Interpretation under logit link (IV). Exercise

- | bronch | dust | smoke | years |
|--------|------|-------|-------|
| 0 | 0.20 | 1 | 5 |
| 0 | 0.25 | 1 | 4 |
| 0 | 0.25 | 1 | 8 |
| 0 | 0.25 | 1 | 4 |

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	

Interpretation under logit link (V). Exercise

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

```
##
## Call:
## glm(formula = branch ~ dust + smoke, family = binomial, data = dust)
##
## 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
```

Interpretation under logit link (VI). Exercise

- What is the **logodd** of a smoker person with 2.5 mg/cm^3 of dust at his/her working place?

```
## [1] -1.087723
```

- And the **Odd**?

```
## [1] 0.3369828
```

- And the **Probability**?

```
## [1] 0.2520472
```

- And the **Odds Ratio** for the smoker factor?

```
## [1] 1.920617
```

Parameter estimation. Bias

Maximum likelihood estimation

Maximum-likelihood estimates are:

- Unbiased
- With asymptotic variance equal to the inverse Fisher information matrix

Bias

With **small sample sizes**, estimators are **biased**.

Separation and Quasi-Separation in Logistic Regression

What is Separation?

- **Separation:** Occurs when the outcome variable (e.g., 0/1) can be perfectly predicted by a linear combination of the predictor variables.
- **Quasi-Separation:** Happens when the outcome can almost be perfectly predicted, except for a few cases.

Consequences of Separation

- Coefficients and/or their standard errors can become **infinitely large**, leading to non-convergence of the model.
- The logistic regression algorithm may **fail to estimate meaningful parameters**.

Remark

Estimates $\hat{\beta}$ do not converge, but fitted values $\hat{\pi}$ tend to a limiting value:

$$\text{If } x_i = 1 \Rightarrow \pi_i = 1 \Rightarrow \beta_2 \rightarrow \infty \Rightarrow \pi_i = \frac{\exp(\beta_1 + \beta_2 x_i)}{1 + \exp(\beta_1 + \beta_2 x_i)} \rightarrow 1$$

or

$$\text{If } x_i = 0 \Rightarrow \pi_i = 0 \Rightarrow \beta_2 \rightarrow -\infty \Rightarrow \pi_i = \frac{\exp(\beta_1 + \beta_2 x_i)}{1 + \exp(\beta_1 + \beta_2 x_i)} \rightarrow 0$$

Example of Complete Separation

Example

Separation Consider the following example data:

Predictor X_1	Outcome Y
1	1
2	1
3	1
4	0
5	0
6	0

In this case, the predictor variable X_1 **perfectly separate** the two outcome classes ($Y = 1$ and $Y = 0$).

Solution to the Separation problem

Possible solutions

- Use penalized regression techniques to avoid infinite coefficients. E.g: **Firth** regression
- Consider reducing model complexity or **merging categories** in the predictor variables.

Parameter estimation. Penalized Firth regression

- **Penalized logistic regression** was proposed by Firth for reducing the bias of ML estimates
- ML estimates of regression parameters $\beta_r (r = 1, \dots, k)$ are the solutions of the **score equations**:

$$\frac{\partial \ell}{\partial \beta_r} = U(\beta_r) = 0$$

- For reducing the bias due to **small samples** or the **(quasi)separation**, it was suggested the modified score equations:

$$U(\beta_r)^* = U(\beta_r) + 1/2 \cdot \text{trace} \left[I(\beta)^{-1} \left\{ \frac{\partial I(\beta)}{\partial \beta_r} \right\} \right] = 0$$

Goodness of fit. Outline

We will study the following methods to assess the goodness of fit of **a single model**:

- **Numerical tools:**

- Deviance statistic \rightarrow Pseudo- R^2
- Pearson statistic
- Hosmer-Lemeshow test

- **Graphical tools:**

- Calibration plot
- Residual plots

We will study the following methods to **compare models**:

- **Nested models:**

- Deviance test

- **Nested or Non-Nested models:**

- AIC statistic
- BIC statistic

Goodness of fit. Deviance statistic. Single model

Expression of Deviance (D) in the binomial model

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

Statistic to test GoF

Asymptotic distribution of D using **aggregated data** under H_0 is:

$$D_M = D(Y, \hat{\pi}) \sim \chi_{n-p}^2 \quad \text{Remark: not to be confused with } \chi_{N-p}^2$$

$$pvalue = P(\chi_{n-p}^2 > D_M)$$

- If $p \text{ value} < \alpha$, **there is evidence to reject** H_0 and thus, the model does not fit properly data.
- If $p \text{ value} > \alpha$, **there is no evidence to reject** H_0 , i.e., there is no evidence that the model does not fit properly the data.

- Deviance statistic in R can be obtained in two ways:

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

Deviance statistic. Comparison of nested models (I)

Setting

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 (β_1) to M_A and those specific (β_2), i.e., $\beta_B^T = (\beta_1^T, \beta_2^T)$ and $\beta_A^T = \beta_1^T$.

Hypothesis test

$$\begin{cases} H_0 : M_A \text{ and } M_B \text{ are equivalent} \\ H_1 : M_A \text{ and } M_B \text{ are not equivalent} \end{cases}$$

Statistic

The test for GzLM equivalent to classical F Test in linear regression compares the scaled deviances between 2 hierarchical (nested) models through their difference:

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

Deviance statistic. Comparison of nested models (II)

Hypothesis test for a single coefficient

We are testing:

$$\begin{cases} H_0 : \beta_2 = 0 \text{ (i.e., } M_B \text{ does not provide additional information to } M_A) \\ H_1 : \beta_2 \neq 0 \text{ (i.e., } M_B \text{ provides additional information to } M_A) \end{cases}$$

Then:

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

- It can be a contrast for multiple coefficients: **large values of the statistic indicate non-equivalent models**

Deviance statistic. *pseudo-R*²

Role

Deviance for a GzLM plays a similar role to the Residual Sum of Squares in classical regression.

Definition

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

where:

$$G(y, \pi_A) = D(y, \pi_0) - D(y, \pi_A)$$

- **Null model** $\rightarrow R^2 = 0$
- **Saturated model** $\rightarrow R^2 = 1$

Goodness of fit. Pearson statistic

Definition

The **Generalized Pearson Statistic** (X^2) is defined as:

$$X^2 = \sum_{i=1}^n \frac{(y_i - m_i \hat{\pi}_i)^2}{m_i \hat{\pi}_i (1 - \hat{\pi}_i)} = \sum_{i=1}^n \frac{m_i (y_i - \hat{\mu}_i)^2}{\hat{\mu}_i (m_i - \hat{\mu}_i)}$$

It has a χ^2 distribution:

$$X^2 \sim \chi_p^2$$

- When disaggregated data is used ($m_i = 1$) asymptotic theory does not apply (see next slide).
- The X^2 Statistic is the sum of the squares of Pearson's residuals:

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

Goodness of fit. Sparseness.

Definition

Sparseness refers to the presence of a large proportion of small observed counts in the sample.

- The effect of sparseness is noticed mainly on the **Deviance** and **Pearson's** statistics: they fail to have the properties required for testing the GoF.

Pearson and Deviance statistics with disaggregated data

If $Y_i \sim B(1, \pi)$, then the statistics reduce to ($\hat{\pi} = \bar{y}$):

$$\text{Pearson} \rightarrow X^2 = \sum \frac{(y_i - \bar{y})^2}{\bar{y}(1 - \bar{y})} = n \rightarrow \text{Sample size } (n) \text{ is not useful}$$

$$\text{Deviance} \rightarrow D = -2n\{\bar{y}\log(\bar{y}) + (1 - \bar{y})\log(1 - \bar{y})\} \rightarrow \text{Depends on } \hat{\pi}!!!$$

- It is a good statistical practice not to rely on either D or X^2 as an absolute measure of GoF with disaggregated data.
- When m_i are small but greater than one, we may use D or X^2 as test statistics.

Goodness of fit. Hosmer-Lemeshow statistic.

- The **Hosmer-Lemeshow Statistic** is another measure of goodness of fit.

Procedure to calculate the statistic

- 1 Partitioning the observations into **G** (e.g, 10) **equal sized** groups according to their predicted probabilities.
 - Do not choose equispaced probabilities, i.e 0–0.1, 0.1–0.2, ..., 0.9–1.
 - It is possible to perform the test with $G < 10$ for small sample size.
- 2 For each group, the observed and expected number of positive/negative responses under independence are compared by the Pearson Statistic applied to the $2 \times G$ frequency table:

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

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

Goodness of fit. Compare unnested models. AIC and BIC.

- **AIC** (Akaike Information Criteria) and **BIC** (Bayesian Information Criteria) are defined as a trade-off between:
 - The **goodness of fit** of the model
 - The number of parameters p (**model complexity**)

AIC & BIC

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

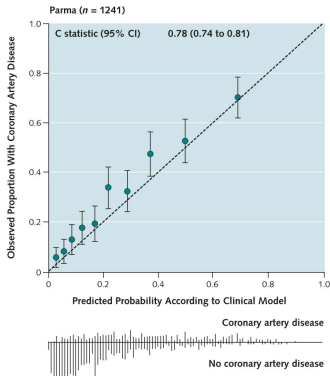
Remarks

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

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


Goodness of fit. Calibration plot.

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



Pearson Residuals. Recap

Pearson Residuals in logistic regression

- The **Pearson residuals** are used to assess how well the model fits the data.
- With disaggregated data, they are computed as:

$$r_i = \frac{y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1 - \hat{p}_i)}}$$

where:

- y_i is the observed outcome (0 or 1).
- \hat{p}_i is the predicted probability of the outcome.
- They measure the standardized difference between observed and expected values.

Key Characteristics

- Residuals **near 0** indicate **good fit**.
- **Large residuals** suggest the model is **poorly fitting** those observations.

Pearson Residuals vs Predicted Values Plot with binary outcome

Disaggregated Data

- In disaggregated data, the plot of Pearson residuals versus predicted values often appears **scattered**.
- Points may show **high variability for cases with predicted probabilities near 0 or 1**, indicating potentially influential observations.
- **More difficult** to assess the goodness of fit

Aggregated Data

- In aggregated data, **Pearson residuals** are plotted against group-level **predicted probabilities**.
- **Patterns** in aggregated data indicate potential **issues** with over- or under-fitting in certain groups.
- A **fan-shaped pattern suggests model misfit**, especially for extreme predicted probabilities.

```
residualPlots(model)
```

Cook's Distance

What is Cook's Distance?

- It measures the **influence** of a single point on the fitted model.
- It quantifies how much the estimated **regression coefficients change** when an observation is removed from the dataset.

Cook's Distance (D_i) for observation i

$$D_i = \frac{\sum_{j=1}^n (\hat{y}_j - \hat{y}_{j(-i)})^2}{p \cdot \hat{\sigma}^2}$$

- \hat{y}_j is the predicted value with all observations.
- $\hat{y}_{j(-i)}$ is the predicted value without observation i .
- p is the number of parameters in the model.
- $\hat{\sigma}^2$ is the estimated variance of the residuals.

Interpreting Cook's Distance

- **Larger values** indicate **higher influence** on the model.
- Rule of thumb: if $D_i > 1$, the observation is considered **highly influential**.
Alternative: graphical tools

Prediction. Confusion matrix

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

Prediction. Measures (I)

Several measures of predictive capability might be obtained from the confusion matrix.

- **Sensitivity (Sens)** or **True positive rate (TPR)** or **Recall**. Proportion of predicted as positive ($\hat{y}_i = 1$) among observed positive outcomes ($Y = 1$).

$$Sens = a / (a + c)$$

- **Specificity (Sp)** or **True negative rate (TPR)**. Proportion of predicted as negative ($\hat{y}_i = 0$) among observed negative outcomes ($Y = 0$).

$$Sp = d / (b + d).$$

- **Positive predictive value (PPV)** or **Precision**. Proportion of positive outcomes ($Y = 1$) among those ones predicted as positive. ($\hat{y}_i = 1$).

$$PPV = a / (a + b).$$

- **Negative predictive value (NPV)**. Proportion of observed negative outcomes ($Y = 0$) among those ones predicted as negative ($\hat{y}_i = 0$).

$$NPV = d / (c + d).$$

Prediction. Measures (II)

- **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 prevalence of positive response. On the other hand, **VPN** and **NPV** depend on that prevalence.
- To interpret the results is more useful the predictive values (PPV/NPV) because they provide the probability of an specific response given the prediction of the model, which is the information we have.

Prediction. Measures. Exercise

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

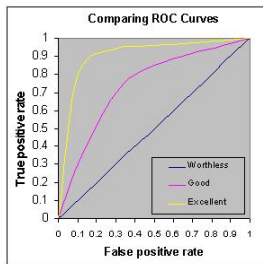
<i>Predicted Probability</i>	Response
0.08	0
0.27	0
0.34	1
0.36	0
0.44	0
0.52	0
0.53	1
0.54	0
0.80	1
0.96	1

Prediction. ROC curve (I)

ROC curve

For each threshold s , ROC Curve represents the **true positive rate (Sens)** vs. the **false negative rate ($1-Sp$)**.

- If the ROC curve rises rapidly towards the upper left-hand corner of the graph, equivalent to say that the value of the Area Under the Curve (AUC) is large, the model performs well:



Prediction. ROC curve (II)

Points interpretation

- Point **(0,1)**. Perfect classifier: all individuals correctly classified.
 - Point **(0,0)**. Classifier that predicts all cases to be negative.
 - Point **(1,1)**. Classifier that predicts all cases to be positive.
-
- **Website** to understand ROC curves.
 - **Video** to understand ROC curves.
 - A guideline for interpreting the *AUC* of a ROC curve is:
 - 0.90 - 1.00 = excellent
 - 0.80 - 0.90 = very good
 - 0.70 - 0.80 = good
 - 0.60 - 0.70 = bad
 - 0.50 - 0.60 = very bad
 - **rms** R package 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* . Also implemented in the *fmsb* package. Other expressions for *pseudo- R^2* can be found **here**.

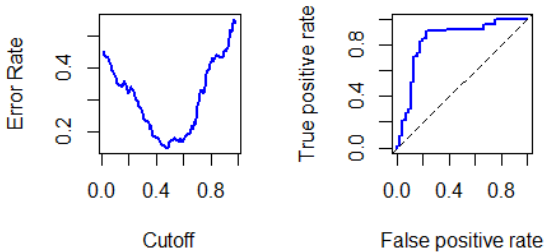
Prediction. ROC curve (III)

AUC interpretation

Given a couple of individuals, one with positive response and one with negative one, the **AUC** is the probability that the individual with positive response has higher predicted probability.


- ROC curves and other performance plots are available in *ROCR* R package.

```
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)
```



Prediction. Exercise (I)

Comment on this tweet:



Karandeep Singh
 @kdpsinghlab

Suivre

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

11 Retweets 53 Me gusta



4 11 53




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](#)

1 4



Karandeep Singh @kdpsinghlab · 19 sept.

Oh, you're asking about the sensitivity of the model? Please refer to our "no refunds" policy.

1 9

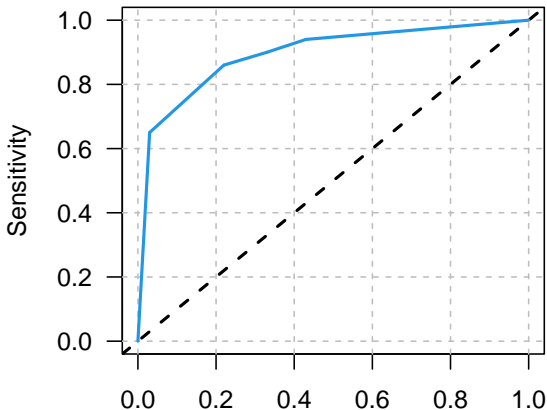
Prediction. Exercise (II)

- Build the ROC curve associated to this data:

	<i>Non Disease</i>		<i>Disease</i>	
<i>cutpoint</i>	<i>Test +</i>	<i>Test -</i>	<i>Test +</i>	<i>Test -</i>
1	25	33	48	3
2	19	39	46	5
3	13	45	44	7
4	2	56	33	18

Prediction. Exercise (II). Solution

<i>cutpoint</i>	<i>1-Specificity</i>	<i>Sensitivity</i>
-	1	1
1	0.43	0.94
2	0.33	0.90
3	0.22	0.86
4	0.03	0.65
-	0	0



Prediction. Types.

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

```
# Linear predictor for observations to fit the model  
predict(model, type='link')  
# Probabilities for observations to fit the model  
predict(model, type='response')  
  
# Linear predictor for new observations  
predict(model,newdata, type='link')  
# Probabilities for new observations  
predict(model,newdata, type='response')
```

- Data of **1000** emails.
- **Explanatory variables:**
 - *Frequency of certain keywords (e.g., "win", "prize")* (x_1)
 - *Length of the email* (x_2)
 - *Presence of links in the email* (x_3)
- **Response:**
 - If an email is **Spam** or **Not Spam**.

Example. Model output

```
sum_model
```

```
##
## Call:
## glm(formula = Spam ~ Keyword_Frequency + Email_Length + Has_Links,
##      family = binomial, data = data)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -2.3317923   0.2937970   -7.937 2.08e-15 ***
## Keyword_Frequency  0.8564552   0.0673942   12.708 < 2e-16 ***
## Email_Length    -0.0020215   0.0003976   -5.084 3.69e-07 ***
## Has_Links       1.8309803   0.1645855   11.125 < 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: 1384  on 999  degrees of freedom
## Residual deviance: 1010  on 996  degrees of freedom
## AIC: 1018
##
```

Example. Model Interpretation

- β_0 : Intercept, representing the baseline probability of spam.
- β_1 : Increase in the odds of the email being spam for each unit increase in keyword frequency.
- β_2 : Effect of email length on the likelihood of being spam.
- β_3 : Impact of the presence of links on the probability of the email being spam.
- The odds ratio e^{β_i} helps quantify these effects.

Example. Prediction and Evaluation

- Given the model, we predict whether an email is spam or not:

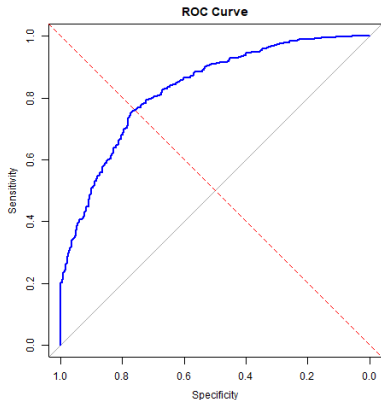
$$\hat{Y} = \begin{cases} 1 & \text{if } P(Y = 1|X) > 0.5 \\ 0 & \text{otherwise} \end{cases}$$

- The performance of the model can be evaluated using a confusion matrix:

	Predicted Spam
Actual Spam & True Positives (TP) & False Negatives (FN)	
Actual Not Spam & False Positives (FP) & True Negatives (TN)	

Example. ROC Curve and AUC

- ROC (Receiver Operating Characteristic) curve plots the True Positive Rate (TPR) vs. False Positive Rate (FPR) for different thresholds.
- The Area Under the Curve (AUC) measures the model's ability to distinguish between spam and non-spam emails.
- A model with an AUC close to 1 indicates a good classifier.



Example. Conclusion

- Logistic regression is a powerful tool for binary classification problems in AI, such as spam detection.
- By modeling the probability of an event (e.g., email being spam) as a function of input features, we can make informed predictions.
- Performance can be evaluated using confusion matrices and ROC curves.

Example: Agresti (2002)

Example

- Data of **68,694** accidents occurred in the state of Maine.
- **Explanatory variables:**
 - *Use of the belt* (D): Yes/No
 - *Gender of the driver* (C): Male/Female
 - *Location* (A): Urban/Rural
- **Response:**
 - The severity of the accident.
 - To study the influence of the factors in the presence of injured people, a dichotomous factor (**Y**) was created: **Without** or **With** wounds

Example: Agresti (2002)

```
kable(summary(acc_des),align='r')
```

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

```
kable(table(acc_des$Y),align='r',col.names=c('Y','Freq'))
```

Y	Freq
0	62420
1	6274

- There are 6,274 accidents with injured people out of 68,694 accidents. The **probability** is 0.0913; the **odds** is $6,274/62,420 = 0.1005$; and the **logodds** is $\log(0.1005) = -2.30$
- First, it is proposed to compare the presence of wounded according to the *belt use* (factor with 2 levels: No-Yes).

Example: Agresti (2002)

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

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

- There are only 2 possible models: the **null model** (M1) that assumes *homogeneity* between the groups defined by the factor and the **saturated model** (M2) that proposes different probability of injuries in 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)
##
## 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)

- Null model

```
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:

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

- The deviance for $M1$ has the expression:

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

- Both statistics are highly significant, implying that the *null model* does not properly fit the data.

Example: Agresti (2002). Model with Seat Belt

```
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)
##
## 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: -3.2094e-12  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 $\hat{\eta} = \text{logit}(3865/30902) = -1.95$ is the logit of the proportion of wounded in the reference group (those that do not use belt),
- The effect of the use of the belt on the logit of injured is $\text{logit}(2409/37792) - \text{logit}(3865/30902) = -0.742$. Thus, the Odds Ratio (OR) is:

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

- Interpretation: the odds of having injuries because of not using the belt are more than double than the odds of having injuries using the belt.

Example: Agresti (2002)

- We are going to analyze the influence of the *gender* of the driver (reference: *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
```

- There are only 2 possible models: the **null model** ($M1$) that assumes *homogeneity* between both genders 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

```
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)
##
## 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)

- Null model

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

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

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

- The deviance for $M1$ has the expression:

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

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

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)
##
## 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 $\hat{\eta} = \text{logit}(3485/31739) = -2.09$ is the logit of the proportion of wounded in women and the effect of the gender on the logit is $\text{logit}(2789/36955) - \text{logit}(3485/31739) = -0.41$.
- The Odds Ratio (OR) is:

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

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

Example: Agresti (2002). Model with Location.

- The last 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. Thus, the odds of accident with injuries in urban area are $\exp(-0.72) = 0.49$ in reference to the odds of non-urban area.

```
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)
##
## 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
##      <fctr>   <fctr> <num> <num>
## 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 *M1* to *M5* with *Seat Belt* and *Location* as explanatory variables. Their deviances 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(li
```

```
## data = acc_LS)
```

```
##
```

```
## 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 depending on belt use. The odds of injuries using belt is less than half the odds when the belt is not used.
- The final attempt is to consider all available explanatory variables, that is, consider three factors: Location (A), Belt (C) and Gender (D).

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 G, & Schemper M. 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 McCullagh P. & Nelder JA. (1989). Generalized Linear Models. Chapman & Hall: CRC