Module de Régression Régularisé.
Régression logistique. Logistic regression
Scoring

Mathilde Mougeot

ENSIIE

2018-2019

Logistic regression

- **Extremely used** in many area, for many applications
- The first applications were in medicine, for medical applications
- Use also for Credit Scoring in Banques, Insurances :
- The "logistic regression " function is available in any Statistical Software

Plan

- Applications
- Logistic regression model
- Model interpretation
- Performances criteria
- Model selection

APPLICATIONS

Health application. Heart attacks

chd : target variable ; 8 covariables

chd	Coronary heart disease binary response
sbp	systolic blood pressure (integer)
tobacco	cumulative tobacco (kg) (real)
ldl	low density lipoprotein cholesterol (real)
adiposity	(real)
famhist	family history of heart disease (Present, Absent)
typea	type-A behavior (integer)
obesity	(real)
alcohol	current alcohol consumption (real)
age	age at onset (real)

A retrospective sample of males in a heart-disease high-risk region of the Western Cape, South Africa.
data described in Rousseauw et al, 1983, South African Medical Journal.
Elementss of Statistical Learning, Hastié, Tibshirani, Friedman.

`http://statweb.stanford.edu/~tibs/ElemStatLearn/`

Data - Cardiac disease

nř	sbp	tobacco	ldl	adiposity	famhist	typea	obesity	alcohol	age	chd
1	160	12.00	5.73	23.11	Present	49	25.30	97.20	52	1
2	144	0.01	4.41	28.61	Absent	55	28.87	2.06	63	1
3	118	0.08	3.48	32.28	Present	52	29.14	3.81	46	0
4	170	7.50	6.41	38.03	Present	51	31.99	24.26	58	1
5	134	13.60	3.50	27.78	Present	60	25.99	57.34	49	1
6	132	6.20	6.47	36.21	Present	62	30.77	14.14	45	0
7	142	4.05	3.38	16.20	Absent	59	20.81	2.62	38	0
8	114	4.08	4.59	14.60	Present	62	23.11	6.72	58	1
9	114	0.00	3.83	19.40	Present	49	24.86	2.49	29	0
10	132	0.00	5.80	30.96	Present	69	30.11	0.00	53	1
...	...	...	...	...	...					

Remark on the volumetry :

$n = 462$, $p(chd = 1) = 34\%$.

Problems

- To understand the key factors linked to an Heart attack
  - Significativity (yes/not : result of a test)
  - Strength
  - Positive or negative effect

  $\rightarrow$ Preventive Health care

- Quality of the model
  What are the model performances on new data ?

- **Decision making process**
  $\rightarrow$To be able to evaluate for a new patient the risk of an heart attack.
  $\rightarrow$ To better take care of a patient showing a high risk level
- A Sparse model is appreciate
  to better understand and follow few risk factrors
  to improve the Predictive power (improving the generalization power)

Bank and Assurance

Incident	Banking problem
revenu	(numerical value)
depnaiss	département de naissance (variable qualitative)
datenaiss	année de naissance
duree	durée du crédit en cours
montcred	montant du crédit en cours
situfam	situtation familiale
ancienn	nombre de mois d'ancienneté
cb	possession d'une carte bleue (1) ou non (0)
numero	numéro du client dans la base

Data description

Real data $n = 50000$ clients, $p_{Incident=1} = 2\%$ (2/1000)

Main objectives

- Find a "Good" model
  $\rightarrow$ being able to predict the correct answer on new data
- Evaluate the performances on the data
  $\rightarrow$ The different error may be differently re-weighted.
- Understand the variables which have an influence on the target
  $\rightarrow$ This may be done only given the variations of the different variables in the data set
  $\rightarrow$ A variable which does not vary in the data set will be considered as non influent

Plan

- Applications
- logistic regression model
- Model interpretation
- Performances criteria
- Model selection

LOGISTIC REGRESSION MODEL

BINARY CASE

(ordinal)

Binary logistic regression

Variables :

- Y Binary target variable $\{0, 1\}$
- $X_1, \dots X_d$ Quantitative or binary explanatory variables (modality indicators)
    - X_1 : Simple logistic regression
    - $X_1, X_2, \dots$: Multiple logistic regression

The data :

- Sample (n, d) of numerical data (ex. SAS table, R dataframe R)
- $\mathcal{D}_n = \{(x_i, y_i)\ \ 1 \leq i \leq n, x_i \in \mathbb{R}^d\ y_i \in \{0, 1\}\}$
- Notations : d variables including the intercept

Logistic regression

Variables :

- Y Binary target variable $\{0, 1\}$ (dependant variable)
- X multivariate explanatory variable (independant variable)

the aim is to modelized : $\mathbb{E}(Y/X = x)$

For a binary target variable Y taking 0 or 1 values :

$$\begin{aligned}\mathbb{E}(Y/X = x) &= Prob(Y = 1/X = x)\\ &= \eta(x)\end{aligned}$$

Remarks :

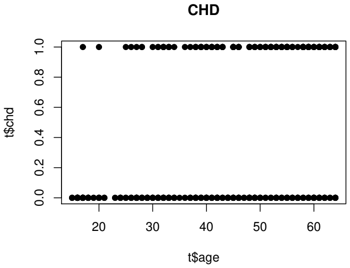
- $Prob(Y = 1/X = x)$: is the Posteriori probability
- A regular linear model is not appropriate in this case
  $\eta(x) = \beta_0 + \beta_1 X_1 + ...$

Simple logistic regression

Simple model (one variable) to explain :

- chd (Coronary Heart Disease, $\{0, 1\}$)
- function of the age (real value)

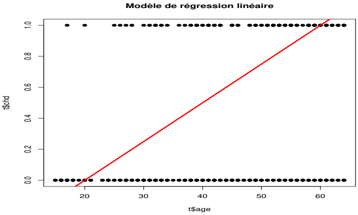
Illustration : raw data

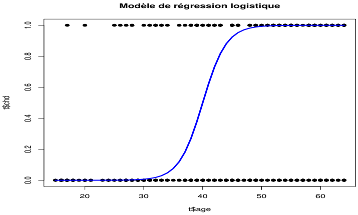


Simple logistic regression

Simple linear model to explain :

- chd (Coronary Heart Disease, $\{0, 1\}$)
- function of the age (real value)





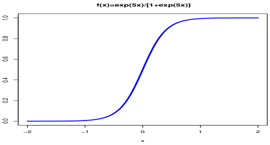
$\rightarrow$ Linear model is not appropriate $\eta(x) = \beta_0 + \beta_1 X_1 + ...$ (left)
$\rightarrow$ logistic regression model (right)

Logistic regression model

Transfer function : $\eta(x)$

with $z = \beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d$

$$\eta(z) = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}$$



$$\eta(z) \ : \mathbb{R} \rightarrow [0,1]$$

- link function **Logit** : $\log(\frac{\eta(x)}{1-\eta(x)}) = \beta^T x$

GLM :Generalized Linear Models

Underlaying statistical model

Logistic regression :

- $Y \in \{0,1\}$, X multivariate variable.

- Statistical model :
  - $\mathcal{L}(X,Y) \equiv (\mathcal{P}_X, \eta)$
    avec $\eta(x) = \mathbb{E}(Y/X = x)$
    with η known and $\mathcal{P}_X$ is not specified.

Discriminant Analysis (reminder) :

- $\mathcal{L}(X,Y) \equiv (p, \mathcal{L}(X/Y))$ with $\mathcal{L}$ Gaussian.

Link functions

Several Transfer function have been proposed :

$$\mathbb{E}(Y/X = x) = Prob(Y = 1/X = x) = \eta(x)$$

It is a modeling choice, "Expert" choice. linked functions :

- Logit model : $\eta(z) = e^z/(1 + e^z) \leftrightarrow g(\eta) = \log(\frac{\eta}{1-\eta})$
- Probit (normit) model : $\eta(z) = \Phi(z) \leftrightarrow g(\eta) = \Phi^{-1}(\eta)$
  cher Φ is the Gaussian repartition function $\mathcal{N}(0, 1)$
- "Log-Log model" : $g(\eta) = \log(-\log(1 - \eta))$
  (epidemiology, toxicology)

$\rightarrow$ **this implies a modification of the statistical model**

$$\mathcal{L}(X, Y) \equiv (\mathcal{P}_X, \eta) \text{ and } \eta = \mathbb{E}(Y/X = x) \text{ are modified}$$

Estimation of the parameters of the logistic regression model

- **Variables :**
  - Y binary target variable $\{0, 1\}$
  - X real values $X \in \mathbb{R}^d$ ($d = 1$ simple logistic regression)

- **Sample of data i.i.d.**
  - n iid observations
  - $\mathcal{D}_n = \{(x_i, y_i), \ \ 1 \leq i \leq n, \ \ y_i \in \{0, 1\}\}$

- **For on x_i observation, the model is :**
  - $\eta(x_i) = Prob(Y = 1/X = x_i)$
    $= \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}$

- The β parameters are estimated by the maximum Likelihood (minimisation of the - log-likelihood)

Conditional likelihood

data sample iid : $\mathcal{D}_n = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$
The probability to observe $\mathcal{D}_n$:

$$\mathcal{L}(\beta, (x_1, y_1), \dots, (x_n, y_n))$$

$$\begin{aligned}
\mathcal{L}_{\beta,\mathcal{D}_n} &= \prod_{i=1}^n Prob(Y = y_i / X = x_i) \\
&= \prod_{i=1}^n \eta(x_i)^{y_i}(1 - \eta(x_i))^{1-y_i} \\
&= \prod_{i=1}^n \left(\frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}\right)^{y_i}\left(\frac{1}{1+e^{\beta^T x_i}}\right)^{1-y_i} \\
&= \mathcal{L}_{\mathcal{D}_n}(\beta)
\end{aligned}$$

with $y_i = 1$ ou $y_i = 0$
$\eta(x_i) = Prob(Y = 1/X = x_i)$ et $1 - \eta(x_i) = Prob(Y = 0/X = x_i)$

Conditional Log-likelihood

$$
\begin{aligned}
\mathcal{L}_{\mathcal{D}_n}(\beta) \quad &= \textstyle\prod_{i=1}^{n} Prob(Y = y_i / X = x_i) \\
\\
\ell_{\mathcal{D}_n}(\beta) \quad &= \log(\mathcal{L}_{\mathcal{D}}(\beta)) \\
&= \log(\textstyle\prod_{i=1}^{n} \eta(x_i)^{y_i}(1 - \eta(x_i))^{1-y_i}) \\
\\
&= \textstyle\sum_{i=1}^{n} y_i \log(\frac{\eta(x_i)}{1-\eta(x_i)}) + \log(1 - \eta(x_i)) \\
\\
&= \textstyle\sum_{i=1}^{n} \{y_i \beta^T x_i - \log(1 + e^{\beta^T x_i}\}
\end{aligned}
$$

note : the intercept are in the "x_i" term ($x_i = 1$).

- We compute $\hat{\beta}$ to maximize the log-likelihood $\ell_{\mathcal{D}}(\beta)$
- La The log-likelihood is a convex function, the $\hat{\beta}$ are then unique

Maximization of the conditional log-likelihood

To maximize the log-likelihood :

- It is necessary to cancel the d derivatives $\beta = (\beta_1, \ldots, \beta_d)^T$
  $\forall j \ \beta_j, \ \frac{\delta \ell}{\delta \beta_j} = 0$
- the (d) score equations are : $\frac{\delta \ell(\beta)}{\delta \beta_j} = \sum_{i=1}^n x_{i,j}(y_i - \eta(x_i, \beta)) = 0$ with $\eta(x_i) = \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}$
- With matrix notation : $\frac{\delta \ell(\beta)}{\delta \beta} = \sum_{i=1}^n x_i(y_i - \eta(x_i, \beta)) = 0$

In particular : for $x_i = 1$ (intercept), we have :
$\sum_{i=1}^n y_i/n = \mathbb{E}\, \eta(x_i = 1, \beta_1)$,
le nombre moyens d'observations dans la classe 1 est égale au nombre attendu, à son espérance.

Maximization of the conditional log-likelihood

We look for $\beta = (\beta_1, .., \beta_d)$ to cancel the score equations :

$\frac{\delta \ell(\beta)}{\delta \beta} = \sum_{i=1}^{n} x_{i,j}(y_i - \eta(x_i, \beta)) = 0$
avec $\eta(x_i) = \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}$

No direct analytical expression
but single solution...

Maximization of the conditional log-likelihood

Cancel the derivates : $\boxed{\frac{\delta \ell}{\delta \beta} = \sum_{i=1}^{n} x_i (y_i - \eta(x_i, \beta)) = 0}$

Taylor developpement of a fonction $f(x)$ at first order :

$$f(x) \sim f(x_0) + f^{'}(x_0)(x - x_0)$$

Solve the affine equation $0 = f(x_0) + f^{'}(x_0)(x - x_0)$,
The first solution is x_1, and using an iterative process

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$$

The Newton-Raphson algorithm used here asks to compute the second derivative, (Hessaian matrix) of the log likekihood, ℓ

The updating values of β are the computed :

$$\beta^{new} = \beta^{old} - (\frac{\delta^2 \ell(\beta^{old})}{\delta \beta \delta \beta^T})^{-1} \frac{\delta \ell(\beta^{old})}{\delta \beta}$$

avec $\frac{\delta^2 \ell(\beta)}{\delta \beta \delta \beta^T} = - \sum_{i=1}^{n} x_i x_i^T \eta(x_i, \beta)(1 - \eta(x_i, \beta))$

Maximization of the conditional log-likelihood

Matrix Notation :

- X : matrix $(n \times p)$
- Y : vector $(n \times 1)$
- η : vector $(n \times 1)$, for i observation, $\eta(x_i, \beta^{old})$

$$\boxed{\frac{\delta \ell(\beta)}{\delta \beta} = X^T(Y - \eta)}$$

- W : diagonal matrix $(n \times n)$,
- $W(i,i) = \eta(x_i, \beta^{old})(1 - \eta(x_i, \beta^{old})$

$$\boxed{H = \frac{\delta^2 \ell(\beta)}{\delta \beta \delta \beta^T} = -X^T W X}$$

H : Hessian matrix

Maximization of the conditional log-likelihood

The estimation method is Newton-Raphson and IRLS

$$\begin{aligned}\beta^{new} &= \beta^{old} + (X^T W X)^{-1} X^T (Y - \eta) \\ &= (X^T W X)^{-1} X^T W (X\beta^{old} + W^{-1}(Y - \eta)) \\ &= (X^T W X)^{-1} X^T W z\end{aligned}$$

With $z = X\beta^{old} + W^{-1}(Y - \eta)$

IRLS : Iteratively Reweighted Least Square

$\beta^{new} \leftarrow ArgMin(z - X\beta)^T W (z - X\beta)$

$\beta = 0$ good initial choice. The convergence is not guarantee.

3 steps : Estimation $\rightarrow$ Prediction $\rightarrow$ Decision

At the beginning, a n-Sample : $\mathcal{D}_n$ and the choice of a model.

1. Estimation of the parameters of the model (β)
   - Use the data to compute/estimate $\hat{\beta}$
   - Parameter Selection (eventually see later)

2. Prediction (using the calibrated model - $\hat{\beta}$)
   - For an new observation x_{new}, $x_{new} \notin \mathcal{D}_n$.
   - Estimation (computation) of the Probability : $\boxed{\hat{\eta}(x, \hat{\beta}) = \frac{e^{\hat{\beta}^T x_{new}}}{1+e^{\hat{\beta}^T x_{new}}}}$

3. Decision

   Given the value of a chosen Threshold S, $S \in [0, 1]$
   - $\hat{Y} = 1$ if $\hat{\eta}(x, \hat{\beta}) > S$
   - $\hat{Y} = 0$ otherwise

   MAP (Maximum A Posteriori) choice $S = 0.5$,
   but other thresholds may be more appropriate (see later).

Plan

- Applications
- Logistic regression model
- Model Interpretation
- Performances criteria
- Model Selection

Estimation and confidence intervals

$\hat{\beta}$ is an estimator computed by the MLE.
Properties.

- It is asymptotically no biaised
- The variance is minimal
- It is asymptotically Gaussian

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \rightarrow_{n\rightarrow\infty} \mathcal{N}(0, \Sigma_0)$$

$\Sigma_0 \simeq -H_n^{-1}$ avec $H_n = -X^T WX/n$ ($H_n = H$ Hessian matrix)

Knowing the asymptotic law helps to compute :

- Confidence interval on β
- Significativity tests ($\neq 0$) on β

Significativity on the β coefficients

- The model

$$\eta_\beta = \frac{e^{\beta_0+\beta_1 X_1+\ldots+\beta_d X_d}}{1+e^{\beta_0+\beta_1 X_1+\ldots+\beta_d X_d}}$$

- Test de Wald (with a α risk) :
  - $H_0 : \beta_j = 0$
  - $H_1 : \beta_j \neq 0$

- The test statistics (Wald) : $Z_j = \frac{\hat{\beta}_j}{S_{\beta_j}}$

  follows asymptotically a Gaussian law (R)

- Decision
  - depends on the p-value and the value of the α risk

Remark : Be careful to the collinearity variables (impact on the computation of S_{β_j})

The global model

The logit model :

$$\eta_\beta = \frac{e^{\beta_1+\beta_2 X_2 \ldots+\beta_d X_d}}{1+e^{\beta_1+\beta_2 X_2+\ldots+\beta_d X_d}}$$

- With no explanatory variable. Model M_0, $\beta_2 = \ldots = \beta_d = 0$.
  - Estimation : $\hat{\beta}_1 = \ln \frac{\bar{y}}{1-\bar{y}} = \ln \frac{n_+}{n_-}$
  - Log-likelihood : $\begin{aligned} \ell_0(\hat{\beta}_1, \mathcal{D}_n) &= \sum y_i \ln(\bar{y}) + (1-y_i)\ln(1-\bar{y}) \\ &= n_+ \ln(\bar{y}) + n_- \ln(1-\bar{y}) \end{aligned}$
  - **Deviance :** $D_0 = -2 \times \ell_0$
- For the full model the deviance : $D_M = -2\ell(\hat{\beta}, \mathcal{D}_n)$
- under the H_0 assumption (all the coefficients are zero) $(D_0 - D_M) \sim \chi^2(d-1)$

Model M_0

Estimation of the β parameter :

$$
\begin{aligned}
\mathcal{L}_{\mathcal{D}_n}(\beta) &= \prod_{i=1}^{n} Prob(Y = y_i / X = x_i) \\
\ell_{\mathcal{D}_n}(\beta) &= \log(\mathcal{L}_{\mathcal{D}}(\beta)) = \log(\prod_{i=1}^{n} \eta(x_i)^{y_i}(1-\eta(x_i))^{1-y_i}) \\
&= \sum_{i=1}^{n} y_i \log(\tfrac{\eta(x_i)}{1-\eta(x_i)}) + \log(1-\eta(x_i)) \\
&= \sum_{i=1}^{n} \{y_i \beta^T x_i - \log(1+e^{\beta^T x_i}\} \\
&= \sum_{i=1}^{n} \{y_i \beta - \log(1+e^{\beta}\}
\end{aligned}
$$

We look for $\hat{\beta}$ to cancel the derivative of $\ell_{\mathcal{D}_n}(\beta)$:

- $\hat{\beta} = \ln \frac{\bar{y}}{1-\bar{y}} = \ln \frac{n_+}{n_-}$
- Log-likelihood : $\begin{aligned} \ell_0(\hat{\beta}_1, \mathcal{D}_n) &= \sum y_i \ln(\bar{y}) + (1-y_i)\ln(1-\bar{y}) \\ &= n_+ \ln(\bar{y}) + n_- \ln(1-\bar{y}) \end{aligned}$

Medical Health -cardiac disease

chd : target variable ; 8 covariables

chd	Coronary heart disease response
sbp	systolic blood pressure (integer)
tobacco	cumulative tobacco (kg) (real)
ldl	low density lipoprotein cholesterol (real)
adiposity	(real)
famhist	family history of heart disease (Present, Absent)
typea	type-A behavior (integer)
obesity	(real)
alcohol	current alcohol consumption (real)
age	age at onset (real)

A retrospective sample of males in a heart-disease high-risk region of the Western Cape, South Africa.
data described in Rousseauw et al, 1983, South African Medical Journal.
Elements of Statistical Learning, Hastié, Tibshirani, Friedman.

http://statweb.stanford.edu/~tibs/ElemStatLearn/

Logistic Regression, application

- sous R

```
res=glm(chd~. ,family=binomial,data=tab);
summary(res)
```

- sous SAS

```
proc logistic data=tab;
class chd(desc);
model chd=age; run;
```

Model interpretation

Outputs of R :

```
res=glm(chd...,family=binomial,data=tab);summary(res)
```

$n = 468$, $p = 7$, $Y \leftarrow chd$

Coefficients :	Estimate	Std. Error	z value	$Pr(> \|z\|)$	
(Intercept)	-4.129	0.964	-4.283	1.84e-05	***
sbp	0.005	0.005	1.023	0.30643	
tobacco	0.079	0.026	3.034	0.00242	**
ldl	0.184	0.057	3.219	0.00129	**
famhistPresent	0.939	0.224	4.177	2.96e-05	***
obesity	-0.034	0.029	-1.187	0.23529	
alcohol	0.000	0.004	0.136	0.89171	
age	0.042	0.010	4.181	2.90e-05	***

Odd-ratio

Indicator used to characterized the negative or positive influence of a co-variable on the target Y.
It measures the ratio of the probability of event $Y = 1$ over $Y = 0$, when X_j increases of 1 unit $(x_j) \rightarrow (x_j) + 1$.

- For a real value variable X :

$$OR = \frac{\eta(x_j + 1)/(1 - \eta(x_j + 1))}{\eta(x_j)/(1 - \eta(_j x))} = e^{\beta_j}$$

- For a binary variable X $\{0, 1\}$:

$$OR = \frac{P(Y = 1/X_j = 1)/(1 - P(Y = 1/X_j = 1)}{P(Y = 1/X_j = 0)/(1 - P(Y = 1/X_j = 0)} = e^{\beta_j}$$

$\rightarrow$ OR < 1 means a negative influence of X_j on Y.
$\rightarrow$ OR $>$ 1means a positive influence of X_j on Y.

Confidence interval for the OR with a confidence of 95% :

Odd-ratio and confidence interval

Résultats SAS :

```
Procédure LOGISTIC (variable CHD):
Estimations par l'analyse du maximum de vraisemblance
Valeur           Erreur              Khi-2
Paramètre  DDL  estimée     type     de Wald    Pr > Khi-2
Intercept   1    3.5212     0.4160   71.6469    <.0001
age         1   -0.0641     0.00853  56.4428    <.0001

Estimations des rapports de cotes:
Valeur estimée    Intervalle de confiance
Effet    du point         de Wald à 95 %
age         0.938      0.922         0.954
```

Odd-ratio

Impact of the binary label on the coefficients

R outputs :

```
Coefficients (CODAGE Y=1 CHD=OUI/1):
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.521710   0.416031   8.465  < 2e-16 ***
age         -0.064108   0.008532  -7.513 5.76e-14 ***
exp(res2$coeff)
(Intercept)         age
 33.8422607   0.9379037


Coefficients (CODAGE Y=1 CHD=NON/0):
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.521710   0.416031  -8.465  < 2e-16 ***
age          0.064108   0.008532   7.513 5.76e-14 ***

exp(res$coeff)
(Intercept)         age
 0.02954885  1.06620758
```

Sortie R Régression logistique sur données CHD

```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)    -4.1295997  0.9641558  -4.283 1.84e-05 ***
sbp             0.0057607  0.0056326   1.023  0.30643
tobacco         0.0795256  0.0262150   3.034  0.00242 **
ldl             0.1847793  0.0574115   3.219  0.00129 **
famhistPresent  0.9391855  0.2248691   4.177 2.96e-05 ***
obesity        -0.0345434  0.0291053  -1.187  0.23529
alcohol         0.0006065  0.0044550   0.136  0.89171
age             0.0425412  0.0101749   4.181 2.90e-05 ***
Null deviance: 596.11  on 461  degrees of freedom
Residual deviance: 483.17  on 454  degrees of freedom
AIC: 499.17
```

Plan

- Applications
- Logistic regression model
- Model interprétation
- Performance criteria
- Model selection

Performances criteria

confusion matrix& co.

Matrice de confusion

- Une observation i est affectée à la classe $Y = 1$

  si $\hat{\eta}(x_i) > S$, (S : seuil, par exemple =0.5)

- Performances sur un ensemble de données étiquetées

$g(x) = \hat{y}$	$y = 0$	$y = 1$
$g(x) = 0$	diag. correcte	Faux Négatif
$g(x) = 1$	**Faux Positif**	diag. correcte
	n_0	n_1

- Notions de Performance, Erreur globale.
- Sensibilité : Capacité à diagnostiquer les $\hat{Y} = 1$ parmi les $Y = 1$
- Spécificité : Capacité à diagnostiquer les $\hat{Y} = 0$ parmi les $Y = 0$
- Faux Positifs : diagnostic $\hat{Y} = 1$ à tort.
- Faux Négatifs : diagnostic $\hat{Y} = 0$ à tort

Trouver un compromis acceptable
entre forte sensibilité et forte spécificité

Standard Error for binary classification

Reality / Decision	$y = 0$	$y = 1$
$\hat{y} = 0$	TN	FN
$\hat{y} = 1$	FP	TP

- Accuracy $= \frac{\text{TP} + \text{TN}}{\text{TP+TN+FP+FN}}$
- Recall $= \frac{\text{TP}}{\#(\text{real P})} = \frac{\text{TP}}{\text{FN+TP}}$
- Precision $= \frac{\text{TP}}{\#(\text{predicted P})} = \frac{\text{TP}}{\text{FP+TP}}$
- F-score$= 2\frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$

Rem. : Recall= sensitivity.
False-Discovery Rate (FDR)= 1-Precision.

Matrice de confusion

Performances du modèle sur base de test (2 classes)
Problématique de risque de crédit (1 : défaillance crédit).
Base de données : $n = 200$, $n_0 = 120$ $\{0\}$, $n_1 = 80$ $\{1\}$ (pb de crédit)

$g(x) = \hat{y}$	$\{0\}$	$\{1\}$	TOTAL
prédiction $\{0\}$	110	**10**	120
prédiction $\{1\}$	**10**	70	80
TOTAL	120	80	200

- Performance : $\frac{110+70}{200} = \frac{180}{200}$. Taux d' Erreur $= \frac{10+10}{200} = \frac{20}{200} = 10\%$
- Sensibilité $= 70/80$ (capacité à diag. les incidents / les incidents)
- Spécificité $= 110/120$ (capacité à reconnaître les "0" parmi les "0")
- Taux de Faux Positifs $= \frac{10}{120} = 8,33\%$ (risque diag. incident / "0")
- Taux de Faux Négatifs $= \frac{10}{80} = 12,5\%$

Plan

- Applications
- Logistic regression model
- Model interpretation
- Perfromances criteria
- Model selection

PERFORMANCE CRITERIA

ROC curve