Régression logistique
[Logistic regression]

Titouan Vayer,Laetitia Chapel

STID Vannes
STATISTIQUE ET INFORMATIQUE DÉCISIONNELLE

STID 2

2019 – 2020

Sommaire

Sommaire

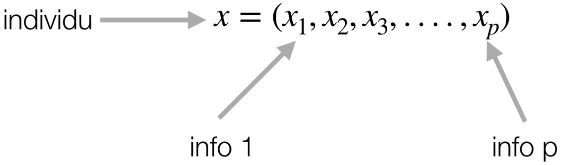
Introduction : la classification supervisée binaire

Organisation du cours

- 2 CMs
- 6 TDs : papier & ordinateur
- Evaluation sur table papier crayon 1h30 avec anti-sèche autorisé

Objectif de la classification supervisée binaire

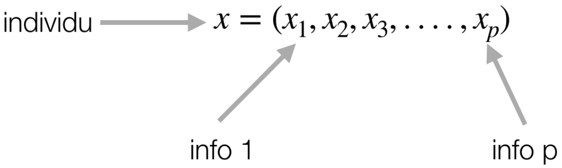
Apprendre à un ordinateur **à classer** selon **deux catégories** des individus x en fonction de leurs **données descriptives** $(x_1, ..., x_p)$



Introduction : la classification supervisée binaire

Objectif de la classification supervisée binaire

Apprendre à un ordinateur **à classer** selon **deux catégories** des individus x en fonction de leurs **données descriptives** $(x_1, ..., x_p)$



Par exemple :

- Donner un crédit en fonction de données d'un individu (âge,csp,...)
- Détecter une maladie en fonction de données d'une image
- Proposer un médicament adapté en fonction des données médicales d'un individu
- Détecter un obstacle pour la voiture autonome
- Trier des CVs

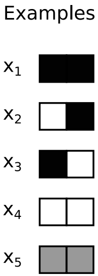
Introduction : la classification supervisée binaire

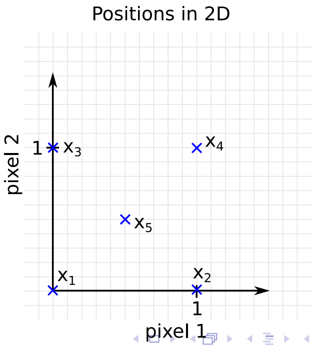
Comment apprendre f ?

- Généralement on dispose d'un ensemble de plusieurs individus (''un dataset''), de leurs données descriptives et de leur catégorie respective (''la classe de l'individu'')
- On essaye d'apprendre une ''règle'' de classification qui sépare nos données selon les deux catégories
- La règle est une fonction f qui s'appelle un classifeur

Comment apprendre f ?

- Généralement on dispose d'un ensemble de plusieurs individus ("un dataset"), de leurs données descriptives et de leur catégorie respective ("la classe de l'individu")
- On essaye d'apprendre une "règle" de classification qui sépare nos données selon les deux catégories
- La règle est une fonction f qui s'appelle un classifeur

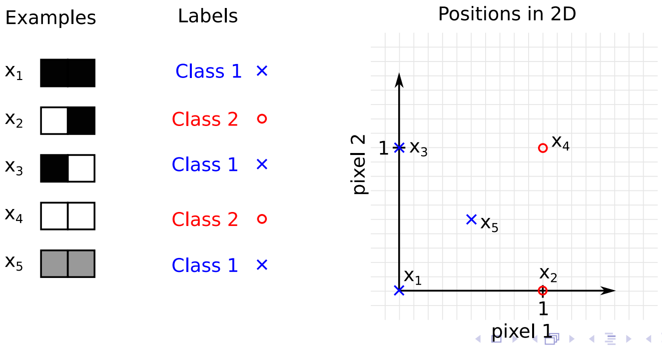




Introduction : la classification supervisée binaire

Comment apprendre f ?

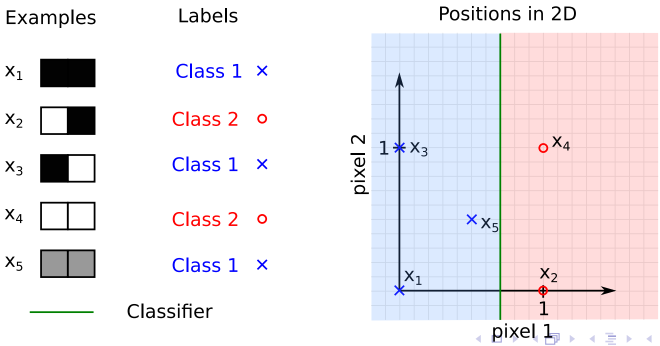
- Généralement on dispose d'un ensemble de plusieurs individus ("un dataset"), de leurs données descriptives et de leur catégorie respective ("la classe de l'individu")
- On essaye d'apprendre une "règle" de classification qui sépare nos données selon les deux catégories
- La règle est une fonction f qui s'appelle un classifeur



Introduction : la classification supervisée binaire

Comment apprendre f ?

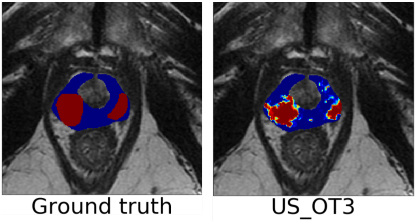
- Généralement on dispose d'un ensemble de plusieurs individus ("un dataset"), de leurs données descriptives et de leur catégorie respective ("la classe de l'individu")
- On essaye d'apprendre une "règle" de classification qui sépare nos données selon les deux catégories
- La règle est une fonction f qui s'appelle un classifeur



Introduction : la classification supervisée binaire

Comment apprendre f ?

- Généralement on dispose d'un ensemble de plusieurs individus (''un dataset''), de leurs données descriptives et de leur catégorie respective (''la classe de l'individu'')
- On essaye d'apprendre une ''règle'' de classification qui sépare nos données selon les deux catégories
- La règle est une fonction f qui s'appelle un classifeur



Introduction : la classification supervisée binaire

La classification supervisée binaire

L'objectif de la classification supervisée binaire est :

- Apprendre à un ordinateur (avec un algorithme codé dans un langage informatique comme R)
- A trouver une règle de classification (f)
- En fonction de données d'individus (x et leur classe)
- De sorte que la règle de décision satisfasse un certain critère (qu'elle classifie au mieux tous nos points)

Introduction : la classification supervisée binaire

- Entrée : un ensemble de N points décrits par p attributs, où $\{\mathbf{x}_i\}_{i=1\cdots p}$ contient les valeurs prises par la $i^{\text{ème}}$ variable pour l'ensemble des individus

  $\mathbf{x}_i = \begin{bmatrix} x_i^1 \\ x_i^2 \\ \cdots \\ x_i^N \end{bmatrix}$ et on note $\mathbf{x}^n$ les attributs du $n^{\text{ème}}$ individu

  $\mathbf{x}^n = [x_1^n, x_2^n, \cdots, x_p^n]$

- Des classes $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_N \end{bmatrix} \in \mathcal{Y}$, avec y_n la classe du $n^{\text{ème}}$ individu

Introduction : la classification supervisée binaire

- Entrée : un ensemble de N points décrits par p attributs, où $\{\mathbf{x}_i\}_{i=1\cdots p}$ contient les valeurs prises par la $i^{\text{ème}}$ variable pour l'ensemble des individus
  $\mathbf{x}_i = \begin{bmatrix} x_i^1 \\ x_i^2 \\ \cdots \\ x_i^N \end{bmatrix}$ et on note $\mathbf{x}^n$ les attributs du $n^{\text{ème}}$ individu
  $\mathbf{x}^n = [x_1^n, x_2^n, \cdots, x_p^n]$
- Des classes $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_N \end{bmatrix} \in \mathcal{Y}$, avec y_n la classe du $n^{\text{ème}}$ individu
- Apprentissage supervisé : la sortie désirée est connue
  But : trouver une fonction f telle que

$$f([\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p]) + \epsilon = \mathbf{y}$$

.

  - si $\mathcal{Y} = \mathbb{R}$: problème de régression
  - si $\mathcal{Y} \in \mathcal{S}$, avec $\mathcal{S}$ un ensemble fini : problème de classification
  - si $\mathcal{S} = \{0, 1\}$ contient 2 éléments : classification binaire

Sommaire

Objectifs du cours

- Comprendre le principe de la régression logistique
- Etre capable de mettre en oeuvre une régression logistique avec R
- Savoir construire un modèle de régression logistique
- Etre capable d'interpréter les résultats d'une régression logistique

Sommaire

Le modèle de régression linéaire

$$\mathbf{y} = f(\mathbf{X}) + \boldsymbol{\epsilon} = \beta_0 + \beta_1 \mathbf{x_1} + \cdots + \beta_p \mathbf{x_p} + \boldsymbol{\epsilon}$$

et pour une observation n :

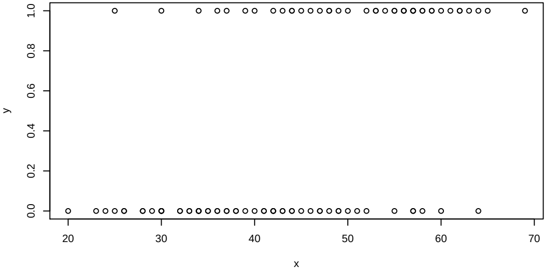
$$y_n = \beta_0 + \beta_1 x_1^n + \cdots + \beta_p x_p^n + \epsilon_n \in \mathbb{R}$$

- Estimations $b_0, b_1, \cdots, b_p$ des coefficients du modèle $\beta_0, \beta_1, \cdots, \beta_p$ par la méthode des MC
- Test sur les coefficients
- Mesure de qualité du modèle

Le modèle de régression logistique

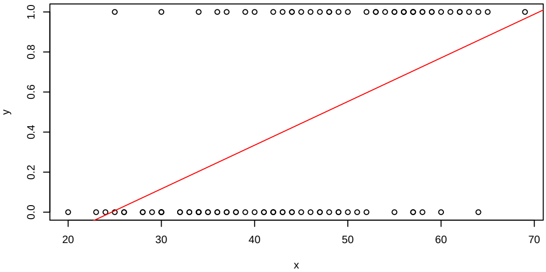
Régression logistique

- Modéliser $\mathbf{y} \in \{0, 1\}$



Régression logistique

- Modéliser $\mathbf{y} \in \{0, 1\}$



- Le modèle linéaire ne convient pas !

Le modèle de régression logistique

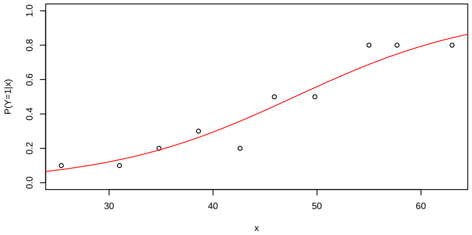
Régression logistique

- Dans le cadre de la régression logistique on cherche à modéliser les probabilités :

$$\pi(\mathbf{x}^n) = \mathbb{P}\left(y_n = 1 \,|\, \mathbf{x}^n\right) \text{ ou } 1 - \pi(\mathbf{x}^n) = \mathbb{P}\left(y_n = 0 \,|\, \mathbf{x}^n\right)$$

  et pour l'ensemble des individus

$$\pi(\mathbf{X}) = \mathbb{P}\left(\mathbf{y} = 1 \,|\, \mathbf{X}\right) \text{ ou } 1 - \pi(\mathbf{X}) = \mathbb{P}\left(\mathbf{y} = 0 \,|\, \mathbf{X}\right)$$



- Une fois qu'on connait la probabilité qu'a un individu d'appartenir à une classe il est facile de deviner à quelle classe il appartient.

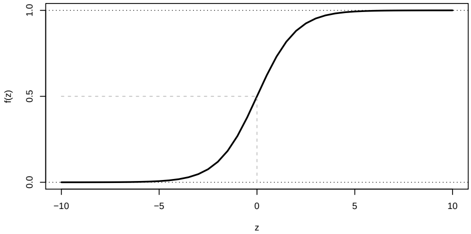
Régression logistique

Comment modéliser $\pi(x^n)$?

- La régression logistique est basée sur la fonction logistique

$$f(z) = \frac{\exp(z)}{1 + \exp(z)} = \frac{1}{1 + \exp(-z)}$$

avec $f(z) \in [0, 1]$ et $z \in \mathbb{R}$



Le modèle

- Le modèle de régression logistique s'écrit :

$$\pi(\mathbf{X}) = \frac{\exp(\beta_0 + \beta_1 \mathbf{x}_1 + \cdots + \beta_p \mathbf{x}_p)}{1 + \exp(\beta_0 + \beta_1 \mathbf{x}_1 + \cdots + \beta_p \mathbf{x}_p)}$$

  avec $\pi(\mathbf{x}^n) \in [0,1]$ et $\beta_0 + \beta_1 x_1^n + \cdots + \beta_p x_p^n \in \mathbb{R}$
- ou de façon équivalente :

$$\text{logit}(\pi(\mathbf{X})) = \log\left(\frac{\pi(\mathbf{X})}{1 - \pi(\mathbf{X})}\right) = \beta_0 + \beta_1 \mathbf{x}_1 + \cdots + \beta_p \mathbf{x}_p$$

Le modèle

- Le modèle de régression logistique s'écrit :

$$\pi(\mathbf{X}) = \frac{\exp(\beta_0 + \beta_1 \mathbf{x}_1 + \cdots + \beta_p \mathbf{x}_p)}{1 + \exp(\beta_0 + \beta_1 \mathbf{x}_1 + \cdots + \beta_p \mathbf{x}_p)}$$

  avec $\pi(\mathbf{x}^n) \in [0,1]$ et $\beta_0 + \beta_1 x_1^n + \cdots + \beta_p x_p^n \in \mathbb{R}$

- ou de façon équivalente :

$$\text{logit}(\pi(\mathbf{X})) = \log\left(\frac{\pi(\mathbf{X})}{1 - \pi(\mathbf{X})}\right) = \beta_0 + \beta_1 \mathbf{x}_1 + \cdots + \beta_p \mathbf{x}_p$$

- avec $\mathbf{x}_i$ variables quantitatives ou binaires représentant les données des individus.

Les variables explicatives

- **Variable explicative qualitative à 2 modalités** : les modalités sont recodées $x_i^n \in \{0, 1\}$, la modalité 0 étant appelée *modalité de référence*.
- **Variable explicative qualitative à** $m > 2$ **modalités** : on créé $m - 1$ *variables design* (indicatrices associées à chaque modalité).
- **Variable explicative quantitative** : on vérifie l'hypothèse de *linéarité du logit*.

Le modèle de régression logistique

Hypothèse de linéarité du logit

- Lorsqu'une variable passe de la valeur x_1 à $x_1 + 1$, la valeur $\text{logit}(\pi(x))$ augmente de β_1, quelle que soit la valeur de x_1 $\Rightarrow$ le logit est linéaire.
- On doit vérifier cette hypothèse pour pouvoir intégrer une variable quantitative dans le modèle, et la mettre en classe sinon.
- Pour vérifier l'hypothèse, on peut notamment mettre la variable en classe, puis tracer l'évolution en fonction du logit.

Sommaire

Estimation des paramètres du modèle

Comment trouver les paramètre $\beta_0, ..., \beta_p$?

L'objectif est de trouver $\beta_0, ..., \beta_p$ de sorte que notre modèle "colle" le mieux à nos données.

Estimation des paramètres du modèle

Comment trouver les paramètre $\beta_0, ..., \beta_p$?

L'objectif est de trouver $\beta_0, ..., \beta_p$ de sorte que notre modèle "colle" le mieux à nos données.

- On ne peut pas utiliser la méthode des moindres carrés (on ne modélise pas directement y_n) : lorsque $y_n \in \mathbb{R}$, on peut écrire

$$\min \sum_{n=1}^{N} e_n^2 = \min \sum_{n=1}^{N} (y_n - \hat{y}_n)^2 = \min \sum_{n=1}^{N} (y_n - (\beta_0 + \beta_1 x_1^n + \cdots + \beta_p x_p^n))^2$$

- On utilise la méthode du maximum de vraisemblance

$$\max \prod_{n=1}^{N} \pi(x^n)^{y_n} \times (1 - \pi(x^n))^{1-y_n}$$

Estimation des paramètres du modèle

Comment trouver les paramètre $\beta_0, ..., \beta_p$?

L'objectif est de trouver $\beta_0, ..., \beta_p$ de sorte que notre modèle "colle" le mieux à nos données.

- On ne peut pas utiliser la méthode des moindres carrés (on ne modélise pas directement y_n) : lorsque $y_n \in \mathbb{R}$, on peut écrire

$$\min \sum_{n=1}^{N} e_n^2 = \min \sum_{n=1}^{N} (y_n - \hat{y}_n)^2 = \min \sum_{n=1}^{N} (y_n - (\beta_0 + \beta_1 x_1^n + \cdots + \beta_p x_p^n))^2$$

- On utilise la méthode du maximum de vraisemblance

$$\max \prod_{n=1}^{N} \pi(x^n)^{y_n} \times (1 - \pi(x^n))^{1-y_n}$$

1. si $y_n = 1$, on veut que $\pi(x^n)$ soit proche de 1 $\Rightarrow \pi(x^n)^{y_n}$ proche de 1 et $(1 - \pi(x^n))^{1-y_n} = 1$
2. si $y_n = 0$, on veut que $\pi(x^n)$ soit proche de 0 $\Rightarrow \pi(x^n)^{y_n} = 1$ et $(1 - \pi(x^n))^{1-y_n}$ proche de 1

Estimation des paramètres du modèle

- On cherche donc des estimations $b_0, b_1, \cdots, b_p$ des paramètres inconnus $\beta_0, \beta_1, \cdots, \beta_p$ telles que la vraisemblance $\mathcal{L}$ soit maximimum :

$$\max \mathcal{L} = \max \prod_{n=1}^{N} \pi(x^n)^{y_n} \times (1 - \pi(x^n))^{1-y_n}$$

avec

$$\pi(x^n) = \frac{\exp(\beta_0 + \beta_1 x_1^n + \cdots + \beta_p x_p^n)}{1 + \exp(\beta_0 + \beta_1 x_1^n + \cdots + \beta_p x_p^n)}$$

- ce qui est équivalent à minimiser la déviance $-2 \times \log(\mathcal{L})$

$$\min -2 \times \log(\mathcal{L})$$

Estimation des paramètres du modèle

- On cherche donc des estimations $b_0, b_1, \cdots, b_p$ des paramètres inconnus $\beta_0, \beta_1, \cdots, \beta_p$ telles que la vraisemblance $\mathcal{L}$ soit maximimum :

$$\max \mathcal{L} = \max \prod_{n=1}^{N} \pi(x^n)^{y_n} \times (1 - \pi(x^n))^{1-y_n}$$

avec

$$\pi(x^n) = \frac{\exp(\beta_0 + \beta_1 x_1^n + \cdots + \beta_p x_p^n)}{1 + \exp(\beta_0 + \beta_1 x_1^n + \cdots + \beta_p x_p^n)}$$

- ce qui est équivalent à minimiser la déviance $-2 \times \log(\mathcal{L})$

$$\min -2 \times \log(\mathcal{L})$$

- On utilise des méthodes d'optimisation pour résoudre le problème

Estimation des paramètres du modèle

- On cherche donc des estimations $b_0, b_1, \cdots, b_p$ des paramètres inconnus $\beta_0, \beta_1, \cdots, \beta_p$ telles que la vraisemblance $\mathcal{L}$ soit maximimum :

$$\max \mathcal{L} = \max \prod_{n=1}^{N} \pi(x^n)^{y_n} \times (1 - \pi(x^n))^{1-y_n}$$

avec

$$\pi(x^n) = \frac{\exp(\beta_0 + \beta_1 x_1^n + \cdots + \beta_p x_p^n)}{1 + \exp(\beta_0 + \beta_1 x_1^n + \cdots + \beta_p x_p^n)}$$

- ce qui est équivalent à minimiser la déviance $-2 \times \log(\mathcal{L})$

$$\min -2 \times \log(\mathcal{L})$$

- On utilise des méthodes d'optimisation pour résoudre le problème
- *Note* : pour N fixé, le modèle 1 sera meilleur que le modèle 2 si $\mathcal{L}_1 > \mathcal{L}_2$

Test du rapport de vraisemblance : test de significativité globale du modèle

- On cherche à savoir si il y a un "lien" entre au moins une variable explicative $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p$ et la variable à expliquer Y

$$\mathcal{H}_0 : \beta_1 = \cdots = \beta_p = 0$$

$$\mathcal{H}_1 : \exists i \text{ tel que } \beta_i \neq 0$$

Estimation des paramètres du modèle

Test du rapport de vraisemblance : test de significativité globale du modèle

- On cherche à savoir si il y a un "lien" entre au moins une variable explicative $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p$ et la variable à expliquer Y

$$\mathcal{H}_0 : \beta_1 = \cdots = \beta_p = 0$$

$$\mathcal{H}_1 : \exists i \text{ tel que } \beta_i \neq 0$$

- Sous $\mathcal{H}_0$,

$$D = -2 \log \left(\frac{\mathcal{L}_0}{\mathcal{L}_p} \right) \sim \chi^2_p$$

Test de Wald : test de significativité individuelle des variables

- On teste la significativité individuelle d'une variable $\mathbf{x}_i$

$$\mathcal{H}_0 : \beta_i = 0$$

$$\mathcal{H}_1 : \beta_i \neq 0$$

Test de Wald : test de significativité individuelle des variables

- On teste la significativité individuelle d'une variable $\mathbf{x}_i$

$$\mathcal{H}_0 : \beta_i = 0$$

$$\mathcal{H}_1 : \beta_i \neq 0$$

- Sous $\mathcal{H}_0$,

$$W^2 = \left(\frac{b_i^2}{V\hat{A}R(b_i)} \right) \sim \chi_1^2$$

where $V\hat{A}R(b_i)$ is an estimation of the variance of b_i

Sommaire

Odds et odds-ratio

odds

- La régression logistique est basée sur l'estimation d'un odds ou cote

$$\text{odds} = \frac{\mathbb{P}\left(y_n = 1 \,|\, \mathbf{x}^n\right)}{\mathbb{P}\left(y_n = 0 \,|\, \mathbf{x}^n\right)} = \frac{\pi(\mathbf{x}^n)}{1 - \pi(\mathbf{x}^n)}$$

Exemple

En Bretagne, il y a 25% de chances de pleuvoir demain. En Irlande, il y a 75 % de chances de pleuvoir demain.
odds(Bretagne) ? odds(Irlande) ?

Odds et odds-ratio

Interprétation de b_0

- Soit un modèle qui ne contient que l'intercept (pas de variable explicative)

$$\text{logit}(\pi(\mathbf{x}^n)) = \beta_0$$

- Une estimation b_0 du paramètre inconnu β_0 correspond au logarithme de l'odds

$$b_0 = \log \frac{\pi(\mathbf{x}^n)}{1 - \pi(\mathbf{x}^n)}$$

et on a donc

$$\exp(b_0) = \frac{\pi(\mathbf{x}^n)}{1 - \pi(\mathbf{x}^n)}$$

- $\exp(b_0)$ correspond donc au nombre de fois de chances de plus que l'on a d'observer $y = 1$ par rapport à $y = 0$.

Odds et odds-ratio

Odds-ratio : cas d'une variable explicative binaire $x \in \{0,1\}$

- Odds-ratio : rapport entre 2 odds
- Soit une variable x qui prend 2 modalités : 0 ou 1

$$OR(x) = \frac{\pi(1)}{1-\pi(1)} \Big/ \frac{\pi(0)}{1-\pi(0)}$$

- Indique ainsi quelle est la quantité de chance en plus d'être $y=1$ dans le groupe $x=1$ par rapport au groupe $x=0$.

Exemple

En Bretagne, il y a 25% de chances de pleuvoir demain. En Irlande, il y a 75 % de chances de pleuvoir demain.
odds-ratio(Irlande vs Bretagne) ?

Odds et odds-ratio

Odds-ratio : cas d'une variable explicative binaire $x \in \{0, 1\}$

- Dans ce cas, on a (preuve...)

$$OR(x) = \exp(\beta_1)$$

- Vocabulaire :
  - si OR > 1 : facteur de risque
  - si OR < 1 : facteur de protection
- On peut également calculer des IC autour des OR

Odds et odds-ratio

Odds-ratio : cas d'une variable explicative continue $x \in \mathbb{R}$

- Dans ce cas, on a (preuve...)

$$OR(x) = \exp(\beta_1)$$

  qui correspond à l'évolution de la chance d'être $y = 1$ lorsque la variable x augmente de 1 unité (passe de la valeur a à la valeur $a + 1$)
- Attention à l'hypothèse de linéarité du logit !

Sommaire

Evaluation et sélection du “meilleur” modèle

- Approche traditionnelle : chercher le modèle le plus parsimonieux qui explique au mieux les données
- Idée générale : on cherche le modèle qui maximise la vraisemblance. MAIS, la vraisemblance augmente avec la complexité du modèle : on cherche donc un compromis entre la qualité de l’ajustement et la complexité du modèle.
- 2 critères classiques :
  - Critère d’Akaike
    $$AIC = -2\mathcal{L} + 2p$$
  - Critère Bayes Information Criterion
    $$BIC = -2\mathcal{L} + p\log(N)$$

Evaluation et sélection du “meilleur” modèle

Méthodes de sélection

- Il est coûteux voire impossible de tester tous les modèles et choisir celui qui minimise le critère BIC ou AIC : on préfère des méthodes pas à pas
- Méthode ascendante : on ajoute à chaque itération une variable dans le modèle – forward
- Méthode descendante : on supprime à chaque itération une variable du modèle – backward
- Méthode stepwise

Evaluation et sélection du “meilleur” modèle

Evaluation d’un modèle

- On peut également comparer des modèles en fonction de leur performance en prédiction
- parcours école : cf module introduction à l’apprentissage automatique (calcul de l’estimation du risque réel, de l’aire sous la courbe ROC, TVP, TVN)
- parcours STID : cf module classification