Big Data

Introduction

benoit.lardeux@isen-ouest.yncrea.fr

Inspiré des notes de cours de M. Saumard, ISEN Brest

Plan du cours

- Introduction:

  1- Corrélations

  2- L'apprentissage statistique

- Régression linéaire

- Régression logistique

- Analyse en composantes principales

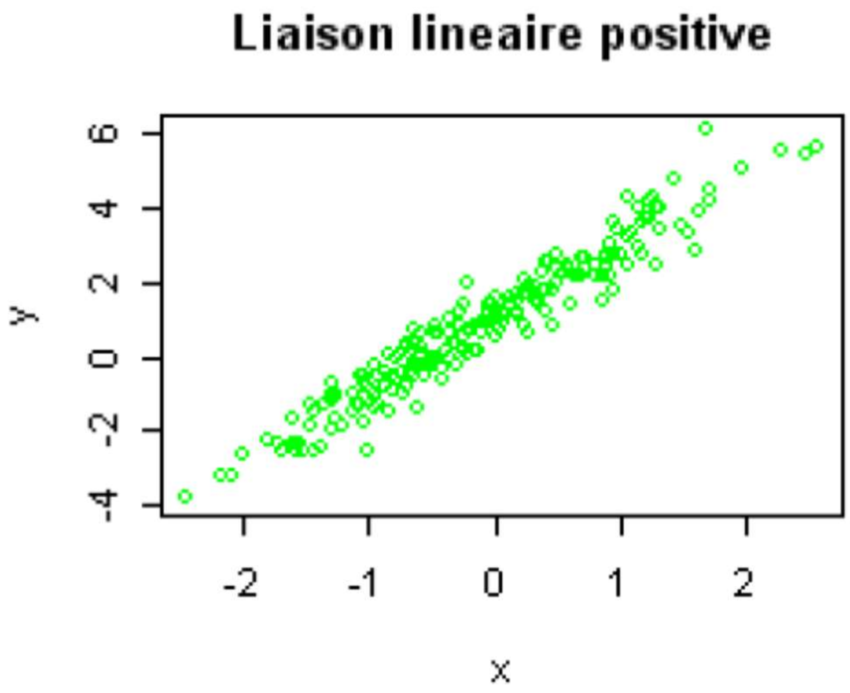
Corrélation

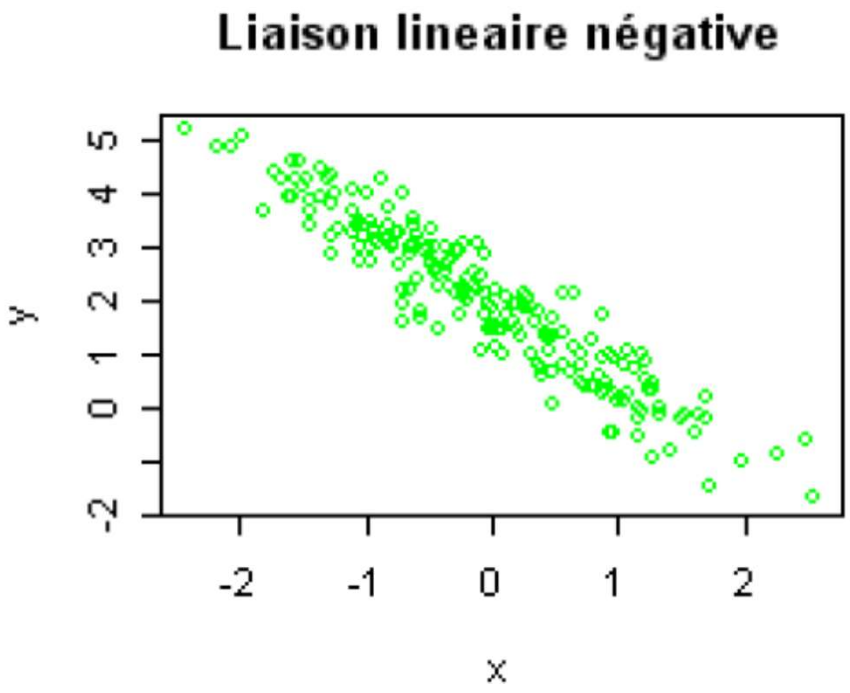
ISEN ALL IS DIGITAL! OUEST | yncréa

Corrélation

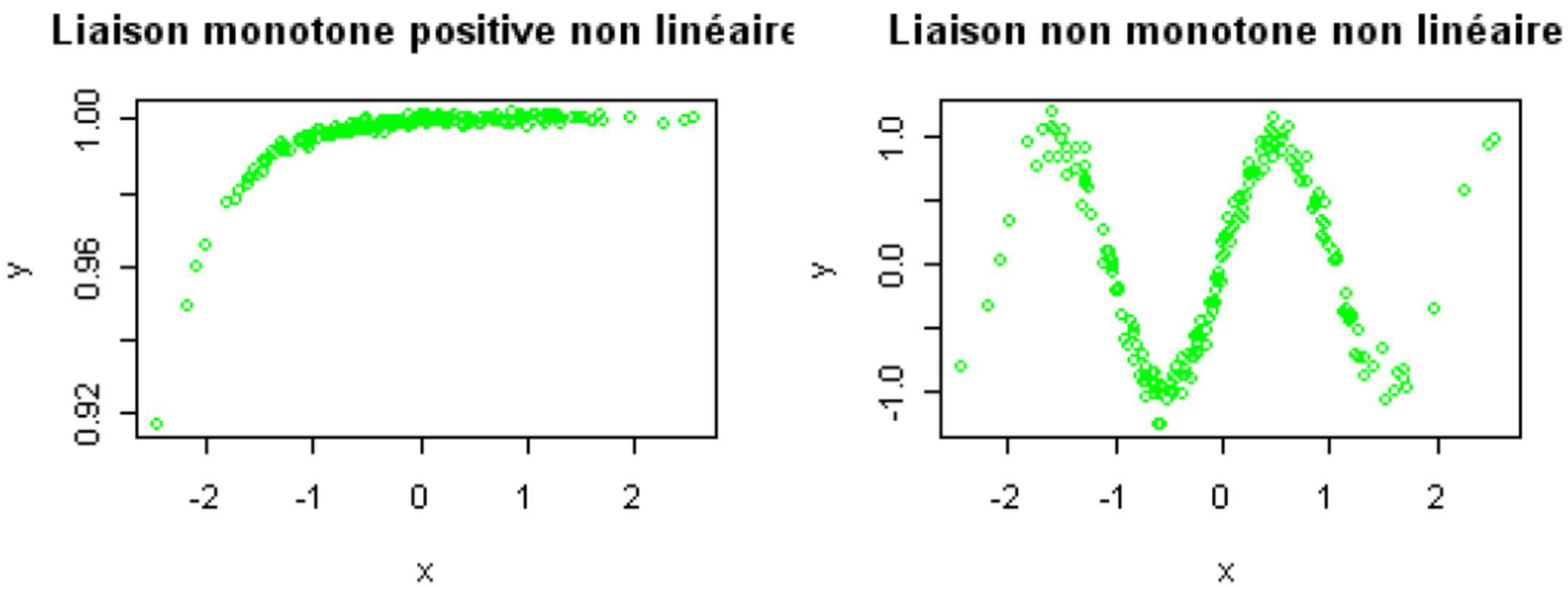
- Objectif: analyser la liaison
- Soient X et Y deux grandeurs statistiques quantitatives observées. On souhaite
  - Déterminer s'il existe une relation entre X et Y
  - Caractériser la forme de la liaison (la relation) entre X et Y (positive ou négative, linéaire ou non linéaire, monotone ou non monotone)
  - Tester si la liaison est statistiquement significative
  - Quantifier l'intensité de la liaison
  - Valider la liaison identifiée. N'est-elle pas le fruit d'un simple artefact ou le produit d'autres informations sous-jacentes dans les données?

Exemples de liaisons linéaires

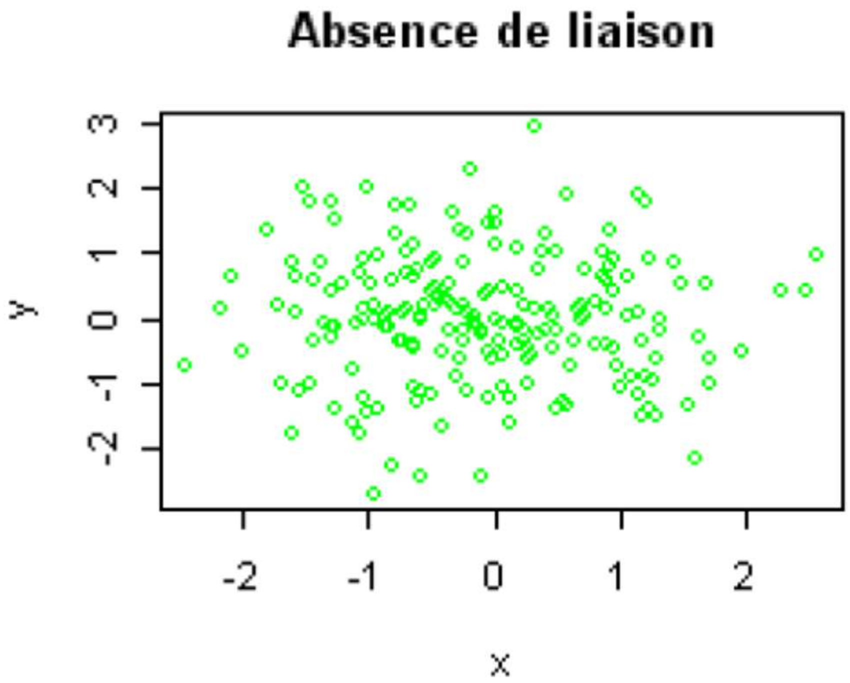


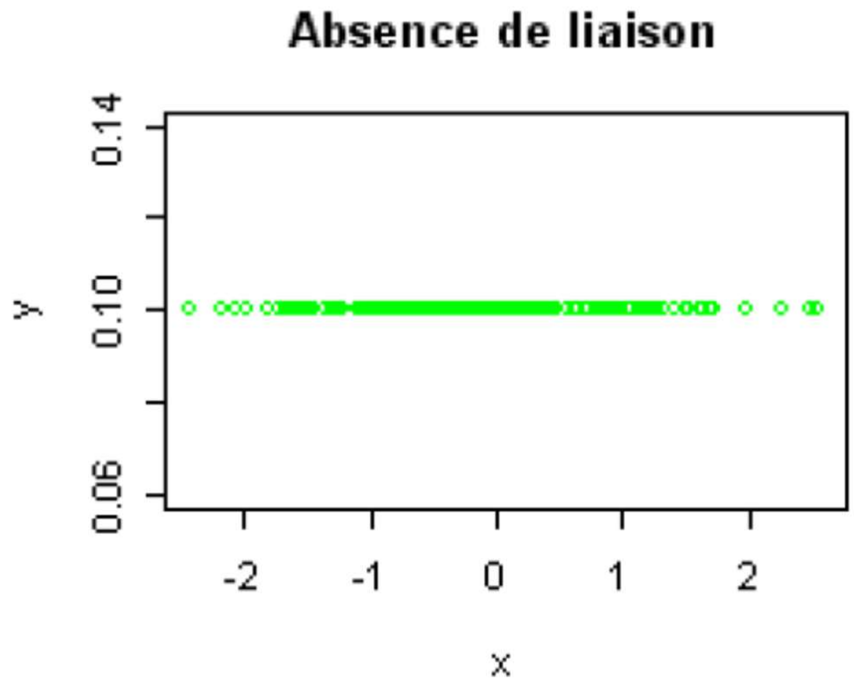


Exemples de liaisons non-linéaires



Absence de liaisons





ISEN ALL IS DIGITAL! OUEST | yncréa

Covariance

- Objectif de la covariance
  - Quantifier la liaison entre deux variables X et Y
  - De manière à mettre en évidence le sens de la liaison
  - Et son intensité

- Définition de la covariance
  - Soient X et Y deux variables
    - $Cov(X,Y) = \mathbb{E}[(\mathrm{X} - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$
    - $Cov(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$

Covariance: Interprétation

- On peut maintenant quantifier le sens de la liaison:
  - $Cov(X, Y) > 0$: la relation est positive, c'est-à-dire lorsque X est plus grand que son espérance, Y a tendance à l'être également
  - $Cov(X, Y) = 0$: absence de relation monotone
  - $Cov(X, Y) < 0$: la relation est négative, c'est-à-dire lorsque X est plus grand que son espérance, Y a tendance à être plus petit que sa propre espérance

Covariance: propriétés

- Symétrie
  - $Cov(X,Y) = Cov(Y,X)$
- Distributivité
  - $Cov(X,Y+Z) = Cov(X,Y) + Cov(X,Z)$
- Covariance avec une constante
  - $Cov(X,a) = 0$
- Covariance avec une variable transformée (transformation affine)
  - $Cov(X,a+bY) = bCov(X,Y)$
- Variance de la somme de deux variables aléatoires
  - $\mathbb{V}(X+Y) = \mathbb{V}(X) + \mathbb{V}(Y) + 2Cov(X,Y)$
- Covariance de deux variables indépendantes
  - X, Y indépendants => $Cov(X,Y) = 0$

Estimation de la covariance

- Définition (Covariance empirique)
  - Sur un échantillon de taille n,
    - $s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$
    - Où $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Estimation de la covariance

- La covariance empirique est un estimateur biaisé de la covariance
  - $\mathbb{E}[s_{xy}] = \frac{n-1}{n} Cov(X, Y)$

ISEN ALL IS DIGITAL! OUEST | yncréa

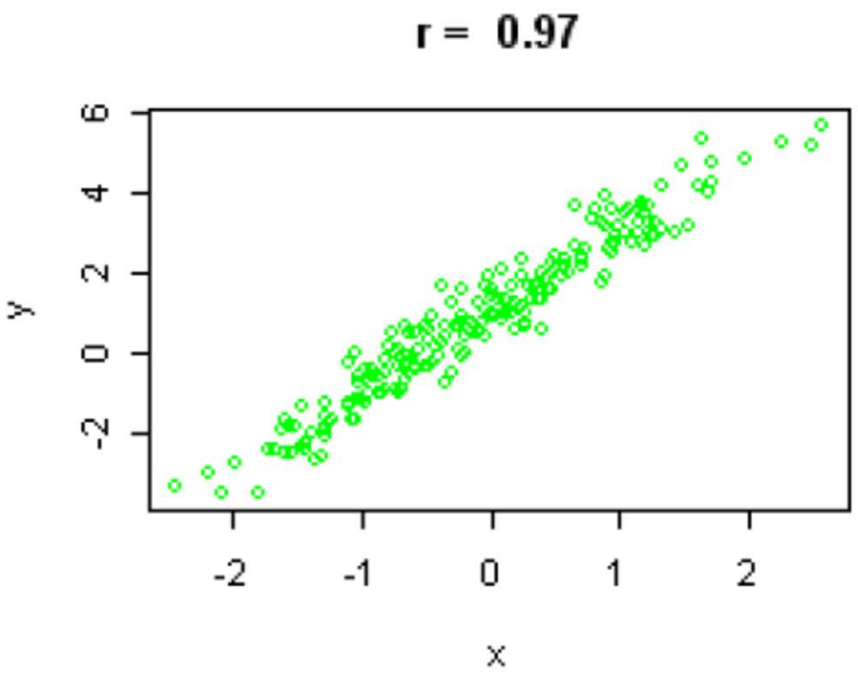
Coefficient de corrélation de Pearson

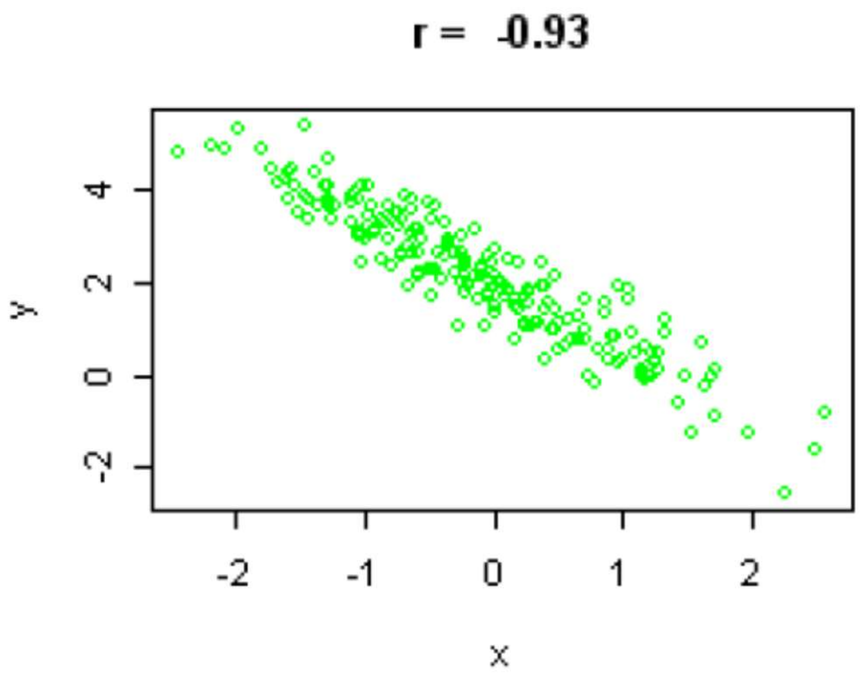
- Définition (coefficient de corrélation)
  - $\rho_{xy} = \frac{Cov(X,Y)}{\sqrt{\mathbb{V}(X)\mathbb{V}(Y)}}$

Coefficient de corrélation de Pearson: propriétés

- ρ_{xy} est de même signe que la covariance (avec les mêmes interprétations)
- X et Y sont indépendants, alors $\rho_{xy} = 0$. (réciproque fausse en général)
- Lorsque le couple de variables (X,Y) suit une loi normale bi-variée, et uniquement dans ce cas là, nous avons l'équivalence $\rho_{xy} = 0 \Leftrightarrow$ X et Y sont indépendants
- Le coefficient de corrélation constitue une mesure de l'intensité de liaison entre 2 variables. Il peut être égal à zéro alors qu'il existe une liaison fonctionnelle entre les variables. C'est le cas lorsque la liaison est non monotone
- $\rho_{xx} = 1$

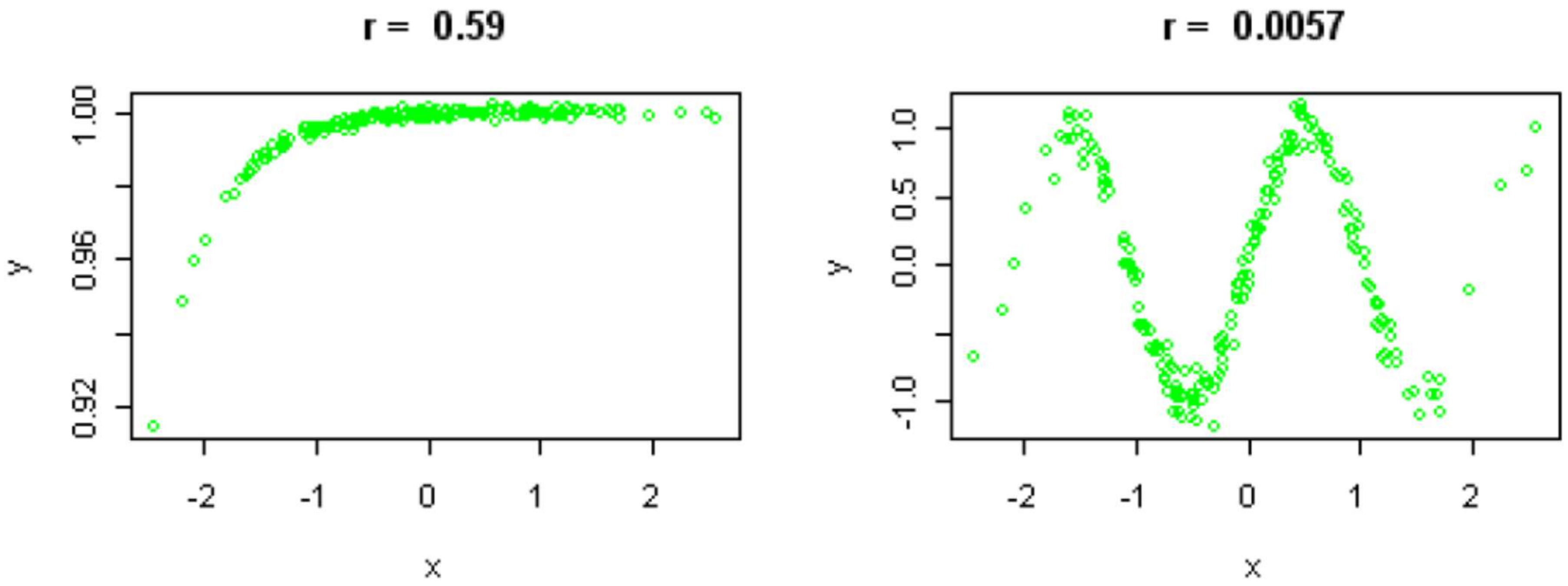
Corrélation: liaisons linéaires



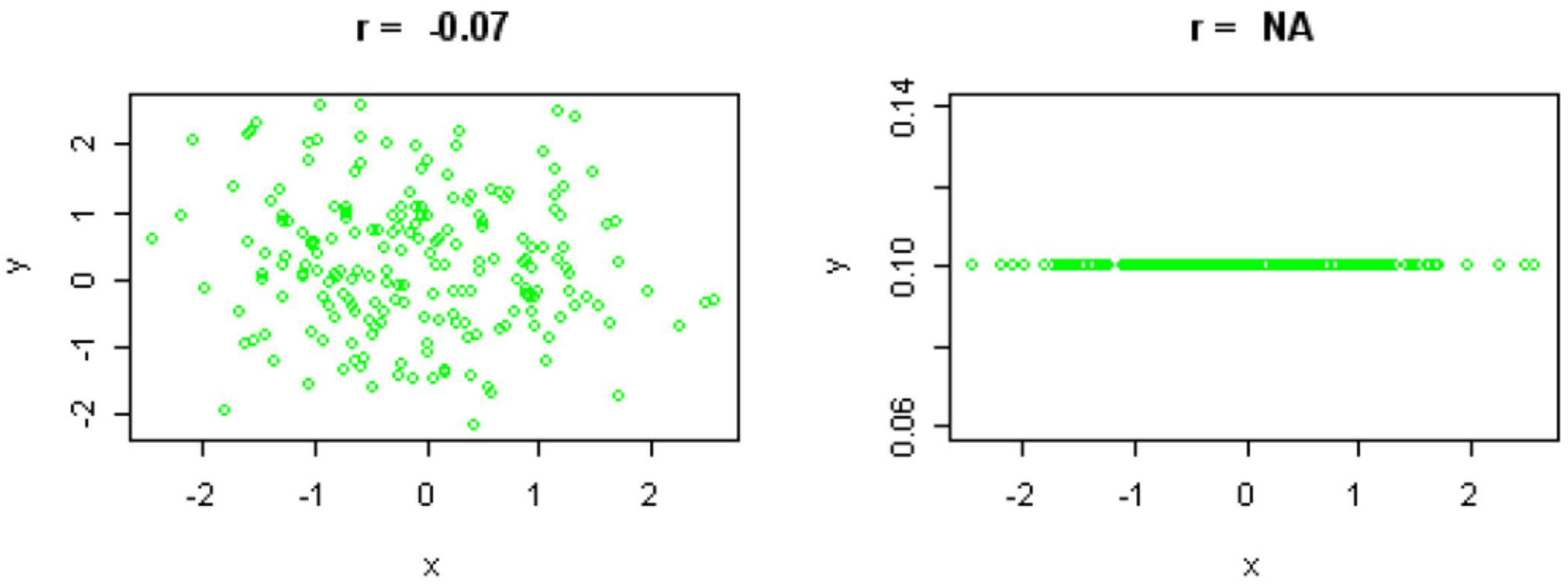


ISEN ALL IS DIGITAL! OUEST | yncréa

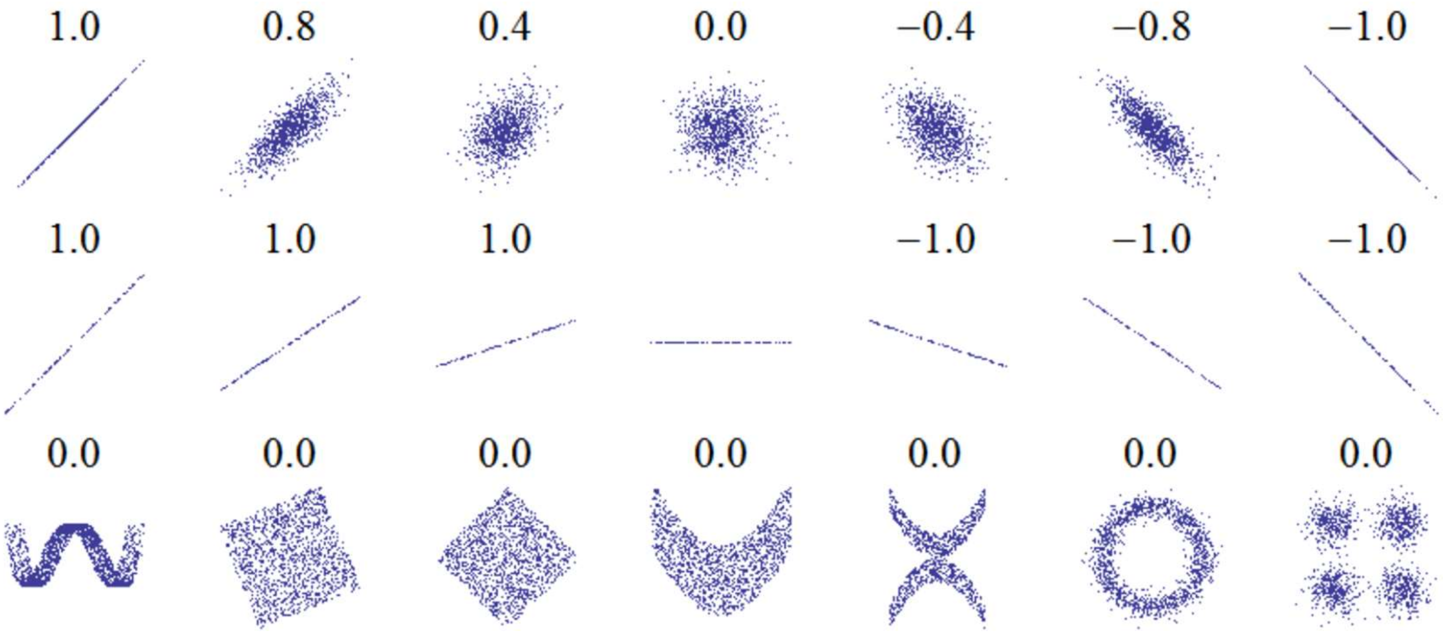
Corrélation: liaisons non linéaires



Corrélation: absence de liaisons



Exemples de corrélation (wikipedia)



L’apprentissage statistique

Qu'est ce que l'apprentissage statistique?

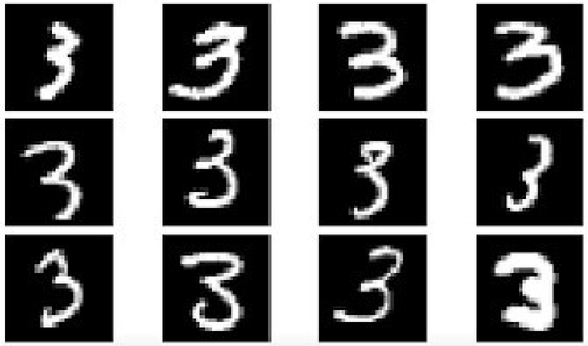
- Exemple:

Problème de reconnaissance automatique des chiffres manuscrits

Qu’est ce que l’apprentissage statistique?

- Solution:

Apprendre à partir d’exemples



- Propriété attendue:
  - Capacité à généraliser sur de nouvelles données

Exemples de questions pouvant être traitées par apprentissage statistique

- Quels sont les gènes impliqués dans une maladie?
- Peut-on prévoir un taux de pollution en fonction de conditions météo?
- Quel pourrait-être le prix d'une maison en fonction de ces caractéristiques?
- Peut-on prévoir les défaillances d'un procédé industriel?

- L'objectif dans tous ces exemples est de minimiser une erreur de prévision ou risque

Apprentissage supervisé: bases mathématiques

- Soit une observation X appelée prédicteur (ou covariable, feature)
- On lui associe une autre variable Y qui est la variable à expliquer, prédire

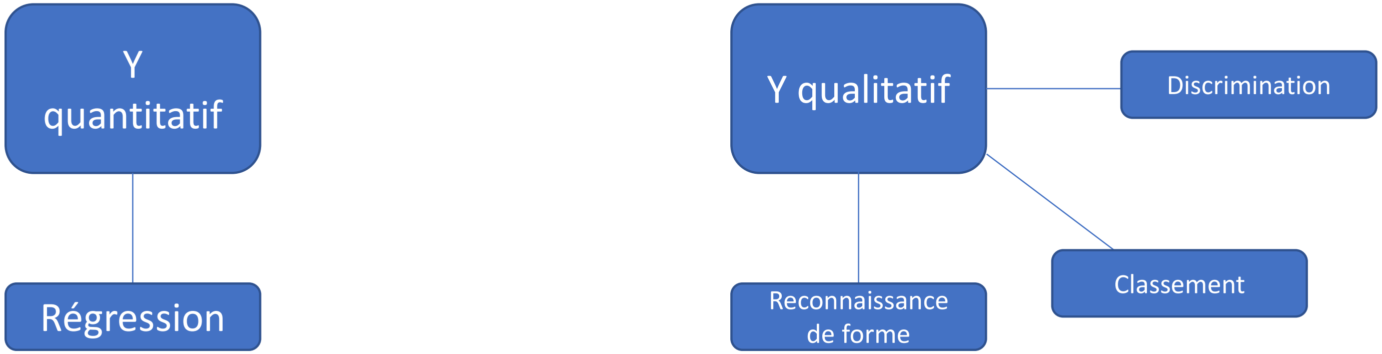
$$Y = f(X) + \varepsilon$$

- Objectif: Trouver une fonction f optimale, au sens d'un critère à définir, qui reproduit aux mieux la variable Y ayant observé X.
- ε est l'erreur associé au modèle (ou erreur de mesure)

Mise en place du problème

- Echantillon d'apprentissage: $D_{app} = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$

- Avec $X_i \in \boldsymbol{\mathcal{X}}$, où $\boldsymbol{\mathcal{X}}$ est quelconque, en général $\mathbb{R}^p$
- Et $Y_i \in \boldsymbol{\mathcal{Y}}$, où Y peut être qualitatif (c'est-à-dire prend des valeurs comme {Homme, Femme} ou {Vert, Jaune, Rouge} ou {0,1} ou quantitatif (c'est-à-dire $\mathbb{R}^p$)
- Les (X_i, Y_i) sont des variables aléatoires indépendantes identiquement distribuées (iid)



ISEN ALL IS DIGITAL! OUEST | yncréa

Fonction de coût

- Définition: (fonction de coût, appelée aussi de perte)

Une fonction $c: \boldsymbol{\mathcal{Y}} \times \boldsymbol{\mathcal{Y}} \longrightarrow \mathbb{R}$ est une fonction de coût si $c(y, y) = 0, \forall y \in \boldsymbol{\mathcal{Y}}$ et $c(y, y') > 0$ pour $y \neq y'$

- Exemples de fonction de coût:

1- Perte quadratique $\quad c(y, y') = |y - y'|^2$

2- Perte L^p avec $p \geq 1$ $\quad c(y, y') = |y - y'|^p$

3- En discrimination binaire $y \in \{0,1\}$ $\quad c(y, y') = \mathbb{I}_{\{y \neq y'\}} = |y - y'|$

ISEN ALL IS DIGITAL! OUEST | yncréa

Risque

- Définition: (règle de prévision)

C'est une fonction $f\,:\boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{Y}}$ qui associe la sortie $f(x)$ à l'entrée $x \in \boldsymbol{\mathcal{X}}$. L'ensemble des règles est Ƒ

- Définition: (risque)

C'est le comportement moyen de la fonction de perte choisie.

Le risque d'une règle de prévision f est défini par $\mathcal{R}(f) = \mathbb{E}[c(f(X), Y)]$

- Définition: (algorithme de prévision)

C'est une application qui associe à un échantillon d'apprentissage une règle de prévision $\hat{f}$

Ainsi, le résultat de l'algorithme de prévision est une estimation de f

- Reformulation du problème:

Trouver une règle de prévision telle que son risque soit minimal

Rappels

Maximum de vraisemblance

- Définition
  - On appelle vraisemblance de l'échantillon $X_1, \dots, X_n \sim P_\theta$ en $a \in \Theta$, la variable aléatoire définie par
    - $L_n(a) = f\big((X_1, \dots, X_n), a\big)$
    - f étant la densité de probabilité

  - Si les variables sont indépendantes et identiquement distribuées, on a
    - $L_n(a) = \prod_{i=1}^{n} f(X_i, a)$

Maximum de vraisemblance

- Définition
  - On appelle estimateur de vraisemblance (EMV), la statistique $\widehat{\theta_n}$, telle que
    - $L_n\left(\widehat{\theta_n}\right) = \max_{a \in \Theta} L_n(a)$
  - L'EMV peut être calculé en minimisant la fonction inverse de log-vraisemblance
    - $-\log(L_n\left(\widehat{\theta_n}\right)) = \min_{a \in \Theta} -\log(L_n(a))$
    - Ce minimum peut être calculé analytiquement en $\frac{\delta(\log(L_n(\theta)))}{\delta\theta} = 0$

Maximum de vraisemblance: propriétés

- Convergent: $\widehat{\theta_n} \xrightarrow{p} \theta_0$, où θ_0 désigne la vraie valeur du paramètre, et p la loi de probabilité
- Invariant: Si $\widehat{\theta_n}$ est l'EMV de θ alors $g\left(\widehat{\theta_n}\right)$ est l'EMV de $g(\theta)$
- Asymptotiquement normal:
  - $\frac{\widehat{\theta_n}-\theta}{se} \xrightarrow{d} \boldsymbol{\mathcal{N}}(0,1)$
  - Où se est l'écart type de $\widehat{\theta_n}$. En clair $\widehat{\theta_n} \approx \boldsymbol{\mathcal{N}}(\theta, se)$

Question?