Rappels : Analyse Univariée et Bivariée

Paul Chapron [1] & Yann Méneroux [1]

2021-2022

[1] IGN-ENSG-UGE

ÉCOLE NATIONALE DES SCIENCES GÉOGRAPHIQUES

Introduction

Dans les cours précédents ...

Notions pour manipuler les variables aléatoires, et estimer certains descripteurs

- co-variance
- intervalle de confiance
- bootstrap
- . . .

Motivation

L'analyse univariée permet de décrire la forme et de quantifier les caractéristiques de la répartition des valeurs d'une variable.

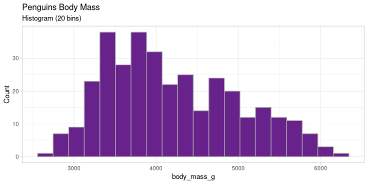
- Notion de distribution
- Visualisation (Histogramme, densité, boxplots, . . .)
- Moments, Quantiles, CV

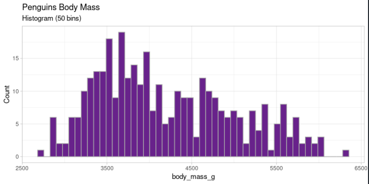
Analyse Univariée

Histogramme d'une variable

Représentation graphique des effectifs associés à des classes de valeurs d'une variable numérique

Le nombre de classes peut varier !



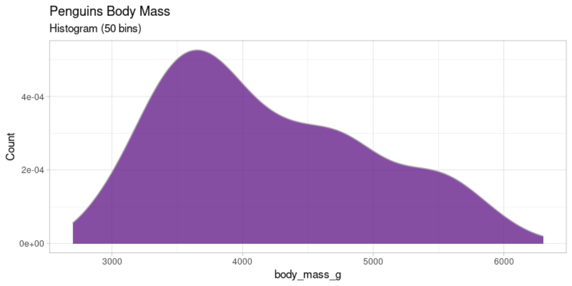


Distribution

Synonymes: distribution empirique, distribution des fréquences, distribution statistique

Tableau ou graphique qui associe les (classes de) valeurs à leur fréquence d'apparition

≈ « Histogramme des fréquences en continu»



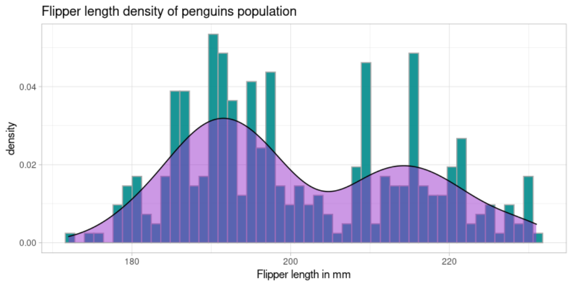
La distribution peut être définie comme une fonction qui donne la probabilité qu'un individu x pris au hasard ait la valeur V_x pour la variable V :

$$distribution(V) \equiv P(V = V_x), \forall V_x \in \Omega_V$$

Avec Ω_V l'ensemble des valeurs que peut prendre V : l'univers de V

Lorsque la variable prend des valeurs réelles, on parle de densité de probabilité, c'est pourquoi on retrouve ce terme "density" sur les axes des ordonnées dans les graphiques de distribution.

Distribution et histogramme



N.B. En toute rigueur, représenter une courbe de distribution de probabilité par dessus un histogramme est impropre : il faudrait deux graphiques distincts, ou au moins deux axes des ordonnées: un pour l'histogramme, représentant un effectif, l'autre pour la distribution , représentant une probabilité

Parfois , les distributions empiriques ressemblent à celles de lois de probabilités connues.

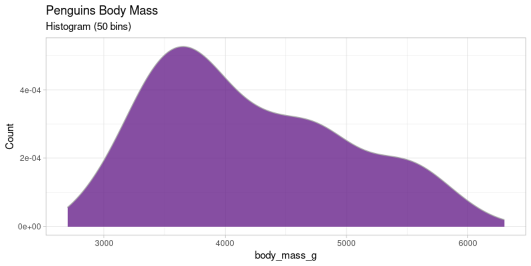
$\rightarrow$ on peut alors modéliser la variable par une variable aléatoire de loi fixée

$\rightarrow$ les paramètres de cette loi doivent être déterminés (ajustement).

Décrire une distribution

La forme d'une distribution donne beaucoup d'informations :

- "pics" : valeurs les plus représentées dans la population
- présence de valeurs extrêmes : la courbe de la distribution est tirée à gauche ou à droite du graphique
- symétrie : les individus se répartissent équitablement de part et d'autre du pic
- aplatissement : la population est plus ou moins resserrée, ou autour de certaines valeurs
- ...



Décrire une distribution : mesures de tendance centrale

Tendance Centrale

La tendance centrale est une valeur qui résume une série de valeurs (quantitative)

- Moyenne
- Médiane
- Mode

Moyenne

$$\bar{x} = \frac{1}{n}\sum_{i=0}^{n} x_i$$

Avantage : chaque valeur compte
Inconvénients :

- sensibilité aux valeurs extrêmes
- pas de signification sur les valeurs discrètes (e.g. 2.5 enfants par foyer)

Pour y remédier (parfois):
→ exclure les outliers
→ utiliser un autre estimateur (médiane)
→ étudier la distribution des valeurs (e.g. cas bimodal) et opérer une classification

ENSG Géomatique

$$\bar{x}_{geom} = \sqrt[n]{\prod_{i=0}^{n} x_i}$$

Moins sensible que la moyenne classique aux valeurs extrêmes.

Mode

Le mode d'une variable est la valeur la plus fréquente (d'effectif maximum) d'une variable.

Avantages :

- peu sensible aux valeurs extrêmes
- interprétation simple : cas le plus fréquent

Inconvénient : la valeur du mode ne dépend pas de toutes les observations, la modification d'une valeur n'entraîne pas la modification du mode (ce qui explique sa robustesse aux valeurs extrêmes)

Calcul du mode

Si la variable est quantitative et continue :

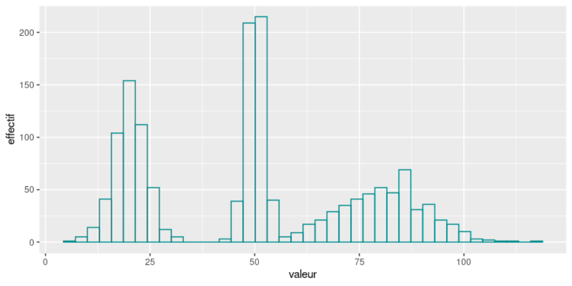
- découper l'étendue de la variable ($max - min$) en intervalle égaux
- compter les effectifs de chaque intervalle
- le mode est la moyenne des valeurs des bornes de l'intervalle de plus grand effectif.

(C'est exactement ce que fait un histogramme graphiquement !)

Modes

Par définition, le mode est unique, mais on peut appeler modes les valeurs des autres pics d'une distribution.
On parle de distribution bi-modale ou tri-modale lorsqu'une distribution présente deux ou trois pics.

Les valeurs modales d'une distribution sont les valeurs correspondant à ces pics.



Médiane

La médiane est la valeur qui sépare une population en deux classes d'égal effectif.

C'est la valeur la plus proche de toutes les autres.

Avantages :

- Souvent plus pertinente que la moyenne
- les valeurs extrêmes ne modifient pas sa valeur
- interprétation facile: un individu sur deux a une valeur inférieure (respectivement supérieure) à la médiane.

Inconvénient : Comme le mode , la médiane ne dépend pas de toutes les observations

N.B. la robustesse de la médiane est bien utile dans le cas de distribution particulièrement asymétriques, où la moyenne est dégradée par les valeurs extrêmes, à droite (valeurs très élevées) ou à gauche (valeurs très faibles).

Moyenne et médiane

Que peut on dire d'une population dont la médiane est inférieure à la moyenne ?

Exemple : revenus mensuels en équivalent temps plein en France en 2016.

Revenu mensuel net moyen 2 238 €
Revenu mensuel net médian 1 789 €

source https://www.insee.fr/fr/statistiques/4277680?sommaire=4318291

Moyenne et médiane

Un salaire mensuel net équivalent temps plein de 2000€ est-il un bon salaire ?

- 2000€ < moyenne : on peut le considérer comme trop bas pour être «bon»
- 2000€ > médiane : supérieur à (au moins) la moitié des salaires du pays, on peut le considérer comme un «bon» salaire.

Double interprétation & «instinctivement» on imagine une dispersion symmétrique, où la moyenne est proche de la médiane

Décrire une distribution : mesures de dispersion

Dispersion

La dispersion décrit la tendance des valeurs d'une variable à se disperser plus ou moins largement autour des valeurs des tendances centrales.

Variance et écart type

La variance est la somme des écarts carrés à la moyenne rapporté à l'effectif

$$var_X = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Avec :

- X une variable
- x_i les valeurs de la variables
- $\bar{x}$ la moyenne de X
- n l'effectif

Variance et écart type

L'écart type est la racine carrée de la variance

$$\sigma_X = \sqrt{var_X}$$

Variance et écart-type sont sensibles aux valeurs extrêmes et toujours positifs.
Si $var_X = 0$ ou $\sigma_X = 0$, alors X est constante.
Un écart-type faible indique que les valeurs sont réparties de façon homogène autour de la moyenne.

Quantiles

La médiane sépare une population en deux classes d'égal effectif

Les quantiles séparent une population en n classes d'égal effectif.

Quartiles

Les quartiles d'une population selon une variable X sont trois valeurs, Q_1, Q_2, Q_3 qui séparent la population en quatre classes d'égal effectif.

- 25% des valeurs de X sont strictement inférieures à Q_1
- 50% des valeurs de X sont strictement inférieures à Q_2 (médiane)
- 75% des valeurs de X sont strictement inférieures à Q_3

Déciles

Les déciles sont les 9 quantiles $Q_1, Q_2, \ldots, Q_9$ qui séparent une population 10 classes d'égal effectif.

Écarts inter-quartiles et inter-déciles

Écart inter-quartile: $Q_3 - Q_1$, capture 50% des valeurs de la population les plus proches de la médiane

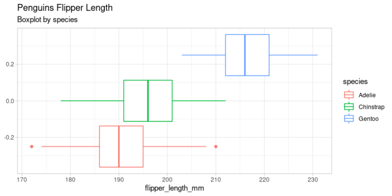
Écart inter-décile: $Q_9 - Q_1$, capture 80% des valeurs de la population les plus proches de la médiane

Quantiles

Avantages	Inconvénients
Peu sensibles aux distributions aplaties et aux valeurs extrêmes	Parfois délicat pour les variables quantitatives discrètes
L'écart inter-quantile est plus robuste que l'écart-type	Les écarts inter-quantiles négligent l'influence des valeurs extrêmes sur la distribution

Les boîtes à moustaches (boxplots)

Représentation courante de la dispersion d'une variable à l'aide de quartiles



- La marque centrale de la boîte est la médiane
- Les bords de la boîte sont les quartiles Q_1 et Q_3
- Les moustaches vont jusqu'à la plus grande (resp. la plus petite) valeur inférieure (resp. supérieure) à 1.5 fois l'écart interquartile
- Les valeurs qui dépassent les moustaches sont affichées sous formes de points

Le coefficient de variation

Le coefficient de variation (CV) est une autre mesure de dispersion.

C'est le ratio entre l'écart-type σ_x et la moyenne $\bar{x}$ d'une variable quantitative X.

$$CV(X) = \frac{\sigma_x}{\bar{x}}$$

Plus il est important , plus la dispersion est grande.

Plus il est proche de 0, plus les données sont homogènes.

Inconvénients similaires à ceux de $\bar{x}$ et σ_x : sensibilité aux valeurs extrêmes.

Exemple : deux communes versent des aides aux entreprises locales, qu'on suppose distribuées suivant une loi normale.

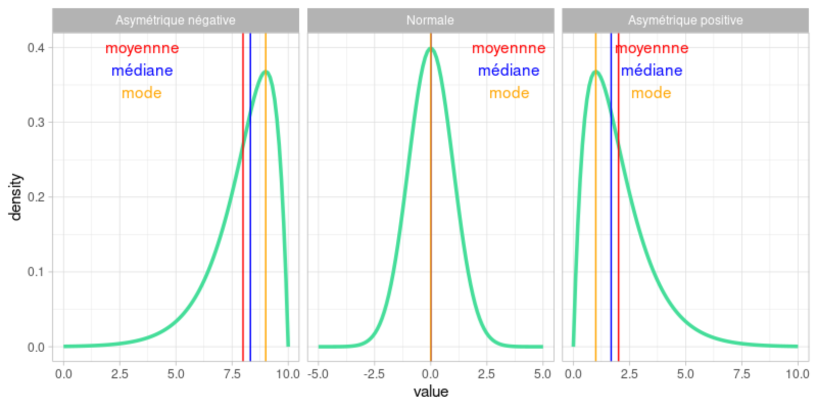
Commune A : moyenne = 390 euros, $\sigma = 30$ euros

Commune B : moyenne = 152 euros, $\sigma = 8$ euros

Pour quelle commune les aides sont les plus homogènes?

Décrire une distribution : asymétrie et aplatissement

Asymétrie (ou skewness)



Coefficients d'asymétrie de Pearson

Deux moyens simples d'estimer l'asymétrie

$$C_1 = \frac{\bar{x} - mode(X)}{\sigma_x}$$

$$C_2 = \frac{3(\bar{x} - mediane(X))}{\sigma_x}$$

- coefficient nul : la distribution est symétrique
- coefficient négatif : la distribution est déformée à gauche de la médiane (sur-représentation de valeurs faibles, à gauche)
- coefficient positif : la distribution est déformée à droite de la médiane (sur-représentation de valeurs fortes, à droite)

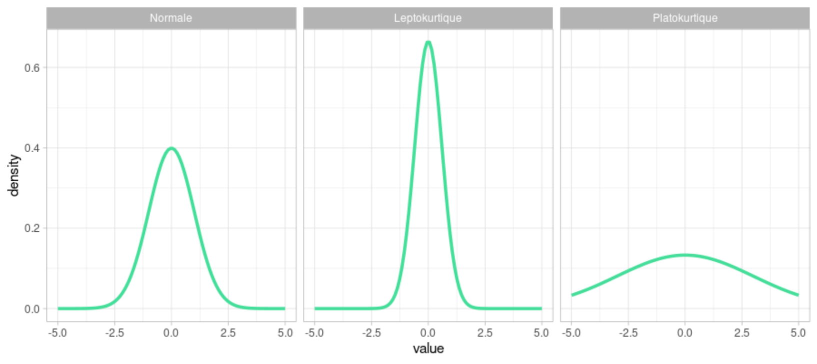
Coefficient d'asymétrie de Fischer

Ce coefficient est le moment d'ordre 3 de la variable X (de moyenne μ et d'écart-type σ) centrée réduite

$$skewness' = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\sum_{i=0}^{n}(x_i - \bar{x})^3}{n\sigma^3}$$

Interprétation similaire aux coefficients de Pearson

Aplatissement (ou Kurtosis)



Courbe piquée: Peu de variation, distribution relativement homogène, beaucoup de valeurs égales ou proches de la moyenne.

Courbe aplatie: Variations importantes, distribution relativement hétérogène, beaucoup de valeurs s'éloignent de la moyenne.

Coefficient de Kurtosis

$$K = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{n\sigma^4}$$

Si la distribution est normale, $K = 3$

Si $K > 3$, la distribution est plus aplatie

Si $K < 3$, la distribution est moins aplatie

On normalise parfois en considérant $K' = K - 3$ (quantifie l'excès d'aplatissement)

Analyse Bivariée

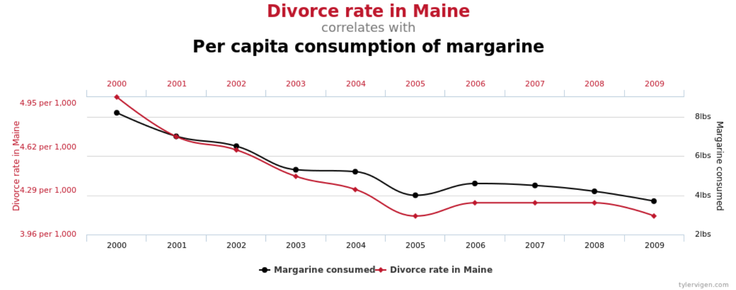
Étude de la relation entre deux variables :

- quantitatives : corrélation, régression linéaire
- qualitatives : test d'indépendance du «Chi deux» / χ^2

Pour le lien entre une variable quantitative et une variable qualitative, on fera simplement un graphique.

Corrélation ≠ Causalité

Une liaison, même très forte, entre deux variables, n'indique pas la causalité.



Erreur très courante , très tentante.

ENSG Géomatique

Données «spatiales»

Individus restreints spatialement (sélection spatiale)

Variables "géographique" (e.g. lieu de résidence) renseignées pour les individus

Prise en compte des distances $\rightarrow$ modèle(s) gravitaire(s) (hors programme)

Données localisées (hors programme pour nous)

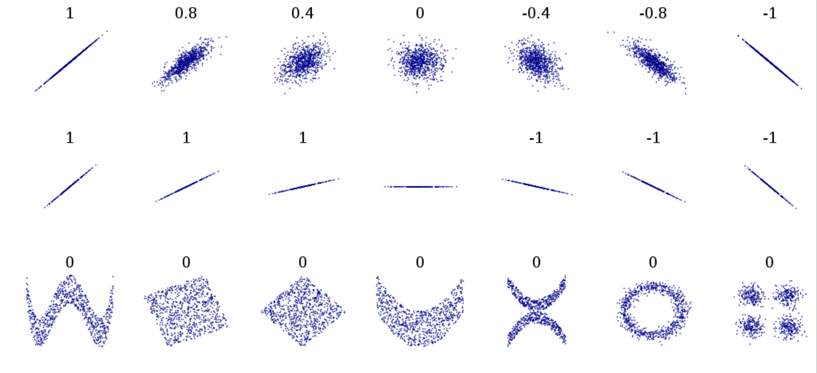
Auto-corrélation spatiale (Moran's I, Geary Index)

Geographically Weighted Regression (GWR) $\approx$ régression linéaire avec prise en compte de la distance entre individus

Variogrammes

Corrélation

Toujours afficher les données, avant de faire quoi que ce soit.



Corrélation de Pearson

La corrélation indique l'intensité du lien linéaire entre deux variables quantitatives.

$$cor(x, y) \in [-1; 1]$$

- $cor(x, y) \approx 0$: pas de relation (linéaire) entre les deux variables
- $cor(x, y) < 0$: les deux variables ont des sens de variations opposés
- $cor(x, y) > 0$: les deux variables varient conjointement
- $cor(x, y) = \pm 1$: variables parfaitement linéairement (anti-)corrélées, i.e. fonction affine l'une de l'autre.

Formule du coefficient de corrélation de Pearson

$$r = \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{E\big[(x - E(x))(y - E(y))\big]}{\sigma_x \sigma_y}$$

Avec :

- r (parfois ρ) le coefficient de corrélation
- x et y deux variables quantitatives
- $E(x)$ l'espérance d'une variable x
- σ_x l'écart-type d'une variable x
- $cov(x,y)$ la covariance de deux variables x et y

Corrélation et indépendance

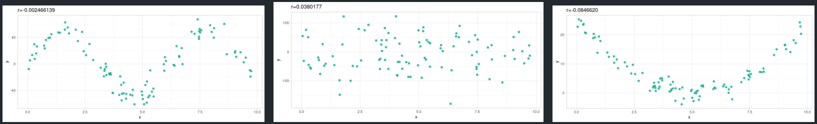
Deux variables indépendantes ont un coefficient de corrélation nul :

$$x \perp\!\!\!\perp y \implies cor(x, y) = 0$$

MAIS une corrélation nulle n'implique pas l'indépendance des variables !

$$cor(x, y) = 0 \;\not\Longrightarrow\; x \perp\!\!\!\perp y$$

D'autres liaisons sont possibles :



Corrélation avec R

Fonction `cor(x,y)` pour obtenir la valeur du coefficient,

Fonction `cor.test(x,y)` pour obtenir la p-value et l'intervalle de confiance.

Résultat :

```
##
##  Pearson's product-moment correlation
##
## data:  iris$Petal.Length and iris$Petal.Width
## t = 43.387, df = 148, p-value < 2.2e-16
## alternative hypothesis: true correlation
## is not equal to 0
## 95 percent confidence interval:
##  0.9490525 0.9729853
## sample estimates:
##       cor
## 0.9628654
```

R donne le coefficient de Pearson par défaut, l'argument `method` de la fonction cor() permet de spécifier deux autres coefficients : Kendall et Spearman.

Matrice de Corrélation

Fonction `cor()` appliquée à plusieurs variables de type `numeric`

e.g. `cor(iris[,1:4])`

Résultat:

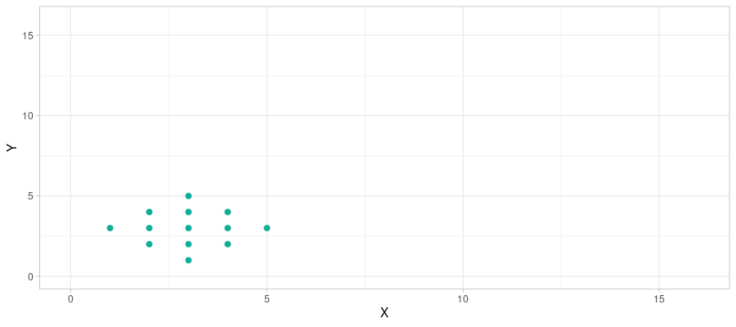
```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
## Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
## Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
## Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
```

Présentation des corrélations entre les variables quantitatives d'un tableau, pour tous les couples de variables.

La matrice de corrélation est symétrique, et sa diagonale est constituée de 1.

Sensibilité aux outliers

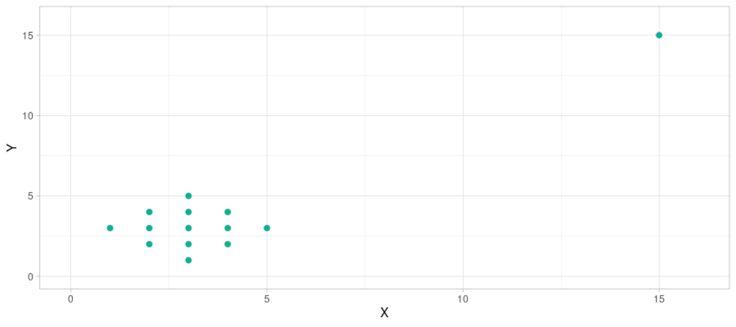
```
X <-  c(3,2,3,4,1,2,3,4,5,2,3,4,3)
Y <-  c(1,2,2,2,3,3,3,3,3,4,4,4,5)
plot(X, Y, xlim = c(0,16), ylim= c(0,16))
```




```
>cor.test(X,Y)$estimate
## cor
##   0
```

Sensibilité aux outliers

```
X <-  c(3,2,3,4,1,2,3,4,5,2,3,4,3,15)
Y <-  c(1,2,2,2,3,3,3,3,3,4,4,4,5,15)
plot(X, Y, xlim = c(0,16), ylim= c(0,16))
```




```
>cor.test(X,Y)$estimate
## cor
##   0.9052224
```

Outliers

Outlier : observation “anormale”, par ses valeurs extrêmes, comparées aux autres.

La corrélation (et la régression linéaire)) sont très sensibles aux outliers.

→ s’interroger sur la nécessité de nettoyer/filtrer les données et les conséquences

→ ne pas faire d’épuration brutale et aveugle

Coefficient de corrélation de Spearman

ENSG Géomatique

Quand les deux variables semblent corrélées , de façon monotone mais non linéaire,

$\rightarrow$ utiliser le coefficient de Spearman, basé sur le rang des individus.

$$\rho_S = \frac{cov(rg_x, rg_y)}{\sigma_{rg_x} \sigma_{rg_y}}$$

Avec :

- rg_x le rang des individus selon la variable x (en cas d'ex-aequo on prend le rang moyen)
- $cov()$ la fonction de covariance
- σ_{rg_x} l'écart-type du rang rg_x

Régression linéaire

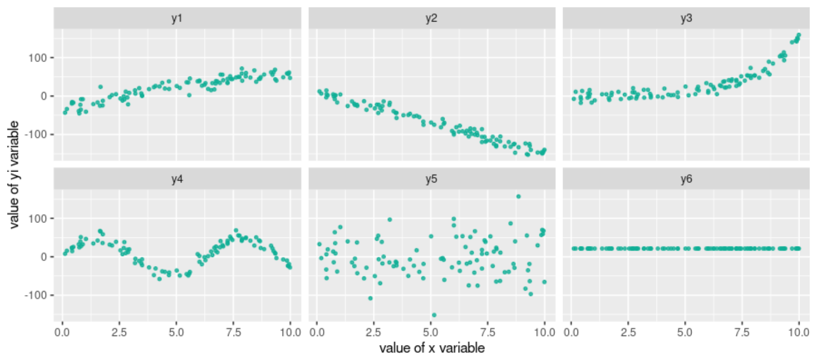
Rappel (encore): Toujours afficher les données, avant de faire quoi que ce soit.

Quand le nuage de points semble «suffisamment» linéaire , on peut tenter de décrire la relation statistique linéaire en proposant un modèle linéaire

$$\hat{y} = \alpha x + \beta$$

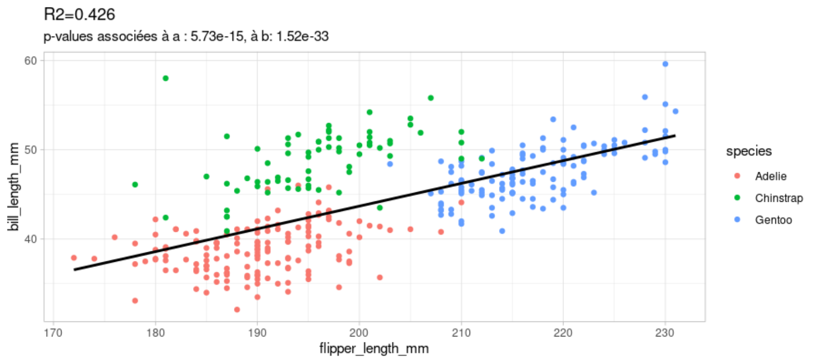
Le modèle retenu doit passer au mieux (i.e. en minimisant une certaine erreur) dans le nuage de points.

Diverses formes de dépendances



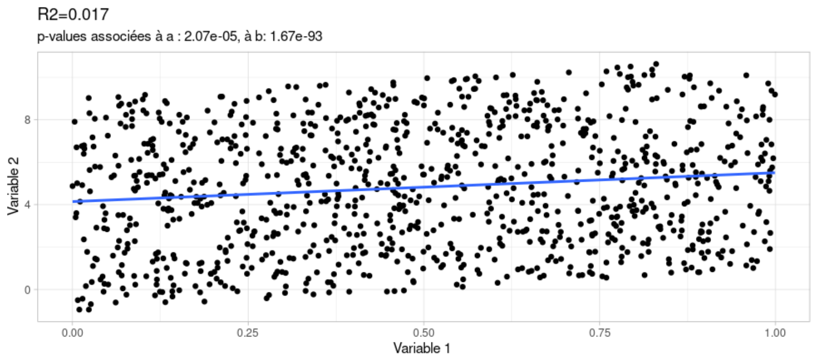
En pratique , les formes sont beaucoup moins régulières.

Exemple



R2=0.426
p-values associées à a : 5.73e-15, à b: 1.52e-33
60
50
40
bill_length_mm
170
180
190
200
210
220
230
flipper_length_mm
species
Adelie
Chinstrap
Gentoo

Exemple


R2=0.017
p-values associées à a : 2.07e-05, à b: 1.67e-93
Variable 2
8
4
0
0.00
0.25
0.50
0.75
1.00
Variable 1

1. Tracer le nuage de points
2. Existe-t-il une relation ?
3. Si oui , Est-elle de forme linéaire ?
   - 3.1 Si oui $\rightarrow$ faire une régression linéaire
   - 3.2 Si non, la liaison est-elle monotone ou de forme connue ?
     - 3.2.1 Si oui $\rightarrow$ Proposer un modèle e.g. polynomial
     - 3.2.2 Alternative: Réaliser un modèle LOESS avec prudence (uniquement descriptif , aucun pouvoir de généralisation) cf le blog de Lise Vaudor `http://perso.ens-lyon.fr/lise.vaudor/regression-loess/`

Réaliser une régression linéaire avec R

La fonction `lm()` réalise une régression linéaire entre deux (ou plusieurs) vecteurs numériques de même taille.

L'objet résultat comporte plusieurs attributs, notamment :

- `$coefficients` les coefficients du modèle linéaire
- `$residuals` les résidus

La fonction `summary()` sur l'objet synthétise les résultats

Format des résultats

le R^2

$$SSE = \sum_i (prediction_i - observation_i)^2$$

$$SSE = \sum_i (ax_i + b - x_i)^2$$

$$SSE = \sum_i (\hat{y}_i - x_i)^2$$

Le **coefficient de détermination linéaire** , noté R^2 est une valeur qui décrit la **qualité de prédiction** de la régression, c'est-à-dire à quel point la droite de régression estime correctement les valeurs de la variable expliquée.

Il est défini par :