



Statistiques & Régression

Niveau Master 1

Plan du cours

- 1- Révision statistiques (Jour 1 - Matin)
- 2- Régression linéaire (Jour 1 - Aprem)
- 3- Régression linéaire multiple (Jour 2 - Matin)
- 4- Régression linéaire généralisée (Jour 2 - Aprem)
- 5- Régression pénalisée et régularisation (Jour 3 - Matin)
- 6- Projet



Jour 1
(Partie 1)

Révision des fondamentaux

Population

Ensemble complet des individus, objets ou événements partageant des **caractéristiques communes** et pertinents pour une étude particulière.

La taille peut être **finie** (ex: tous les employés d'une entreprise) ou **infinie** (ex: tous les lancers possibles d'un dé).

Exemples:

- Tous les étudiants en France.



Population

Ensemble complet des individus, objets ou événements partageant des **caractéristiques communes** et pertinents pour une étude particulière.

La taille peut être **finie** (ex: tous les employés d'une entreprise) ou **infinie** (ex: tous les lancers possibles d'un dé).

Exemples:

- Tous les étudiants en France.
- Tous les clients d'une entreprise.



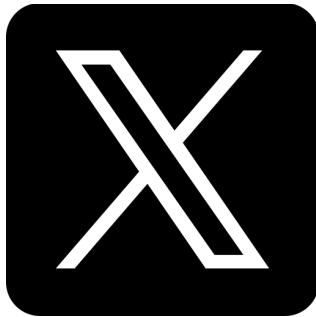
Population

Ensemble complet des individus, objets ou événements partageant des **caractéristiques communes** et pertinents pour une étude particulière.

La taille peut être **finie** (ex: tous les employés d'une entreprise) ou **infinie** (ex: tous les lancers possibles d'un dé).

Exemples:

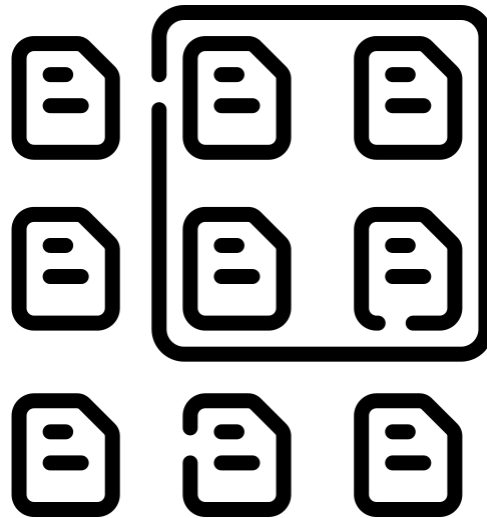
- Tous les étudiants en France.
- Tous les clients d'une entreprise.
- Tous les tweets publiés sur un sujet donné.



Échantillon

Sous-ensemble d'individus, d'objets ou d'événements sélectionnés à partir d'une population plus large, dans le but de la représenter de manière adéquate.

Objectif: Obtenir des informations sur la population entière en étudiant une partie plus petite et plus gérable, permettant ainsi de gagner du temps, de réduire les coûts et de simplifier l'analyse.



Échantillon

Sous-ensemble d'individus, d'objets ou d'événements sélectionnés à partir d'une population plus large, dans le but de la représenter de manière adéquate.

Objectif: Obtenir des informations sur la population entière en étudiant une partie plus petite et plus gérable, permettant ainsi de gagner du temps, de réduire les coûts et de simplifier l'analyse.

- Un échantillon **représentatif** reflète fidèlement les caractéristiques importantes de la population.
- Plus l'échantillon est **grand**, plus les estimations statistiques sont **précises**, mais cela augmente aussi le coût et le temps de l'étude.

Variables

Caractéristique ou attribut qui peut prendre différentes valeurs.

Exemples : âge, sexe, couleur des yeux, température, revenu...

Variables

Caractéristique ou attribut qui peut prendre différentes valeurs.

Exemples : âge, sexe, couleur des yeux, température, revenu...

Variables qualitatives

- satisfaction client
- couleur des yeux
- nationalité
- ...

Variables quantitatives

Variables

Caractéristique ou attribut qui peut prendre différentes valeurs.

Exemples : âge, sexe, couleur des yeux, température, revenu...

Variables qualitatives

- satisfaction client
- couleur des yeux
- nationalité
- ...

Variables quantitatives

- température
- poids
- taille
- ...

Mesures de tendance centrale

Les mesures de tendance centrale résument l'ensemble d'un jeu de données en une valeur unique qui représente son "centre".

Elles permettent de décrire la position centrale des données et d'avoir une idée de la valeur "typique" dans l'ensemble.

Mesures de tendance centrale

Les mesures de tendance centrale résument l'ensemble d'un jeu de données en une valeur unique qui représente son "centre".

Elles permettent de décrire la position centrale des données et d'avoir une idée de la valeur "typique" dans l'ensemble.

Types de mesures de tendance centrale

- **Moyenne** (arithmétique)
 - Somme de toutes les valeurs divisée par le nombre de valeurs.
 - Représente le point d'équilibre de l'ensemble de données.
 - Sensible aux valeurs extrêmes (valeurs aberrantes).

Mesures de tendance centrale

Les mesures de tendance centrale résument l'ensemble d'un jeu de données en une valeur unique qui représente son "centre".

Elles permettent de décrire la position centrale des données et d'avoir une idée de la valeur "typique" dans l'ensemble.

Types de mesures de tendance centrale

- **Moyenne** (arithmétique)
- **Médiane**
 - Valeur du milieu lorsque les données sont ordonnées par ordre croissant.
 - Moins sensible aux valeurs extrêmes que la moyenne.
 - Préférée à la moyenne lorsque les données sont biaisées ou comportent des valeurs aberrantes.

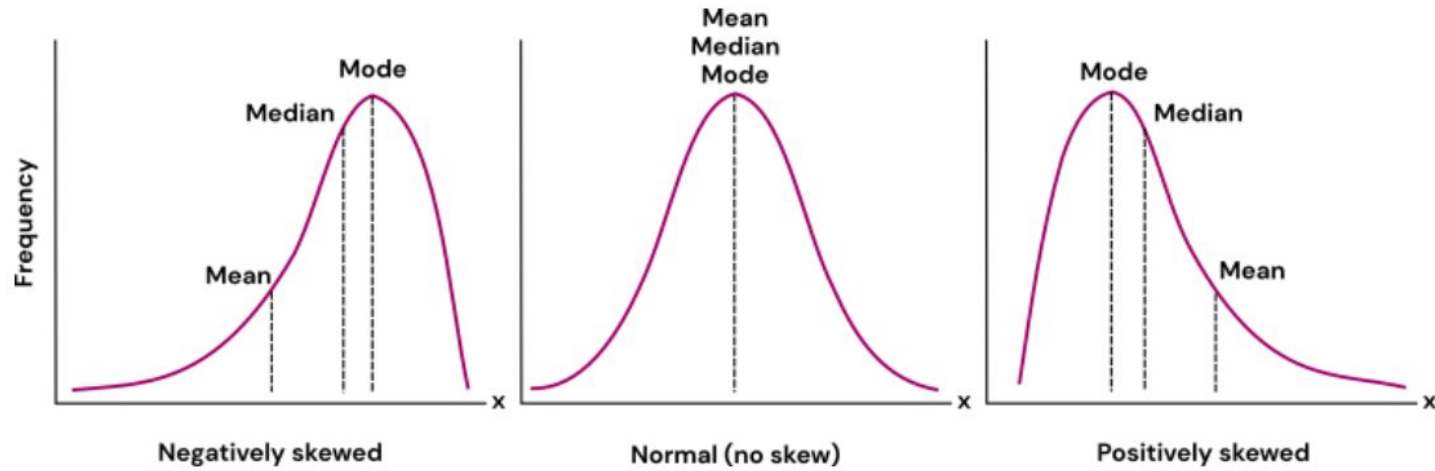
Mesures de tendance centrale

Les mesures de tendance centrale résument l'ensemble d'un jeu de données en une valeur unique qui représente son "centre".

Elles permettent de décrire la position centrale des données et d'avoir une idée de la valeur "typique" dans l'ensemble.

Types de mesures de tendance centrale

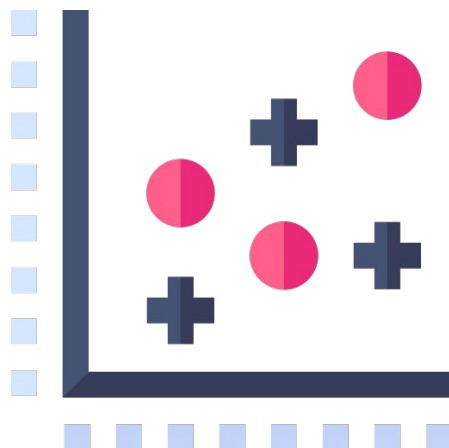
- **Moyenne** (arithmétique)
- **Médiane**
- **Mode**
 - Valeur la plus fréquente dans l'ensemble de données.
 - Utile pour les données qualitatives et les données quantitatives avec des valeurs discrètes.
 - Peut y avoir plusieurs modes ou aucun mode.



Mesures de dispersion

Les mesures de dispersion quantifient l'étalement ou la variabilité d'un ensemble de données.

Elles indiquent à quel point les valeurs sont éloignées les unes des autres et de la mesure de tendance centrale.



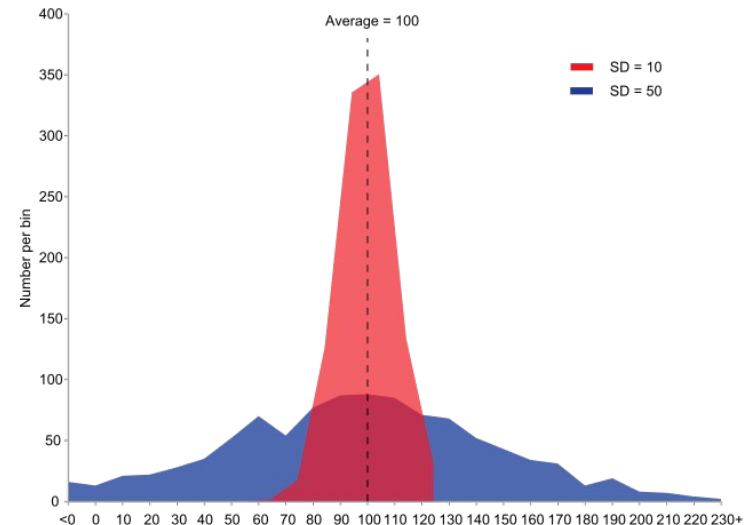
Variance

Moyenne des carrés des écarts à la moyenne

Mesure la dispersion globale des données

Unité de mesure au carré par rapport aux données originales

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



Écart-type

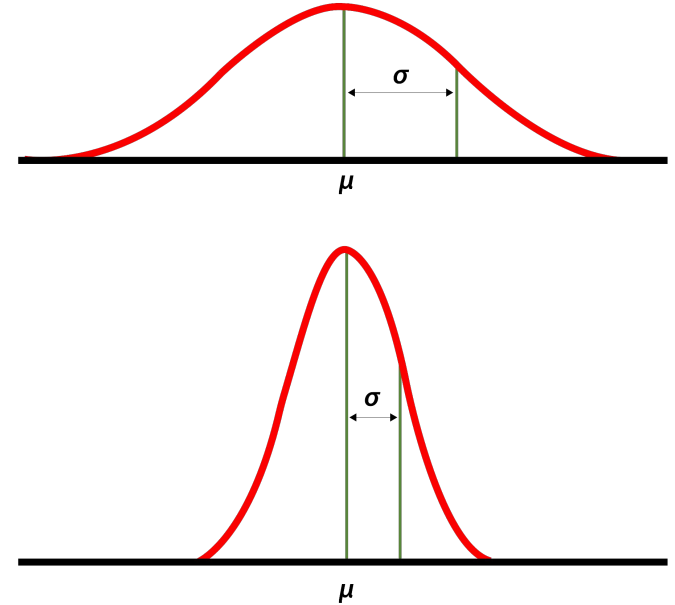
Racine carrée de la variance

Mesure la dispersion moyenne des données

Même unité de mesure que les données originales

Plus facile à interpréter que la variance

$$\sigma = \sqrt{V} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} = \sqrt{\frac{1}{n} \left(\sum_{i=1}^n x_i^2 \right) - \bar{x}^2}$$

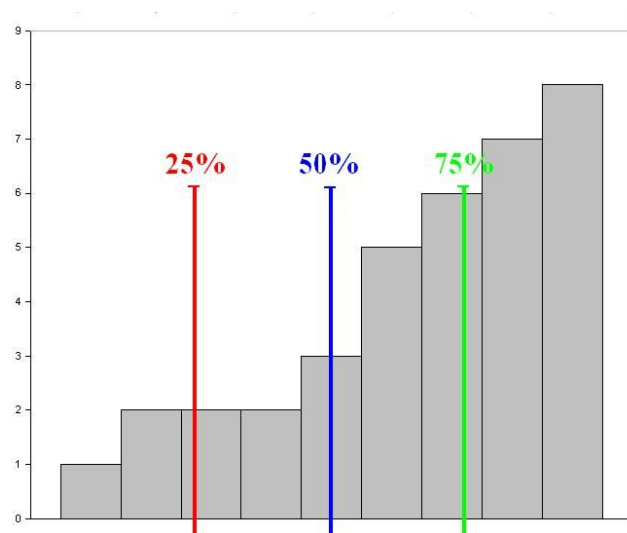


Quartiles

Divisent les données ordonnées en quatre parties égales

Q1 (25e percentile), **Q2** (médiane, 50e percentile), **Q3** (75e percentile)

L'écart interquartile (**IQR = Q3 - Q1**) mesure la dispersion de la moitié centrale des données



Représentations graphiques

Pourquoi visualiser les données ?

- Faciliter la compréhension de la distribution et des tendances des données.
- Identifier des structures, des relations et des anomalies.
- Comparer efficacement différents ensembles de données
- Communiquer les résultats de manière claire et percutante

Histogramme

Représente la distribution de **données continues** en les regroupant en classes (barres).

La hauteur de chaque barre indique la fréquence des valeurs dans la classe.

Permet de visualiser la forme de la distribution (symétrie, asymétrie, pics).

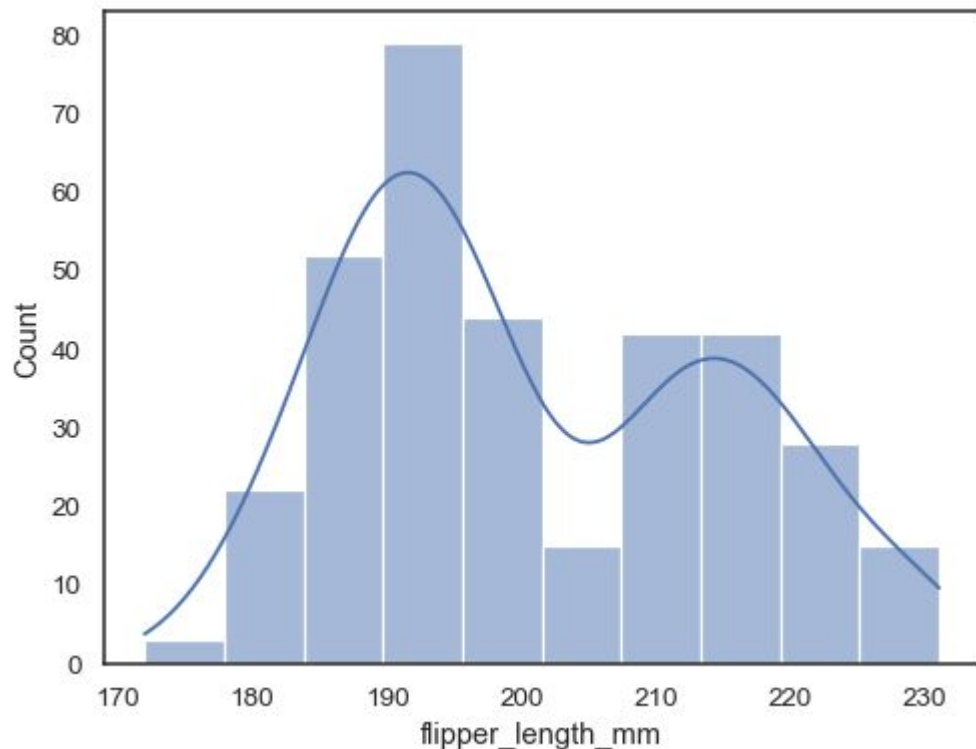
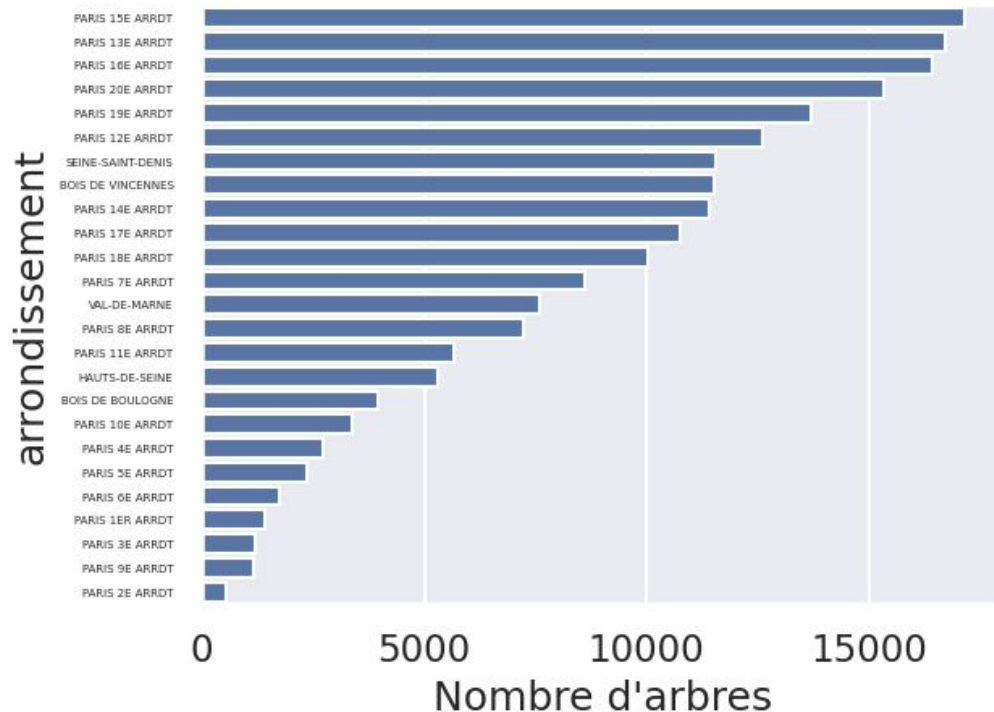


Diagramme en barre

Représentation graphique de données **catégorielles** ou **discrètes**.

Utilise des barres rectangulaires dont la hauteur est proportionnelle à la fréquence ou à la valeur de la variable pour chaque catégorie.



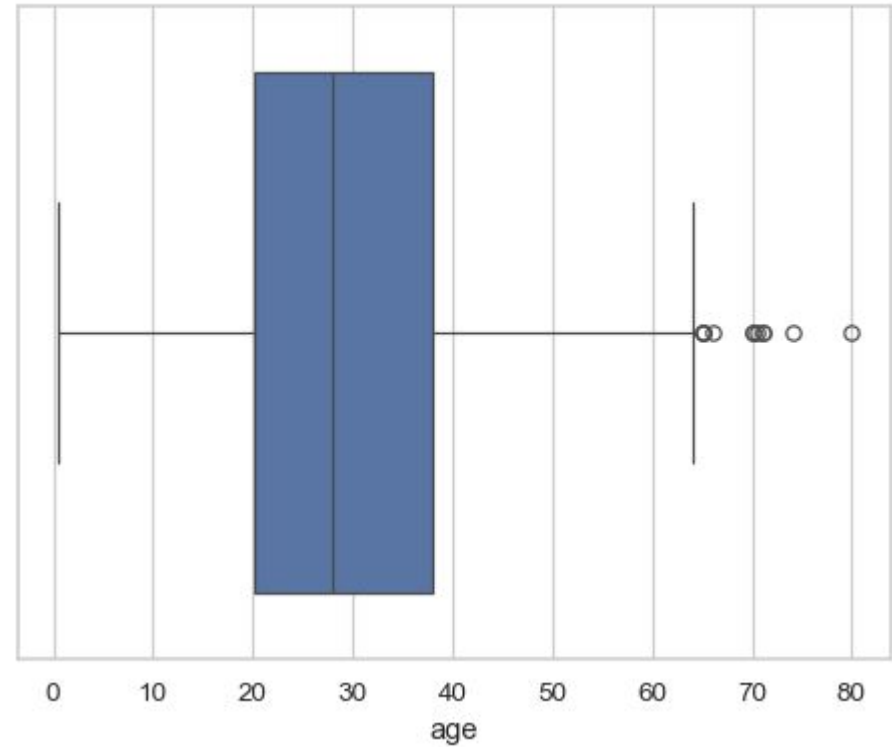
Boxplot

Représente la distribution de données en utilisant les quartiles (**Q1, médiane, Q3**) et les **valeurs extrêmes**.

La boîte montre l'écart interquartile ($IQR = Q3 - Q1$), contenant 50% des données.

Les moustaches indiquent la dispersion des données en dehors de la boîte.

Utile pour comparer la distribution et détecter les valeurs aberrantes entre différents groupes.



Nuage de points

Représente la relation entre deux variables *continues* en plaçant chaque observation comme un point dans un plan.

Permet d'identifier des tendances, des corrélations ou des relations non linéaires entre les variables.

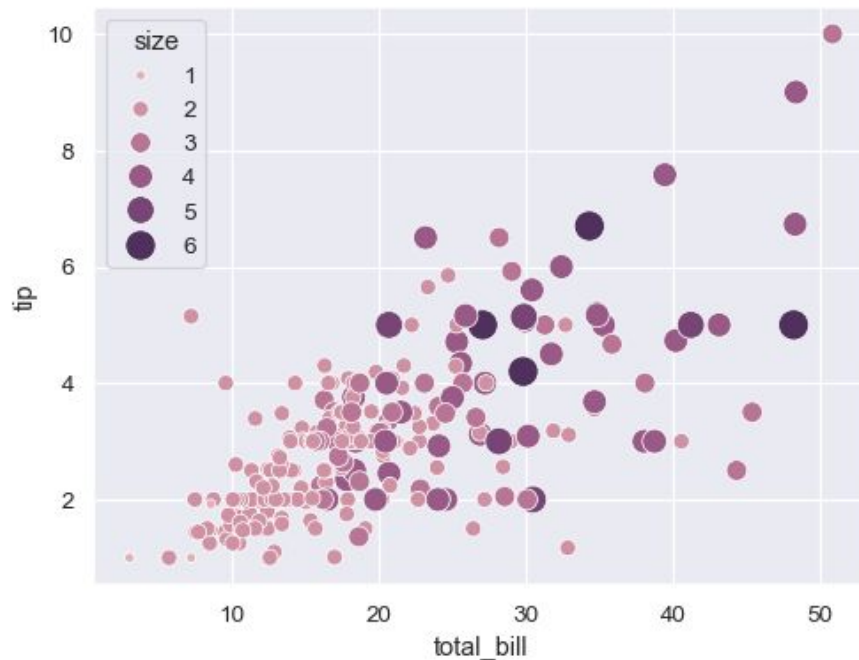
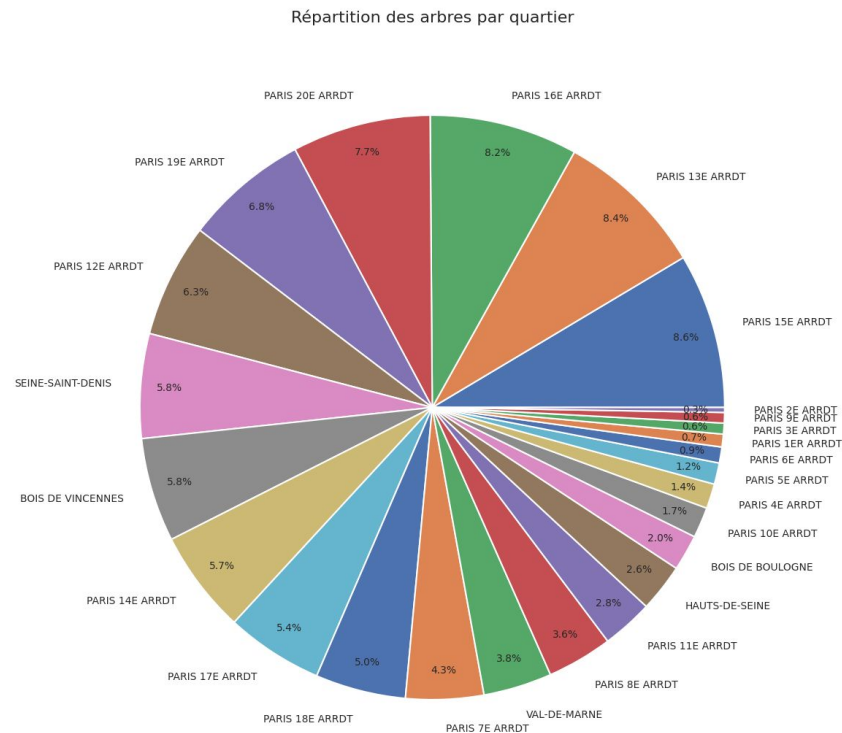


Diagramme circulaire

Représente la répartition de **données catégorielles** en divisant un cercle en secteurs proportionnels à la fréquence de chaque catégorie

Utile pour visualiser les parts relatives de chaque catégorie





Jour 1
(Partie 2)

Régression linéaire

Introduction

Régression linéaire

- Outil fondamental en statistique pour modéliser la relation entre une **variable dépendante** (cible) et une ou plusieurs **variables indépendantes** (prédicteurs).
- Permet de comprendre comment les changements dans les prédicteurs influencent la cible.
- Utilisée pour la prédiction, l'explication et le contrôle.

Types de régression linéaire

- Régression linéaire simple : un seul prédicteur
- Régression linéaire multiple : plusieurs prédicteurs

Modèle de régression linéaire simple

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Y : Variable dépendante (ce que l'on cherche à prédire)

X : Variable indépendante (prédicteur)

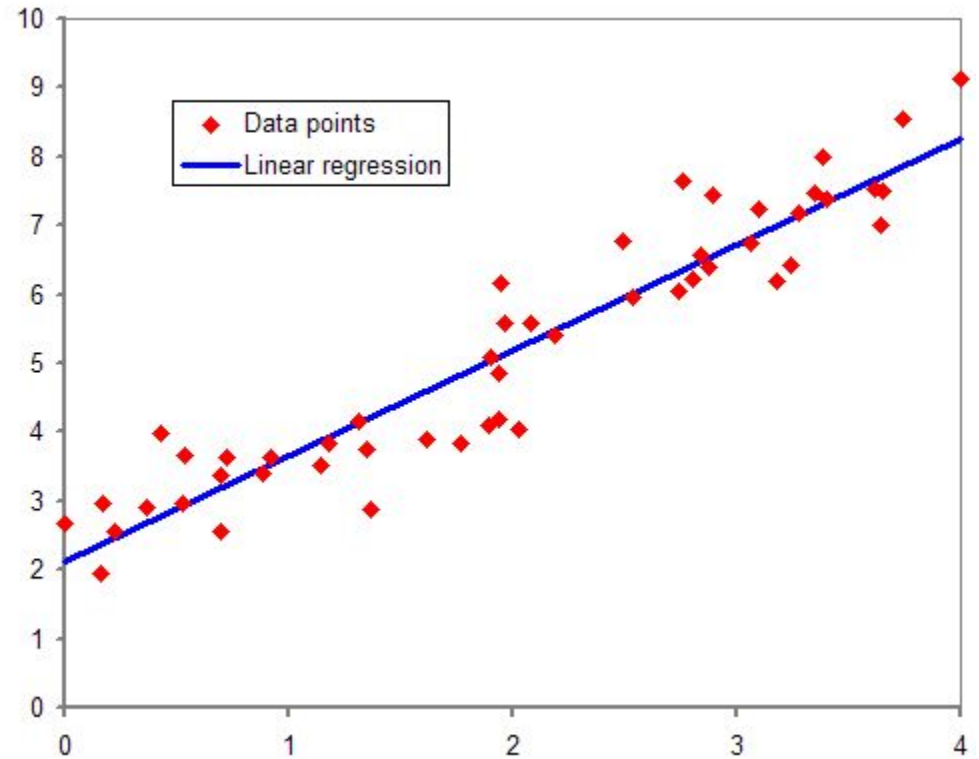
β_0 : Ordonnée à l'origine (valeur de Y quand X = 0)

β_1 : Coefficient de régression (pente de la droite, indique l'effet de X sur Y)

ε : Erreur aléatoire (résidu), représente la part de Y non expliquée par X

*50 points aléatoires dans
une distribution
gaussienne autour de la
ligne de régression qui
correspond le mieux à ce
nuage :*

$$y = 1,5x + 2,129333$$



Modèle de régression linéaire simple

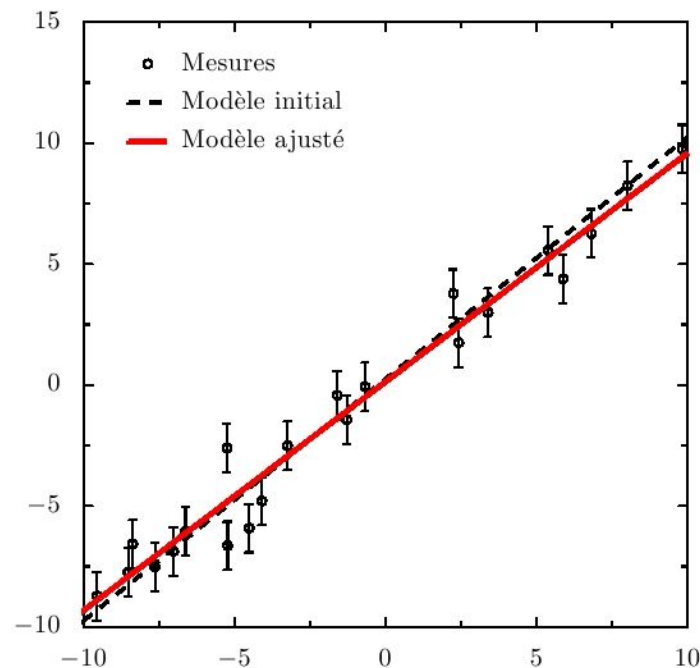
Hypothèses :

- Linéarité de la relation entre X et Y
- Indépendance des erreurs
- Homoscédasticité (variance constante des erreurs)
- Normalité des erreurs (pour l'inférence statistique)

Estimation des paramètres

Méthode des moindres carrés ordinaires (MCO)

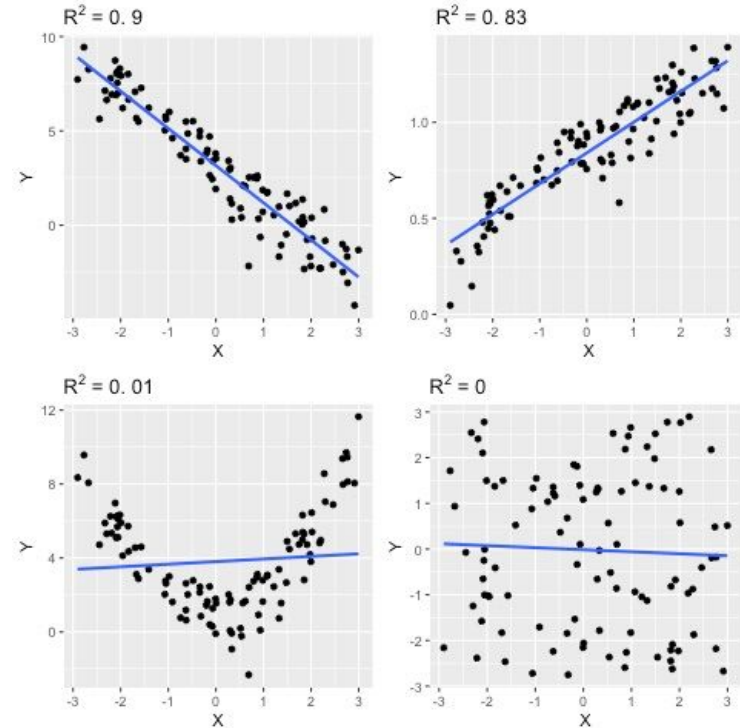
- Trouve la droite qui minimise la somme des carrés des distances verticales entre les points de données et la droite de régression (**les résidus**).
- Les résidus représentent l'erreur entre la valeur prédite par le modèle et la valeur réelle.



Évaluation de la qualité de l'ajustement

Coefficient de détermination (R^2)

- Proportion de la variance de Y expliquée par le modèle
- Varie entre 0 et 1 (plus R^2 est proche de 1, meilleur est l'ajustement)



Évaluation de la qualité de l'ajustement

Coefficient de détermination (R^2)

- Proportion de la variance de Y expliquée par le modèle
- Varie entre 0 et 1 (plus R^2 est proche de 1, meilleur est l'ajustement)

Erreur quadratique moyenne (MSE)

- Moyenne des carrés des résidus
- Plus MSE est petit, meilleur est l'ajustement

$$\text{MSE} = \overset{\text{Mean}}{\frac{1}{n}} \sum_{i=1}^n \left(\overset{\text{Error}}{Y_i - \hat{Y}_i} \right) \overset{\text{Squared}}{^2}$$

Régression Linéaire Multiple

Régression linéaire simple vs. multiple

La régression linéaire multiple comme une extension naturelle de la régression simple, où **plusieurs variables explicatives** sont utilisées pour prédire **une variable dépendante**.

Régression linéaire simple vs. multiple

Régression Linéaire Simple :

- Modèle simple avec une seule variable explicative

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Régression Linéaire Multiple :

- Plusieurs variables explicatives :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \epsilon$$

Régression linéaire simple vs. multiple

Régression Linéaire Simple :

- Modèle simple avec une seule variable explicative

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Régression Linéaire Multiple :

- Plusieurs variables explicatives :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \epsilon$$

*Expliquer que la régression linéaire multiple permet de prendre en compte **plusieurs facteurs** qui influencent la variable dépendante Y*

Forme Générale du Modèle de Régression Linéaire Multiple

La formule du modèle est :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \epsilon$$

Forme Générale du Modèle de Régression Linéaire Multiple

La formule du modèle est :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$



variable dépendante

Forme Générale du Modèle de Régression Linéaire Multiple

La formule du modèle est :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \epsilon$$

variable dépendante

variables explicatives

Forme Générale du Modèle de Régression Linéaire Multiple

La formule du modèle est :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \epsilon$$

variable dépendante

coefficients

variables explicatives

Forme Générale du Modèle de Régression Linéaire Multiple

La formule du modèle est :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \epsilon$$

variable dépendante

coefficients

erreur

variables explicatives

Estimation des Coefficients β

On estime les β en minimisant la somme des carrés des écarts entre les valeurs observées et prédites de \mathbf{Y} .

Plus exactement, on minimise :

$$S(\beta_0, \beta_1, \dots, \beta_k) = \sum_{i=1}^n (Y_i - (\beta_0 + \beta_1 X_{i1} + \dots + \beta_k X_{ik}))^2$$

Interprétation des Coefficients

Chaque coefficient β_j représente l'**effet marginal** de X_j sur Y c'est-à-dire la variation de Y pour une unité de variation de X_j , en gardant les autres variables constantes.

Exemple concret :

dans un modèle où Y est le prix d'une maison, β_1 pourrait représenter l'effet d'une chambre supplémentaire, toutes choses égales par ailleurs.



Qualité du Modèle : R^2 et R^2 Ajusté

Le **R^2** mesure la proportion de la variance de Y expliquée par le modèle. Il est donné par :

$$R^2 = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2}$$

Le **R^2 ajusté** tient compte du nombre de prédicteurs et pénalise l'ajout de variables non pertinentes

$$R^2_{ajusté} = 1 - \left(\frac{(1-R^2)(n-1)}{n-k-1} \right)$$

Multicolinéarité

La multicolinéarité survient lorsque les variables explicatives sont fortement corrélées entre elles, rendant difficile l'estimation précise des coefficients.

Le **Variance Inflation Factor (VIF)** mesure cette colinéarité :

$$VIF_j = \frac{1}{1-R_j^2}$$

Si **VIF_j > 10**, cela indique une forte colinéarité et un problème potentiel dans le modèle.

Sélection de Variables (Feature Selection)

Méthodes :

- Sélection pas à pas : ajouter ou retirer des variables explicatives.
- Critères **AIC/BIC** : critères de sélection basés sur l'ajustement du modèle et la pénalisation de la complexité.

$$AIC = 2k - 2 \ln(L)$$

$$BIC = k \ln(n) - 2 \ln(L)$$

Sélection de Variables (Feature Selection)

Méthodes :

- Sélection pas à pas : ajouter ou retirer des variables explicatives.
- Critères **AIC/BIC** : critères de sélection basés sur l'ajustement du modèle et la pénalisation de la complexité.

Le **maximum de vraisemblance** estime les paramètres qui rendent les données observées les plus probables.

$$AIC = 2k - 2\ln(L)$$

$$BIC = k \ln(n) - 2\ln(L)$$

L : Maximum de vraisemblance

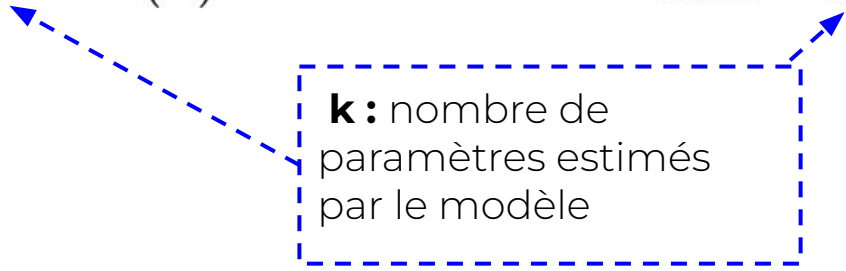
Sélection de Variables (Feature Selection)

Méthodes :

- Sélection pas à pas : ajouter ou retirer des variables explicatives.
- Critères **AIC/BIC** : critères de sélection basés sur l'ajustement du modèle et la pénalisation de la complexité.

$$AIC = 2k - 2\ln(L)$$

$$BIC = k\ln(n) - 2\ln(L)$$



k : nombre de
paramètres estimés
par le modèle

The diagram consists of a dashed blue rectangular box containing the text 'k : nombre de paramètres estimés par le modèle'. Two dashed blue arrows originate from the box: one points to the '2k' term in the AIC formula on the left, and the other points to the 'k ln(n)' term in the BIC formula on the right.

Sélection de Variables (Feature Selection)

Méthodes :

- Sélection pas à pas : ajouter ou retirer des variables explicatives.
- Critères **AIC/BIC** : critères de sélection basés sur l'ajustement du modèle et la pénalisation de la complexité.

$$AIC = 2k - 2\ln(L)$$

$$BIC = k \ln(n) - 2\ln(L)$$

n : taille de l'échantillon



Régression Linéaire Généralisée

Introduction

Limites : ne fonctionne pas bien avec des variables **dépendantes non continues** (par exemple, variables binaires ou comptages).

Exemple : modélisation du nombre d'événements, ou probabilité d'un événement (régression logistique).

⇒ Introduction de la **régression généralisée** pour traiter ces cas.

Structure Générale des Modèles Linéaires Généralisés

GLM a trois composantes principales :

- **Composante aléatoire** : distribution des erreurs.
- **Composante systématique** : relation linéaire entre les prédicteurs et une fonction des espérances.
- **Fonction de lien** : relie la moyenne de la variable dépendante à une combinaison linéaire des prédicteurs.

Forme du Modèle GLM

Formule générale : $g(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$

$g(\mu)$ est la fonction de lien.

μ est la moyenne de la variable dépendante \mathbf{Y} , et les β_j sont les coefficients des variables explicatives \mathbf{X}_j

Exemples de Distributions dans GLM

Exemples de distributions utilisées dans les GLM :

- **Régression logistique** : variable binaire, distribution de Bernoulli.
- **Régression de Poisson** : comptage, distribution de Poisson.
- **Régression Gamma** : données positives continues, distribution Gamma.

Fonctions de Lien Courantes

La fonction de lien transforme la moyenne μ pour qu'elle soit linéairement reliée aux prédictors :

- Lien logit (régression logistique) : $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$
- Lien logarithmique (régression de Poisson) : $g(\mu) = \log(\mu)$
- Lien identité (modèle linéaire classique) : $g(\mu) = \mu$

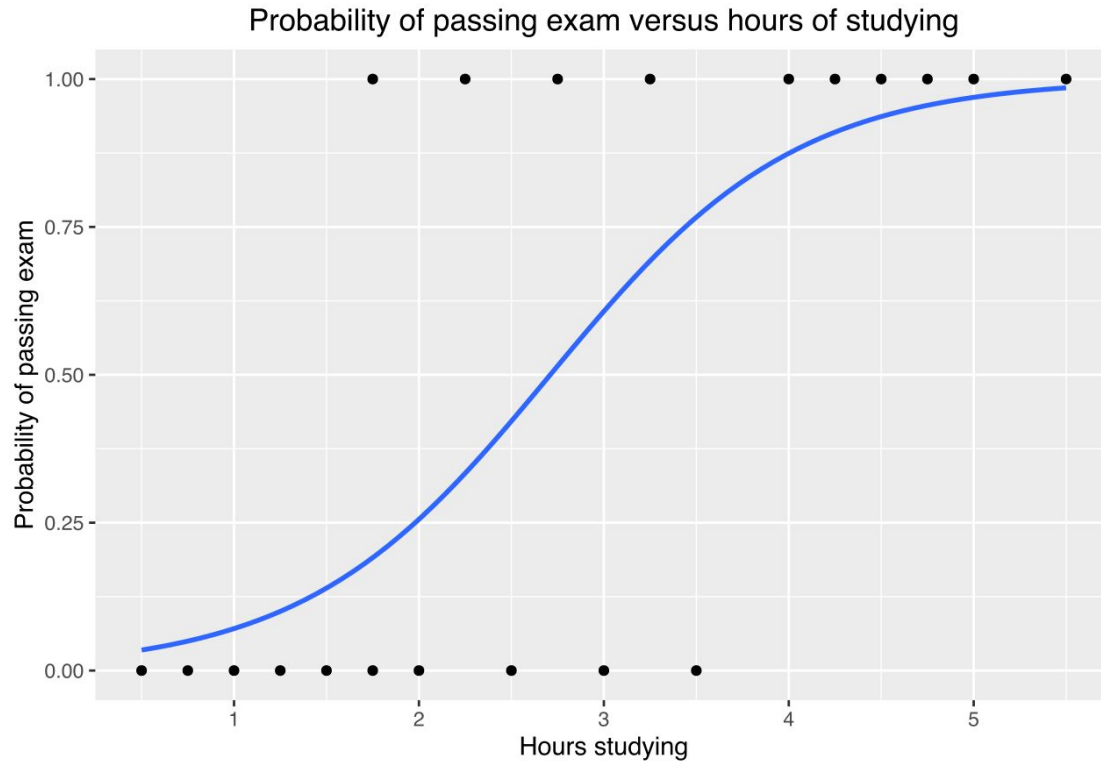
Régression Logistique

Problème : modéliser la probabilité d'un événement (succès ou échec).

Modèle : $\log \left(\frac{p}{1-p} \right) = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k$

p est la probabilité que **Y=1**

⇒ Expliquer comment prédire des probabilités avec des variables explicatives.



Régression entre le nombre d'heures de travail et la réussite à l'examen

Régression de Poisson

Problème : modéliser des données de comptage (par exemple, nombre d'accidents).

Modèle : $\log(\mu) = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k$

μ est le nombre attendu d'événements.

⇒ Cas d'application : comptage du nombre d'occurrences d'un événement.

Estimation des Coefficients dans GLM

Méthode de l'estimation des **moindres carrés généralisés** pour minimiser la divergence entre les valeurs observées et les valeurs prédites.

L'estimation se fait via la **vraisemblance** : $L(\beta) = \prod_{i=1}^n f(y_i | \mu_i)$

Maximisation de la fonction de vraisemblance pour obtenir les coefficients β

Diagnostic des Modèles GLM

Analyser les **résidus déviance** et les **résidus Pearson** pour vérifier la qualité du modèle.

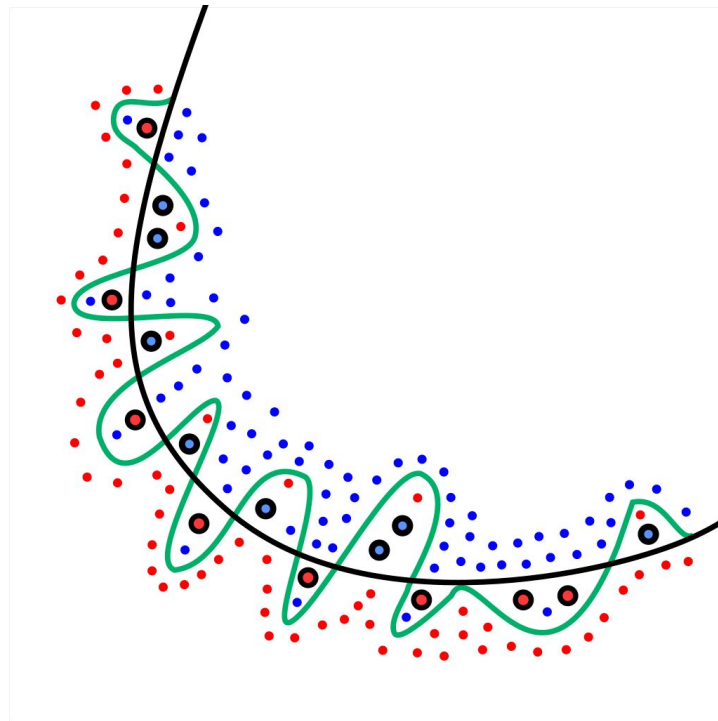
Vérification de l'ajustement du modèle : analyse des graphes des résidus, tests de déviance.

Régression Pénalisée et Régularisation

Problème du Surajustement (Overfitting)

Lorsque le modèle est **trop complexe** et s'ajuste trop bien aux données d'entraînement, au détriment de sa capacité à généraliser.

Exemple : un modèle avec trop de prédicteurs qui capte le bruit des données plutôt que les tendances sous-jacentes.



Principe de la Régularisation

L'idée est d'ajouter une **pénalité** au modèle pour éviter que les coefficients ne prennent des valeurs trop grandes (qui refléteraient un surajustement).

On modifie la fonction à minimiser en ajoutant un terme de régularisation qui pénalise la complexité du modèle.

Régression Lasso (ou L1)

La régression Lasso (*Least Absolute Shrinkage and Selection Operator*) ajoute une pénalité basée sur la **somme des valeurs absolues des coefficients**.

Minimiser :

$$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \lambda \sum_{j=1}^k |\beta_j|$$

Régression Lasso (ou L1)

La régression Lasso (*Least Absolute Shrinkage and Selection Operator*) ajoute une pénalité basée sur la **somme des valeurs absolues des coefficients**.

Minimiser :
$$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \lambda \sum_{j=1}^k |\beta_j|$$

Certains coefficients peuvent être **exactement nuls**, ce qui permet de sélectionner automatiquement les variables.

Utile pour la sélection de variables **lorsque k est grand**.

Régression Ridge (ou L2)

La régression Ridge ajoute une pénalité sur la **somme des carrés des coefficients**

On minimise :

$$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \lambda \sum_{j=1}^k \beta_j^2$$

Régression Ridge (ou L2)

La régression Ridge ajoute une pénalité sur la **somme des carrés des coefficients**

On minimise :

$$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \lambda \sum_{j=1}^k \beta_j^2$$

Contrôle la force de la régularisation :

- Si $\lambda=0$, c'est la régression linéaire ordinaire.
- Si λ est grand, les coefficients seront fortement contraints vers 0.

Régression Élastique (Elastic Net)

L'Elastic Net combine les pénalités L1 (Lasso) et L2 (Ridge) :

Minimiser :
$$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \lambda_1 \sum_{j=1}^k |\beta_j| + \lambda_2 \sum_{j=1}^k \beta_j^2$$

⇒ Peut résoudre les limitations du Lasso en présence de **variables corrélées**.

Choix de l'Hyperparamètre λ

Importance du choix de λ :

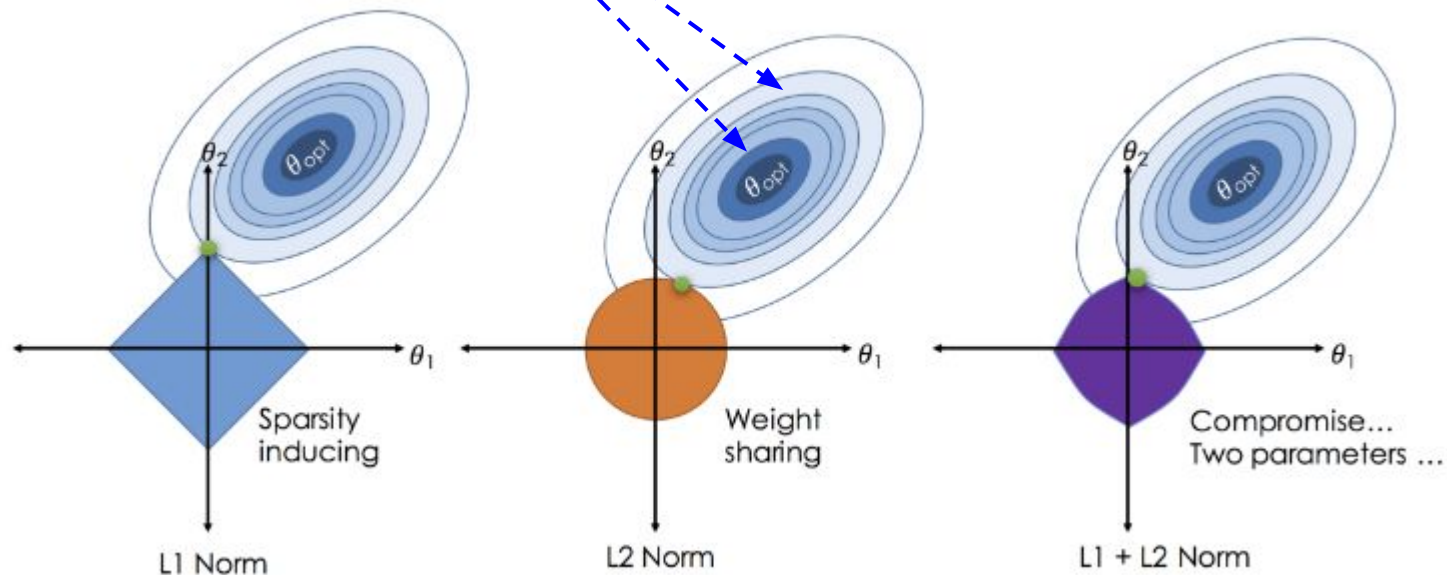
- trop petit, peu d'effet sur la régularisation
- trop grand, le modèle est sous-ajusté.

Utilisation de la **validation croisée** pour choisir la valeur optimale de λ

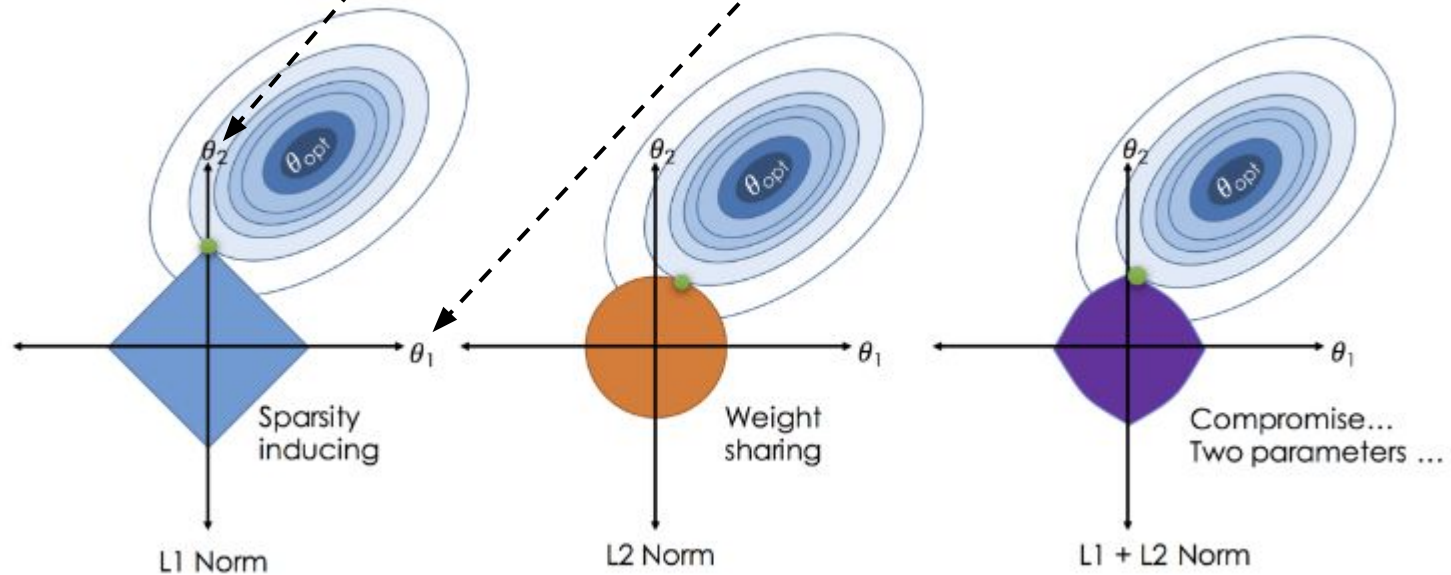
Méthode graphique : graphe de validation croisée montrant l'erreur en fonction de λ

Ridge vs Lasso vs ElasticNet

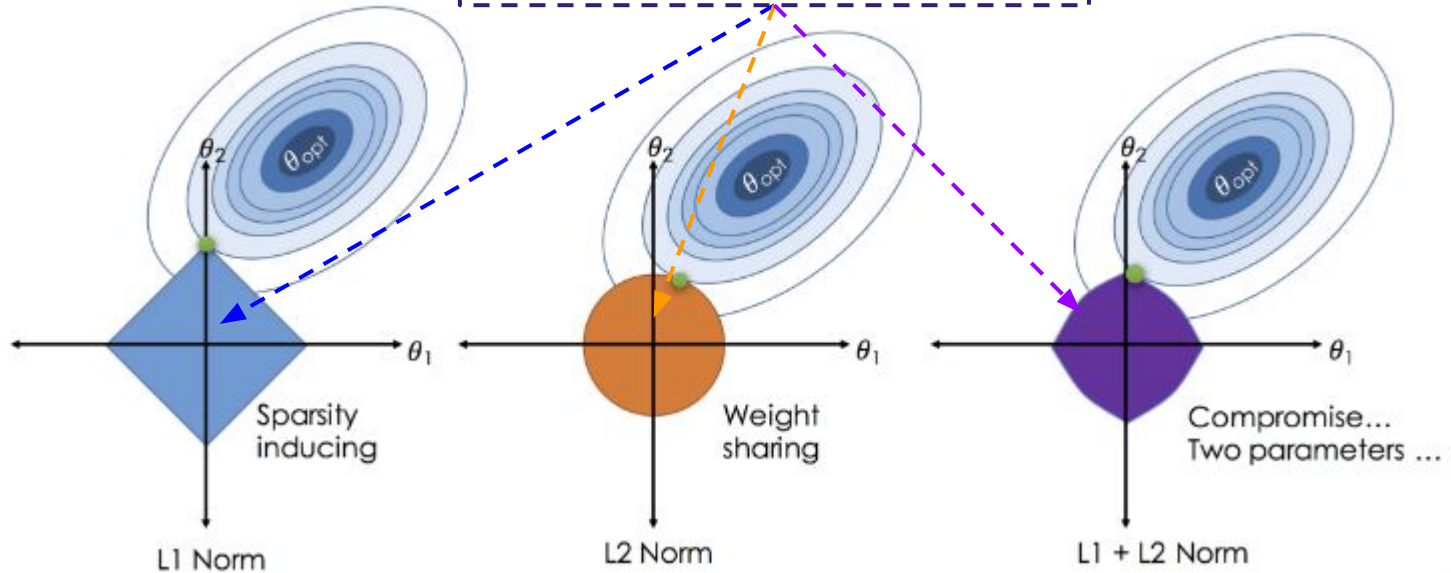
Représentent les courbes de niveau d'une **fonction d'erreur**. Le but est de trouver le point où cette fonction est minimale (le centre des ellipses).



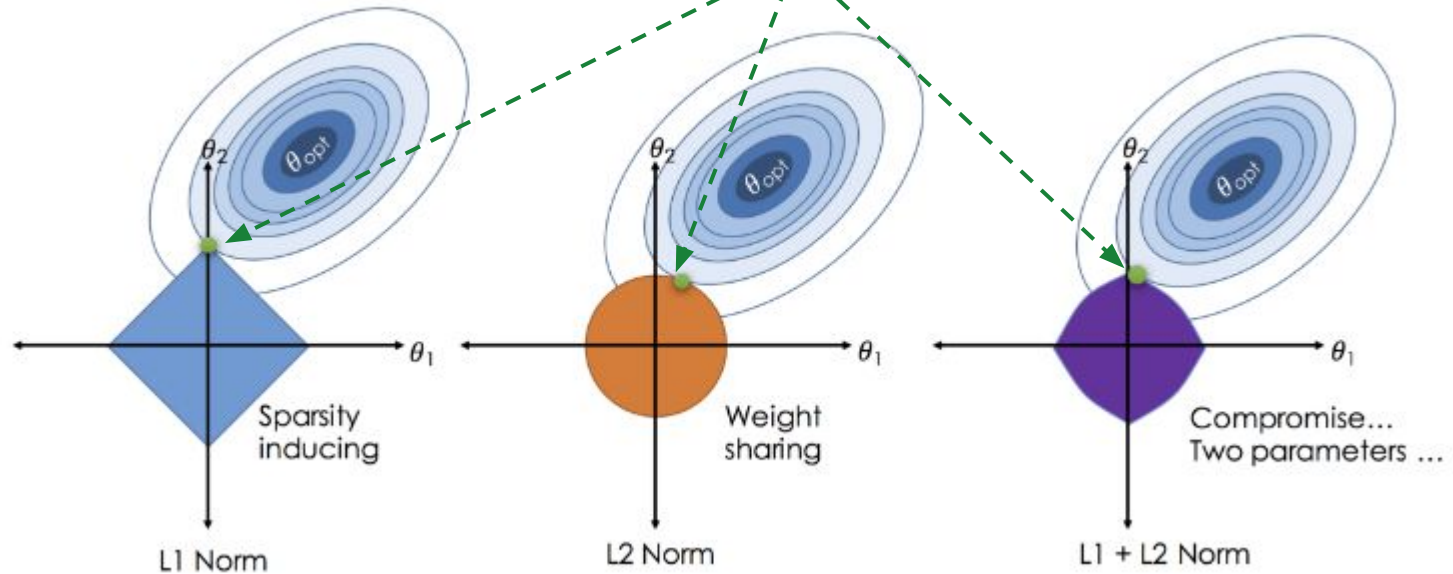
Les deux **paramètres** (ou poids) du modèle que l'on cherche à optimiser



Contraintes imposées par les normes de régularisation.

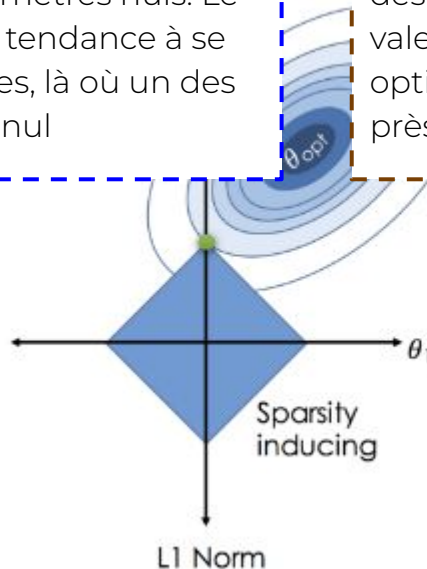


Le point optimal recherché, minimisant la fonction de coût tout en respectant les contraintes de régularisation.



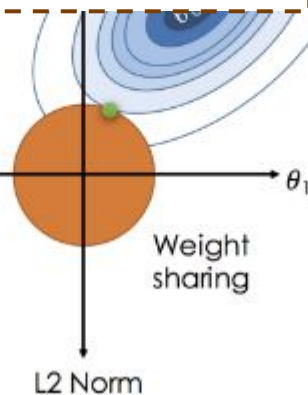
Lasso :

Favorise la **sparsité**, c'est-à-dire des solutions avec de nombreux paramètres nuls. Le point optimal a tendance à se situer sur les axes, là où un des paramètres est nul



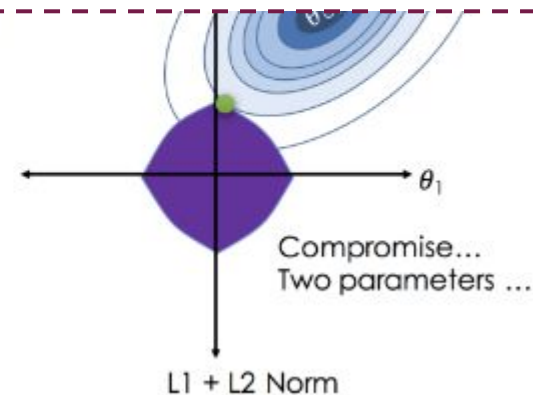
Ridge :

Favorise le **weight sharing**, c'est-à-dire des solutions avec des paramètres de petite valeur mais non nuls. Le point optimal a tendance à se situer près de l'origine



ElasticNet :

Un **compromis** entre les deux normes précédentes, permettant à la fois une certaine sparsité et des poids de petite valeur





Merci pour votre attention !