FIP - Analyse des données

Régression linéaire

Michel Crucianu

(prenom.nom@cnam.fr)

Département Informatique

Conservatoire National des Arts & Métiers, Paris, France

15 mai 2025

Plan du cours

2 Régression linéaire

- Définition, illustration
- Estimation du modèle : une seule variable explicative
- Validation du modèle
- Prédiction avec le modèle
- Estimation du modèle : plusieurs variables explicatives
- Validation et diagnostic du modèle
- Contrôle de la complexité du modèle

Régression : données, objectifs

- Statistique inférentielle, construction de modèle prédictif
- Données : n observations $\{(\mathbf{x}_i, y_i)\}, 1 \leq i \leq n$, caractérisées par $m \geq 1$ variables quantitatives explicatives X et une variable quantitative expliquée Y
- Objectif : prédire la valeur de la variable expliquée Y à partir des valeurs prises par les variables explicatives X

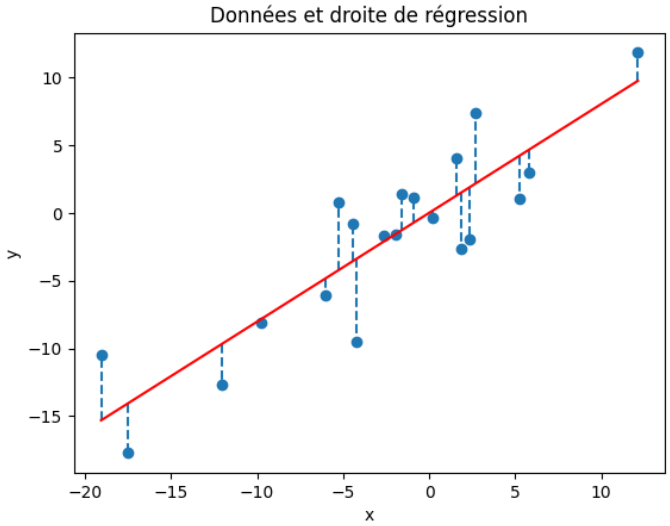


FIG. 1 – Illustration de la régression linéaire : variable explicative X, variable expliquée Y

Régression : dans quelles situations est-ce utile

1. La variable expliquée est difficile/coûteuse à mesurer alors que les variables explicatives peuvent être mesurées facilement
   - Ex. : estimer la résistance à la traction d'un polymère (dont la mesure destructive est coûteuse) à partir de variables caractérisant le monomère et la polymérisation
2. Les valeurs des variables explicatives peuvent être connues avant celle de la variable expliquée et une estimation en avance de cette dernière est utile
   - Ex. : prédire le volume d'algues vertes sur des plages à partir de la quantité d'engrais utilisés dans le bassin hydrographique correspondant et de la température moyenne de l'eau lors du trimestre précédent
3. On souhaite contrôler les variables explicatives pour obtenir une valeur désirée pour la variable expliquée
   - Ex. : contrôler la concentration d'un produit d'une réaction chimique à partir des quantités initiales de réactifs, de la température et de la pression

◀▶ Remarque : une relation causale entre variable(s) explicative(s) et variable expliquée n'est pas nécessaire, la simple corrélation suffit

Régression linéaire simple

- Variables aléatoires : variable explicative X, variable expliquée Y
- Régression : le modèle recherché fait dépendre un indicateur de Y (en général l'espérance) de la valeur prise par X à travers une fonction f :

$$\mathbb{E}(Y|X = x) = f(x), \text{ ou } Y = f(x) + \epsilon$$

  ϵ étant une variable aléatoire d'espérance nulle
- Régression linéaire : f est une fonction affine $f(x) = \omega_0 + \omega_1 x$

$$Y = \omega_0 + \omega_1 X + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$$

  - « simple » : une seule variable explicative
- Intérêt de la régression linéaire :
  - Simplicité d'estimation et d'utilisation
  - Méthodes bien fondées de validation
  - Lisibilité : la forme additive rend explicite la dépendance de chaque variable explicative

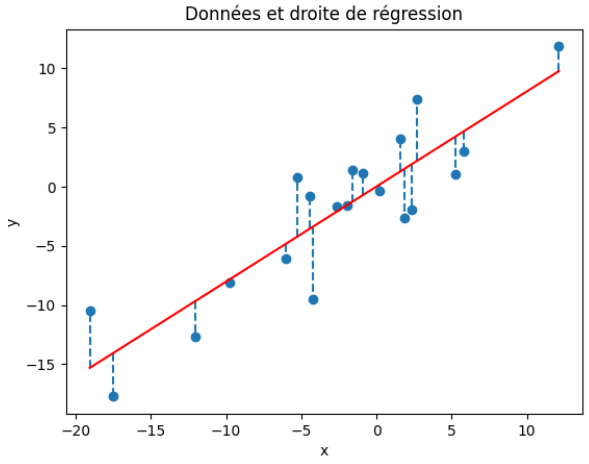
Estimation du modèle

- Revient à estimer les paramètres ω_0, ω_1 à partir d'observations $\{(x_i, y_i)\}, 1 \leq i \leq n$
- Soit w_0 l'estimation de ω_0 et w_1 l'estimation de ω_1, nous avons alors

$$y_i = w_0 + w_1 x_i + \epsilon_i, \quad \epsilon_i \text{ étant les résidus}$$

- Méthode d'estimation : critère des moindres carrés

$$\arg\min_{w_0, w_1} \sum_{i=1}^{n} \epsilon_i^2, \text{ c'est à dire } \arg\min_{w_0, w_1} \sum_{i=1}^{n} (y_i - w_0 - w_1 x_i)^2$$



Estimation du modèle (2)

- Résultats pour w_1 (pente de la droite) et w_0 (ordonnée pour $x = 0$ ou *intercept*) :

$$\begin{aligned} w_1 &= \frac{s_{xy}}{s_x^2} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2} \\ w_0 &= \overline{y} - w_1\overline{x} \end{aligned}$$

- Estimation de σ_ϵ^2 : $\frac{1}{n-2}\sum_{i=1}^{n}(y_i - w_0 - w_1 x_i)^2$
- Propriétés des estimateurs :
  - Sans biais pour ω_0 et ω_1
  - Sans biais pour σ_ϵ^2
  - Variance de l'estimateur de ω_1 : $s_b^2 = \frac{\sigma_\epsilon^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}$
  - Variance de l'estimateur de ω_0 : $s_a^2 = \sigma_\epsilon^2\left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right)$

La relation trouvée est-elle significative ?

- Les paramètres estimés à partir de l'échantillon $\{(\mathbf{x}_i, y_i)\}_{1 \leq i \leq n}$ peuvent-ils être assimilés à 0, ce qui voudrait dire que Y ne dépend pas de X ?

→ Tests sur les paramètres individuels :

- Test sur la pente (ω_1) : test sur la liaison entre X et Y
  - $H_0 : \omega_1 = 0$ (absence de liaison entre X et Y), $H_1 : \omega_1 \neq 0$
  - Statistique de test : $t_{w_1} = \frac{w_1}{s_{w_1}}$; H_0 vraie $\Rightarrow$ t_{w_1} suit une loi de Student de paramètre $n-2$
  - → Si $|t_{w_1}| \geq t_{1-\frac{\alpha}{2};(n-2)}$ on rejette H_0 au risque α
  - → Si on ne peut pas rejeter H_0, on ne peut pas se servir du modèle
- Test sur l'ordonnée pour $x = 0$ (ω_0) :
  - $H_0 : \omega_0 = 0$, $H_1 : \omega_0 \neq 0$
  - Statistique de test : $t_{w_0} = \frac{w_0}{s_{w_0}}$; H_0 vraie $\Rightarrow$ t_{w_0} suit une loi de Student de paramètre $n-2$
  - → Si $|t_{w_0}| \geq t_{1-\frac{\alpha}{2};(n-2)}$ on rejette H_0 au risque α
  - → Si on ne peut pas rejeter H_0, on cherche un modèle $Y = \omega_1 X + \epsilon$

Le modèle trouvé est-il utile ?

- Objectif de la modélisation : prédire les valeurs de Y à partir de X, $\hat{y} = w_0 + w_1 x$
- Décomposition de la variance :

$$\underbrace{\sum_{i=1}^{n}(y_i - \overline{y})^2}_{\text{variabilité totale}} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}_{\text{expliquée par le modèle}} + \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\text{résiduelle}}$$

- Qualité prédictive d'un modèle : coefficient de détermination R^2 défini par

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

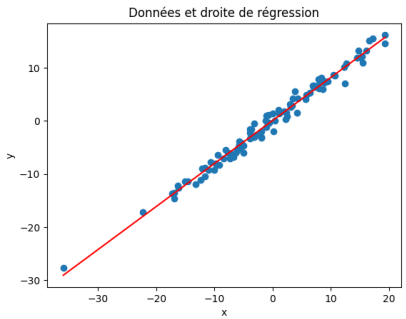
ou

$$R^2 = \frac{\text{variabilité de } y \text{ expliquée par le modèle}}{\text{variabilité totale de } y}$$

- $R^2 \geq 0$ ($\Leftarrow$ rapport de sommes de carrés), $R^2 \leq 1$ ($\Leftarrow$ décomposition de la variance)

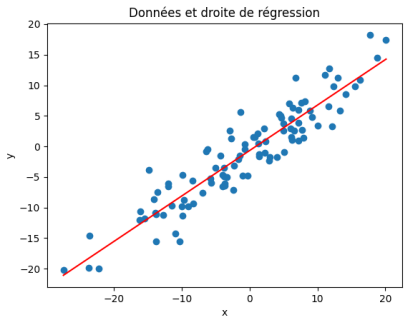
Le modèle trouvé est-il utile ? (2)

- Plus R^2 est proche de 1, meilleur est le modèle
- Illustration : 4 modèles estimés à partir d'échantillons ($n = 100$) correspondant à une même pente $\omega_1 = 0.8$ et *intercept* $\omega_0 = 0$, mais obtenus avec des valeurs croissantes pour σ_ϵ :



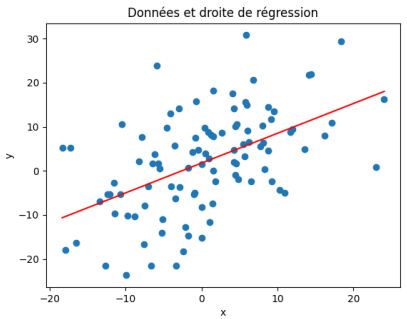
$R^2 = 0.98$

$w_1 = 0.79$



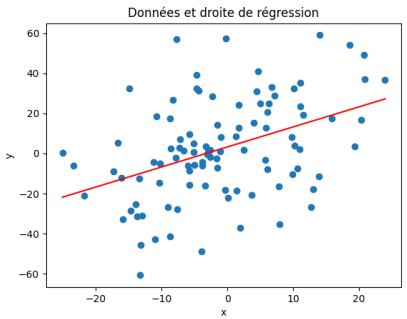
$R^2 = 0.88$

$w_1 = 0.81$



$R^2 = 0.37$

$w_1 = 0.72$



$R^2 = 0.19$

$w_1 = 0.69$

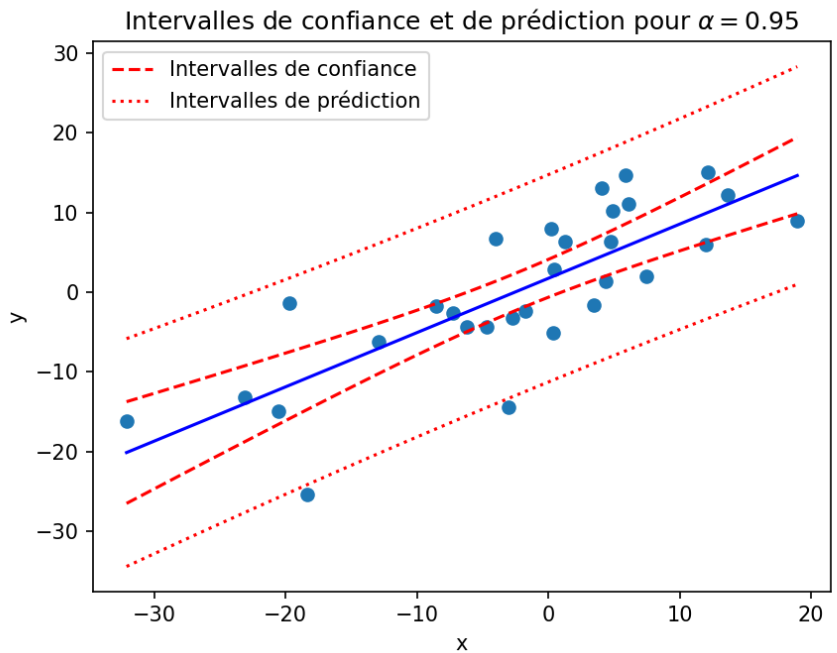
- Observation : pour σ_ϵ élevé, la pente estimée (w_1) s'éloigne de la vraie pente (ω_1)

Prédiction avec le modèle linéaire

- Modèle $(w_0, w_1) \Rightarrow$ pour une valeur x de la variable explicative X on peut estimer

$$\hat{y} = w_0 + w_1 x$$

- Sources de variation :
  1. Variabilité de l'estimation de (ω_0, ω_1) par (w_0, w_1) issus de l'échantillon $\{(x_i, y_i)\}_{1 \leq i \leq n} \rightarrow$ intervalles de confiance pour $\mathbb{E}[Y|X = x]$
  2. Variabilité de $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2) \rightarrow$ intervalles de prédiction pour $\hat{y}$



Régression linéaire multiple

- Prédire la valeur de la variable expliquée Y à partir des valeurs de plusieurs ($m > 1$) variables explicatives X_j :

$$Y = \omega_0 + \sum_{j=1}^{m} \omega_j X_j + \epsilon, \quad 1 \leq j \leq m, \quad \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$$

- Soit w_j l'estimation de ω_j, $1 \leq j \leq m$, nous avons alors les relations suivantes :

$$y_i = w_0 + \sum_{j=1}^{m} w_j x_{ij} + \epsilon_i, \quad 1 \leq i \leq n, 1 \leq j \leq m, \quad \epsilon_i \text{ étant les résidus (inconnus)}$$

- Estimation des ω_j à partir de $\{(\mathbf{x}_i, y_i)\}, 1 \leq i \leq n$, sur la base du même critère des moindres carrés :

$$\arg \min_{w_0 \ldots w_m} \sum_{i=1}^{n} \epsilon_i^2$$

Régression linéaire multiple : expression matricielle

- On introduit une variable constante $X_0 = 1$ que multiplie ω_0

$\rightarrow$ L'expression de Y devient

$$Y = \sum_{j=0}^{m} \omega_j X_j + \epsilon, \quad 0 \leq j \leq m, \quad \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$$

- Le modèle estimé s'écrit alors $\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\epsilon}$

  $\mathbf{y}$ est le vecteur des y_i (n composantes)

  $\mathbf{X}$ est la matrice des x_{ij} (matrice $n \times (m+1)$)

  $\mathbf{w}$ est le vecteur des w_j ($m+1$ composantes)

  $\boldsymbol{\epsilon}$ est le vecteur des ϵ_i (n composantes)

- Le critère des moindres carrés devient

$$\arg\min_{\mathbf{w}} \|\boldsymbol{\epsilon}\|^2$$

Régression linéaire multiple : les paramètres estimés

- La solution obtenue par la minimisation de $\|\epsilon\|^2$ est

$$\mathbf{w} = \mathbf{X}^+ \mathbf{y}$$

où $\mathbf{X}^+$ (matrice $(m+1) \times n$) est la pseudo-inverse Moore-Penrose de la matrice $\mathbf{X}$

1. Si $\mathbf{X}^\top \mathbf{X}$ est inversible et bien conditionnée ($\frac{|\lambda_{\max}|}{|\lambda_{\min}|}$ pas très élevé), alors

$$\mathbf{X}^+ = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top$$

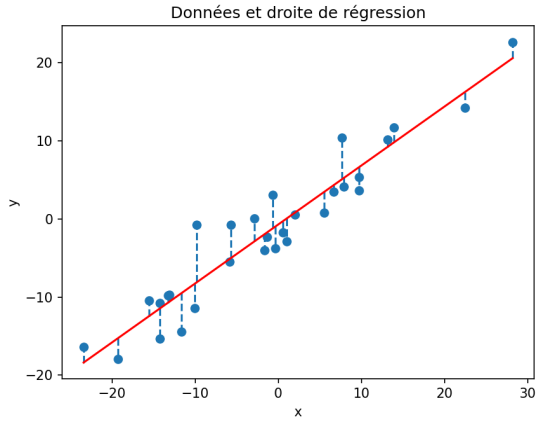
2. Si $\mathbf{X}^\top \mathbf{X}$ est inversible mais mal conditionnée ($\frac{|\lambda_{\max}|}{|\lambda_{\min}|}$ très élevé $\rightarrow$ l'inversion introduit de fortes erreurs d'arrondi) alors on régularise : $\mathbf{X}^+ = \left(\mathbf{X}^\top \mathbf{X} + r\mathbf{I}_{m+1}\right)^{-1} \mathbf{X}^\top$, avec $r > 0$ la constante de régularisation et $\mathbf{I}_{m+1}$ la matrice identité d'ordre $m+1$
3. $\mathbf{X}^\top \mathbf{X}$ n'est pas inversible dans les cas de multicolinéarité (dépendance linéaire entre plusieurs variables explicatives) $\rightarrow$ on vérifie les données (par exemple variables redondantes ?), on applique d'autres méthodes : régression sur composantes principales, moindres carrés partiels (*Partial Least Squares*), etc.

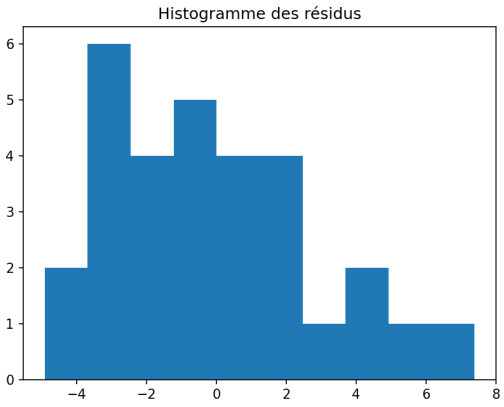
Hypothèses et vérifications

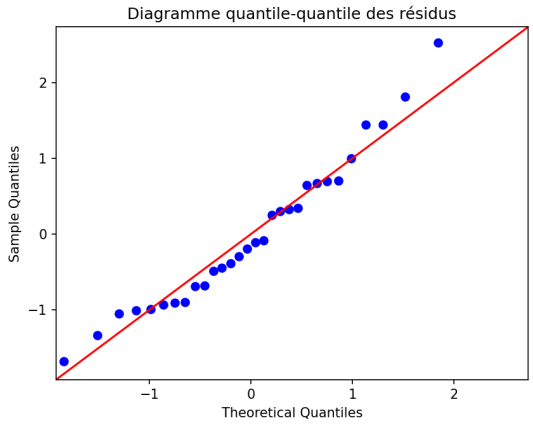
- Estimation du modèle et prédiction avec le modèle reposent sur des hypothèses :
  - Relation linéaire (ou affine) entre variable expliquée Y et variable(s) explicative(s) X
  - Résidus $\epsilon_i = y_i - \hat{y}_i$: suivent une loi normale, $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$
  - Résidus : de même variance σ_ϵ^2 sur tout le domaine de variation de Y

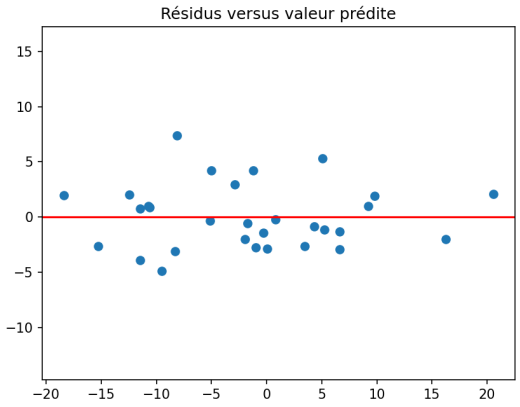
- Vérifications : tests possibles, mais la vérification graphique des résidus est usuelle
  - Graphique 2D des observations (uniquement avec une seule variable explicative) : linéarité de la relation entre variables, variance des résidus
  - Aspect de l'histogramme des résidus : loi normale, valeurs extrêmes
  - Diagramme quantile-quantile (*Q-Q plot*) des résidus : comparaison à la loi $\mathcal{N}(0, 1)$
  - Résidu en fonction de la valeur prédite : linéarité relation entre variables, variance constante (homoscédasticité)

Diagnostic du modèle : illustration cas normal

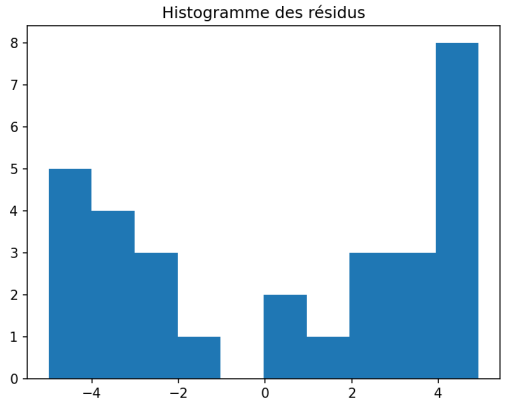


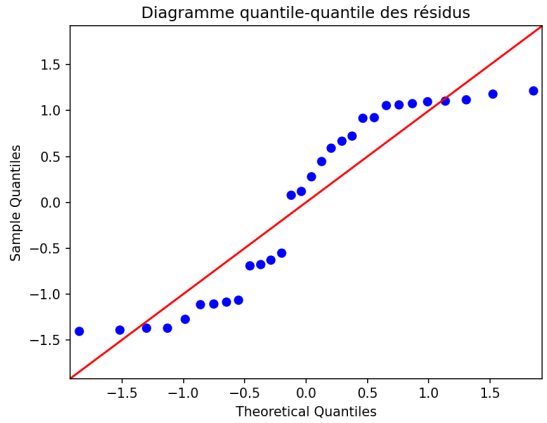


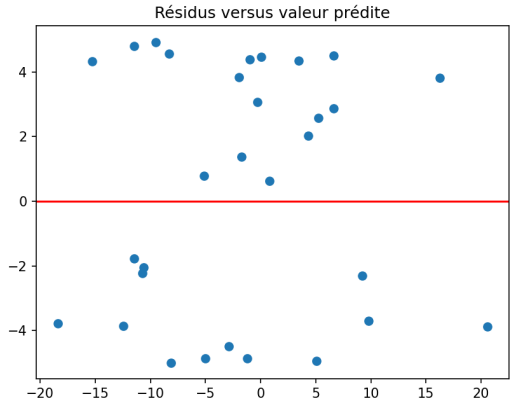


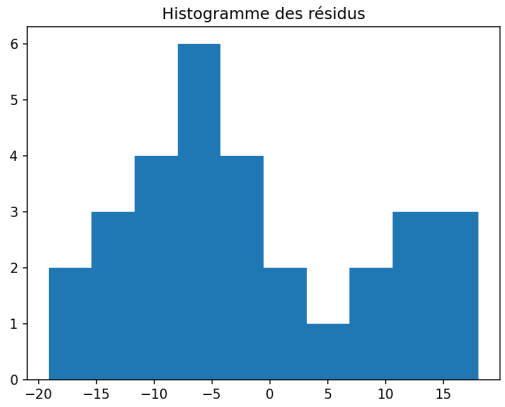


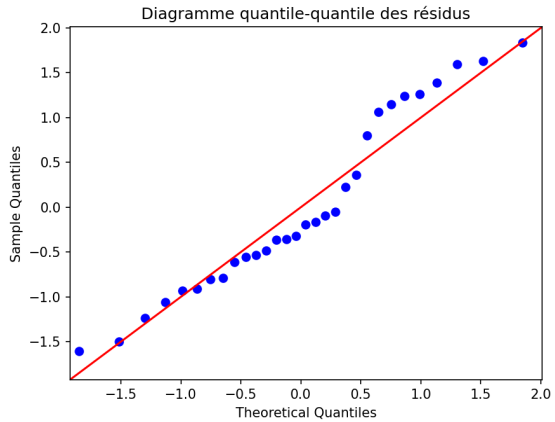
Diagnostic du modèle : illustration cas pathologiques

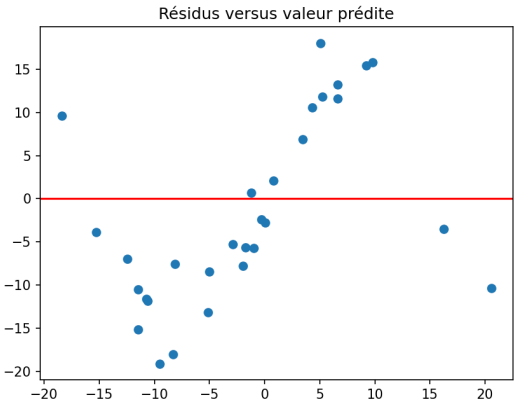












Régularisation de la régression linéaire

- Moindres carrés : minimisation de la somme des carrés des résidus $\arg\min_{\mathbf{w}} \|\boldsymbol{\epsilon}\|^2$
  - Fragilités : mauvais conditionnement de $\mathbf{X}^\top\mathbf{X}$, sensibilité aux valeurs extrêmes dans les données, certaines variables peu explicatives, etc.

→ L'inclusion d'un terme de régularisation peut améliorer la robustesse :

$$\arg\min_{\mathbf{w}} \left(\|\boldsymbol{\epsilon}\|^2 + \mathcal{R}(\mathbf{w})\right)$$

1. Régularisation L_1 ou LASSO : $\mathcal{R}(\mathbf{w}) = r\|\mathbf{w}\|_1$ ($\|\cdot\|_1$: somme des valeurs absolues)
   → Annule les coefficients w des variables trop peu explicatives (suivant la valeur de r)
   - Il n'y a pas de solution explicite pour régression linéaire multiple (sauf cas particuliers)
2. Régularisation L_2 ou *ridge* : $\mathcal{R}(\mathbf{w}) = r\|\mathbf{w}\|_2^2$, $r > 0$ est la pondération
   → Solution explicite pour régression linéaire multiple : $\mathbf{X}^+ = \left(\mathbf{X}^\top\mathbf{X} + r\mathbf{I}_{m+1}\right)^{-1}\mathbf{X}^\top$
3. Régularisation *elastic-net* : combinaison de *ridge* et LASSO