À la découverte des prédictions conformes

(et des statistiques)

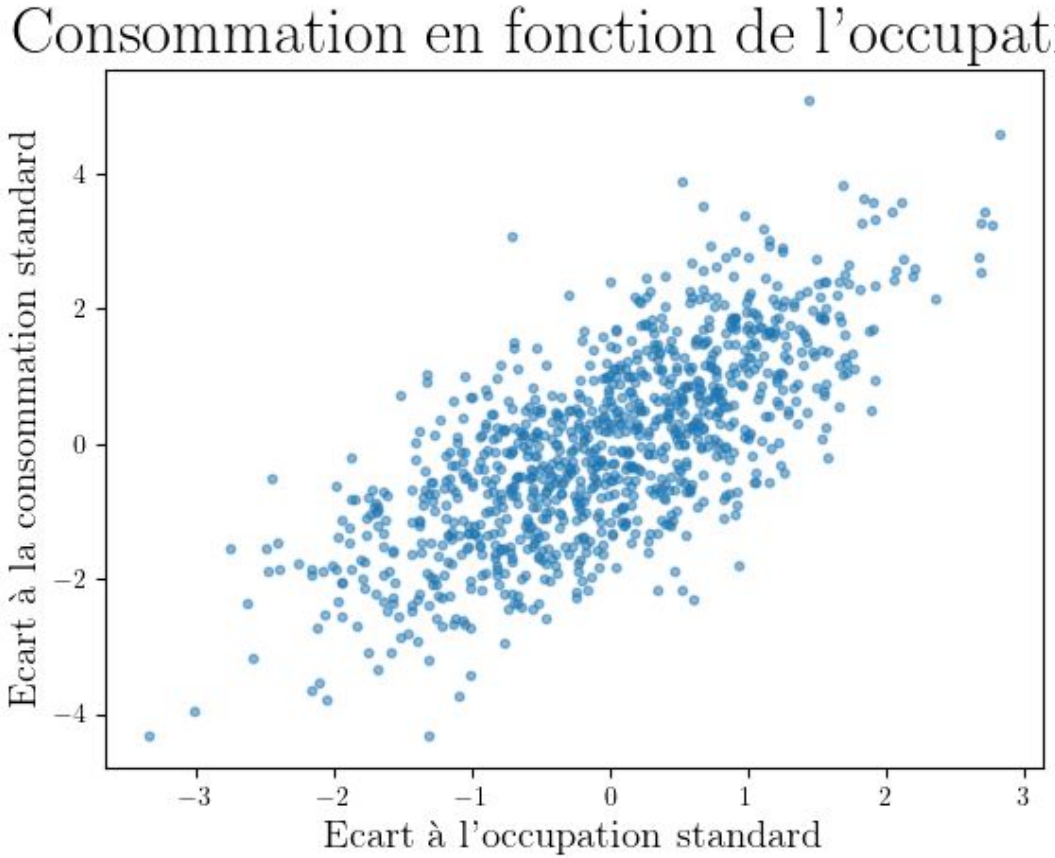
Contexte ? La consommation électrique de l’IF

👥 Taux d’occupation
🌡️ Température extérieure
🕗 Heure de la journée
…

Variables explicatives 👀

⚡ Consommation électrique

Variable expliquée 🎯

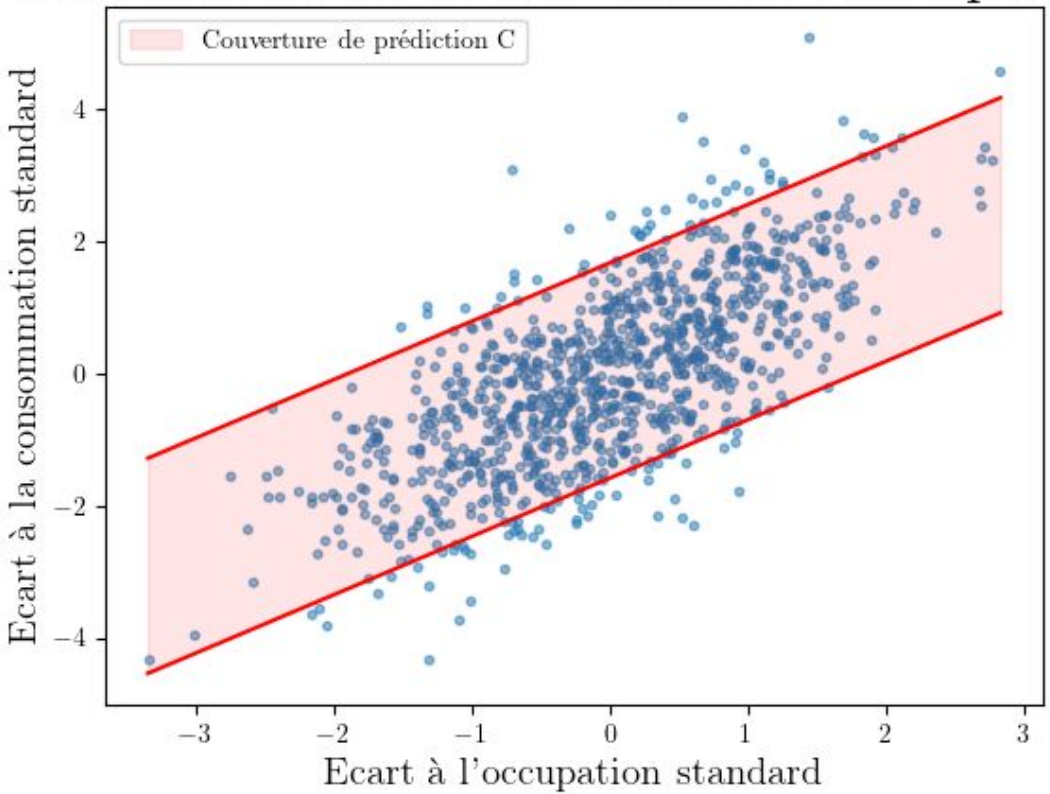


$$X_i \sim \mathcal{N}(0, 1)$$

$$\epsilon_i \sim \mathcal{N}(0, 1)$$

$$Y_i = X_i + \epsilon_i$$

Consommation en fonction de l'occupation



$$\alpha \in]0, 1[$$

$$\mathbb{P}[Y \in C(x)|X = x] \geq 1 - \alpha$$

Nouvelle observation

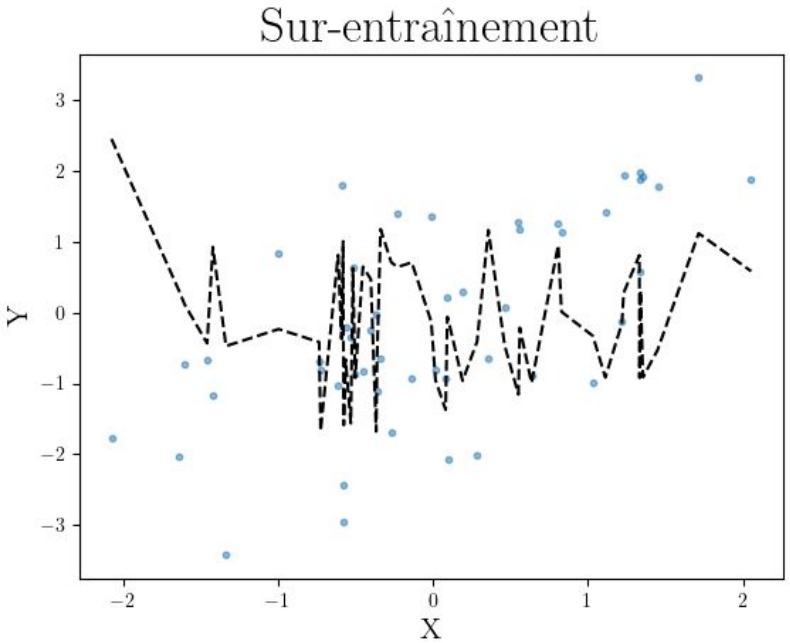
Définition (Statistique d'ordre)**.** *Soient $X_1, \dots, X_n$ des variables aléatoires réelles. On appelle* statistiques d'ordre *les variables aléatoires $X_{(1)}, \dots, X_{(n)}$ obtenues en réarrangeant les X_i par ordre croissant :*

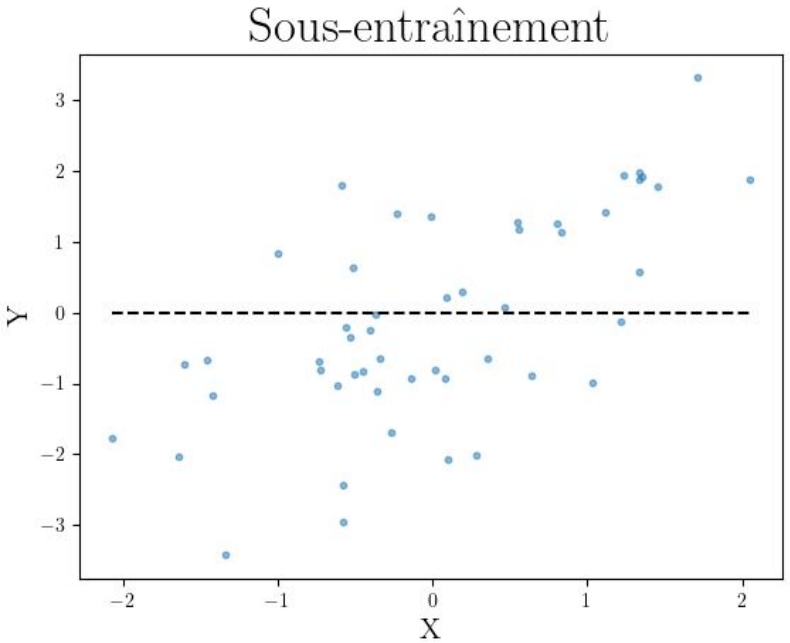
$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$$

Ainsi, pour tout $k \in [\![1; n]\!]$, $X_{(k)}$ représente la k-ième plus petite valeur parmi les X_i.

Proposition 1. *Soient $U_1, \dots, U_{n+1}$ des variables aléatoires réelles échangeables. On note F leur fonction de répartition commune qu'on suppose continue (pas de cas d'égalité) et f leur densité. On pose $U_{(1)}, \dots, U_{(n+1)}$ leur statistique d'ordre. Le rang de U_{n+1} parmi $U_1, \dots, U_{n+1}$ suit la loi uniforme sur $[\![1; n+1]\!]$*

La problématique





$\hat{\mu}$ la fonction de régression empirique

Données i.i.d.

Les données observées $Z_1, \ldots, Z_d$ sont i.i.d à valeur dans $\mathbb{R}^d \times \mathbb{R}$ issues d'une distribution P avec une fonction moyenne $\mu : x \mapsto \mathbb{E}[Y|X = x]$.

Bruit symétrique et indépendant)

Le bruit est symétrique et indépendant. C'est-à-dire que pour, $(X, Y) \sim P$ on a $\epsilon = Y - \mu(X)$ indépendant de X. De plus, la fonction densité de ϵ est symétrique et décroissante sur $[0, \infty[$.

Stabilité de l'échantillonage

Soit $\hat{\mu}_n$ la fonction de régression approchée entraînée sur n observations. Il existe (η_n) et (ρ_n) deux suites réelles convergeant vers 0 et une fonction $\tilde{\mu} : \mathbb{R}^d \to \mathbb{R}$ telles qu'à partir d'un certain rang n :

$$\mathbb{P}[\|\hat{\mu}_n - \tilde{\mu}\|_\infty \geq \eta_n] \leq \rho_n$$

Sensibilité à la perturbation

Pour $\mathcal{Y}$ un segment, en notant pour $y \in \mathcal{Y}$, $\hat{\mu}_{n,(X,y)}$ la fonction de régression entraînée sur n observation agrémentée du point (X, y). Il existe (η_n) et (ρ_n) deux suites réelles convergeant vers 0 telles que

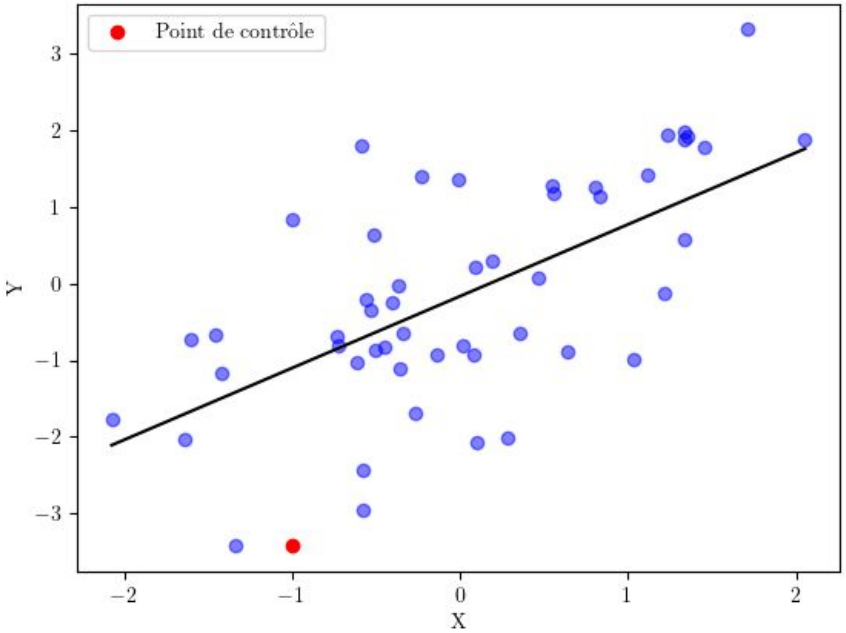
$$\mathbb{P}[\sup_{y \in \mathcal{Y}} \|\hat{\mu}_n - \hat{\mu}_{n,(X,y)}\|_\infty \geq \eta_n] \leq \rho_n$$

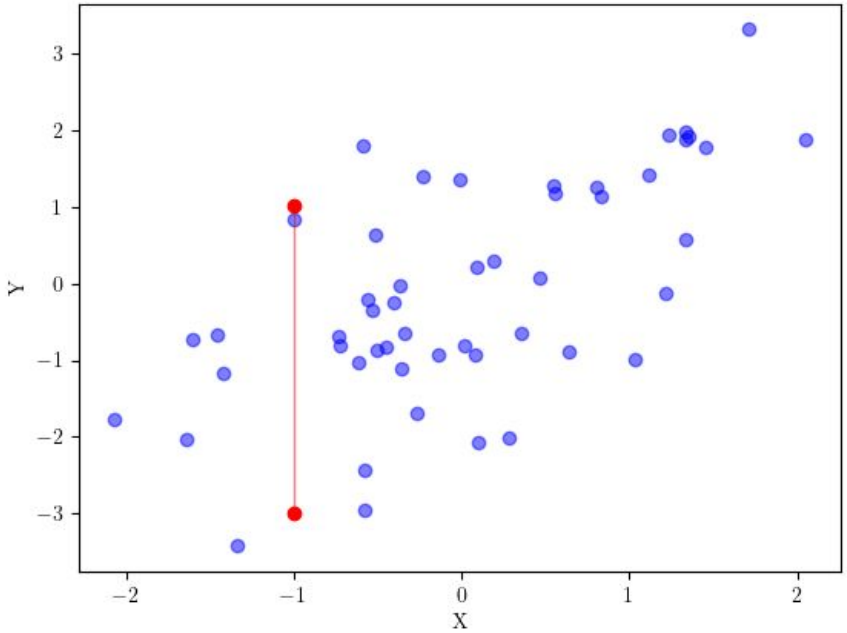
Cohérence de l'estimateur

Il existe (η_n) et (ρ_n) deux suites réelles convergentes vers 0 telles qu'à partir d'un certain rang

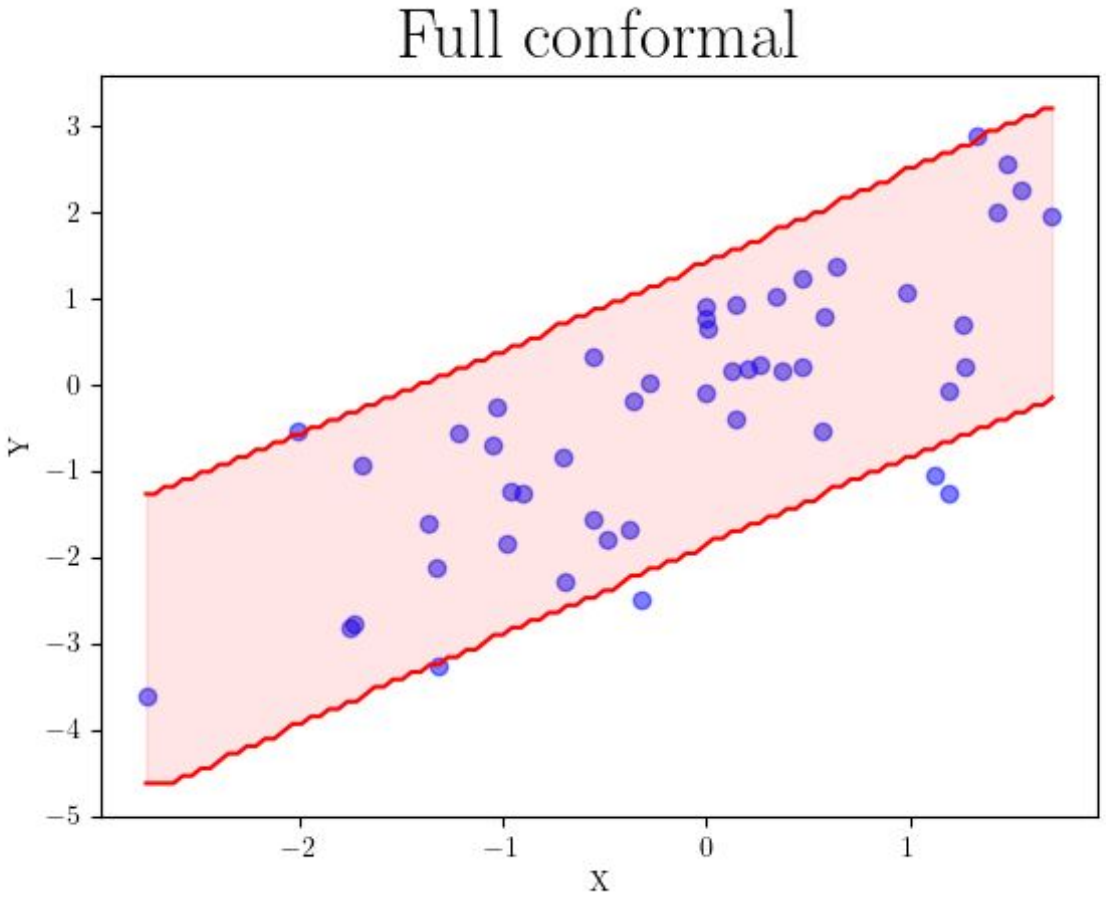
$$\mathbb{P}[\mathbb{E}[(\hat{\mu}_n(X) - \mu(X))^2 | \hat{\mu}_n] \geq \eta_n] \leq \rho_n$$

Full conformal





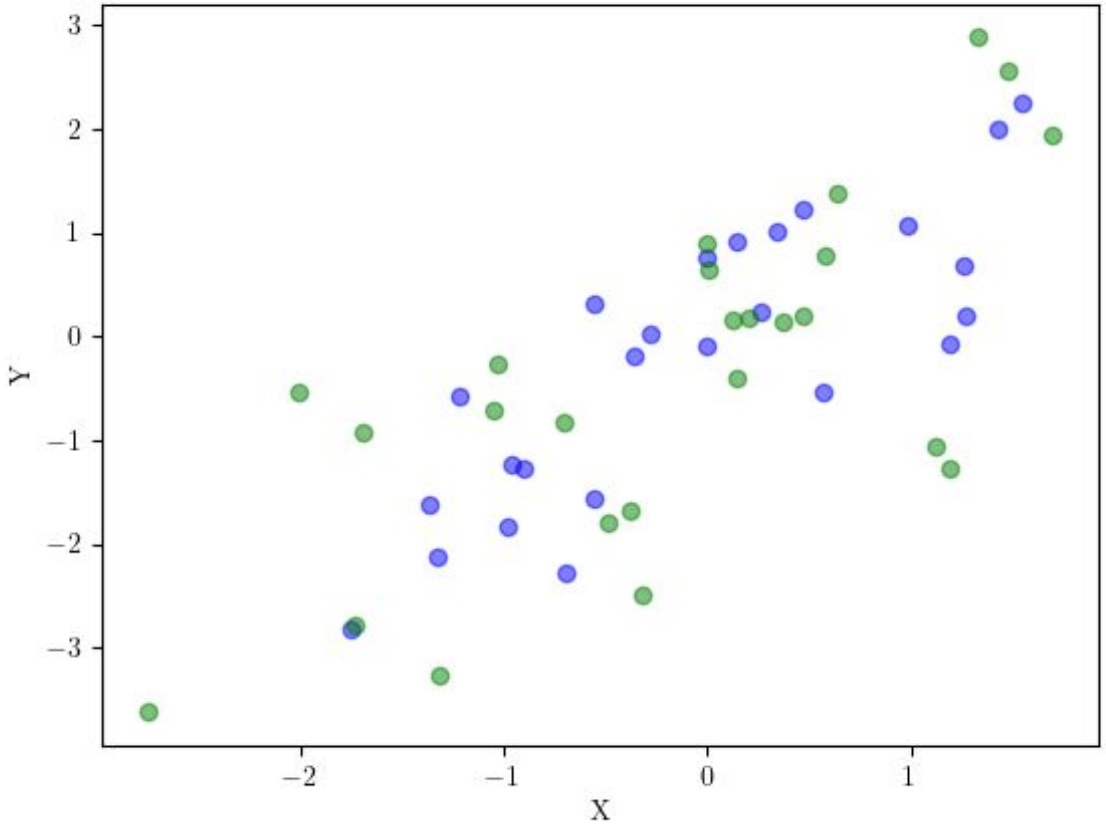
$R_{i,y} := |Y_i - \hat{\mu}_y(X_i)|$, le résidu absolu ajusté



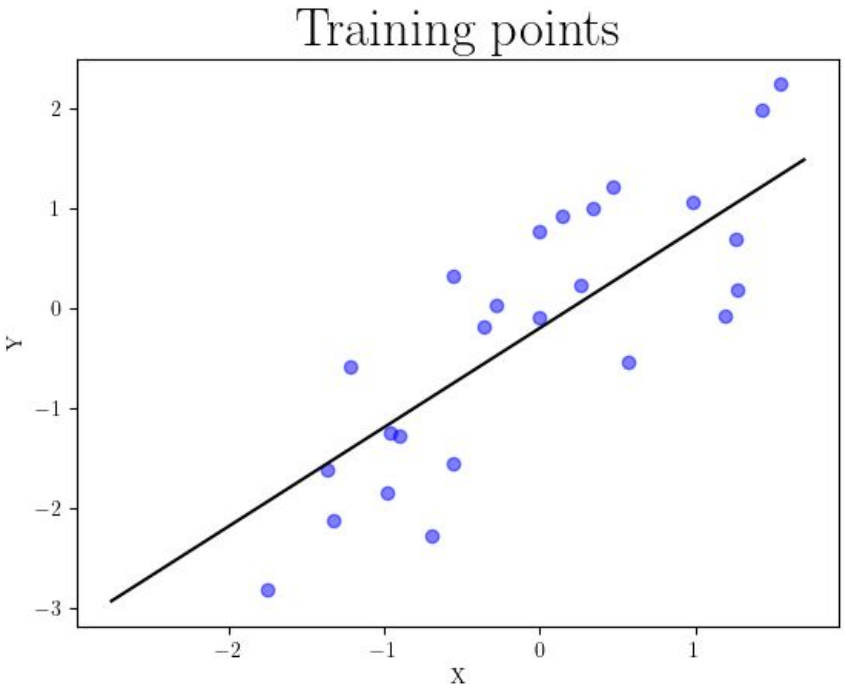
$$\#Y_{trial} = 100$$

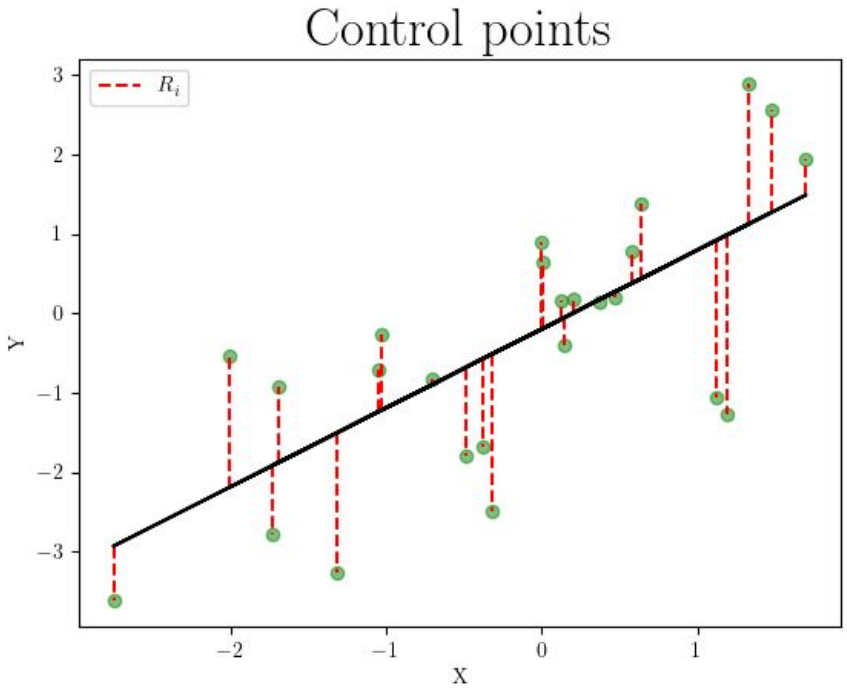
$$\#X_{conf} = 100$$

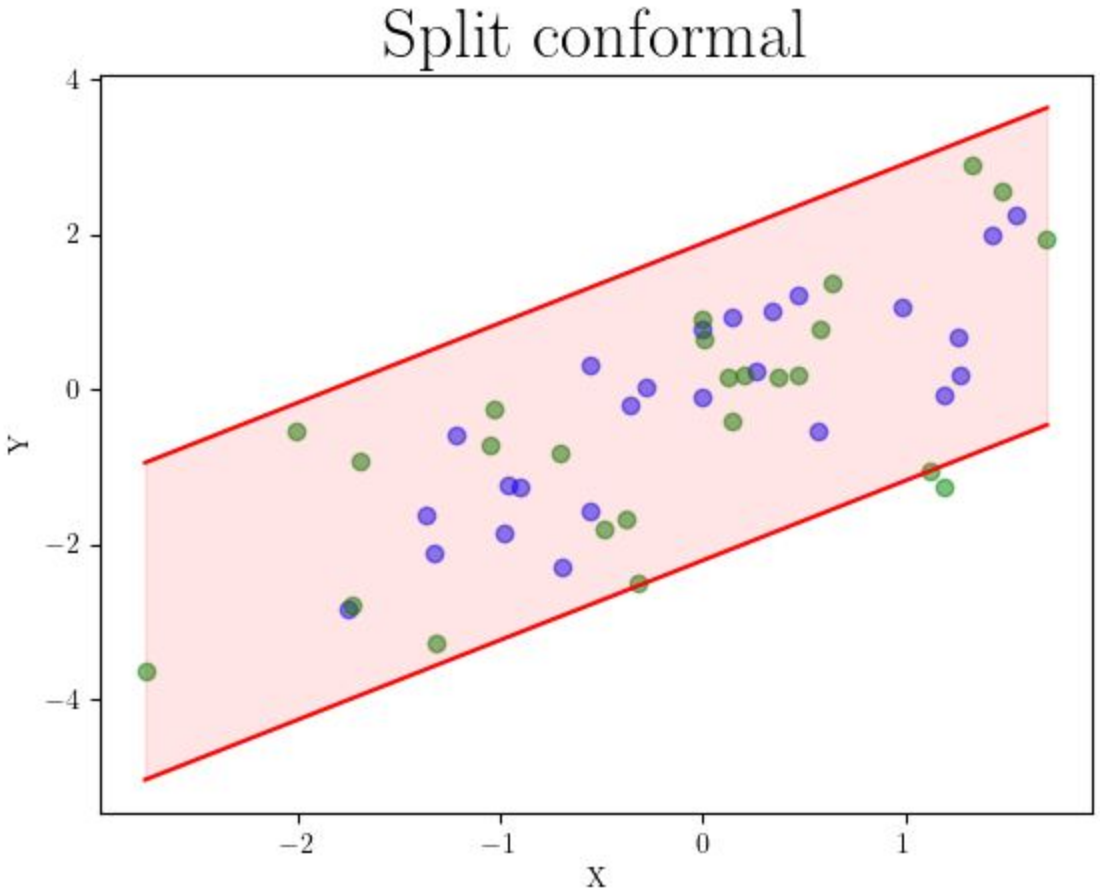
Split Conformal

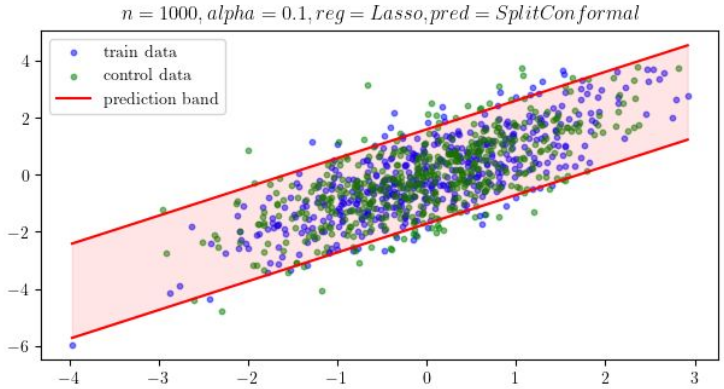


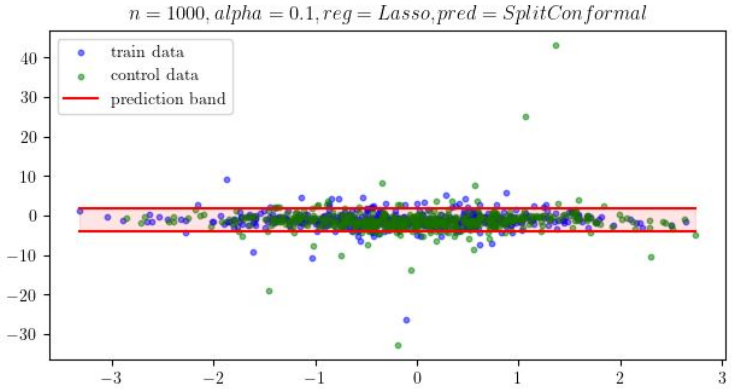
Split Conformal

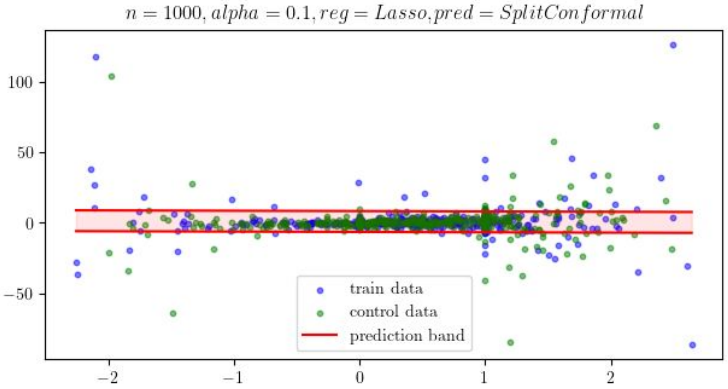


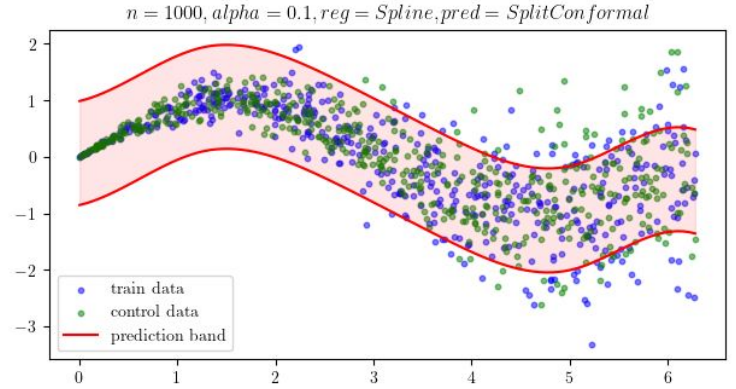



Split conformal
4
2
0
−2
−4
Y
−2
−1
0
1
X









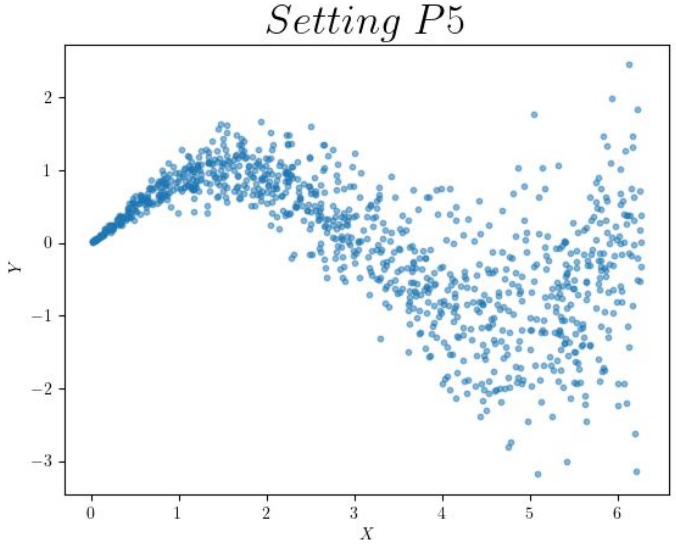
	Setting A	Setting B	Setting C	Setting Partie 5
Coverage	89.89	89.90	90.02	89.93

Définition (Hétéroscédastcité). *Soit un modèle de régression*

$$Y_i = \mu(X_i) + \varepsilon_i, \quad i = 1, \dots, n,$$

où les erreurs ε_i *vérifient* $\mathbb{E}[\varepsilon_i] = 0$. *On dit que le modèle présente de l'*hétéroscédasticité *s'il existe une fonction non constante* f *telle que*

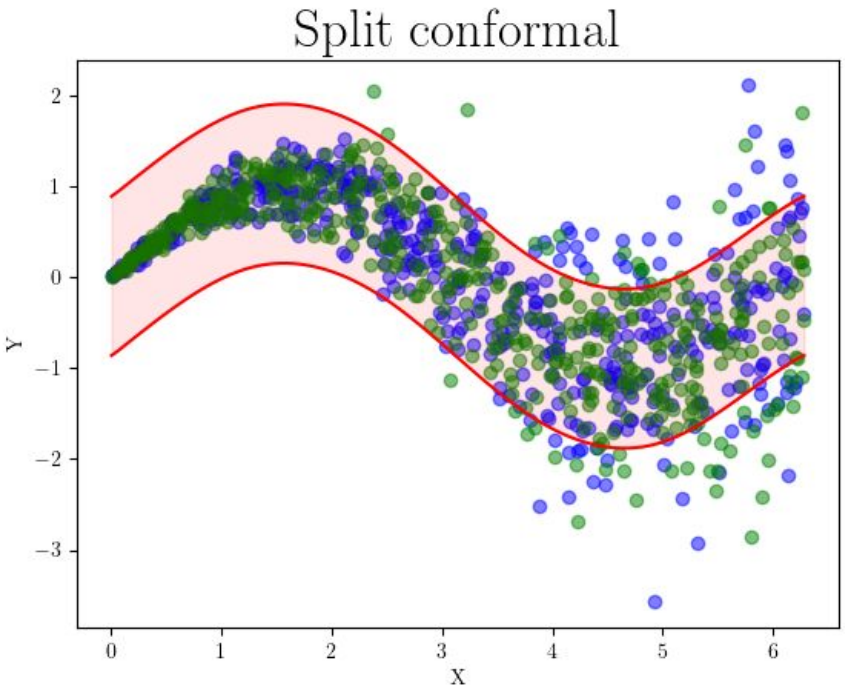
$$\mathrm{Var}(\varepsilon_i) = f(X_i) \quad \textit{pour tout } i.$$

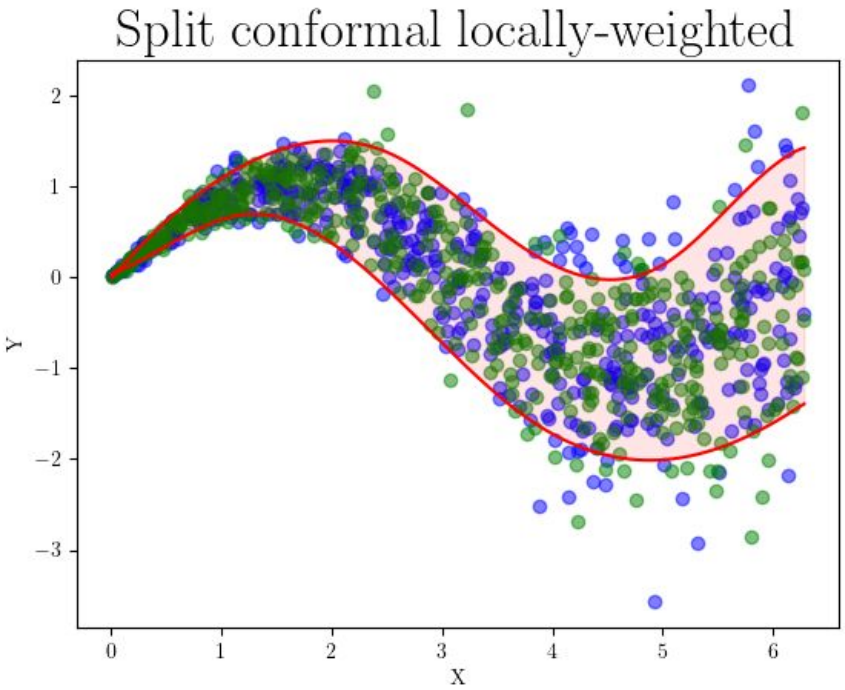


Exemple :

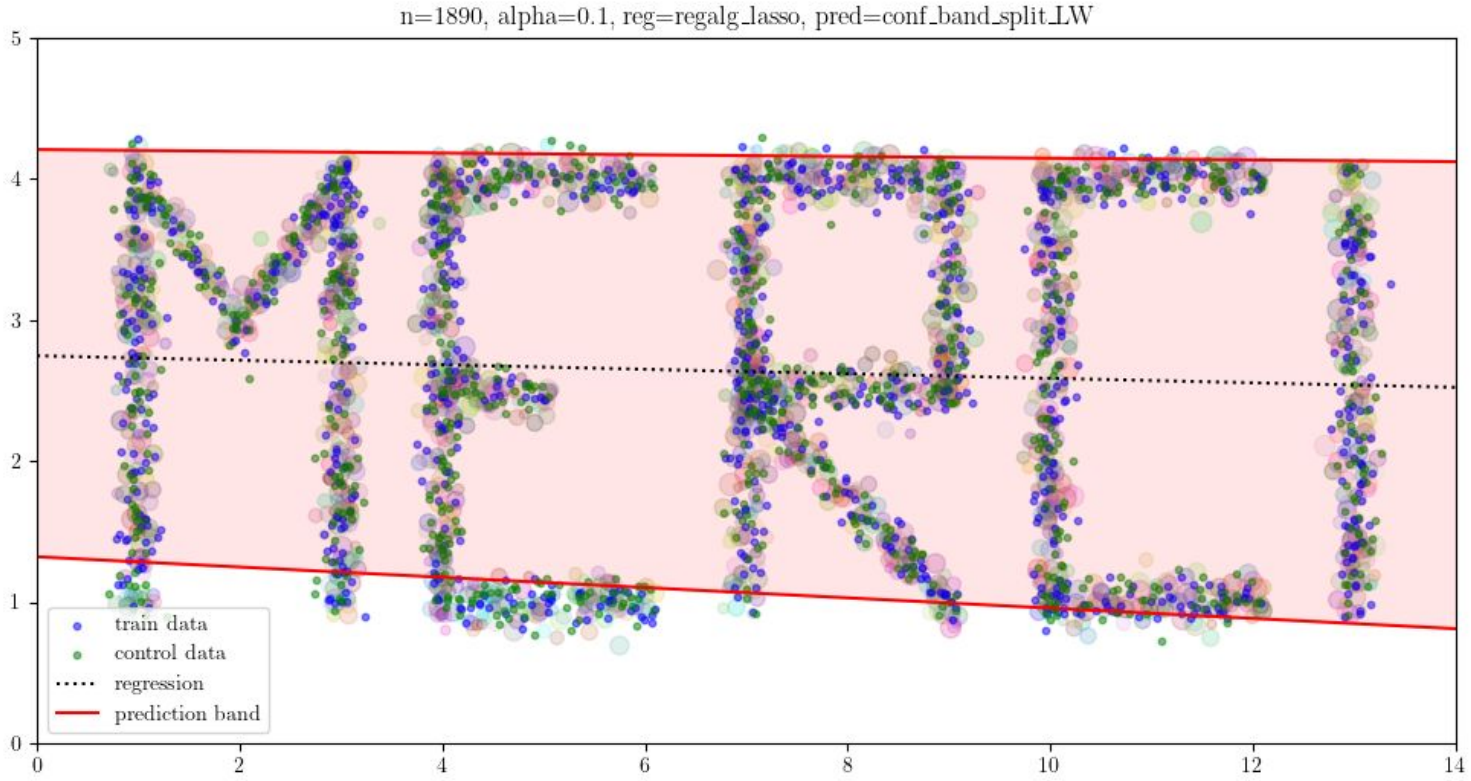
$$X_i \sim \mathcal{U}(0, 2\pi) \qquad \epsilon_i \sim \mathcal{N}(0, 1)$$

$$Y_i = \sin(X_i) + \frac{\pi |X_i|}{20} \epsilon_i$$





Hétéroscédasticité


n=1890, alpha=0.1, reg=regalg_lasso, pred=conf_band_split_LW
train data
control data
regression
prediction band