Prédiction Conforme I

Université Lyon 2

2025/2026

Rémi Vaucher

HALIAS & ERIC Lab.

Table des matières

1. Introduction: quantification de l'incertitude
   - C'est quoi la quantification d'incertitude
   - La quantification d'incertitude de prédiction en Statistiques
2. La régression quantile

Table of Contents

1. Introduction: quantification de l'incertitude
   - C'est quoi la quantification d'incertitude
   - La quantification d'incertitude de prédiction en Statistiques
2. La régression quantile

Les deux types de quantification d'incertitudes

- La quantification d'incertitude **de mesure**.
- La quantification d'incertitude **de prédiction.**

Intervalle de confiance

- **Question :** Qu'est ce qu'un intervalle de confiance?

Intervalle de confiance

- **Question :** Qu'est ce qu'un intervalle de confiance?

$\Rightarrow$ Un intervalle de confiance sert à encadrer, en probabilité, une valeur réelle, généralement la moyenne. L'objectif est donc de trouver $[a; b]$ tel que $\mathbb{P}[a \leq v \leq b] \simeq 0,95$.

Intervalle de confiance

- **Question :** Qu'est ce qu'un intervalle de confiance?

$\Rightarrow$ Un intervalle de confiance sert à encadrer, en probabilité, une valeur réelle, généralement la moyenne. L'objectif est donc de trouver $[a; b]$ tel que $\mathbb{P}[a \leq v \leq b] \simeq 0,95$.

- **Comment construit on l'intervalle de confiance pour la moyenne?**

Intervalle de confiance

- **Question :** Qu'est ce qu'un intervalle de confiance?

  $\Rightarrow$ Un intervalle de confiance sert à encadrer, en probabilité, une valeur réelle, généralement la moyenne. L'objectif est donc de trouver $[a; b]$ tel que $\mathbb{P}[a \leq v \leq b] \simeq 0,95$.

- **Comment construit on l'intervalle de confiance pour la moyenne?**

  $\Rightarrow$ Cet intervalle est construit sur la base du théorème central limite: Une somme de variable aléatoire S_n converge en loi vers une loi normale $\mathcal{N}(n\mu, \sigma\sqrt{n})$.

Intervalle de confiance

On obtient facilement (du moment que l'on connait assez bien les quantiles de la $\mathcal{N}(0,1)$) un intervalle dans lequel $\frac{S_n}{n}$ est contenu a hauteur de 95%.

Intervalle de confiance

On obtient facilement (du moment que l'on connait assez bien les quantiles de la $\mathcal{N}(0,1)$) un intervalle dans lequel $\frac{S_n}{n}$ est contenu a hauteur de 95%.

Question : Comment exploiter cette notion dans une régression linéaire?

Intervalle de confiance

On obtient facilement (du moment que l'on connait assez bien les quantiles de la $\mathcal{N}(0,1)$) un intervalle dans lequel $\frac{S_n}{n}$ est contenu a hauteur de 95%.

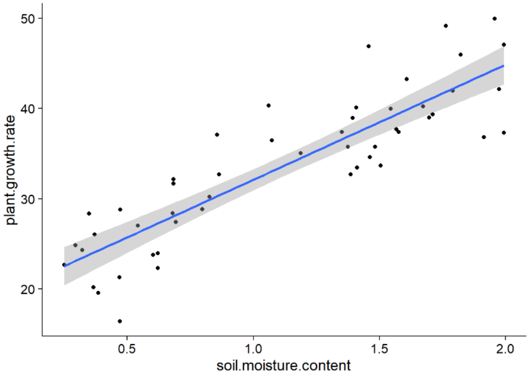
Question : Comment exploiter cette notion dans une régression linéaire?

Comme l'intervalle de confiance **ne permets que d'encadrer une "moyenne"**, nous ne pouvons qu'encadrer la moyenne des données $\overline{x}$. **Sauf que cette quantité intervient dans le calcul des coefficients β_0 et β_1 de la droite de régression D.** On obtient donc un encadrement de β_0 et β_1.

Intervalle de confiance

Finalement, on obtient un ensemble de droite de régressions $\mathcal{R}$ tel que

$$\mathbb{P}[D \in \mathcal{R}] \simeq 0,95$$



Intervalle de prédiction

⇒ L'intervalle de confiance permets (dans ce cas ci) de créer un ensemble de modèles possibles. Mais généralement, on utilise la droite de régression "*principale*" pour la suite.

Intervalle de prédiction

$\Rightarrow$ L'intervalle de confiance permets (dans ce cas ci) de créer un ensemble de modèles possibles. Mais généralement, on utilise la droite de régression "*principale*" pour la suite.

Il est évident que la réalité ne marche pas selon un modèle linéaire, nous devons donc maintenant quantifier **l'erreur de prédiction**.

Intervalle de prédiction

⇒ L'intervalle de confiance permets (dans ce cas ci) de créer un ensemble de modèles possibles. Mais généralement, on utilise la droite de régression "*principale*" pour la suite.

Il est évident que la réalité ne marche pas selon un modèle linéaire, nous devons donc maintenant quantifier **l'erreur de prédiction**.

Pour cela rappelons nous donc les conditions/hypothèses du **modèle linéaire**:

Intervalle de prédiction

Intervalle de prédiction

- Concernant les résidus :

Intervalle de prédiction

- Concernant les résidus :
  - Indépendance
  - Normalité
  - Même variance

Intervalle de prédiction

- Concernant les résidus :
  - Indépendance
  - Normalité
  - Même variance
- Il existe une relation linéaire entre les prédicteurs et la réponse moyenne.

Intervalle de prédiction

- Concernant les résidus :
  - Indépendance
  - Normalité
  - Même variance
- Il existe une relation linéaire entre les prédicteurs et la réponse moyenne.
- L'erreur de mesure des prédicteurs est négligeable.

Intervalle de prédiction

Question : En pratique, combien de ces conditions sont vraies ?

Question : En pratique, combien de ces conditions sont vraies ?

Et en admettant que ce soit vrai, comment calculer l'erreur de prédiction ?

Intervalle de prédiction

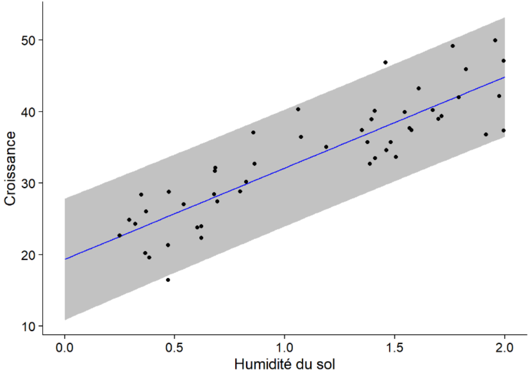
Question : En pratique, combien de ces conditions sont vraies ?

Et en admettant que ce soit vrai, comment calculer l'erreur de prédiction ?

$\Rightarrow$ On considère (en faisant une petite simplification) que $Y_{rec} = Y_{pred} + 1,96\sigma_{err}$.

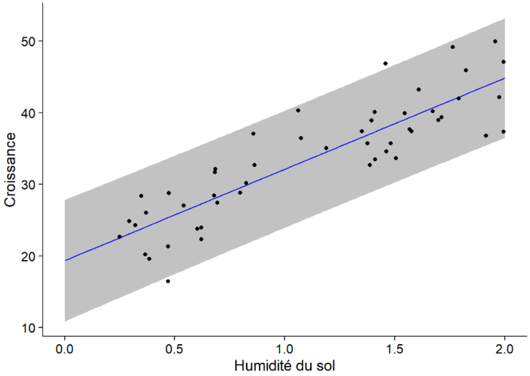
Intervalle de prédiction

Comme σ_{err} est partout le même, on obtient deux nouvelles droites encadrant les valeurs:



Slide ou je me repose.

En s'appuyant sur ce qui précède et cette illustration, lister les avantages et les inconvénients de cette méthode (selon vous).



Conclusion de l'introduction

- Problème de généralisation

Conclusion de l'introduction

- Problème de généralisation
- Des hypothèses trop théoriques.

Conclusion de l'introduction

- Problème de généralisation

- Des hypothèses trop théoriques.
- Un intervalle trop rigoureux.

Tous ces inconvénients motive l'introduction de nouvelles méthodes de quantification d'incertitude (si vous n'êtes pas convaincu, je ne peux plus rien pour vous).

Conclusion de l'introduction

- Problème de généralisation
- Des hypothèses trop théoriques.
- Un intervalle trop rigoureux.

Tous ces inconvénients motive l'introduction de nouvelles méthodes de quantification d'incertitude (si vous n'êtes pas convaincu, je ne peux plus rien pour vous).

$\Rightarrow$ C'est là qu'intervient **la régression quantile.**

Table of Contents

1. Introduction: quantification de l'incertitude
   - C'est quoi la quantification d'incertitude
   - La quantification d'incertitude de prédiction en Statistiques
2. La régression quantile

Rappel

Définition

On considère une variable aléatoire Y de fonction de répartition F_Y, et un seuil $\tau \in]0;1[$.
Le **quantile d'ordre τ pour Y** est:

Rappel

Définition

On considère une variable aléatoire Y de fonction de répartition F_Y, et un seuil $\tau \in]0;1[$.
Le **quantile d'ordre τ pour Y** est:

$$q_\tau(Y) = \inf\{y : F_Y(y) \leq \tau\} = F_Y^{-1}(\tau)$$

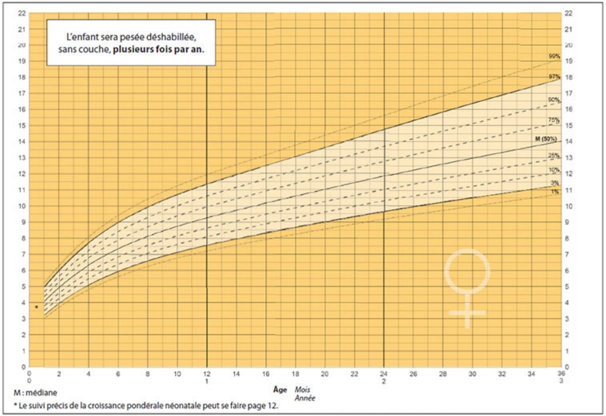
Le but d'une régression quantile

L'objectif d'une régression quantile est de déterminer comment les quantiles conditionnels

$$q_\tau(Y|X)$$

se comportent selon X. Ainsi, la régression quantile "peut" sortir une régression correspondante à chaque niveau voulu.

Exemple présent dans le carnet de santé



Courbes de croissance AFPA – CRESS/INSERM - CompuGroup Medical, 2018 [enfants nés à plus de 2 500 g et suivis par des médecins sur le territoire métropolitain].

Le contexte

Le **modèle quantile** a été créé par Roger Koenker et George Bassett en 1978

Le contexte

Le **modèle quantile** a été créé par Roger Koenker et George Bassett en 1978

L'objectif est de déterminer les quantiles d'une variable aléatoire Y conditionellement à X

La régression quantile standard

On considère ici que la fonction quantile conditionnelle est linéaire en X:

$$q_\tau(Y|X) = X^T \beta_\tau$$

La régression quantile standard

On considère ici que la fonction quantile conditionnelle est linéaire en X:

$$q_\tau(Y|X) = X^T\beta_\tau$$

Ce qui équivaut écrit autrement à

$$Y = X^T\beta_\tau + \varepsilon_\tau, \quad \text{avec} \quad q_\tau(\varepsilon_\tau|X) = 0$$

Remarques :

- Dans ce modèle, on obtient des modèles "propres" à chaque quantile.

$\Rightarrow$ Pour obtenir un intervalle, il faut donc réaliser deux régressions.

Remarques :

- Dans ce modèle, on obtient des modèles "propres" à chaque quantile.

  $\Rightarrow$ Pour obtenir un intervalle, il faut donc réaliser deux régressions.

- Forcément, selon la définition d'ε_τ le modèle se comportera différemment.

Détermination des coefficients β

Dans le modèle de régression quantile standard (donc linéaire) on démarre de l'estimateur du quantile d'ordre τ:

$$\hat{q}_\tau(Y|X) = \arg\min_b \frac{1}{N} \sum_{i=1}^{n} \rho_\tau(Y_i - b)$$

avec $\rho_\tau(u) = (\tau - \mathbb{1}_{\mathbb{R}_-}(u))u$. Ces fonctions loss (dépendantes de τ) s'appellent les fonctions **pinball**.

Détermination des coefficients β

Dans le modèle de régression quantile standard (donc linéaire) on démarre de l'estimateur du quantile d'ordre τ:

$$\hat{q}_\tau(Y|X) = \arg\min_b \frac{1}{N} \sum_{i=1}^{n} \rho_\tau(Y_i - b)$$

avec $\rho_\tau(u) = (\tau - \mathbb{1}_{\mathbb{R}_-}(u))u$. Ces fonctions loss (dépendantes de τ) s'appellent les fonctions **pinball**.

On pourra remarquer que pour $\tau = 0,5$, nous retrouvons bien l'estimateur de la médiane.

De l'estimateur à une fonction loss

En considérant que l'on dispose d'une approximation de la forme $q_\tau(Y|X)$, on obtient:

$$\beta_\tau = \arg\min \mathbb{E}[\rho_\tau(Y - X^T\beta)]$$

De l’estimateur à une fonction loss

En considérant que l’on dispose d’une approximation de la forme $q_\tau(Y|X)$, on obtient:

$$\beta_\tau = \arg\min \mathbb{E}[\rho_\tau(Y - X^T\beta)]$$

Attention tout de fois: la fonction ρ_τ n’est pas convexe, et elle n’est pas différentiable en 0!

Les fonctions python/R

- **Sur R :** On utilisera la librairie *quantreg* et la fonction *rq* associée.
- **Sur python :** On utilisera la fonction *QuantileRegressor* de *SciKitLearn*