Conformal prediction

Rémi Vaucher

Doctorant, et prof à ses heures perdues
Laboratoire ERIC UR 3083

Un certain jour de l'année

Contents

En gros: quoiqu'on attends de vous à la fin

- Savoir mettre en place (from scratch et sous python/R) une régression quantile sur un jeu de donnée au choix.
- Savoir mettre en place (from scratch) un processus de prédiction conforme sur un prédicteur de régression ET de classification sur des jeux de données (classique) au choix.
- Savoir interpréter les résultats de manière critique.

Commençons par un exemple simple pour bien comprendre.

Commençons par un exemple simple pour bien comprendre. Prenons un mur et mesurons le (vous hein, pas moi). Que peut il se passer?

Commençons par un exemple simple pour bien comprendre. Prenons un mur et mesurons le (vous hein, pas moi). Que peut il se passer?

Nous allons trouver plusieurs mesures différentes. Alors que le mur est le même. D'où peuvent venir ces différences?

Commençons par un exemple simple pour bien comprendre. Prenons un mur et mesurons le (vous hein, pas moi). Que peut il se passer?

Nous allons trouver plusieurs mesures différentes. Alors que le mur est le même. D'où peuvent venir ces différences?

Si maintenant, je demande à chacun d'entre vous de mesurer 10 murs, chacun mesurant 1m, que va-t-il se passer?

Deux types d'incertitudes

On distingue deux types d'incertitudes de mesure:

Deux types d'incertitudes

On distingue deux types d'incertitudes de mesure:

- Les erreurs systématiques:

Deux types d'incertitudes

On distingue deux types d'incertitudes de mesure:

- Les erreurs systématiques: Ce sont les erreurs entraînées par l'instrument de mesure (ou l'humain qui mesure). Cette erreur sera "minime" (de l'ordre de l'arrondi)

Deux types d'incertitudes

On distingue deux types d'incertitudes de mesure:

- Les erreurs systématiques: Ce sont les erreurs entraînées par l'instrument de mesure (ou l'humain qui mesure). Cette erreur sera "minime" (de l'ordre de l'arrondi)
- Les erreurs aléatoires:

Deux types d'incertitudes

On distingue deux types d'incertitudes de mesure:

- Les erreurs systématiques: Ce sont les erreurs entraînées par l'instrument de mesure (ou l'humain qui mesure). Cette erreur sera "minime" (de l'ordre de l'arrondi)
- Les erreurs aléatoires: Ce sont les erreurs dont les facteurs sont inconnus. Cette erreur peut être grande: on parlera alors d'anomalie.

Et nous dans tout ça?

Quantifier l'incertitude de mesure, c'est un truc de physicien (enfin de métrologue aussi \o/). Est ce vraiment ce que l'on souhaite faire ici?

Et nous dans tout ça?

Quantifier l'incertitude de mesure, c'est un truc de physicien (enfin de métrologue aussi \o/). Est ce vraiment ce que l'on souhaite faire ici?

Si l'on regarde de manière purement technique, notre modèle est l'instrument dont la mesure est le score. Quantifier l'incertitude dans ce cas là reviendrait donc à quantifier **l'erreur systématique**.

Et nous dans tout ça?

Quantifier l'incertitude de mesure, c'est un truc de physicien (enfin de métrologue aussi \o/). Est ce vraiment ce que l'on souhaite faire ici?

Si l'on regarde de manière purement technique, notre modèle est l'instrument dont la mesure est le score. Quantifier l'incertitude dans ce cas là reviendrait donc à quantifier **l'erreur systématique**.

Si l'on adopte une vision un peu plus statistique, notre objectif est de donner un intervalle/ensemble (le plus petit possible) dans lequel la valeur de la mesure véritable est contenu. Cela ressemblerait à s'y méprendre à un intervalle de confiance.

La différence entre l'incertitude de mesure et l'incertitude de prédiction

Il y a (à mon humble avis) une différence fondamentale entre l'incertitude de mesure et l'incertitude de prédiction :

La différence entre l'incertitude de mesure et l'incertitude de prédiction

Il y a (à mon humble avis) une différence fondamentale entre l'incertitude de mesure et l'incertitude de prédiction :

La calibration de l'instrument de mesure.

La différence entre l'incertitude de mesure et l'incertitude de prédiction

Il y a (à mon humble avis) une différence fondamentale entre l'incertitude de mesure et l'incertitude de prédiction :

La calibration de l'instrument de mesure. Ou plutôt, les données sur lesquelles sont calibrées les instruments de mesures.

Ne tournons pas les talons au problème

En métrologie, toutes les mesures sont basées sur un **étalon**: une valeur considérée de référence. La calibration d'un modèle est donc "facile", car le principe recherché est l'overfitting (on cherche à reproduire a tout prix cet étalon).

Ne tournons pas les talons au problème

En métrologie, toutes les mesures sont basées sur un **étalon**: une valeur considérée de référence. La calibration d'un modèle est donc "facile", car le principe recherché est l'overfitting (on cherche à reproduire a tout prix cet étalon).

Dans notre cas, nos données de calibration (donc d'entraînement) ne sont en aucun cas des valeurs références, ce sont des échantillons (et donc des réalisations de variables aléatoires).

De plus, les données sont elles mêmes des mesures (donc soumises aux mêmes incertitudes de mesure), dont l'incertitude n'a pas été quantifiées.

Mais du coup on fait quoi?

Pour commencer, nous allons rappeler les deux principaux outils de quantifications d'incertitudes en Statistiques (et pour illustrer ces deux principes, nous allons les appliquer à une régression linéaire):

Mais du coup on fait quoi?

Pour commencer, nous allons rappeler les deux principaux outils de quantifications d'incertitudes en Statistiques (et pour illustrer ces deux principes, nous allons les appliquer à une régression linéaire):

- **L'intervalle de confiance**

Mais du coup on fait quoi?

Pour commencer, nous allons rappeler les deux principaux outils de quantifications d'incertitudes en Statistiques (et pour illustrer ces deux principes, nous allons les appliquer à une régression linéaire):

- **L'intervalle de confiance**
- **L'intervalle 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?

Un intervalle de confiance sert à encadrer une valeur rélle, 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?

Un intervalle de confiance sert à encadrer une valeur rélle, 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?

Un intervalle de confiance sert à encadrer une valeur rélle, 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?

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%.

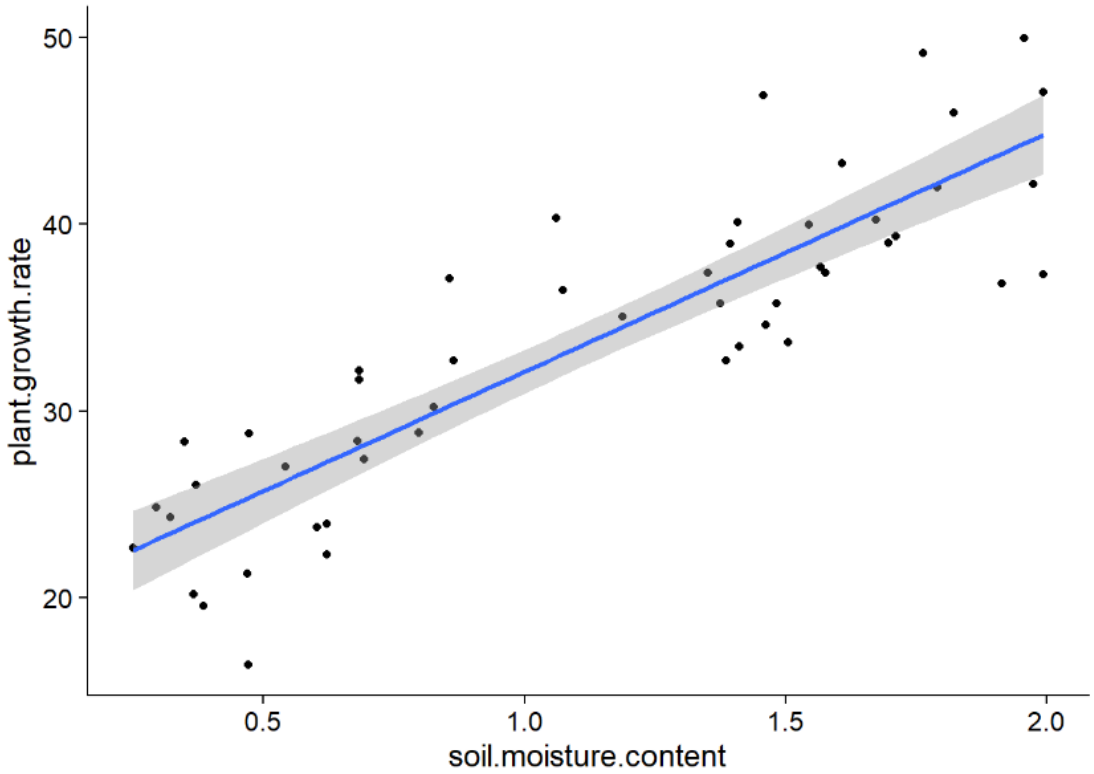
Question : Comment exploiter cette notion dans une régression linéaire?

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

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:
  - 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?

Intervalle de prédiction

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

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

Intervalle de prédiction

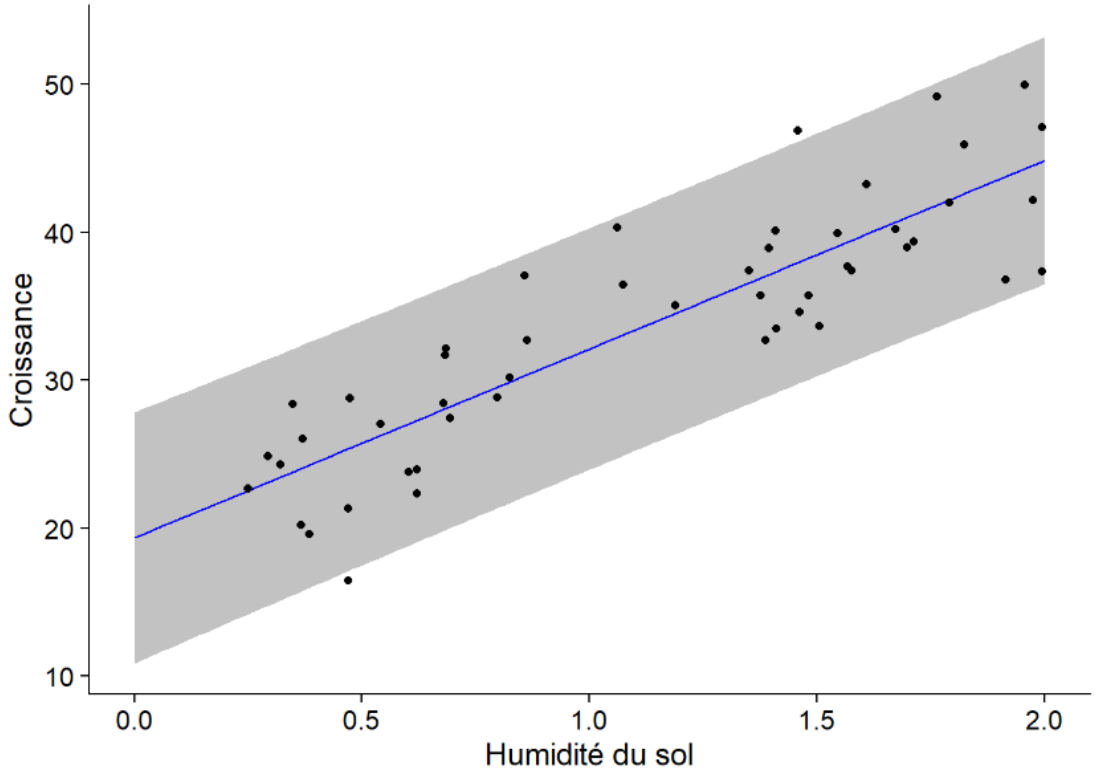
Question : En pratique, combien de ces conditions sont vraies?

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

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:



Conclusion de l'introduction

Est ce vraiment ce que l'on veut?

Conclusion de l'introduction

Est ce vraiment ce que l'on veut?

Peut on vraiment rentrer dans ces conditions si... drastiques?

Conclusion de l'introduction

Est ce vraiment ce que l'on veut?

Peut on vraiment rentrer dans ces conditions si... drastiques?

Dans les deux cas, la réponse est non (ou alors, je ne peux plus rien pour vous).

Conclusion de l'introduction

Est ce vraiment ce que l'on veut?

Peut on vraiment rentrer dans ces conditions si... drastiques?

Dans les deux cas, la réponse est non (ou alors, je ne peux plus rien pour vous).

Toutefois, il faut savoir qu'il existe des méthodes permettant de quantifier l'erreur plus précisément et de manière plus adaptable.

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:

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

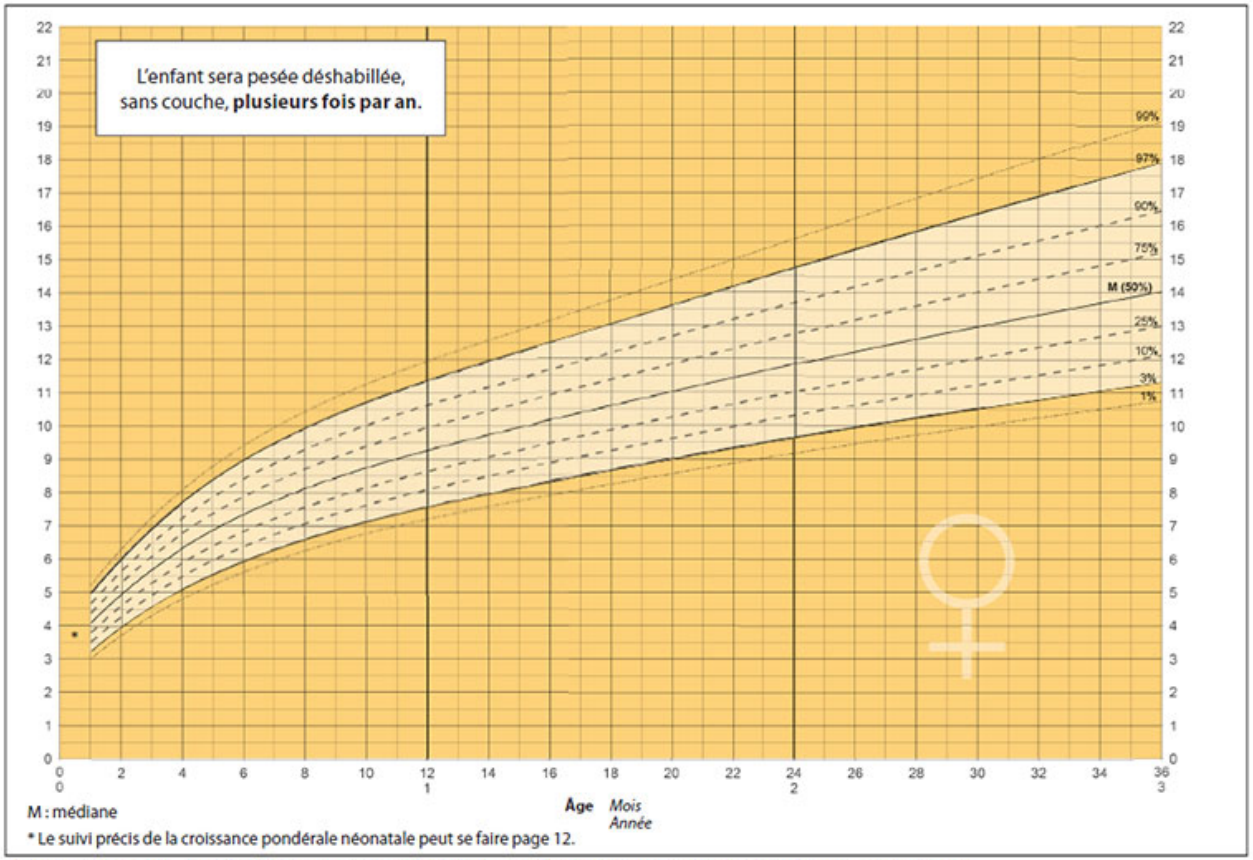
Le but d'une régression quantile

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].

Ca paraît trop beau pour être vrai

Sur cet exemple, les courbes sont plutôt belles, dans un premier temps nous n'arriverons pas à cela...

Ca paraît trop beau pour être vrai

Sur cet exemple, les courbes sont plutôt belles, dans un premier temps nous n'arriverons pas à cela...

Nous allons voir deux méthodes qui permettent de mettre en oeuvre la régression quantile.

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 :

Remarques :

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

Remarques :

- Dans ce modèle, on obtient des coefficients "propres" à chaque quantile.
- Nous pourrons voir au cours de plusieurs exercices que, 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 $\rho_\tau(u) = (\tau - \mathbb{1}_{\mathbb{R}_-}(u))u$

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 $\rho_\tau(u) = (\tau - \mathbb{1}_{\mathbb{R}_-}(u))u$
On pourra remarquer que pour $\tau = 0,5$, nous retrouvons bien l'estimateur de la médiane. (Démonstration sous réserve de demande et de motivation du prof)

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!

Exercices :

- Déterminer les fonctions ρ_τ pour $\tau \in \{0.05, 0.25, 0.5, 0.75, 0.95\}$
- Interpréter les différences de comportement pour les deux modèles de régressions quantiles suivant:
  - $Y = X^T\beta + \varepsilon$
  - $Y = X^T\beta + X^T\theta\varepsilon$

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*