Détection d'individus atypiques en régression SIR

Hadrien Lorenzo, Jérôme Saracco

Equipe ASTRAL, Inria

Equipe OptimAI, IMB

Dataquitaine 2022, Bordeaux, jeudi 10 février

Qu'est-ce qu'une observation **normale** ?

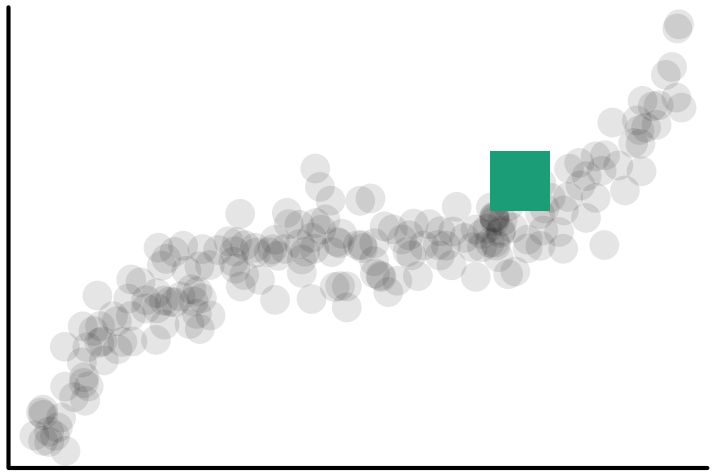
Avant toute chose, qu'est-ce qu'une observation normale ?

Une observation qui ressemble aux autres. Elles sont en majorité.

Qu'est-ce qu'une observation **normale** ?

Avant toute chose, qu'est-ce qu'une observation normale ?

Une observation qui ressemble aux autres. Elles sont en majorité.



Qu'est-ce qu'une observation **outlier** ?

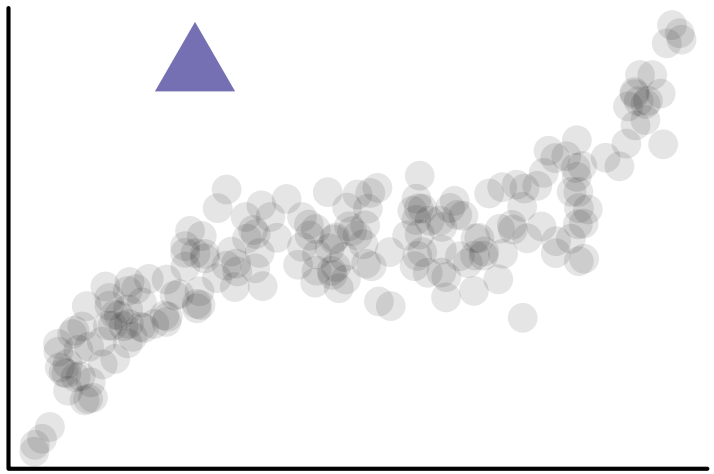
Quid d'une observation **outlier** ?

Une observation qui ne ressemble pas aux autres et qu'il serait absurde de comparer aux autres.

Qu'est-ce qu'une observation **outlier** ?

Quid d'une observation **outlier** ?

Une observation qui ne ressemble pas aux autres et qu'il serait absurde de comparer aux autres.

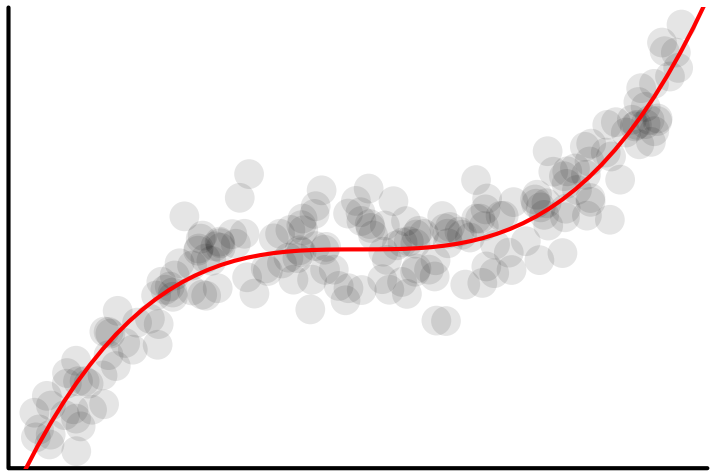


Notion de modèle

Un modèle régit la structure des données. Dans notre cas

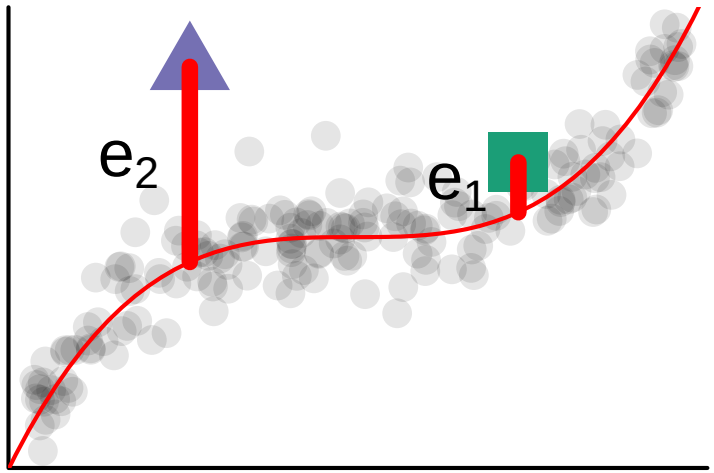
$$y = x^3 + \varepsilon/7,$$

où $x \sim \mathcal{U}_{[-1,1]}$ et $\varepsilon \sim \mathcal{N}(0,1)$.



Notion de distance au modèle

On peut mesurer à quel point une observation est loin du modèle:



Ici $e_2 \gg e_1$.

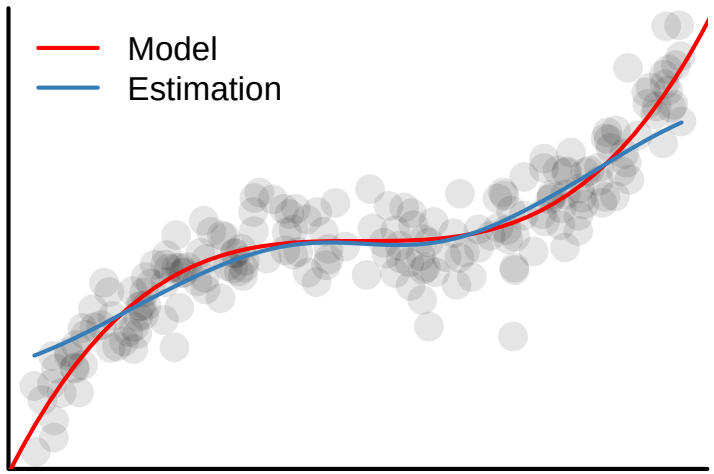
e_1 est à peu près la valeur prise par tous les points, le point est dans la norme.

Le point **violet** est certainement un **outlier**.

Problème: le modèle est inconnu

Estimation

Le modèle doit être estimé sur un jeu de données d'**entraînement**, noté $\mathcal{D}_n$, de taille finie n.
Dans notre exemple $n = 200$:



Bonne estimation au centre et mauvaise estimation sur les bords.

$\Longrightarrow$ Moins d'observations et donc moins d'information.

Implication

Les conclusions sur les bords sont à surveiller.

Qu'est-ce qu'un borderline ?

Une observation pouvant être classée **normale**/**outlier** selon l'estimation du modèle qui dépend du jeu de données disponible.

Notre travail

Certaines zones sont sensibles

Ceci à cause :
- Du modèle supposé.
- Du faible échantillonnage (n faible)

Notre travail

- Perturber l'échantillon pour enrichir le jeu de données.
- Estimer à chaque fois un modèle.
- Tester la pertinence du modèle sur certaines observations (inclues (ou pas!) dans le jeu de données d'entraînement)
- Regrouper tous les résultats.
- Conclure sur la sensibilité de certaines observations (**normale**/**borderlines**/**outlier**?)

Le **bootstrap**

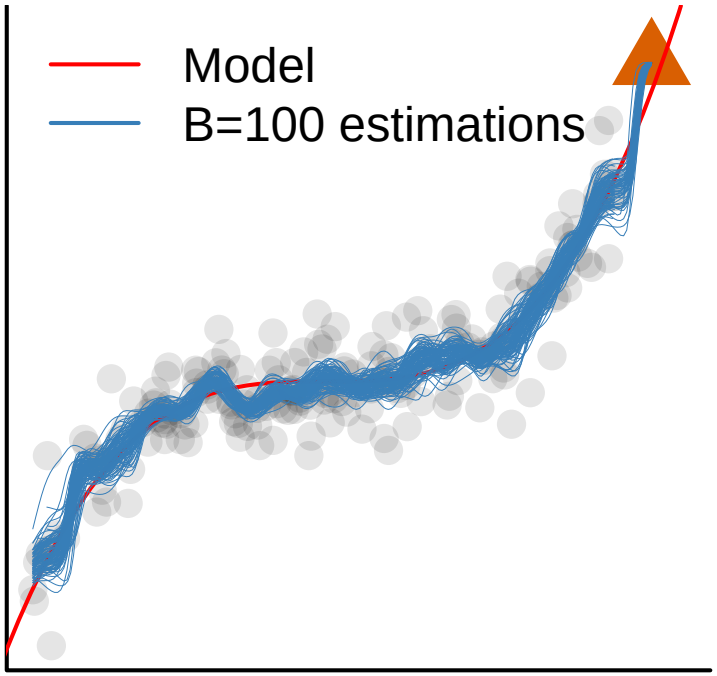
Le **bootstrap**

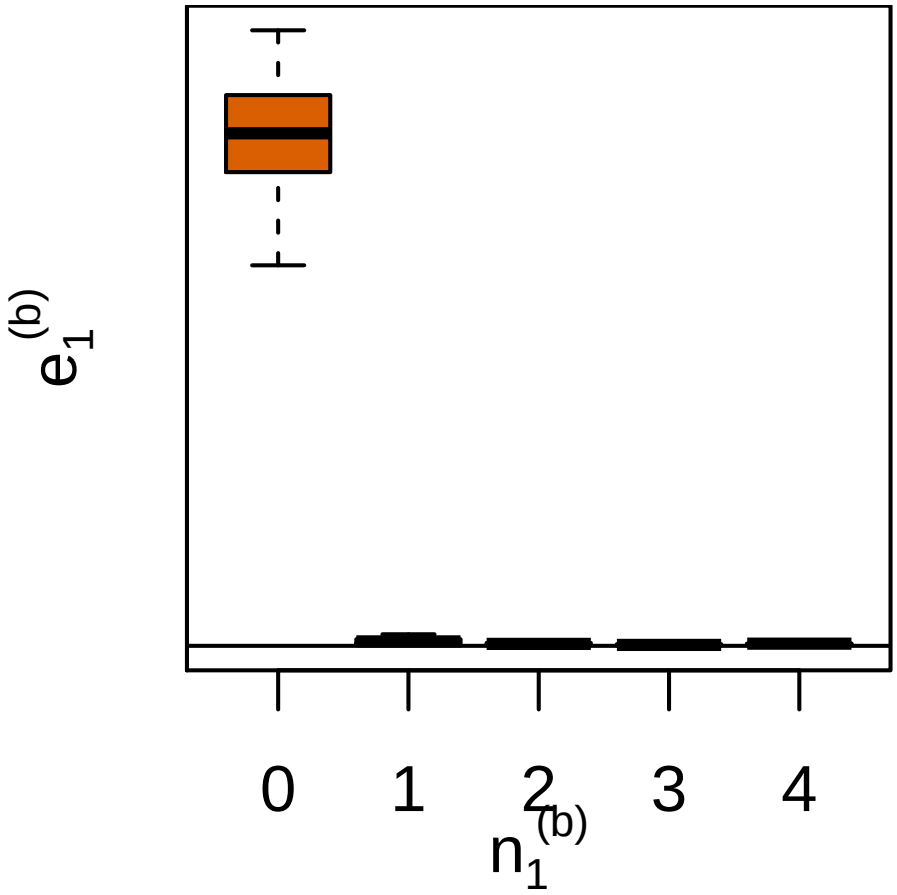
On créé B jeux de données $(\mathcal{D}_b)_{b=1\dots B}$ de taille n en piochant dans $\mathcal{D}_n$ avec remise.

$\Longrightarrow$ Une observation peut être observée plusieurs fois dans un même jeu de donnée $\mathcal{D}_b$, elle n'est donc plus dans une zone dépeuplée.

$\Longrightarrow$ Pour chaque observation i: $e_i^{(b)}$ erreur de modèle pour l'estimation obtenue sur $\mathcal{D}_b$ où l'observation i apparaît $n_i^{(b)}$ fois.

Plus concrètement





Plusieurs solutions imaginées (1-2/3)

MONO

Regarde les erreurs sur le jeu d'entraînement seulement.
Un seul jeu de données,
$\Longrightarrow$ un seul modèle,
$\Longrightarrow$ beaucoup de **Faux Positifs** (**FP**).

TTR

Divise $R \approx 2000$ fois $\mathcal{D}_n$ (de taille 90% $(train)$ et 10% $(test)$ de n).
Estimer le modèle sur le $train$ et évaluer l'erreur e_j sur le $test$.
$\Longrightarrow$ R jeu de données,
$\Longrightarrow$ R modèles de taille $0.9n$,
$\Longrightarrow$ Moins de **Faux Positifs** (**FP**).

Plusieurs solutions imaginées (3/3)

BOOT

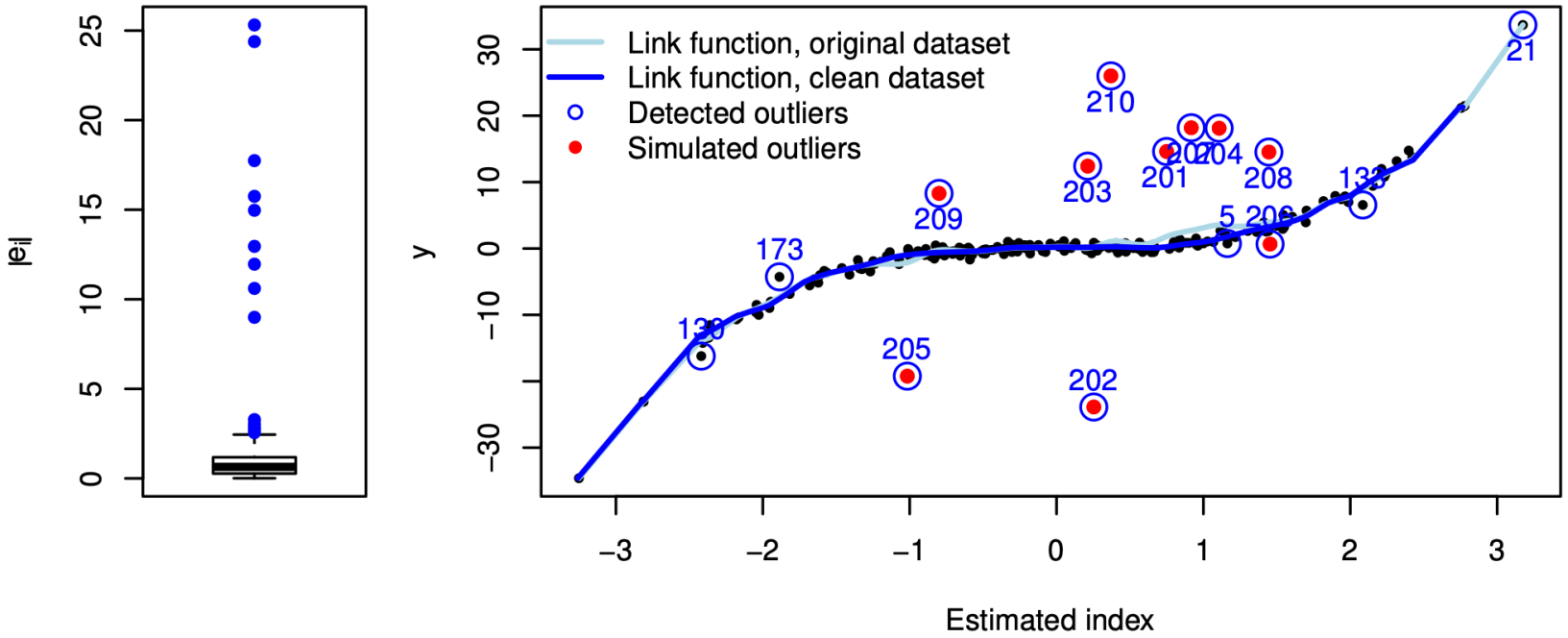
Comme **TTR** mais avec du **bootstrap**, où les individus
- tirés au sort forment les **In Bag** (**IB**),
- non tirés forment les **Out Of Bag** (**OOB**).

Comment décider du type d'observation ?

2 critères possible au regard des e_i

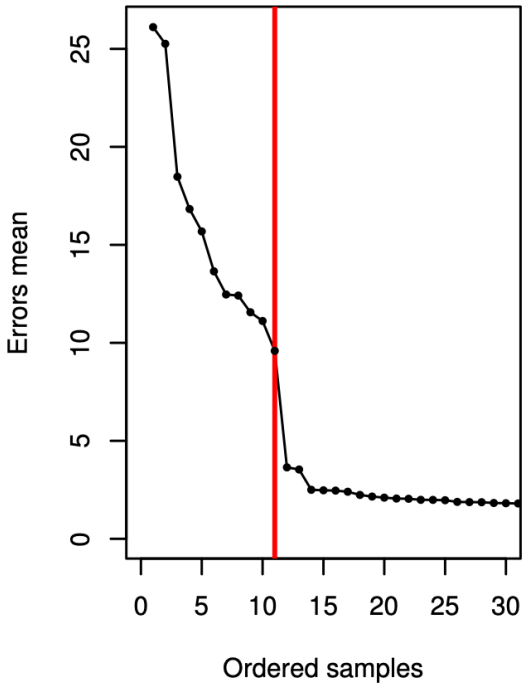
- Critère sur les quantiles, ce qui dépasse de la boîte à moustaches: **MONO** et **BOOT**.
- **TTR** : détection d'une rupture dans la série ordonnée des e_i moyennés.

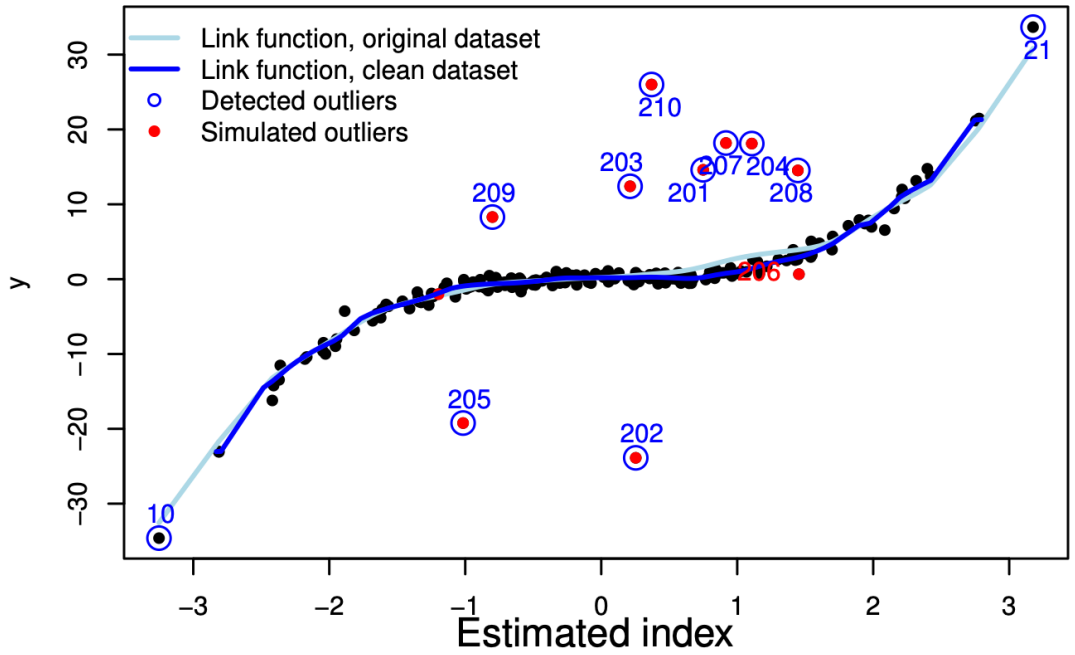
Un exemple pour **MONO**



Effectivement il y a des **FP**...

Voyons avec **TTR**

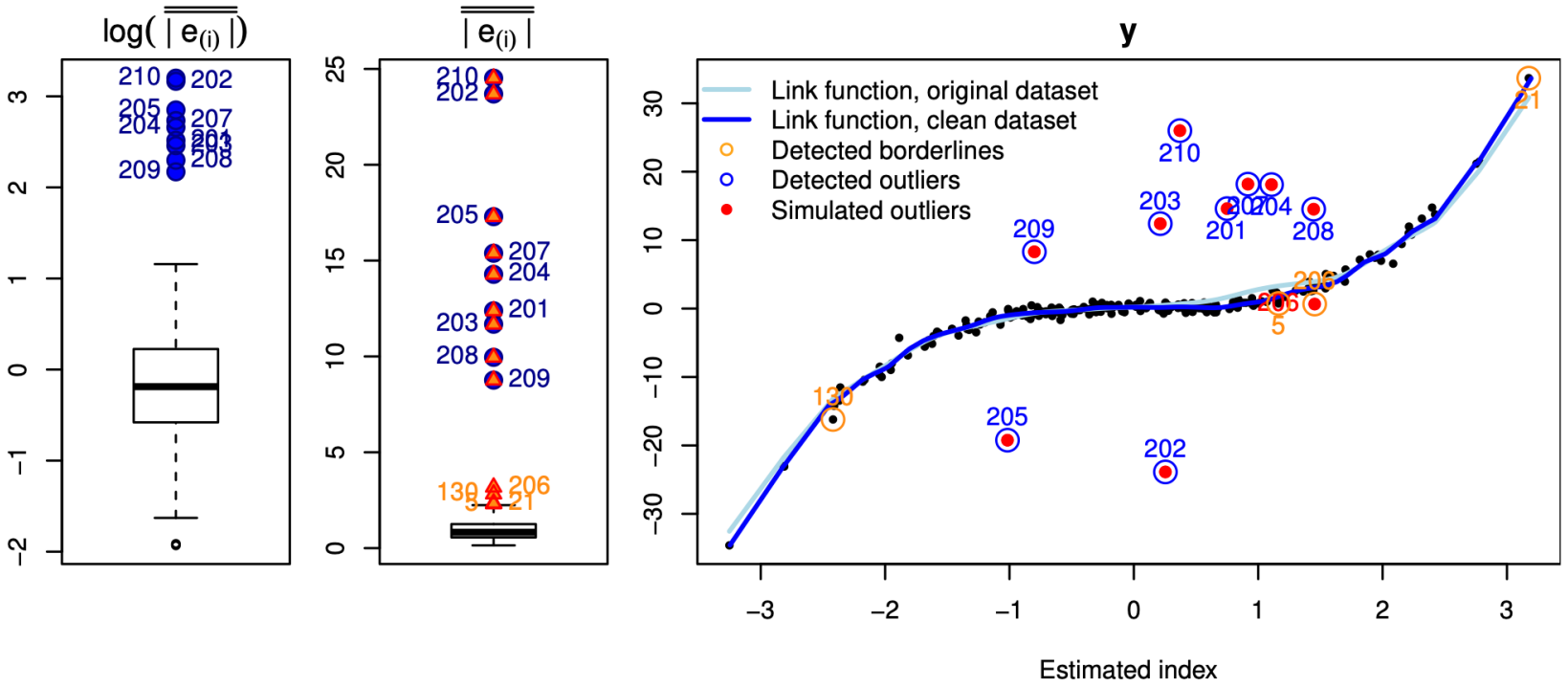




Moins de **FP**, mais des problèmes en **queues** de distribution.

$\Longrightarrow$ Zones des modèles les moins stables (cf plus haut).

Voyons avec **BOOT**



On repère d'abord les **outliers** et ensuite **borderlines**.

Moins de problèmes en queues de distribution et en bords de distribution $\approx$ existence de **borderlines**.

Un modèle de régression

Un modèle semi-paramétrique

$$y = f(\beta^t x) + \epsilon,$$

où
- x sont les variables permettant de prédire.
- y, la variable à expliquer.
- β est estimable par la méthode SIR $\implies$ paramétrique.
- f est inconnue et est estimée localement $\implies$ non-paramétrique.
- ϵ est l'erreur aléatoire du modèle.

Intêrets

L'indice $\beta^t x$ permet de réduire la dimension, ce qui est utile pour la partie non paramétrique.
Rendre compte de relations non linéaires en prédiction sans avoir à choisir de base de représentation (splines, ondelettes, sigmoïdes,...).

Un jeu de données multivarié : `ozone` Cornillon *et al.* (2012)

- $n = 112$ mesures journalières de variables météorologiques à Rennes.
- `max03` : concentration max. en ozone en gr/m^3.
- Les températures à 9h, 12h et 15h.
- Les mesures de nébulosité à 9h, 12h et 15h.
- Les vitesses du vent à 9h, 12h et 15h.
- La concentration max. en ozone mesurée la veille.

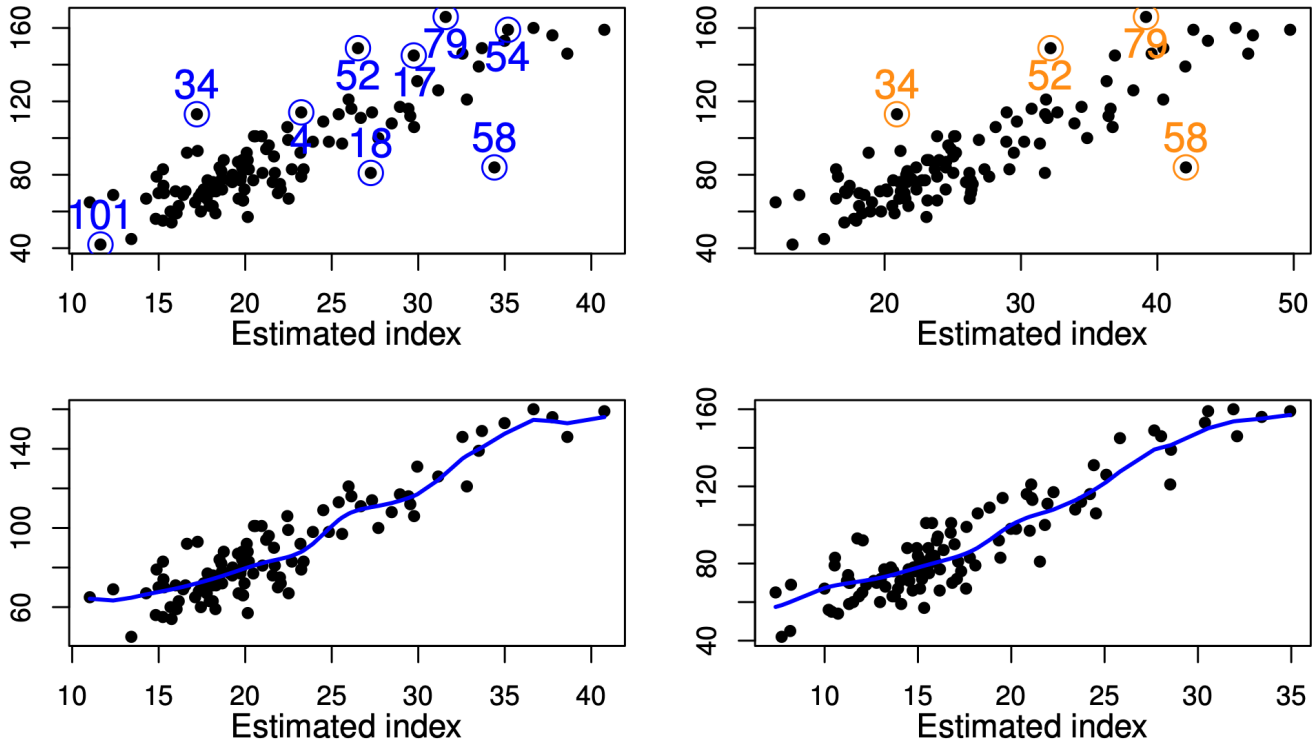
Objectif : Prédire `max03` en fonction des autres variables...

Mais pour nous! Regarder si les observations sont **normales**/**borderlines**/**outliers**

Résultats

A gauche : **TTR**. A droite : **BOOT**.

En bleu : **outliers**. En orange : **borderlines**.



Commentaires

4 **borderlines**

Connus par la littérature : Dates de chassés-croisés!
Ecarts non dues aux conditions météo, mais à un excés de circulation automobile.

$\Longrightarrow$ Le modèle supposé est limite (ne prend pas en compte la circulation automobile).

$\Longrightarrow$ Bien **borderlines** et non pas **outliers**.

Conclusion

Cette approche a pour l'instant été appliquée à la méthode SIR mais peut être élargie à tout modèle de régression.

Elle est à la limite entre détection d'outliers/évènements rares.

Intérêt pratique : Détecter des anomalies en fonctionnement dans une chaîne de production.

Intérêt académique : Chapitre d'un livre, édité chez Springer, voir Lorenzo & Saracco (2021).

References I

Cornillon, P.-A., Guyader, A., Husson, F., Jegou, N., Josse, J., Kloareg, M., Matzner-Lober, E., & Rouvière, L. (2012) *R for statistics*, ed. CRC press.

Lorenzo, H. & Saracco, J. (2021) Computational outlier detection methods in sliced inverse regression. *Advances in contemporary statistics and econometrics.*, ed. (Daouia A., R.-G.A. ed). Springer, Cham.