Détection d'individus atypiques en régression SIR

Lorenzo, Hadrien, hadrien.lorenzo@inria.fr, Inria, IMB, Bordeaux (Orateur)

Saracco, Jérôme, IMB, Inria, Bordeaux, ENSC Bordeaux INP

Thématique : *Thème 4 : Analyse des données*
Résumé : La régression inverse par tranches (*sliced inverse regression*, SIR) considère un modèle semi-paramétrique de régression entre une variable dépendante y et une variable explicative p-dimensionnelle x via un indice $\beta' x$ et une fonction de lien f. Cependant, si des observations atypiques (*outliers*) sont présentes dans les données, cette méthodologie ne va plus fonctionner convenablement. Cette communication présente trois méthodes computationnelles permettant de détecter des individus atypiques, illustrées par des simulations et un exemple sur données réelles.

Mots clés : *outlier, régression semiparamétrique, SIR, erreur out-of-bag, erreur in-bag, bootstrap*

1. Introduction

Trois méthodes sont introduites durant cette présentation afin de détecter les *outliers* dans le contexte du modèle SIR. L'estimation de la fonction de lien se fera au moyen d'un estimateur non-paramétrique à noyau. Un *outlier* correspond à une observation (x_i, y_i) qui ne proviendrait pas du modèle sous-jacent $y_i = f(\beta' x_i) + \varepsilon_i$. Ainsi, un *outlier* est ici une observation ne satisfaisant pas la relation entre x et y définie par le modèle. Un *outlier* peut alors clairement ne pas être atypique si l'on se focalise uniquement sur sa valeur x_i, ou bien uniquement sur sa valeur y_i. Ce type d'*outlier* n'est donc pas détectable en explorant uniquement la distribution des x_i ou des y_i.

En pratique, il est toujours intéressant de détecter des *outliers* (plutôt que de développer seulement des méthodes robustes), de les isoler et de comprendre pourquoi ces observations sont atypiques ou aberrantes (mauvaises valeurs numériques ? individus hors norme ? etc.)

2. Méthodologie

Trois méthodologies sont proposées, MONO, TTR et BOOT. Ces méthodes utilisent l'attrait des erreurs de prédiction IB (*in-bags*) et OOB (*out-of-bags*) et s'appuient sur des approches de type sous-échantillonnage ou ré-échantillonnage afin de discriminer les *outliers* des observation normales (i.e. qui ne sont pas hors norme). Elles ont été implémentées dans R et les codes associés sont disponibles à l'adresse suivante

https://github.com/hlorenzo/outlierSIR

Ces trois méthodes suivent le même schéma :

Étape 1 : Estimation(s) de(b, f) où b est l'estimation de la direction de β.
Étape 2 : Estimation d'erreurs de prédiction.
Étape 3 : Classification « normal » / « *outlier* » (/ « *borderline* »).

Un exemple est fourni à la figure suivante pour la détection d'individus *outliers* mais aussi *atypiques*, appelés « *borderline* ») (i.e. ni complètement *outliers*, ni totalement « normaux ») dans le cas de la méthode BOOT.

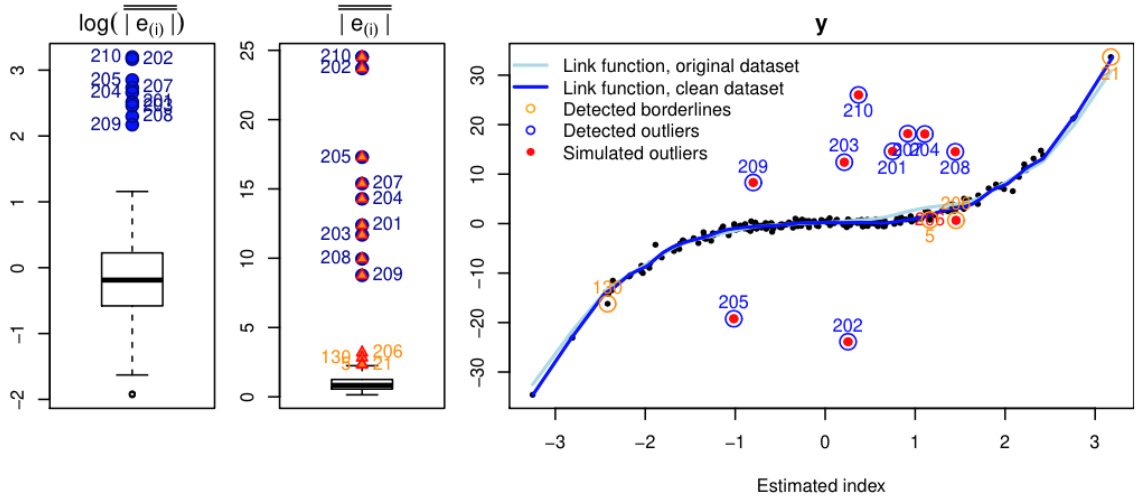
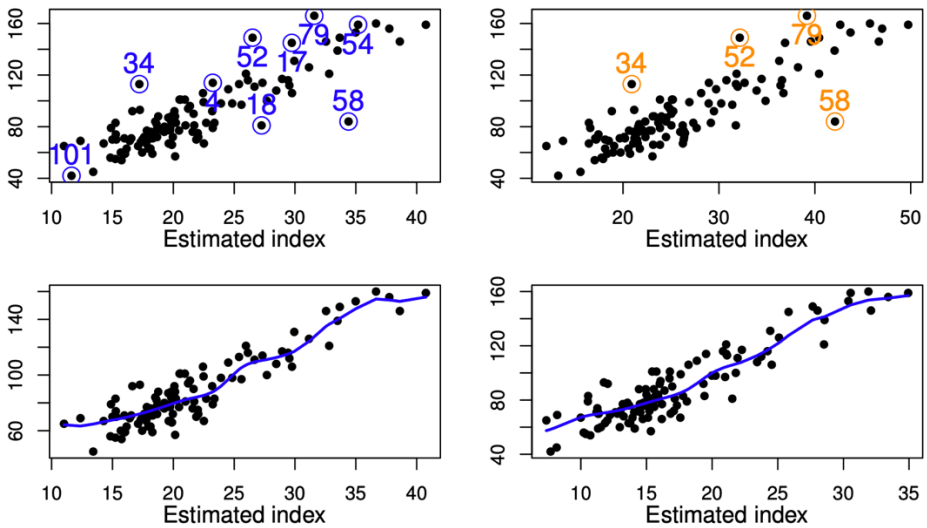


FIGURE – Exemple d'application de la méthode **BOOT**.

En pratique, cet algorithme peut être utilisé pour analyser des jeux de données complexes tels que le jeu de données *Ozone*, contenant 112 mesures journalières de variables météorologiques (vitesse du vent, température, …) enregistrées à Rennes. Parmi ces mesures, 4 sont connues pour être non pas **outliers** mais **borderlines** et correspondent à des chassés-croisé de départ en vacances. La figure qui suit montre la classification **outlier** et **borderlines** pour ce jeu de données et les méthodes TTR (colonne de gauche) et BOOT (colonne de droite).



La méthode BOOT est ici capable de reconnaître 4 mesures comme étant **borderlines**, celles-ci étant bien les mesures reconnues par différentes études postérieurs comme étant atypiques.

3. Originalité / perspective

Ces approches proposent de détecter les *outliers* avec un réel souci de le faire sur la loi jointe alors que beaucoup de méthodes n'utilisent que les distributions marginales. La véritable originalité de cette méthode est d'utiliser ce que l'on a appelé la dynamique d'apprentissage afin de catégoriser les individus. Plus un individu est présent un grand nombre de fois dans le jeu de données d'entraînement (IB), moins il sera difficile à prédire (faible erreur d'apprentissage). La dynamique observée est donc celle de l'erreur d'apprentissage en fonction du nombre d'occurrences dans le jeu de données d'entrainement.

La philosophie des méthodologies proposées peut s'adapter à tout modèle de prédiction (autre que le modèle SIR) puisqu'il utilise seulement les erreurs OOB et IB.

Références

Lorenzo H., Saracco J. (2021) Computational Outlier Detection Methods in Sliced Inverse Regression. In: Daouia A., Ruiz-Gazen A. (eds) *Advances in Contemporary Statistics and Econometrics*. Springer, Cham, 101-122. DOI : https://doi.org/10.1007/978-3-030-73249-3_6