Détection non-paramétrique robuste d'observations aberrantes et à effet de levier

Giuseppe Melfi[1] & Susana Faria[2]

[1] *Groupe de Statistique*
Université de Neuchâtel
CH-2002 Neuchâtel, Suisse

[2] *University of Minho*
DMCT
P-4800-058 Guimarães, Portugal

Résumé: La détection d'observations aberrantes et à effet de levier selon la méthode des moindre carrés est un problème qui a été largement étudié. Le diagnostic utilisant la régression LAD offre des approches alternatives dont la caractéristique principale est la robustesse. Ici une méthode non paramétrique pour détecter les observations aberrantes et les points levier est présentée et comparée avec d'autres méthodes classiques de diagnostic.

Abstract: The detection of influential observations for the standard least squares regression model is a problem which has been extensively studied. LAD regression diagnostics offers alternative approaches whose main feature is the robustness. Here a nonparametric method for detecting influencial observations is presented and compared with other classical diagnostics methods.

Mots clés: Regression LAD, Robustesse, Observations aberrantes, Points levier.

Key words: LAD Regression, Robustness, Outliers, Leverage points.

1 Introduction

La robustesse de la méthode LAD par rapport aux observations aberrantes, et sa susceptibilité aux points levier ont été largement étudiées en littérature (Dodge, 1987; Dodge, 1997).

Nous proposons une méthode non paramétrique pour détecter les observations influentes (aberrantes et à effet de levier) avec la mise au point d'une technique derivée de la régression LAD. Les points levier et les observations aberrantes sont déterminés en considérant des perturbations appropriées de l'ensemble de données de base. Ces méthodes sont ensuite comparées à d'autres méthodes connues.

Soit un ensemble fini d'observations representé par S, ensemble discret de points dans $\mathbb{R}^{p+1}$. On denote les éléments de S par $(x_{i1}, \ldots, x_{ip}, y_i)$, où la dernière variable est expliquée par les précédentes selon un modèle de régression linéaire

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i, \qquad \text{pour } i = 1, \ldots, n,$$

où p est le nombre de variables indépendantes, ε_i sont des termes d'erreur et n est le nombre d'observations.

Le modèle de régression LAD est determiné en minimisant la somme des valeurs absolues des erreurs. En termes plus précis, le vecteur $(\beta_0, \beta_1, \ldots, \beta_p) \in \mathbb{R}^{p+1}$ est le point de minimum de la fonction

$$F(\beta_0, \beta_1, \ldots, \beta_p) := \sum_{i=1}^{n} \left| y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right|.$$

Si S est donné, le modèle linéaire de régression LAD est un hyperplan qui traverse toujours au moins $p+1$ points de S (Arthanari et Dodge, 1993), bien que la solution puisse ne pas être unique. Pour nos buts, nous supposons que, pour l'ensemble de données S et chaque sous-ensemble que nous traitons, l'hyperplan correspondant au modèle linéaire de régression LAD est unique et qu'il existe une seule observation avec la déviation absolue maximale. Ces suppositions sont raisonnables pour les ensembles de données dont la taille est suffisamment grande et/ou dont les données contiennent suffisamment de chiffres significatifs. Nous supposons également que l'ensemble de données est tel que $p+2$ points ne sont jamais dans le même hyperplan. Avec ces hypothèses de travail le modèle linéaire de régression LAD est unique et traverse exactement $p+1$ points. En outre, si $n > p+1$, il y a toujours un point qui n'appartient pas à l'hyperplan de régression, ayant donc une déviation absolue positive.

Considérons les n ensembles de données composés par tous les sous-ensembles de S de taille $n-1$. Sous les hypothèses décrites ci-dessus pour chaque ensemble de donnés, nous avons une solution unique. Pour chacun de ces sous-ensembles, nous assignons un score partiel 1 aux points par lesquels son hyperplan de règression LAD passe et le score partiel 0 aux autres points. Nous définissons le score de chaque point comme étant la somme des scores partiaux sur tous les sous-ensembles de données de S de taille $n-1$. Ces scores sont produits par l'utilisation répétée des mêmes points, chaque fois pour un sous-ensemble différent de l'ensemble de données original, ainsi dans un certain sens *en bootstrapant* le modèle linéaire de régression LAD. La notation 'le point $(x_{k1}, x_{k2}, \ldots, x_{kp}, y_k)$' sera abrégé par 'le point k' et son score sera $L(k)$.

De façon analogue, nous pouvons définir une autre fonction complémentaire, dénotée $O(k)$, de la manière suivante. Considérons à nouveau les n sous-ensembles de données composés par tous les sous-ensembles de S de taille $n-1$. Pour chaque sous-ensemble,

nous considérons son hyperplan de régression LAD et nous donnons un score partiel 1 (selon les hypothèses ci-dessus) au point (unique) qui maximise la déviation absolue de l'hyperplan de régression LAD.

Nous définissons le score $O(k)$ comme étant la somme (sur de tous les sous-ensembles possibles de taille $n-1$ de S) des scores partiaux résultant des hyperplans de régression LAD.

Sous les hypothèses ci-dessus, la somme des scores L sur tous les points est $n(p+1)$, et la somme des scores O est n. Donc, si on voit L et O comme des variables aléatoires, $E(L(k)) = p+1$ et $E(O(k)) = 1$. Supposons maintenant que nous avons un ensemble de données, toutes concentrées dans une région et une observation isolée et horizontalement très loin des autres mais telle que l'hyperplan du modèle de régression LAD la traverse (typiquement un point levier). Il est probable que le score L de cette observation soit élevé. D'autre part, supposons que nous avons un ensemble de données dans lequel tous les points sont grosso modo dans un hyperplan, et un point au dessus lointain des autres (une observation aberrante). Le modèle de régression LAD sera très proche de l'hyperplan et le score L de l'observation aberrante sera probablement zéro. Au même temps le score O auquel il faut s'attendre est $n-1$.

Ces arguments justifient des algorithmes pour la détection des observations aberrantes et des points levier présentés dans la Section 2. Dans la Section 3, nous discutons quelques exemples, et comparons les résultats à ceux obtenus en utilisant d'autres méthodes classiques.

2 Les algorithmes

Dans cette section nous proposons deux algorithmes basés sur les arguments de la section précédente. Le but de l'Algorithme 1 et de l'Algorithme 2 est de détecter les observations aberrantes et les points levier respectivement. Il est important de noter que dans les algorithmes proposés, les ensembles S, A, B, C, et D changent de composition et de taille au cours de l'exécution de l'algorithme. Toutefois, la taille originale de S, n, est fixée et utilisée pour décider quand l'algorithme doit s'arrêter.

Algorithme 1 (Détection de points levier)
1. On considère un ensemble de données S de taille n. Soient A et B des ensembles vides.
2. Soit m la taille de S.
3. On considère tous les m sous-ensembles de S de taille $m-1$ et les hyperplans de régression LAD correspondants.
4. On calcule $L(k)$ pour chaque point de S et on sélectionne le point $k_1 \in S$ qui maximise $L(k)$.
5. Si $L(k_1) \geq \frac{8}{9}(m-1)$ et $L(k_1) \geq \frac{3}{4}(n-1)$ alors on déplace k_1 dans B et on déplace dans S les points se trouvant éventuellement dans A; sinon on déplace k_1 dans A.

6. Si la taille de S ne depasse pas $\frac{9}{10}n$ alors le procèssus s'arrête et B réprésente les points levier; sinon on retourne au pas 2.

Dans cet algorithme, les éléments de S sont transférés dans un ensemble B de points levier, ou dans un ensemble temporaire A où les points qui n'ont pas atteint un score suffisant sont classifiés comme des points qui pourront être reconsidérés après qu'un autre point a été détecté. Cette procédure évite le *masking effect*.

Les valeurs discriminatoires $\frac{8}{9}(m-1)$, $\frac{3}{4}(n-1)$ et $\frac{9}{10}n$ pour la fonction de score L, ont été déterminés à l'aide de simulations, où ont été considérés des ensembles de données type et des variations par rapport au nombre d'éléments et à d'autres propriétés.

Ces valeurs ont, cependant, une interprétation naturelle. Quand il y a un point levier unique, presque tous les m modèles de régression le détectent, donc son score L est près du maximum. Quand il y a plus de points levier, les scores peuvent être très différents, et le *masking effect* peut produire de relativement petits scores. Finalement, nous tenons compte de la taille de S, pour déterminer combien de points levier un ensemble de données peut avoir. Le processus s'arrête quand la taille de l'ensemble des candidats est inférieure ou égale à $\frac{9}{10}n$, donc avec cette méthode nous ne pouvons pas avoir des ensembles de données ayant plus que $\frac{1}{10}n$ de points levier.

Algorithme 2 (Détection d'observations aberrantes)
1. On considère un ensemble de données S de taille n. Soient C et D des ensembles vides.
2. On initialise la variable 'last maximum score' (LMS) à 0.
3. Soit m la taille de S.
4. On considère tous les m sous-ensembles de S de taille $m-1$ et les hyperplans de régression LAD correspondants.
5. On calcule $O(k)$ pour chaque point de S et on sélectionne le point $k_1 \in S$ qui maximise $O(k)$.
6. Si $O(k_1) = m-1$
a) si $O(k_1) = LMS - 1$ ou $LMS = 0$, alors on déplace k_1 dans D, on pose $LMS = O(k_1)$ et on déplace dans S les points se trouvant éventuellement dans C; sinon le procèssus s'arrête et D represente les observations aberrantes.
b) sinon on déplace k_1 dans C.
7. Si la taille de S est inférieur ou égale à $\frac{4}{5}n$ le procédure s'arrête et D représente les observations aberrantes; sinon on retourne au pas 3.

Dans cet algorithme, l'ensemble D contient les points classifiés comme observations aberrantes et l'ensemble C contient les points qui n'ont pas atteint le score pour être classifiés comme observations aberrantes, mais qui peuvent être reconsidérés jusqu'à l'arrêt de l'algorithme.

Les deux algorithmes proposés ont une structure similaire. Cependant, la différence principale est une caractéristique de l'Algorithme 2 : le observations aberrantes ont des

scores de la forme O_1, $O_1 - 1$, $O_1 - 2$, et ainsi de suite. L'algorithme s'arrête quand cette séquence ne peut plus être continuée.

Le processus s'arrête aussi quand la taille de ensemble S est inférieure ou égale à $\frac{4}{5}$ de sa taille originale, donc ici un ensemble de données ne peut pas avoir plus que $\frac{1}{5}n$ d'observations aberrantes.

Les valeurs discriminatoires pour la fonction de score O, ont été aussi empiriquement déterminées comme pour l'Algorithme 1.

3 Quelques exemples

Dans cette section nous illustrons les algorithmes proposés et les comparons avec deux autres méthodes en utilisant plusieurs ensembles de données réels et simulées.

L'une des méthodes est le P-R plot proposé par Hadi (1992) pour la classification des observations comme les points levier, observations aberrantes ou une combinaison des deux. Il est d'usage de dire que les points avec $h_{ii} > \frac{2(p+1)}{n}$, où h_{ii} est le i-ème élément de la matrice diagonale H, peuvent être classifiés comme des points levier et les points avec $\frac{r_i}{\hat{\sigma}\sqrt{1-h_{ii}}} > 2$, où r_i est le résidu de la i-ème observation et $\hat{\sigma}$ est l'estimateur de l'écart-type des erreurs, peuvent être classifiés comme observations aberrantes. Dans ce qui suit, ces critères de classification des observations seront denotés comme *méthodes classiques*.

Le premier ensemble de données 'Telephone' concerne le nombre d'appels téléphoniques internationaux par an de la Belgique (en dizaines de millions de minutes) relatif à une période de 24 ans et peut être trouvé dans (Rousseeuw and Leroy, 1987). Les cas 15 à 20 sont anormalement élevés et sont des observations aberrantes. Le deuxième ensemble de données 'Hawkins' consiste en 75 observations en quatre dimensions, une variable de réponse et trois variables indépendantes, et peut être trouvé dans (Hawkins *et. al.*, 1984) pour l'étude de phénomènes pathologiques spéciaux dans la détection de points levier et observations aberrantes. Les cas 1 à 10 sont des points levier et des observations aberrantes. Les données 'Scottish' décrivent comment les temps record (en secondes) de 35 courses aux Collines Écossaises dépendent de deux variables indépendantes : la longueur du parcours de la course (en miles) et la hauteur (en pieds), et peut être trouvé dans (Hadi, 1992). Les données contiennent deux observations aberrantes claires (les observations 7 et 18). Les deux derniers ensembles de données ont été créé créées par les auteurs. Les données 'twovariables' consistent en 56 observations sur une variables indépendante et une variable de réponse. La variable indépendante a été créé uniforme $(0, 10)$ et la variable de réponse conforme au modèle $Y = X_1 + 4 + \varepsilon$ avec $\varepsilon \sim N(0, 1)$. Trois observations (51 à 53) ont été conçues comme points leviers et trois autres (54 à 56) comme observations aberrantes. Les données 'threevariables' sont l'équivalent en trois variables de l'ensemble précédent.

Les calculs sur ordinateur ont été faits à l'aide d'un script Splus, et les résultats sont resumés dans la Table 1.

Données	Méthode	Points levier	Obs. aberrantes
Telephone	Méthodes Classiques	-	20
	Méthode de Hadi	-	19, 20
	Nos Résultats	-	17 à 20
Hawkins	Méthodes Classiques	12 à 14	7, 11 à 14
	Méthode de Hadi	14	7, 11 à 14
	Nos Résultats	3 à 6, 9, 10, 13	11 à 14
Scottish	Méthodes Classiques	7, 11, 33, 35	7, 18
	Méthode de Hadi	7, 11	7, 18
	Nos Résultats	11, 17, 35	7, 18, 33
Twovariables	Méthodes Classiques	51 à 53	54 à 56
	Méthode de Hadi	51 à 53	54 à 56
	Nos Résultats	52	54 à 56
Threevariables	Méthodes Classiques	18, 51 à 53	54 à 56
	Méthode de Hadi	51 à 53	54 à 56
	Nos Résultats	51 à 53	9, 37, 54 à 56

Table 1: Détection d'observations aberrantes et de points levier selon les différentes méthodes.

4 Conclusion

Cette méthode marche particulièrement bien dans la détection de toutes les observations aberrantes pour l'ensemble de données 'Telephone'. Les autres méthodes échouent parce que les observations 19 et 20 masquent toutes les autres. Concernant l'ensemble de données 'Scottish', la Table 1 montre que les trois méthodes ont identifié correctement les observations aberrantes 7 et 18. Ces observations masquent l'observation 33 détecté par notre méthode.

Le calcul des scores exige la détermination d'un certain nombre de modèles de régression LAD et ceci est computationnellement long. Cependant il est important de noter que le principe est très simple et, de nos jours, l'exécutions ne demande dans la pluspart des cas que quelques secondes.

Bibliographie

[1] Arthanari, T.S. et Dodge, Y., 1993. *Mathematical Programming in Statistics*, Classic edition, John Wiley and Sons, New York.

[2] Dodge, Y. (Ed.) 1987, *Statistical data analysis based on the L_1-norm and related methods*, Elsevier/North-Holland, New York; Amsterdam.

[3] Dodge, Y. (Ed.) 1997, *L_1-statistical procedures and related topics*, Institute of Mathematical Statistics, Hayward.

[4] Hadi, A., 1992. A new measure of overall potential influence in linear regression. Comp. Stat. and Data Analysis, 14, 1–27.

[5] Hawkins, D.M., Bradu, D. et Kass, G.V., 1984. Location of several outliers in multiple regression data using elemental sets. Technometrics, 26, 197–208.

[6] Rousseeuw, P.J. et Leroy, A.M., 1987. *Robust Regression and Outlier Detection*, John Wiley and sons, New York.