Méthodes d'Imputation de Valeurs Aberrantes pour des Données d'Enquête

R. Ren

Laboratoire de Statistique d'Enquêtes
CREST-ENSAI, Campus de Ker Lann
Rue Blaise Pascal. 35170, Bruz.

1. Introduction

- Valeur aberrante représentative : observation ayant grand écart par rapport à sa valeur prévue, existant dans l'échantillon et dans la population non-observée :

$$\boldsymbol{U} = \boldsymbol{U}_0 + \boldsymbol{U}_a$$

- Soit un échantillon $\boldsymbol{s}$ contenant des valeurs aberrantes :

$$\boldsymbol{s} = \boldsymbol{s}_0 + \boldsymbol{s}_a$$

- L'estimateur de Horvitz-Thompson non-résistant :

$$\hat{\boldsymbol{t}}_\pi = \sum_{\boldsymbol{k}\in\boldsymbol{s}} \boldsymbol{d}_k \boldsymbol{Y}_k = \sum_{\boldsymbol{k}\in\boldsymbol{s}_0} \boldsymbol{d}_k \boldsymbol{Y}_k + \sum_{\boldsymbol{k}\in\boldsymbol{s}_a} \boldsymbol{d}_k \boldsymbol{Y}_k$$

- Pour produire un estimateur résistant aux valeurs aberrantes :
  - Estimateur résistant par la modification de poids :

$$\hat{\boldsymbol{t}}_{MP} = \sum_{\boldsymbol{k}\in\boldsymbol{s}} \boldsymbol{d}_k^* \boldsymbol{Y}_k$$

où $\left\{\boldsymbol{d}_k^*, \boldsymbol{k} \in \boldsymbol{s}\right\}$ est un ensemble de poids associé à la variable $\boldsymbol{Y}$.

- Estimateur résistant par la modification de valeurs :

$$\hat{t}_{MV} = \sum_{k \in s} d_k Y_k^*$$

où $\left\{ Y_k^*, k \in s \right\}$ est un ensemble des observations modifiées.

- Pour délivrer un fichier nettoyé et prêt à l'emploi par le grand public, les valeurs aberrantes doivent être modifiées pour que les méthodes classiques puissent être appliquées et produire des résultats raisonnables.
- Les méthodes d'imputation classiques pour les valeurs manquantes s'appliquent à ce cas, mail il y a une différence : ici on a une information complète sur une valeur aberrante. Les méthodes classiques doivent être adaptées au cas actuel.
- L'idée est de modifier les valeurs aberrantes le moins que possible pour :
  - garder au maximum la répartition de données
  - retrouver l'estimation résistante du total par une méthode classique

2. Imputation par régression

- Supposons existence d'une variable auxiliaire $\boldsymbol{X}$ liant la variable d'enquête $\boldsymbol{Y}$ par un modèle linéaire :

$$\boldsymbol{Y_k} = \beta \boldsymbol{X_k} + \varepsilon_{\boldsymbol{k}}$$

où $\{\varepsilon_{\boldsymbol{k}}\}$ sont les résidus, $\boldsymbol{E}(\varepsilon_{\boldsymbol{k}} | \boldsymbol{X_k}) = 0$, $\boldsymbol{Var}(\varepsilon_{\boldsymbol{k}} | \boldsymbol{X_k}) = \sigma^2 [\boldsymbol{v}(\boldsymbol{X_k})]^2$.

- Une imputation classique par la régression :

$$\boldsymbol{Y_k^*} = \hat{\beta}_{\boldsymbol{s_0}} \boldsymbol{X_k}, \quad \boldsymbol{k} \in \boldsymbol{s_a}$$

- Elle n'est plus d'une imputation optimale parce que les valeurs aberrantes n'ont pas de même modèle que les valeurs non-aberrantes. Une adaptation est nécessaire.

- La méthode la plus simple est de rajouter un terme de correction :

$$\boldsymbol{Y_k^*} = \hat{\beta}_{\boldsymbol{s_0}} \boldsymbol{X_k} + \delta_{\boldsymbol{k}}, \quad \boldsymbol{k} \in \boldsymbol{s_a}$$

- Une correction déterministe :

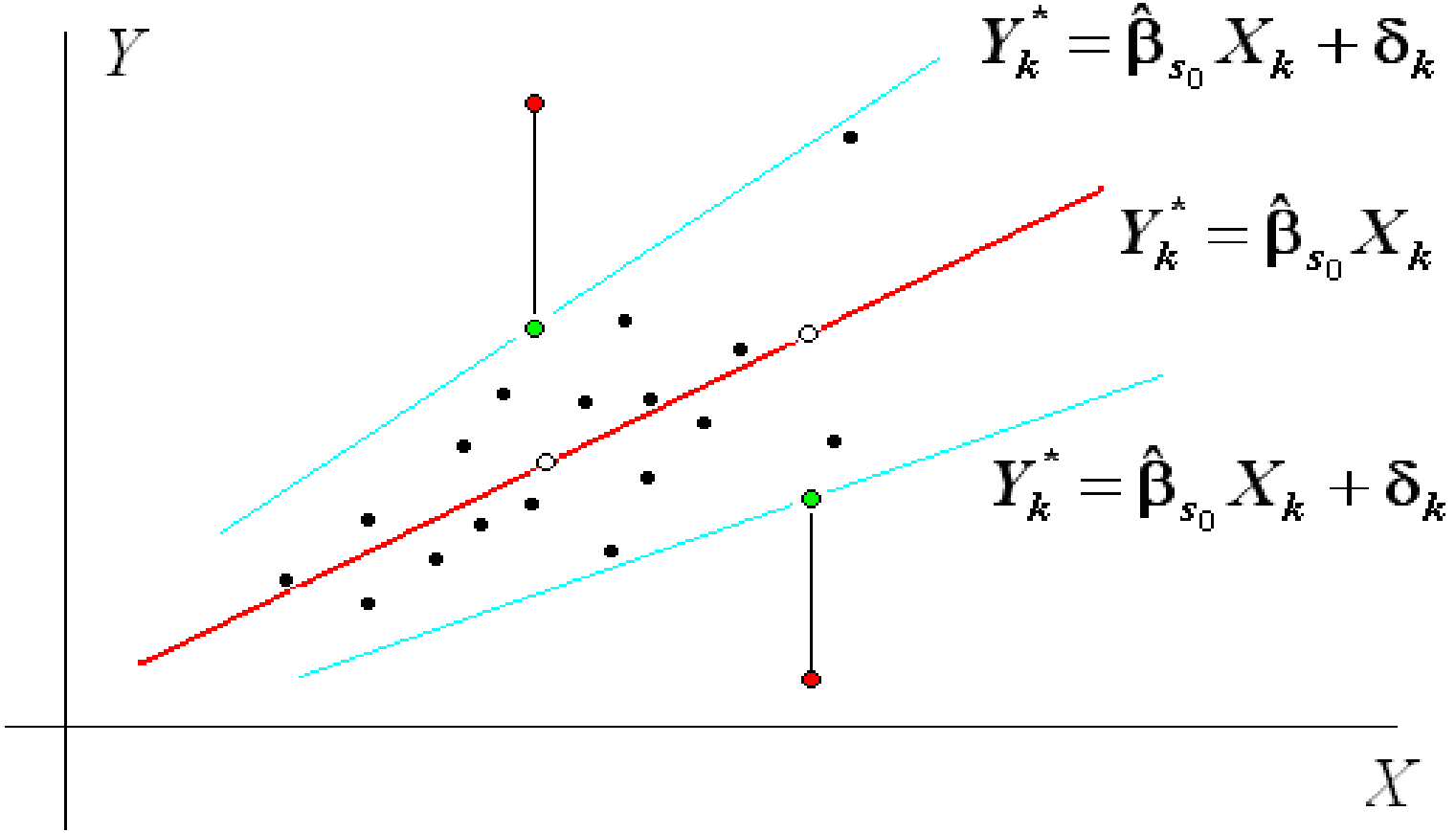
$$\delta_k = \boldsymbol{Sign}\left(Y_k - \hat{\beta}_{s_0} X_k\right) z_{1-\alpha/2} \hat{\sigma}_{s_0} v\left(X_k\right), \quad k \in s_a$$

où $z_{1-\alpha/2}$ est la valeur critique d'une variable $N(0,1)$, $\hat{\sigma}_{s_0}$ est l'écart type du résidu estimé sur échantillon non-aberrante.

- Une correction aléatoire :

$$\delta_k = \boldsymbol{Sign}\left(Y_k - \hat{\beta}_{s_0} X_k\right) \left|\hat{\varepsilon}_k\right| \hat{\sigma}_{s_0} v\left(X_k\right), \quad k \in s_a$$

où $\left\{\hat{\varepsilon}_k, k \in s_a\right\}$ est un échantillon **iid** tiré dans une loi $N(0,1)$.



Légendes : points rouges - valeurs aberrantes ; points blancs - valeurs imputées par régression ; points verres - valeurs imputées par régression modifiée

3. Imputation par le plus proche voisin

- Pour une valeur aberrante donnée $\boldsymbol{Y}_{\boldsymbol{k}_0}$, supposons connue une variable auxiliaire $\boldsymbol{x}$, le plus proche voisin de l'unité $\boldsymbol{k}_0$ au sens classique, noté $\boldsymbol{k}'$, est celui qui minimise une certaine distance parmi les non-aberrants :

$$\boldsymbol{k}' = \mathrm{Arg}\,\underset{\boldsymbol{k}\in\boldsymbol{s}_0}{\boldsymbol{Min}}\left\{\boldsymbol{d}\left(\boldsymbol{X}_{\boldsymbol{k}}, \boldsymbol{X}_{\boldsymbol{k}_0}\right)\right\}$$

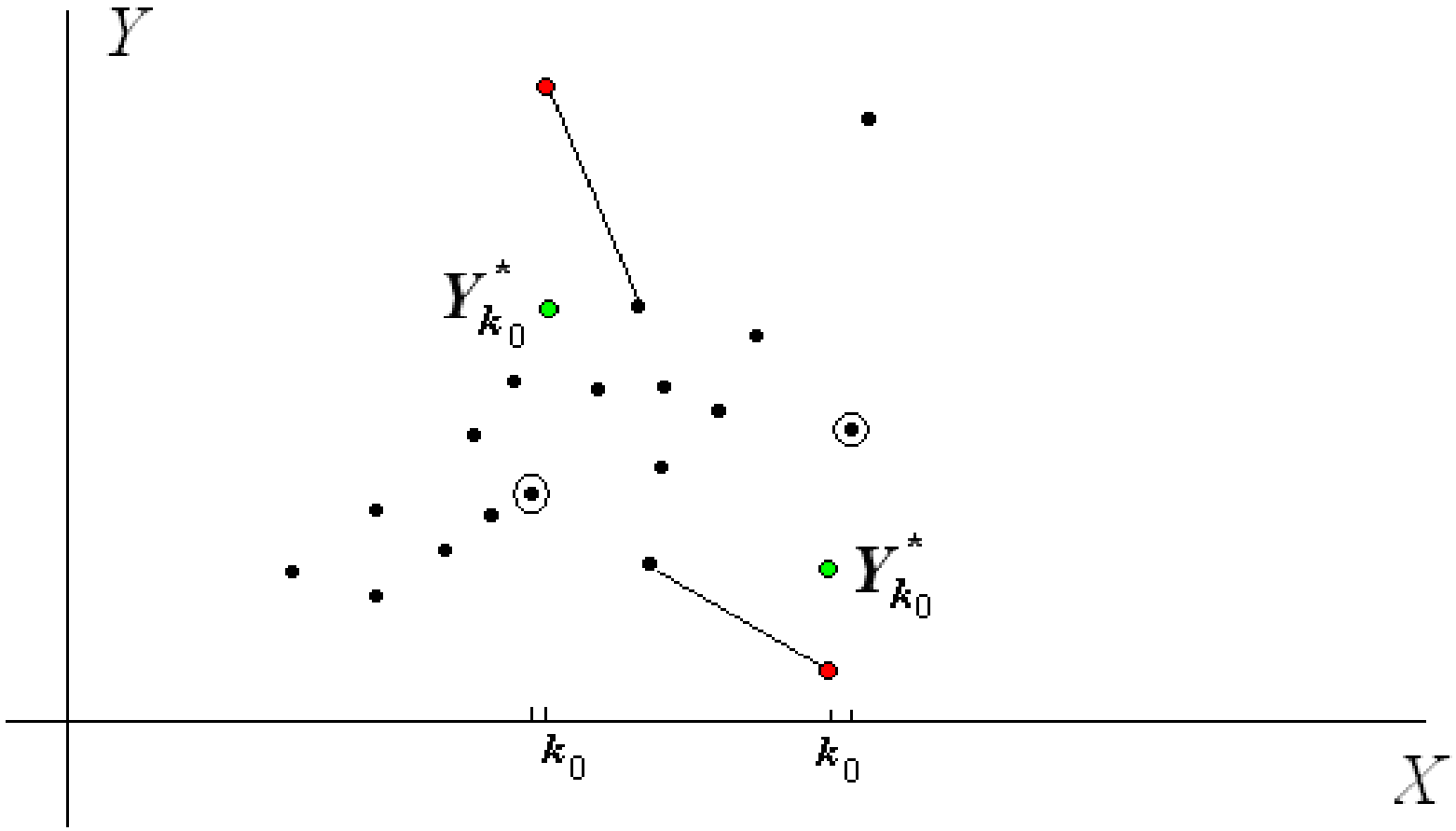
- Une simple modification du plus proche voisin en prenant compte la valeur aberrante consiste à minimiser une distance :

$$\boldsymbol{k}' = \mathrm{Arg}\,\underset{\boldsymbol{k}\in\boldsymbol{s}_0}{\boldsymbol{Min}}\left\{\boldsymbol{d}\left[\left(\boldsymbol{X}_{\boldsymbol{k}}, \boldsymbol{Y}_{\boldsymbol{k}}\right), \left(\boldsymbol{X}_{\boldsymbol{k}_0}, \boldsymbol{Y}_{\boldsymbol{k}_0}\right)\right]\right\}$$

- Par exemple, une mesure de distance usuelle est :

$$\boldsymbol{d}\left(\boldsymbol{k}, \boldsymbol{k}_0\right) = \boldsymbol{Sqrt}\left[\alpha\left(\boldsymbol{Y}_{\boldsymbol{k}} - \boldsymbol{Y}_{\boldsymbol{k}_0}\right)^2 + (1-\alpha)\left(\boldsymbol{X}_{\boldsymbol{k}} - \boldsymbol{X}_{\boldsymbol{k}_0}\right)'\left(\boldsymbol{X}_{\boldsymbol{k}} - \boldsymbol{X}_{\boldsymbol{k}_0}\right)\right]$$

où $0 \le \alpha < 1$ est une pondération choisie par le statisticien.



Légendes : points rouges - valeurs aberrantes ; points enveloppés - les plus proches voisins ; points verres - valeurs imputées par les plus proches voisins modifiés

4. Imputation par calage inverse

- Supposons un estimateur du total $\hat{t}_y$ est obtenu par une méthode résistante.

- Nous cherchons à modifier ou imputer les valeurs aberrantes Y_k , $k \in s_a$, par des valeurs Y_k^*, $k \in s_a$ normales ou moins aberrantes telles que :

$$\hat{t}_y^*\left(Y_k^* \mid k \in s\right) = \hat{t}_y, \quad \hat{t}_y^* \text{ est un estimateur classique.}$$

- Par exemple, lorsque $\hat{t}_y^*$ est un estimateur repondéré :

$$\hat{t}_y^* = \sum_{k \in s} w_k Y_k^* = \sum_{k \in s_0} w_k Y_k + \sum_{k \in s_a} w_k Y_k^* = \hat{t}_y$$

- Cela implique que la contribution des valeurs non-aberrantes, notée $\hat{t}_{1y}$, et la contribution des valeurs aberrantes, notée $\hat{t}_{2y}$, seront :

$$\hat{t}_{1y} = \sum_{i \in s_0} w_i Y_i \,, \quad \hat{t}_{2y} = \sum_{i \in s_a} w_i Y_i^*$$

- Notons que $\hat{t}_{1y}$ est connu, donc $\hat{t}_{2y}$ peut être calculé par :

$$\hat{t}_{2y} = \hat{t}_y - \hat{t}_{1y}, \quad (\text{supposons } \hat{t}_{2y} > 0)$$

- L'objectif de l'imputation est donc d'imputer des valeurs Y_k^*, $k \in s_a$:

$$\sum_{k \in s_a} w_k Y_k^* = \hat{t}_{2y}$$

- Comme les valeurs aberrantes sont des valeurs vraies, nous ne voulons pas que une valeur imputée est trop éloignée de la valeur vraie.
- C'est un problème de calage sur marge :

$$Y_k^* = Y_k F_k(w_k \lambda), \; k \in s_a$$

où $F_k(0) = 1$ et $F_k'(0) = q_k$; λ est une constante à déterminer par :

$$\sum_{k \in s_a} w_k Y_k F_k(w_k \lambda) = \hat{t}_{2y}.$$

- Par exemple, lorsque la mesure de distance est donnée:

$$\rho\left(Y^*, Y\right) = \sum_{k \in s_a} \left(Y_k^* - Y_k\right)^2 / 2 q_k Y_k$$

où $q_k > 0, k \in s_a$ sont des poids de calage choisis par statisticien (elle correspond à la méthode linéaire dans Deville et Särndal (1992)).

- Le calage donne les valeurs imputées :

$$Y_k^* = Y_k + q_k w_k Y_k \left(\sum_{k \in s_a} q_k w_k^2 Y_k \right)^{-1} \left(\hat{t}_{2y} - \sum_{k \in s_a} w_k Y_k \right), \; k \in s_a$$

- Lorsque les poids de calage sont choisis $q_k = w_k^{-1}, k \in s_a$, on a :

$$Y_k^* = Y_k \frac{\hat{t}_{2y}}{\sum_{k \in s_a} w_k Y_k}, \; k \in s_a$$

- On impute séparément deux types de valeurs aberrantes : valeur aberrante extrêmement grande et valeur aberrante extrêmement petite.

Remarques pour l'imputation par calage inverse :

- La variable ***Y*** joue le rôle de la variable de poids, la variable ***w*** joue le rôle d'une variable auxiliaire, d'où vient le nom 'calage inverse'.
- La variance de l'estimateur en utilisant les données imputées est identique à celle de l'estimateur résistant.
- Avantages:
  - L'estimateur du total résistant aux valeurs aberrantes peut être retrouvé par estimateur classique en utilisant les données imputées.
  - Le programme ***CALMAR*** peut être utilisé pour accomplir le calage.
- Problèmes en communs pour les trois méthodes d'imputation :
  - Valeurs imputées non-valides.
  - Estimation de variance.

5. Validations numériques

- Données d'enquête auprès des entreprises sur un secteur spécifique, contenant des valeurs aberrantes. **Nombre d'observations** : 6099.

Nombre de variables :

Nous considérerons deux variables d'intérêt : chiffre d'affaires (*turnover*) et total d'achat (*purtot*) ;

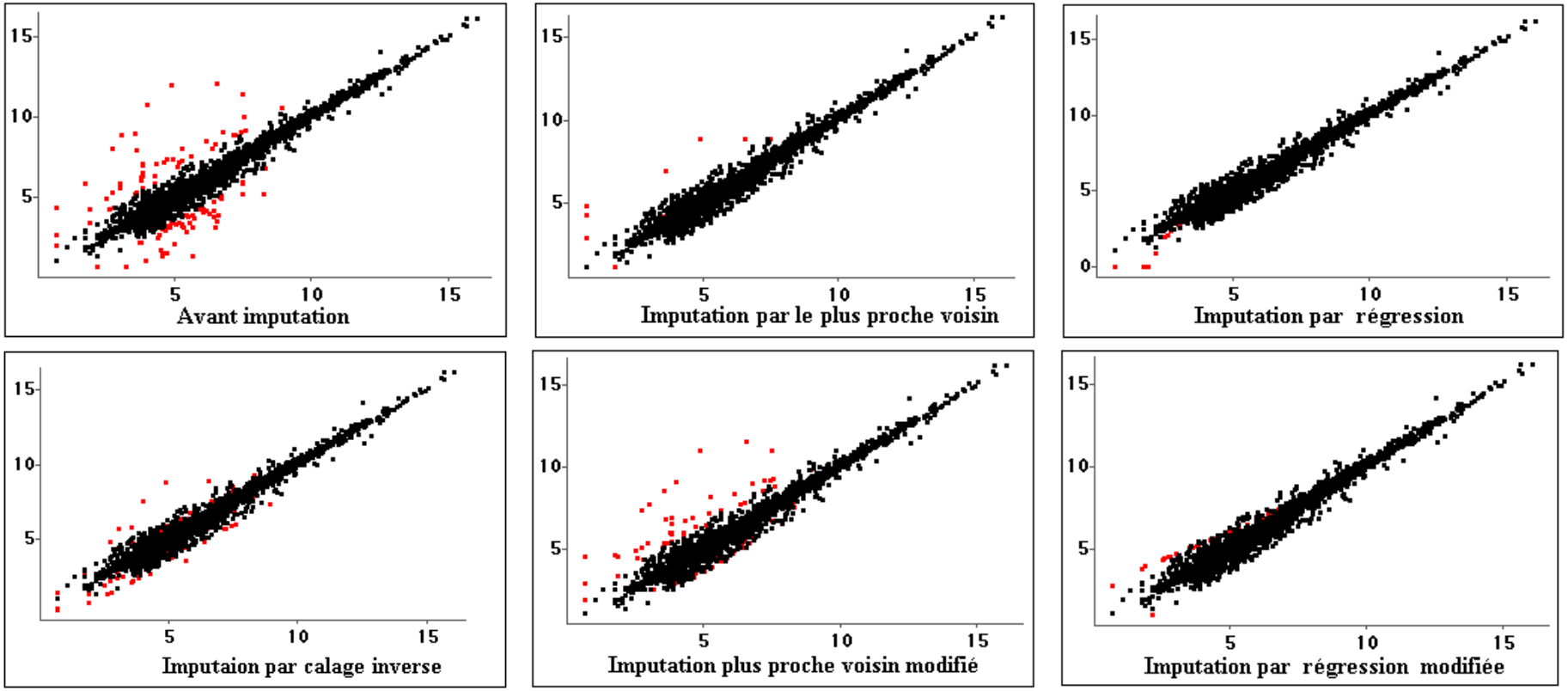
Nous disposons deux variables auxiliaires : chiffre d'affaires enregistré (*turnreg*) et effectif enregistré (*empreg*)

- Estimateur du total : estimateur basé sur un modèle linéaire (Chambers, 1986, JASA)

Nombres de valeurs aberrantes et estimation du total
Par estimateur basé sur un modèle linéaire

	Nombre de valeurs Aberrantes	*Estimation non-résistante du total*	*Estimation résistante du total*
Turnover	106	269545407	252938704
Purtot	111	192575028	180732418

Chiffres d'affaires par rapport aux chiffres d'affaires enregistrés
(En échelle de logarithme)

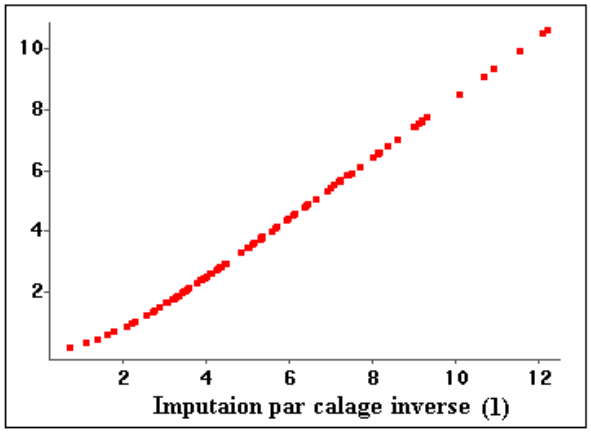


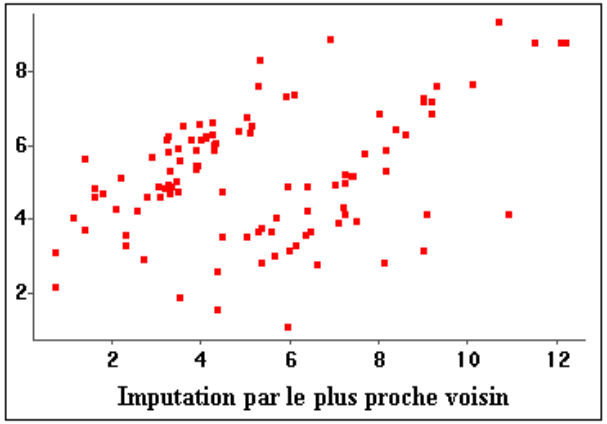
Légendes: Les points rouges sont les valeurs aberrantes ou leur imputation.

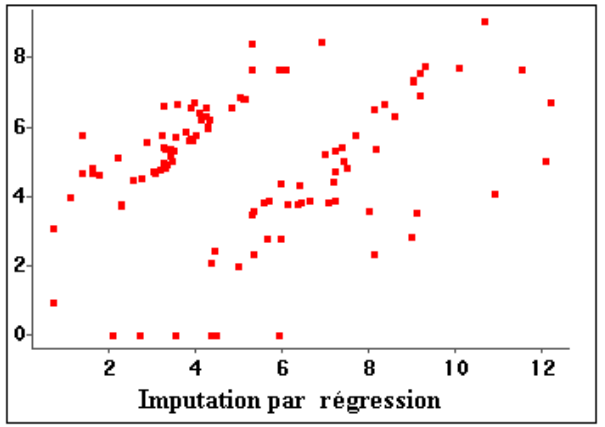
Les huit grosses valeurs aberrantes avant et après imputation (proche voisin modifié)

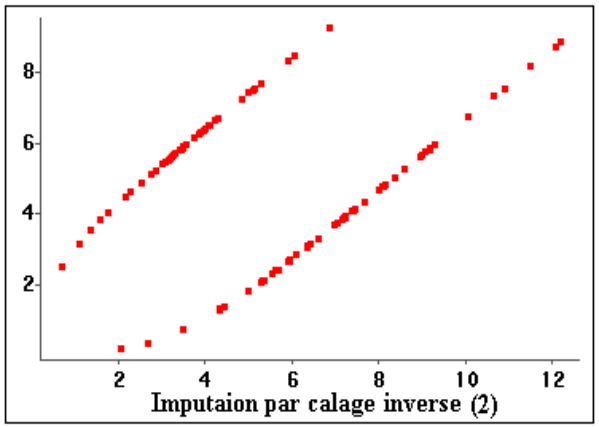
Valeurs Origines	209683	186399	103956	56000	9084	8176	3380
Valeurs Imputées	118879	67231	67231	9738	5683	2544	1699

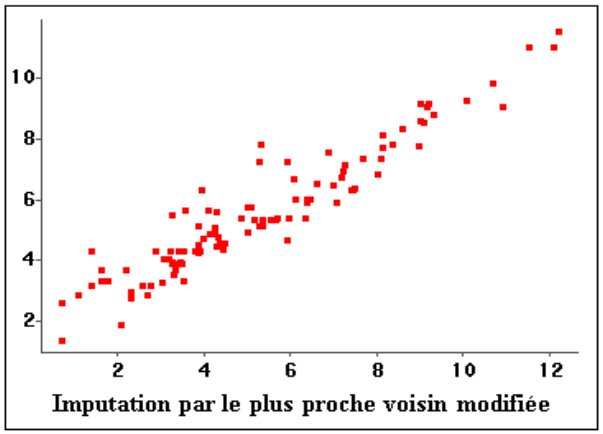
Chiffres d'affaires imputés par rapport aux vrais chiffres d'affaires
(en échelle de logarithme)

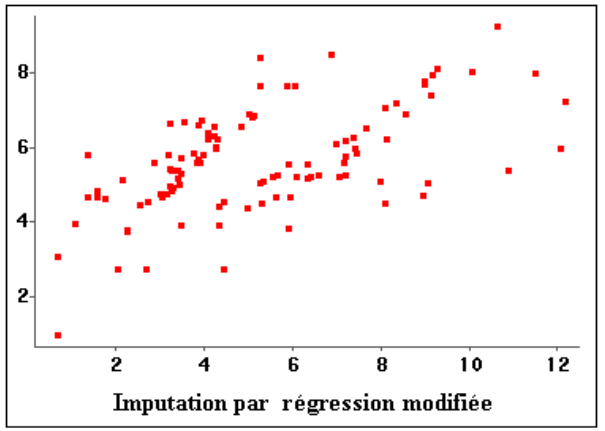












Estimation du total avant et après imputation
par estimateur basé sur un modèle linéaire

	Estimation Résistante	Estimation Classique Après Imputation				
	Avant Imputation	Calage Inverse	Régression	Proche voisin	Régression modifiée	Proche voisin modifié
Turnover	252938707	252808484	252225060	253479240	253005961	259245185
Purtot	180732418	180670463	180483772	181352764	181098312	185556428

Valeur moyenne et coefficient de corrélation
entre les vraies valeurs aberrantes et leurs imputations

	Vraie valeur	Calage inverse	Régression	Proche voisin
Turnover	1456	325 (1,00) 325 (0,42)	214 (0,07) 273 (0,14) (modifiée)	217 (0,33) 547 (0,87) (modifiée)
Purtot	763	182 (1,00) 182 (0,08)	173 (0,03) 209 (0,05) (modifiée)	156 (0,09) 320 (0,89) (modifiée)