risques chroniques émergents

CHROME

EA 7352

Document de recherche #2019-01

La régression Gini : une revue de la littérature

Téa Ouraga

UNÎMES UNIVERSITÉ

La régression Gini : une revue de la littérature

Téa OURAGA *
CHROME
Université de Nîmes

Résumé

Ce papier propose une revue de la littérature de la méthodologie. La méthodologie Gini, dans le cadre des modèles de régression, a été retenue pour ses performances en cas d'erreur de mesures ou d'outliers venant contaminer les données. La régression Gini met en évidence des estimateurs robustes donnant de meilleurs résultats que les estimateurs LAD ou MCO lorsque les outliers sont présents dans les régresseurs.

Mots-clés : Erreurs de mesure ; Outliers ; Régression Gini ; Robustesse

Résumé

This paper provides a revew of the Gini's methodology. The Gini's methodology, within the framework of the models of regression, was retained for its performances in cases of error of measures or outliers coming to contaminate the data. The regression Gini highlights strong estimators giving better results that LAD or MCO estimators when outliers are present in regressors.

Keywords : Measurement errors ; Outliers ; Gini Regression ; Robustness

* *Université de Nîmes - Laboratoire CHROME, 5 Rue du Dr Georges Salan, 30000 Nîmes . E-mail : jeromeouraga@gmail.com*

1 Introduction

La modélisation économétrique est très répandue dans les domaines scientifiques tels que l'économie, la finance, la sociologie, la biologie, etc. La recherche d'estimateurs robustes, peu sensibles aux valeurs extrêmes (outliers) s'avère indispensable afin d'obtenir des estimations et prévisions crédibles permettant de valider les raisonnements et hypothèses issus de modèles théoriques.

La statistique « Gini's Mean Difference » (GMD), ou coefficient de Gini absolu, est une mesure alternative de variabilité introduite par le statisticien italien Corrado Gini en 1912. Elle sera très répandue, dans un premier temps en économie du développement en tant qu'indice d'inégalité, puis progressivement s'étendra vers d'autres domaines tels que la finance en tant qu'indice de dispersion (mesure de risque) et l'économétrie.

Entre 1970 et 1990, de nouvelles méthodes de régressions ont été introduites comme les régressions quantiles de Basset et Koenker (1978), la régression linéaire floue de Tanaka et al. (1982), et la régression Gini de Olkin et Yitzhaki (1992) basée sur la statistique GMD. La régression Gini possède deux particularités : elle permet de s'affranchir de certaines hypothèses standards en économétrie et autorise l'utilisation de données comportant des erreurs de mesures ou des valeurs aberrantes.

La présence d'outliers dans la base de données rend difficile l'utilisation de certaines techniques statistiques et de data mining, car elle donne des résultats fallacieux lorsqu'elle n'est pas correctement traitée en amont. De même, le non-respect des hypothèses de base d'un modèle économétrique comme celui de l'exogénéité des variables explicatives dans la méthode des moindres carrés ordinaires (MCO) conduit à de mauvaises interprétations des coefficients (coefficients non significatifs et parfois avec des signes inversés).

L'objet de cet article est de mettre en évidence, dans le cadre des modèles de régressions généralisés, l'utilisation du GMD qui permet de répondre aux problèmes posés par les outliers et les erreurs de mesure sur les variables explicatives, et de s'affranchir de certaines hypothèses de base en économétrie comme la linéarité. Quand la distribution du processus générateur suit une loi normale univariée, la moyenne et la variation de l'échantillon sont des statistiques suffisantes pour décrire la distribution, rendant l'utilisation du GMD superflu. De même, dans le cas multivarié, lorsque la distribution des régresseurs suit une loi normale multivariée, l'estimateur issu du GMD est équivalent à celui des MCO. Néanmoins, lorsque la distribution n'est pas normale multivariée, le GMD se révèle être un estimateur robuste. Yitzhaki et Schechtman (2013) montrent que le GMD est utile lorsque les relations entre les variables aléatoires sont symétriques ou non, lorsque les rangs des

variables aléatoires sont liés, lorsque la population est stratifiée, lorsque l'hypothèse de linéarité du modèle de régression est soutenue ou non par les données observées. La méthodologie Gini permet à son utilisateur d'éstimer les coefficients d'un modèle de régression en utilisant les rangs des régresseurs, d'utiliser les U-statistiques pour l'inférence des coefficients, de tester la linérité du modèle, et de vérifier si les erreurs de mesure dans les régresseurs peuvent avoir une influence sur les estimations.

Cette revue de la littérature de la régression Gini est structurée comme suit : la première section introduit l'opérateur de Gini covariance basé sur les rangs des variables aléatoires ; la deuxième section décrit les différents types de régression Gini ; la troisième section présente les conséquences des erreurs de mesure sur les estimateurs Gini ; la quatrième section expose l'inférence des estimateurs Gini avec la théorie des U-statistiques ; enfin la dernière section conclut l'article.

2 L'opérateur de Gini covariance

Nous exposons dans cette section l'opérateur de Gini covariance introduit par Schechtman et Yitzhaki (1987), la notion de Gini corrélation qui en découle, et l'intérêt d'utiliser cette corrélation issue des rangs des variables aléatoires dans les modèles de régression, comme l'avait indiqué Durbin (1954).

Définissons tout d'abord les notations utilisées :
$\hookrightarrow$ $\mathbf{y}$ le vecteur $N \times 1$ de la variable à expliquer (à valeur dans $\mathbb{R}$).
et d'écart-type empirique sans biais s_y.
$\hookrightarrow$ $\mathbf{X}$ la matrice $n \times K$ des variables explicatives, avec $\mathbf{x}_k$ une variable explicative (un vecteur), $k = 1 \dots K$, $K < n$ et x_{ik} la i ème observation du vecteur $\mathbf{x}_k$, $i = 1, \dots, n$.
$\hookrightarrow$ $\mathbf{Z}$ la matrice $n \times K$ des variables instrumentales, avec $\mathbf{z}_k$ une variable instrumentale (un vecteur), $k = 1 \dots K$, $K < n$ et z_{ik} la i ème observation du vecteur $\mathbf{z}_k$, $i = 1, \dots, n$.
$\hookrightarrow$ $\widehat{\boldsymbol{\beta}} = \widehat{\beta}_1, \dots, \widehat{\beta}_K$, le vecteur des coefficients estimés.
$\hookrightarrow$ $\boldsymbol{\theta}$ le vecteur de contamination de dimension $n \times 1$.
$\hookrightarrow$ $\widehat{\beta^{\boldsymbol{\theta}}}_G$, le coefficient estimé dans la régression Gini en présence de l'erreur d'observation.
$\hookrightarrow$ $\widehat{V}$ & $\widehat{\sigma}$ respectivement la variance estimée et l'écart-type estimé.
$\hookrightarrow$ $\rho_{X,Y}$ le coefficient de corrélation de Pearson.
$\hookrightarrow$ $Cov\,(X, Y)$ est la covariance entre les variables X et Y.
$\hookrightarrow$ F_X et D_X désignent respectivement la fonction de répartition de X et l'ensemble de définition de F_X.

2.1 Définitions

Le coefficient de Gini, ou indice de Gini, est une mesure homogène de degré zéro qui permet de mesurer des disparités (inégalités) au sein d'une population donnée. Il est compris entre 0 et 1. Pour une population oú il existe une parfaite répartition (des revenus), ce coefficient est égal à 0, et inversement. L'indice de Gini homogène de degré 1, encore appelé « Gini's Mean Difference » ou GMD, est un indice de variabilité d'une variable aléatoire. Il définit la valeur attendue entre deux observations prises au hasard dans une population.

Soient X_1 et X_2 des observations indépendantes issues d'une même variable aléatoire X de fonction de répartition F_X. Le GMD est donné par :

$$G_X = \mathbb{E}|X_1 - X_2| \tag{1}$$

Remarquons que la variance peut aussi s'écrire en terme d'écarts absolus espérés, mais avec une métrique différente :

$$\sigma_X^2 = \frac{1}{2}\,\mathbb{E}|X_1 - X_2|^2 \tag{2}$$

L'indice GMD peut être récrit à l'aide de l'opérateur de Gini covariance mesurant la variabilité entre la variable aléatoire et sa fonction de répartition (Stuart, 1954) :

$$G_X = 4\ Cov(X, F_X) \tag{3}$$

Lorsque X est distribuée selon une loi normale, $G_X = 2\sigma_x/\sqrt{\pi}$, alors l'indice de Gini et la variance deviennent des mesures de dispersion équivalentes, dans la mesure où elles renvoient à un même préordre classant deux alternatives.

L'expression (3) met en évidence la covariance au sens de Gini. Pour deux variables aléatoires X et Y, il existe précisément deux Gini covariances ($GCov$), introduites par Schechtman et Yitzhaki (1987) :

$$GCov(Y, X) = Cov\,(Y, F_X) \tag{4}$$

$$GCov(X, Y) = Cov\,(X, F_Y) \tag{5}$$

La Gini covariance est la mesure de la covariance entre une variable aléatoire et la fonction de répartition (ou le vecteur rang) d'une autre variable aléatoire. Habituellement pour étudier la relation qui existe entre deux variables aléatoires, on utilise la covariance usuelle $Cov(X, Y)$ et le coefficient de corrélation de Pearson dont la métrique est euclidienne (L_2). Afin de changer de métrique, le coefficient de corrélation de Pearson peut être remplacé par celui de Spearman qui nécessite au préalable que les deux variables aléatoires X et Y soient représentées par leur fonction de répartition, $Cov(F_X, F_Y)$. La métrique obtenue est de type L_1 (distance de Manhattan). La Gini covariance peut donc être considérée comme un mélange entre les deux métriques précédentes, L_2 et L_1 respectivement.

2.2 La Gini corrélation

La Gini corrélation, notée GC, est une mesure normalisée de corrélation issue de la Gini covariance. Elle prend ses valeurs dans l'intervalle $[-1;1]$. Dans la mesure où la Gini covariance n'est pas syémtrique, voir Eq.(3), deux Gini corrélations peuvent être définies :

$$GC(Y,X) = \frac{GCov(Y,X)}{GCov(Y,Y)} = \frac{Cov(Y,F_X)}{Cov(Y,F_Y)} \tag{6}$$

$$GC(X,Y) = \frac{GCov(X,Y)}{GCov(X,X)} = \frac{Cov(X,F_Y)}{Cov(X,F_X)} \tag{7}$$

Proposition 2.1 – Schechtman et Yitzhaki (1987) :
(i) *Soit deux variables aléatoires X et Y interchangeables, alors il existe une fonction $h : \mathbb{R} \to \mathbb{R}$ telle que $Xh(Y) = Yh(X)$ et donc $\mathbb{E}(Xh(Y)) = \mathbb{E}(Yh(X))$. L'interchangeabilité implique que $GC(X,Y) = GC(Y,X)$.*
(ii) *Si (X,Y) suit une distribution normale bivariée d'espérance (μ_X, μ_Y), de variances σ_X^2, σ_Y^2, tel que $\rho_{X,Y}$, alors :*

$$GC(X,Y) = GC(Y,X) = \rho_{X,Y} \tag{8}$$

(iii) *Soit X et Y deux variables aléatoires, alors :*

$$GC_{X+Y} = GC(X, X+Y)\, G_X + GC(Y, X+Y)\, G_Y. \tag{9}$$

En contraste avec la nature symétrique de l'opérateur usuel de covariance, les Gini corrélations $GCov(X,Y)$ et $GCov(Y,X)$, peuvent être de signe et d'intensité différents. Cette propriété peut être vue *a priori* comme une limite de la méthode Gini. Cependant, comme l'ont démontré Carcea et Serfling (2015) dans le cadre des séries temporelles, le fait de disposer de deux fonction d'autocorrélation de type Gini permet de mieux identifier les processus ARMA (notamment en présence de valeurs aberrantes).

2.3 La méthode des rangs de Durbin

Afin de comprendre l'intérêt de l'opérateur de Gini covariance dans le cadre de la modélisation économétrique, revenons sur l'apport de Durbin (1954). Ce dernier propose d'utiliser les rangs des variables explicatives comme instruments. Utilisons la statistique d'ordre $\tilde{\mathbf{x}}$ d'une réalisation de la variable aléatoire X telle que $\tilde{x}_1 \leq \tilde{x}_1 \leq \cdots \leq \tilde{x}_n$. Le rang de l'observation i de la variable observée $\mathbf{x}$ issue d'un échantillon est :

$$\mathbf{r}_{x_i} = \sum_{i=1}^{n} \mathbb{1}(x \leq \tilde{x}_i) \tag{10}$$

avec $\mathbb{1}(x \leq \tilde{x}_i)$ la fonction qui retourne la valeur 1 lorsque l'expression $x \leq \tilde{x}_i$ est vraie. Un estimateur basique de la fonction de répartition est ainsi obtenu :

$$\widehat{F}_X = \frac{\mathbf{r_x}}{n} \tag{11}$$

avec $\mathbf{r_x} = (\mathbf{r}_{x_1}, \ldots, \mathbf{r}_{x_n})$. Yitzhaki et Schechtman (2013) indique que les *ex aequo* peuvent induire des biais dans le calcul de l'estimateur de la Gini covariance. Dans la pratique, trois méthodes sont couramment utilisées pour traiter les *ex aequo* :

• utiliser le rang supérieur des deux observations, ce qui conduira à une sur-pondération et donc un biais positif ;

• utiliser le rang inférieur, ce qui aboutira à une sur-pondération et donc un biais négatif ;

• estimer le rang au point moyen, ce qui produira des estimateurs sans biais.

Par exemple, la troisième méthode peut être illustrée de la manière suivante :

$$\mathbf{x} = \begin{pmatrix} 1 \\ 1 \\ 4 \\ 7 \\ 6 \end{pmatrix} \longrightarrow \mathbf{r_x} = \begin{pmatrix} 1,5 \\ 1,5 \\ 3 \\ 5 \\ 4 \end{pmatrix} \tag{12}$$

Les estimateurs de la Gini covariance s'expriment donc comme suit :

$$\widehat{GCov}(X,Y) = \frac{1}{n}\, Cov(\mathbf{x}, \mathbf{r_y}) \tag{13}$$

$$\widehat{GCov}(Y,X) = \frac{1}{n}\, Cov(\mathbf{y}, \mathbf{r_x}) \tag{14}$$

En 1954, Durbin fait remarquer que lorsque des données comportent des valeurs aberrantes, il est utile d'utiliser le vecteur rang dont les valeurs restent relativement stables en cas de contamination. Reprenons l'exemple précédent, en multipliant la plus forte observation par 100, le vecteur rang reste inchangé :

$$\mathbf{x}_1 = \begin{pmatrix} 1 \\ 1 \\ 4 \\ 700 \\ 6 \end{pmatrix} \longrightarrow \mathbf{r_{x_1}} = \begin{pmatrix} 1,5 \\ 1,5 \\ 3 \\ 5 \\ 4 \end{pmatrix} = \mathbf{r_x} \tag{15}$$

Durbin (1954) propose par conséquent d'utiliser la matrice d'instruments $\mathbf{R} = (\mathbf{r_{x_1}}, \ldots, \mathbf{r_{x_K}})$ qui contient en colonne les vecteurs rangs des variables explicatives. Puisqu'un vecteur rang peut raisonnablement être considéré comme indépendant du terme d'erreur, des estimateurs robustes peuvent être obtenus dans le cas d'une régression multiple de type $y = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$:

$$\hat{\boldsymbol{\beta}}_{VI-rank} = (\mathbf{R}'\mathbf{X})^{-1}\, \mathbf{R}'\mathbf{y} \tag{16}$$

En notant $\mathbf{Z}$ la matrice des instruments, telle que $\mathbf{Z} = \mathbf{R}$, on retrouve l'estimateur usuel $(\mathbf{Z}'\mathbf{X})^{-1}\ \mathbf{Z}'\mathbf{y}$ des moindres carrés ordinaires par variables instrumentales dans le cas où il existe autant d'instruments que de régresseurs. Durbin (1954) venait, sans le savoir, de découvrir une forme particulière de la régression Gini.

3 La régression Gini

La régression Gini compte deux approches. La première approche consiste à minimiser une mesure de dispersion des résidus, alternative à la variance. Il s'agit simplement de changer de norme en utilisant l'indice de Gini des résidus, autrement dit passer de la norme L_2 à la norme L_1. Cette première approche, par minimisation, est l'approche paramétrique. La seconde, l'approche semi-paramétrique, consiste à trouver des estimateurs robustes à partir de moyennes pondérées d'estimateurs de tendance centrale comme la médiane. L'idée est de trouver un système de pondération moins sensible aux outliers aboutissant à des propriétés désirables pour les estimateurs obtenus.

3.1 Hypothèses classiques

Revenons sur les hypothèses standards des modèles linéaires usuels $\mathbf{y} = \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\epsilon}$.

Hypothèses 3.1 – **(H1)** – : *Le modèle est linéaire en* $\mathbf{x}_k$ *(ou en n'importe quelle transformation de* $\mathbf{x}_k$*).*

Hypothèses 3.2 – **(H2)** – : *Les valeurs* x_{ik} *sont observées sans erreur (*x_{ik} *non aléatoire).*

Hypothèses 3.3 – **(H3)** – : $\mathbb{E}(\epsilon_i^2) = \sigma_{\boldsymbol{\epsilon}}^2$*, la variance de l'erreur est constante (homoscédasticité) : le risque de l'amplitude de l'erreur est constante quelle que soit la période.*

Hypothèses 3.4 – **(H4)** – : $\mathbb{E}(\epsilon_i \epsilon_{i'}) = 0$ *si* $i \neq i'$*, les erreurs sont non corrélées ou indépendantes*

Hypothèses 3.5 – **(H5)** – : $\mathbb{E}(\mathbf{x}_k' \boldsymbol{\epsilon}) = 0$*, l'erreur est indépendante de la variable explicative* k*, pour tout* $k = 1, \ldots, K$*.*

Il est intéressant de mettre en évidence, dans le cadre des régressions Gini, les hypothèses nécessaires à leur mise en oeuvre et celles qui peuvent être relâchées.

3.2 La regression Gini paramétrique

La régression paramétrique est basée sur la minimisation du Gini des résidus. Pour procéder par minimisation, la forme du modèle doit être spécifiée. L'hypothèse (H1) est donc invoquée.

Considérons le modèle linéaire simple suivant $\mathbf{y} = \alpha\mathbf{1} + \beta\mathbf{x} + \boldsymbol{\epsilon}$. Pour un échantillon de taille n, la valeur estimée de $\mathbf{y}$ est notée $\hat{\mathbf{y}} = \hat{\alpha} + \hat{\beta}\mathbf{x}$. L'indice de Gini du résidu $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ est une fonction de $\hat{\beta}$:

$$G_{\mathbf{e}}(\hat{\beta}) = \frac{1}{n} Cov(\mathbf{e}, \mathbf{r_e}), \tag{17}$$

où $\mathbf{r_e}$ représente le vecteur rang du résidu. Minimiser (17) revient à minimiser $\sum_{i=1}^{n} e_i \mathbf{r}_{e_i}$ qui est la fonction de distance utilisée dans la R-regression de Jurecková (1981), Jaeckel (1972) et McKean & Hettmansperger (1976). L'indice de Gini du résidu se récrit :

$$\begin{aligned} G_{\mathbf{e}}(\hat{\beta}) &= \frac{4}{n}\, Cov(\mathbf{e}, \mathbf{r_e}) \\ &= \frac{4}{n}\, Cov(\mathbf{y} - \hat{\alpha} - \hat{\beta}\mathbf{x}, \mathbf{r_e}) \\ &= \frac{4}{n}\Big[Cov(\mathbf{y}, \mathbf{r_e}) - \hat{\beta}\, Cov(\mathbf{x}, \mathbf{r_e})\Big] \\ &= \frac{4}{n}\Big[Cov(\mathbf{y}, \mathbf{r_e}) - \hat{\beta}\, Cov(\mathbf{x}, \mathbf{r_e})\Big] \end{aligned} \tag{18}$$

Pour un $\hat{\beta}$ donné, calculons l'indice de Gini du résidu en utilisant l'écart espéré entre deux observations prises au hasard (avec remise) :

$$G_{\mathbf{e}}(\hat{\beta}) = \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{|e_i - e_j|}{n^2} \tag{19}$$

Minimiser $G_e(\widehat{\beta})$ revient à minimiser :

$$\begin{aligned} \sum_{i=1}^{n}\sum_{j=1}^{n} |e_i - e_j| &= \sum_{i,j} |(y_i - \widehat{y}_i) - (y_j - \widehat{y}_j)| \\ &= \sum_{i,j} |(y_i - y_j) - (\hat{\alpha} + \hat{\beta}x_i - \hat{\alpha} - \hat{\beta}x_j)| \\ &= \sum_{i,j} |(y_j - y_i) - \hat{\beta}\,(x_j - x_i)| \\ &= 2\sum_{i<j} [\,(y_j - y_i) - \hat{\beta}\,(x_j - x_i)\,] \end{aligned} \tag{20}$$

Proposition 3.1 – Olkin et Yitzhaki (1992) – : *Lorsque l'indice de Gini du résidu est à son minimum,* $G_{\mathbf{e}}(\hat{\beta}) = 2\sum_{i<j}[\,(y_j - y_i) - \hat{\beta}\,(x_j - x_i)\,] = 0$, *la pente de la régression Gini notée* $\hat{\beta}_G$ *est alors :*

$$\widehat{\beta}_G = \frac{\sum_{i<j}(y_j - y_i)}{\sum_{i<j}(x_j - x_i)} \tag{21}$$

Si les résidus sont ordonnés $e_1 \leq e_2 \leq \ldots \leq e_n$, la dérivée de l'indice de Gini en $\hat{\beta}$ est :

$$\begin{aligned} \frac{\partial G_{\mathbf{e}}(\widehat{\beta})}{\partial \widehat{\beta}} &= -2 \sum_{i<j} (x_j - x_i) \\ &= 4 \sum_{i=1}^{n} x_i \left[i - \frac{n+1}{2} \right] \\ &= 4\, n\, Cov\,(\,\mathbf{x}, \mathbf{r_e}\,) \end{aligned} \tag{22}$$

Par construction, l'indice de Gini atteint son minimum lorsque le terme $Cov(\,\mathbf{x}, \mathbf{r_e}\,)$ est minimisé, ce qui est le cas lorsque $Cov(\,\mathbf{x}, \mathbf{r_e}\,) = 0$.

Proposition 3.2 – Olkin et Yitzhaki (1992) – : *Lorsque l'indice de Gini du résidu est minimal, la covariance entre l'estimateur $\hat{y}$ et le rang du résidu est nulle :*

$$\begin{aligned} Cov(\hat{y}, \mathbf{r_e}) &= Cov(\alpha + \hat{\beta}_G \mathbf{x}, \mathbf{r_e}) \\ &= \hat{\beta}_G Cov(\mathbf{x}, \mathbf{r_e}) = 0 \end{aligned} \tag{23}$$

Proposition 3.3 – Olkin et Yitzhaki (1992) – : *Lorsque l'indice de Gini du résidu est minimal, la covariance entre la variable dépendante y et le rang du résidu $\mathbf{r_e}$ est égal à l'indice de Gini du terme d'erreur :*

$$\begin{aligned} Cov(\mathbf{y}, \mathbf{r_e}) &= Cov(\widehat{\mathbf{y}} + \mathbf{e}, \mathbf{r_e}) \\ &= Cov(\alpha + \widehat{\beta}_G \mathbf{x} + \mathbf{e}, \mathbf{r_e}) \\ &= \widehat{\beta}_G Cov(\mathbf{x}, \mathbf{r_e}) + Cov(\mathbf{e}, \mathbf{r_e}) \\ &= Cov(\mathbf{e}, \mathbf{r_e}) \end{aligned} \tag{24}$$

Dans le cadre d'un modèle de régression généralisé,

$$\mathbf{y} = \mathbf{X}\, \boldsymbol{\beta} + \boldsymbol{\epsilon}$$

l'approche paramétrique de la régression Gini consiste à déterminer le vecteur suivant :

$$\hat{\boldsymbol{\beta}}_G = \arg\min \left\{ \sum_{i=1}^{n} \sum_{j=1}^{n} |e_i - e_j| \right\} = \arg\min \{ Cov(\mathbf{e}, \mathbf{r_e}) \} \tag{25}$$

Il n'existe pas dans ce cas de formes fonctionnelles fermées pour l'estimateur $\hat{\boldsymbol{\beta}}_G$, la méthode est simplement numérique.

3.3 La régression Gini non paramétrique

La méthode des moindres carrés ordinaires reste l'une des méthodes les plus utilisées pour estimer la relation entre une ou plusieurs variables explicatives et une variable dépendante. Soit le modèle linéaire simple $\mathbf{y} = \alpha\mathbf{1} + \beta\mathbf{x} + \boldsymbol{\epsilon}$, le coefficient de la pente de la droite de régression des moindres carrés ordinaires peut être obtenu par :

$$\begin{aligned} Cov(\mathbf{y}, \mathbf{x}) &= Cov(\beta\ \mathbf{x} + \boldsymbol{\epsilon}, \mathbf{x}) \\ &= Cov(\beta\ \mathbf{x}, \mathbf{x}) + Cov(\boldsymbol{\epsilon}, \mathbf{x}) \\ &= Cov(\beta\ \mathbf{x}, \mathbf{x}) \\ Cov(\mathbf{y}, \mathbf{x}) &= \beta\ Cov(\ \mathbf{x}, \mathbf{x}) \qquad (\mathbf{H5}) \end{aligned}$$

D'où :

$$\beta = \frac{Cov(y, \mathbf{x})}{Cov(\mathbf{x}, \mathbf{x})}$$

La régression Gini non paramétrique est construite en remplaçant la covariance usuelle par l'opérateur de Gini covariance :

$$\beta_G = \frac{Cov(\mathbf{y}, \mathbf{F}(\mathbf{x}))}{Cov(\mathbf{x}, \mathbf{F}(\mathbf{x}))} \tag{26}$$

Un estimateur du coefficient de la pente est obtenu en utilisant le vecteur rang mesuré au point moyen en cas d'ex aequo :

$$\hat{\beta}_G = \frac{Cov(\mathbf{y}, \mathbf{r_x})}{Cov(\mathbf{x}, \mathbf{r_x})} \tag{27}$$

Il est possible de montrer que l'estimateur est sans biais en supposant que $Cov(\boldsymbol{\epsilon}, \mathbf{r_x}) = 0$.
En remplaçant dans la formule de $\widehat{\beta}_G$, $\mathbf{y}$ par $\beta_G\mathbf{x} + \boldsymbol{\epsilon}$, on obtient :

$$\begin{aligned} \widehat{\beta}_G &= \frac{Cov(\beta_G\mathbf{x} + \boldsymbol{\epsilon}\ ,\ \mathbf{r_x})}{Cov(\mathbf{x}, \mathbf{r_x})} \\ \widehat{\beta}_G &= \frac{Cov(\beta_G\ \mathbf{x}\ ,\ \mathbf{r_x})}{Cov(\mathbf{x}, \mathbf{r_x})} + \frac{Cov(\boldsymbol{\epsilon}, \mathbf{r_x})}{Cov(\mathbf{x}, \mathbf{r_x})} \\ \widehat{\beta}_G &= \beta_G\ \frac{Cov(\mathbf{x}, \mathbf{r_x})}{Cov(\mathbf{x}, \mathbf{r_x})} + \frac{Cov(\boldsymbol{\epsilon}, \mathbf{r_x})}{Cov(\mathbf{x}, \mathbf{r_x})} \\ \widehat{\beta}_G &= \beta_G \quad \text{car } Cov(\boldsymbol{\epsilon}, \mathbf{r_x}) = 0 \end{aligned} \tag{28}$$

L'approche par l'opérateur de Gini covariance est considérée comme non paramétrique puisque la méthode ne nécessite pas de spécifier un modèle, ne nécessite aucune hypothèse sur les distributions, et ne nécessite aucune méthode d'optimsation. Elle est néanmoins similaire dans sa structure à la méthode des moindres carrés ordinaires.

La méthode non paramétrique (Gini et MCO) peut être vue comme une méthode géométrique où les estimateurs sont des moyennes pondérés des pentes de la courbe de régression entre chaque paire d'observations (y_i, x_i). La différence entre la régression Gini et celle des MCO se trouve dans la structure des poids attachés aux coefficients des pentes.

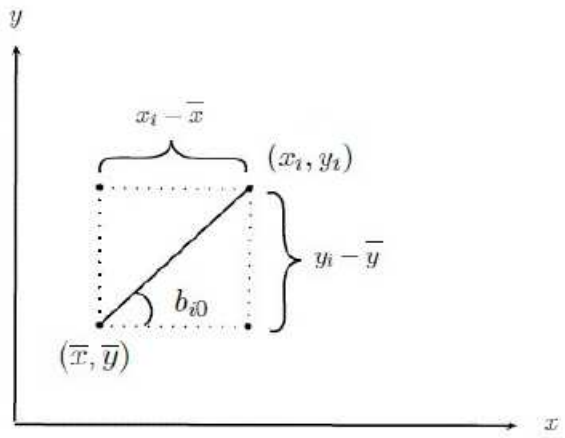

FIGURE 1 – Tangentes

Soit (x_i, y_i) tel que $i = 1, \ldots, n$, un échantillon provenant d'une distribution bivariée de moments d'ordre 1 et 2 finis, tel que $x_1 \leq x_2 \leq \cdots \leq x_n$. Les tangentes formées par les couples d'observations (i, j) sont :

$$m_{ij} = \frac{y_i - y_j}{x_i - x_j}, \tag{29}$$

Construisons maintenant un estimateur du coefficient de pente du modèle de regression par une moyenne pondérée :

$$\hat{\beta}* = \sum_{i,j} p_{ij}\, m_{ij}, \quad \text{avec} \quad \sum_{i,j} p_{ij} = 1 \tag{30}$$

Le schéma (ou choix) de pondération (p_{ij}) déterminera les propriétés de l'estimateur. L'estimateur par MCO se définit comme suit :

$$\widehat{\beta} = \sum_{i>j} \underbrace{\frac{(x_i - x_j)^2}{\sum_{i>j}(x_i - x_j)^2}}_{p_{ij}} \underbrace{\frac{(y_i - y_j)}{(x_i - x_j)}}_{m_{ij}} \tag{31}$$

Avec une pondération quadratique, les observations extrêmes (très éloignées de la moyenne) auront une influence non négligeable même pour des échantillons de grandes tailles. En effet, elles verront leur distance s'amplifier et elles auront donc un poids beaucoup plus important. La présence de ce

point prépondérant dans le cas des MCO peut donner une autre allure à la droite de regression et par conséquent fausser les interprétations.

Il serait donc judicieux de trouver un estimateur avec une pondération moins sensible aux valeurs extrêmes. L'estimateur de la pente du modèle de régression par la méthode de Gini $\widehat{\beta}_G$ peut s'écrire aussi sous forme de pondération, comme le proposent Olkin et Yitzhaki (1992). Néanmoins certains auteurs avaient déjà privilégié la piste des estimateurs obtenus par moyenne pondérée. L'estimateur proposé par Scholz (1977) et Sievers (1978) est défini comme la médiane pondérée des tangentes. Comme le montrent Olkin et Yitzhaki (1992, théorème 1), il a les mêmes propriétés que celles de la régression Gini par minimisation. En utilisant les moyennes pondérées des tangentes à la place des médianes, les auteurs définissent l'estimateur de la régression Gini non paramétrique :

$$\widehat{\beta}_G = \sum_{i>j} \underbrace{\frac{(x_i - x_j)}{\sum_{i>j}(x_i - x_j)}}_{p_{ij}} \underbrace{\frac{(y_i - y_j)}{(x_i - x_j)}}_{m_{ij}} \tag{32}$$

La pondération m_{ij} confère à l'esimateur $\widehat{\beta}_G$ une propriété remarquable, sur laquelle nous reviendrons, il est moins sensible aux valeurs extrêmes que l'estimateur par MCO. Dans le cas d'une régression multiple l'estimateur est donné par :

$$\widehat{\boldsymbol{\beta}}_G = (\mathbf{r_X}^T \mathbf{X})^{-1} \ (\mathbf{r_X}^T y)$$

avec $\mathbf{r_X}$ la matrice rang comportant en colonne les vecteurs rang des régresseus. Si les vecteurs rangs sont non corrélés au terme d'erreur, alors ils peuvent être utilisés comme instruments et $\mathbf{r_X}$ correspond à la matrice $\mathbf{Z}$ d'une régression sur variables instrumentales. La régression Gini peut donc être vue comme une régression par variables instrumentales :

$$\begin{aligned} \widehat{\beta}_{VI-MCO} &= \widehat{\beta}_G \\ (\mathbf{Z}^T \ \mathbf{X})^{-1} \ (\mathbf{Z}^T \ y) &= (\mathbf{r_X}^T \mathbf{X})^{-1} \ (\mathbf{r_X}^T y) \end{aligned} \tag{33}$$

Néanmoins, l'estimateur $\widehat{\beta}_G$ ne nécessite aucune hypthèse particulière contrairement à celui des MCO qui repose sur les hypothèses **H1-H5**.

3.4 Outliers

La présence d'outliers dans les données nécessite un traitement approprié pour éviter de produire des résultats peu fiables (estimateurs biaisés et/ou non convergents).

Toute valeur qui présente des aberrations, valeur trop élevée ou trop faible, et qui ne passe pas inaperçue (pour des échantillons de faible taille) est considérée comme valeur aberrante. Les valeurs aberrantes peuvent être dues par exemple à une mauvaise saisie, ce que l'on appelle communément erreurs de mesure. A ne pas confondre avec les outliers ou valeurs extrêmes

qui peuvent être considérés comme valeurs aberrantes mais avec la particularité d'être des valeurs exactes qui ne doivent pas être retirées de la base de données. Il s'agit ici d'individus qui peuvent présenter des caractéristiques atypiques sur l'une des variables étudiées. La présence de valeurs extrêmes dans la base de données peut changer l'amplitude et les signes des coefficients [1], aboutir à de mauvaises interprétations et demande l'utilisation d'une technique statistique appropriée afin d'obtenir une modélisation économétrique pertinente. En présence d'outliers, le critère des MCO ne procure plus d'estimateur stable puisque l'écart-type estimé de l'erreur tend vers l'infini ($\widehat{\sigma}_\epsilon \longrightarrow \infty$) du fait de la violation de l'hypothèse **H3**. La variance estimée de l'estimateur $\widehat{V}(\widehat{\beta})$ tend à croître de manière disproportionnée si bien que l'estimateur devient peu fiable et s'accompagne d'une statistique de *Student* qui tend vers zéro, augmentant ainsi la probabilité d'accepter à tort l'hypothèse nulle.

Différentes méthodes de détection des outliers existent comme les méthodes statistiques inférentielles qui consistent à créer un intervalle de confiance $[\pm\ x\ ecarts - types]$ autour de la moyenne. Sera considérée comme valeur aberrante ou valeurs extrême, toute valeur qui se trouve hors de cet intervalle. Il est couramment utilisé dans les méthodes de détection non-automatiques des représentations graphiques des données :
1- La boîte à moustache : la détection varie selon l'amplitude x associée à l'écart-type ou à l'étendue interquartile [2] du bord de la boîte.
2- Le nuage de points : sont soupçonnés d'être des valeurs extrêmes, les points éloignés des autres points.
Il est plus prudent de ne pas se fier uniquement à ces représentations graphiques. Il est nécessaire de confirmer nos soupçons par des tests statistiques. L'un des tests les plus connus est celui de Grubbs (Grubbs, 1969 et Stefansky, 1972). Il est est utilisé pour détecter un outlier dans une distribution univariée qui suit approximativement une distribution normale :

$$\left\| \begin{array}{l} H_0 : \text{ Présence d'outliers} \\ H_1 : \text{Absence d'outliers.} \end{array} \right.$$

La statistique de test de Grubbs est :

$$Gb = \frac{max|y_i - \bar{y}|}{s_y},$$

Le test statistique de Grubbs est le plus grand écart absolu à la moyenne de l'échantillon par unité d'écart-type. L'hypothèse d'absence d'outliers dans les données est rejetée si :

$$Gb > \frac{n-1}{\sqrt{n}} \sqrt{\frac{t^2_{\alpha/2}}{n-2+t^2_{\alpha/2}}}$$

1. Knorr et Ng(1998), Ramsawmy et al. (2000), Choi (2009).
2. La différence entre le troisième et le premier quartile (Q3 - Q1).

où $t_{\alpha/2}$ représente la valeur critique de la distribution de Student à $n-2$ degrés de liberté.

La méthode du point de levier quant à elle se base aussi sur une fonction distance. Une observation qui a une valeur extrême sur une variable prédictive est appelée un point avec un effet de levier élevé. Dans un modèle de régression linéaire, le score de levier pour la $i$ème unité de données se définit comme : $h_i = \mathbf{x}_i \ (\mathbf{X}'\mathbf{X})^{-1} \ \mathbf{x}_i'$. Les éléments diagonaux h_i de la matrice $H = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ sont appelés les leviers et déterminent l'influence de la $i$ème observation sur les estimations obtenues par régression. Les valeurs de levier sont comprises entre 0 et 1, et la somme des h_i est égale au nombre de paramètres estimés. L'observation qui a une valeur levier $h_i > 2K/n$ est considérée comme aberrante.

4 Erreurs de mesure sur les variables

Les variables explicatives sont supposées être observées sans erreurs de mesure, ce qui s'apparente au respect de l'hypothèse **H5**. Il peut arriver dans la pratique que cette hypothèse ne soit pas vérifiée lorsque des valeurs aberrantes contaminent l'échantillon. Dans cette section, nous examinons les cas où la variable expliquée et les variables explicatives sont entachées d'erreurs de mesure.Nous dévelopons tout d'abord l'erreur de mesure sur la variable à expliquer, ensuite sur la variable explicative, et enfin le traitement par variables instrumentales.

4.1 Erreurs sur la variable endogène

Soit le modèle linéaire centré suivant $\mathbf{y} = \beta \ \mathbf{x} + \boldsymbol{\epsilon}$ tel que :

$$\begin{aligned} y_i &= \beta \ x_i + \ \epsilon_i \qquad \text{(la } i \text{ ème observation de } y \text{ sans erreur)} \\ y_i^* &= y_i \ + \ \theta_i \qquad \text{(la } i \text{ ème observation de } y \text{ avec erreur)} \\ y_i^* &= \beta \ x_i + \ (\epsilon_i + \theta_i) \qquad (Cov(x_i, \epsilon_i) = 0) \end{aligned} \tag{34}$$

L'estimateur $\widehat{\beta}$ de β est par conséquent égal à :

$$\begin{aligned} \widehat{\beta} &= \frac{\sum \ x_i \ y_i^*}{\sum x_i^2} \\ &= \frac{\sum \ x_i \ [\ \beta \ x_i + \ (\epsilon_i + \theta_i) \ ]}{\sum x_i^2} \\ &= \beta \ \frac{\sum \ x_i^2}{\sum \ x_i^2} \ + \ \frac{\sum \ x_i \ \epsilon_i}{\sum \ x_i^2} \ + \ \frac{\sum \ x_i \ \theta_i}{\sum \ x_i^2} \\ \widehat{\beta} &= \beta \ + \ \frac{\sum \ x_i \ \theta_i}{\sum \ x_i^2} \end{aligned} \tag{35}$$

Si $Cov(x_i, \theta_i) = 0$, alors l'estimateur $\hat{\beta}$ est sans biais puisque que par hypothèse $Cov(x_i, \epsilon_i) = 0$. De même il est convergent, mais il existe une perte d'efficacité car la variance de l'erreur est plus forte comparé au cas où la variable y_i est observée sans erreur. De même la variance de l'estimateur $\widehat{\beta}$ est plus forte lorsque y_i comprend des erreurs de mesures.

Dans la littérature, en présence d'outliers ou d'erreurs d'observations dans la variable dépendante, une méthode préconisée[3] est le *Least Absolute Deviations* (LAD). Pour mettre en évidence la pertinence de l'utilisation du LAD, nous proposons de simples simulations de Monte Carlo où nous déduisons l'erreur quadratique moyenne (MSE) des coefficients estimés $\widehat{\beta}$ lorsque la contamination de y porte sur une seule observation.

Résultat : Méthode robuste du LAD en présence d'outliers dans y
Générer des variables $\mathbf{X} \sim \mathcal{N}$, $\boldsymbol{\varepsilon} \sim \mathcal{N}$;
Déduire la variable $y = \mathbf{1}\alpha + \beta\mathbf{x} + \boldsymbol{\varepsilon}$, en fixant $\beta = 10$;
$\theta = 50$ [θ est la valeur de l'erreur d'observation], $n = 1000$ et ;
$i = 1$ [i est le nombre d'itérations];
répéter
  Déduire la variable y_l en introduisant l'outlier uniquement dans la ligne l de y : $y_l^* = y_l + \theta$;
  Calculer et récupérer le coefficient $\widehat{\beta}$ issu de de la regression $y/\mathbf{x}$ par les méthodes MCO, Gini et LAD;
jusqu'à $i = 1000$ [par pas de 1];
retourner Mean squared Errors (MSE) du coefficient β selon les trois méthodes ;

Algorithme 1 : Simulations de Monte Carlo

Les résultats sont les suivants.

MSE MCO	MSE Gini	MSE LAD
79640.81	78515.66	48054.62

TABLE 1 – Comparaison des Mean Squared Error (MSE)

Ce tableau confirme à travers le calcul des MSE que la méthode LAD se démarque des autres méthodes lorsqu'il existe des erreurs de mesure au niveau de la variable dépendante. Pour illustrer ces résultats, prenons un échantillon de dix observations et introduisons arbitrairement à la dixième observation un outlier dans de y.

Ce graphique illustre la robustesse de l'estimateur LAD mais aussi le fort impact que la présence d'un outlier dans la variable y peut engendrer sur l'estimateur de la regression Gini. En effet soit les modèles centrés sans

3. A titre d'exemple, voir Dodge (1997).

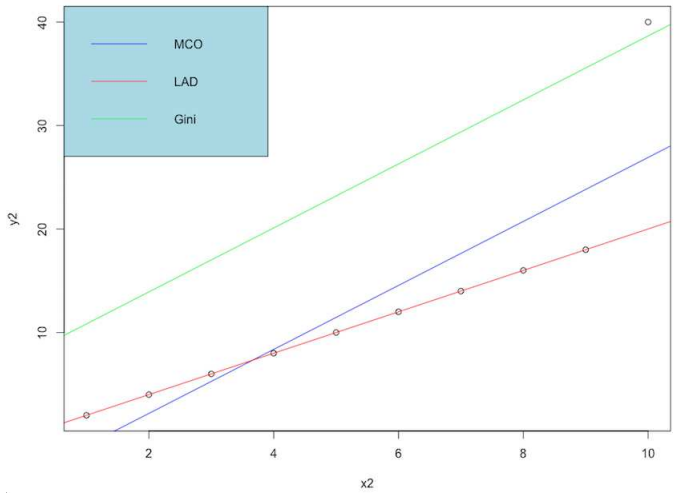

FIGURE 2 –
Illustration graphique de la présence d'outlier dans $\mathbf{y}$

erreur et contaminé suivants où θ est un vecteur de même taille que y :

$$\begin{aligned} y &= \beta_G\, \mathbf{x} + \boldsymbol{\varepsilon} \\ y^* &= y + \theta \end{aligned}$$

L'estimateur non paramétrique de la régression Gini contaminée est :

$$\begin{aligned} \widehat{\beta^{\theta}}_G &= \frac{Cov(y + \theta, \mathbf{r_x})}{Cov(\mathbf{x}, \mathbf{r_x})} \\ \widehat{\beta^{\theta}}_G &= \frac{Cov(y, \mathbf{r_x}) + Cov(\theta, \mathbf{r_x})}{Cov(\mathbf{x}, \mathbf{r_x})} \\ \widehat{\beta^{\theta}}_G &= \frac{Cov(y, \mathbf{r_x})}{Cov(\mathbf{x}, \mathbf{r_x})} + \frac{Cov(\theta, \mathbf{r_x})}{Cov(\mathbf{x}, \mathbf{r_x})} \\ \widehat{\beta^{\theta}}_G &= \widehat{\beta}_G + \frac{Cov(\theta, \mathbf{r_x})}{Cov(\mathbf{x}, \mathbf{r_x})} \end{aligned}$$

L'estimateur est sans biais si, et seulement si, $Cov(\theta, \mathbf{r_x}) = 0$. Lorsque $\widehat{\beta}_G$ est positif il peut donc être biaisé par le bas ou par le haut :

$$\begin{cases} \widehat{\beta^{\theta}}_G > \widehat{\beta}_G & \text{ssi } Cov(\theta, \mathbf{r_x}) > 0 \\ \widehat{\beta^{\theta}}_G < \widehat{\beta}_G & \text{ssi } Cov(\theta, \mathbf{r_x}) < 0 \end{cases}$$

4.2 Erreurs sur les variables explicatives

Soit le modèle avec les erreurs suivantes :

$$\begin{aligned}
y_i &= \beta\ x_i + \ \epsilon_i \qquad \text{(avec } x_i \text{ la } i \text{ ème valeur de } \mathbf{x} \text{ correctement observée)} \\
x_i^* &= x_i\ +\ \theta_i \qquad \text{(la } i \text{ ème observation contaminée de } \mathbf{x}) \\
y_i &= \beta\ x_i^* + \epsilon_i \qquad \text{avec } \mathbb{E}(x_i^*\ \epsilon_i) = 0 \\
y_i &= \beta\ (x_i + \theta_i) + \epsilon_i \\
y_i &= \beta\ x_i + \epsilon_i + \beta\ \theta_i \\
y_i &= \beta\ x_i + \epsilon_i^* \ \text{ avec } \epsilon_i^* = (\epsilon_i + \beta\ \theta_i) \qquad (36)
\end{aligned}$$

Nous faisons les hypothèses usuelles suivantes $\theta_i \rightsquigarrow N(0, \sigma_\theta)$, $\mathbb{E}(\theta_i\ x_i^*) = 0$, et $\mathbb{E}(\theta_i\ \epsilon_i) = 0$. Dans l'expression de y_i (ligne 3), nous avons : $\mathbb{E}(x_i^*\ \epsilon_i) = 0$ mais $\mathbb{E}(x_i\ \epsilon_i^*) \neq 0$.

Preuve

$$\begin{aligned}
\mathbb{E}(x_i\ \epsilon_i^*) &= \mathbb{E}(\ (x_i^*\ -\ \theta_i)(\epsilon_i + \beta\ \theta_i)\ ) \\
&= \mathbb{E}(\ x_i^*\epsilon_i\ +\ \beta\theta_i x_i^*\ -\ \theta_i\epsilon_i\ -\ \beta\ \theta_i^2\ ) \\
&= \mathbb{E}(\ -\beta\ \theta_i^2\ ) \\
&= -\beta\ \mathbb{E}(\ \theta_i^2\ ) \\
\mathbb{E}(x_i\ \epsilon_i^*) &= -\beta\ \sigma_\theta^2 \qquad (37)
\end{aligned}$$

La variable explicative sans erreur de mesure est par conséquent corrélée avec le terme d'erreur, car rappelons que le modèle estimée est : $y_i = \beta\ x_i + \epsilon_i^*$. Par conséquent, nous avons les implications de la corrélation entre la variable explicative et le terme d'erreur (**H5**). L'estimateur ($\hat{\beta}$) du modèle est biaisé et non convergent. En effet, nous avons :

$$\begin{aligned}
\widehat{\beta} &= \frac{\sum\ x_i\ y_i}{\sum\ x_i^2} \\
&= \frac{\sum\ x_i\ (\beta\ x_i + \epsilon_i^*)}{\sum\ x_i^2} \\
&= \frac{\sum\ \beta\ x_i^2\ +\ x_i\epsilon_i^*}{\sum\ x_i^2} \\
&= \beta\ +\ \frac{\sum\ x_i\epsilon_i^*}{\sum\ x_i^2} \\
&= \beta\ +\ \frac{\sum\ (x_i^* - \theta_i)\ (\epsilon_i + \beta\ \theta_i)}{\sum\ x_i^2} \\
&= \beta\ +\ \frac{\sum\ (x_i^*\ \epsilon_i)}{\sum\ x_i^2}\ + \beta\frac{\sum\ (x_i^*\ \theta_i)}{\sum\ x_i^2}\ -\ \frac{\sum\ (\theta_i\ \epsilon_i)}{\sum\ x_i^2} - \beta\ \frac{\sum\ \theta_i^2}{\sum\ x_i^2} \\
\widehat{\beta} &= \beta\ -\beta\ \frac{\sum\ \theta_i^2}{\sum\ x_i^2} \qquad (38)
\end{aligned}$$

La limite en probabilité de $\widehat{\beta}$ est donc :

$$\text{plim}(\hat{\beta}) = \beta \; - \; \beta \; \frac{s^2_{(\theta_i)}}{s^2_{(x_i)}} \tag{39}$$

Avec respectivement $s^2_{(\theta_i)}$ et $s^2_{(x_i)}$ les variances estimées de θ_i et x_i. L'estimateur $\hat{\beta}$ est non convergent et biaisé négativement. Montrons par simulation de Monte Carlo la robustesse de la regression Gini, lorsqu'il existe une erreur de mesure dans une seule observation de la variable explicative.

Résultat : Robustesse du Gini en présence d'erreur de mesure dans $\mathbf{x}$

Générer des variables $\mathbf{X} \sim \mathcal{N}$, $\boldsymbol{\varepsilon} \sim \mathcal{N}$;
Déduire la variable $y = \alpha + \beta\mathbf{x} + \boldsymbol{\varepsilon}$, en fixant $\beta = 10$;
$\theta = 50$ [θ est la valeur de l'erreur d'observation], $n = 1000$ et ;
$i = 1$ [i est le nombre d'itérations];
répéter
  Générer une variable $\mathbf{x}$ en introduisant l'outlier uniquement dans la ligne l de $\mathbf{x}$: $x_l = x^*_l + \theta$;
  Calculer le vecteur rang de $\mathbf{x}$;
  Calculer et récupérer le coefficient $\widehat{\beta}$ issu de de la regression $y/\mathbf{x}$ par les méthodes des MCO, Gini et LAD;
jusqu'à $i = 1000$ [par pas de 1];
retourner Mean squared Errors (MSE) du coefficient $\boldsymbol{\beta}$ selon les trois méthodes ;

Algorithme 2 : Simulations de Monte Carlo

Les résultats sont les suivants.

MSE MCO	MSE Gini	MSE LAD
66.73	12.11	39.24

TABLE 2 – Comparaison des Mean Squared Error (MSE)

Le tableau montre que l'estimateur de la régression Gini non paramétrique a le plus faible MSE et donc semble être une méthode appropriée pour les erreurs de mesures en $\mathbf{x}$. Pour illustrer ce résultat, générons une droite de régression avec erreur de mesure à la dixième et dernière observation.

Ce graphique illustre la robustesse de la méthode Gini grâce à l'utilisation du vecteur rang. Supposons que ce dernier, $\mathbf{r_x}$, reste inchangé après l'introduction d'un vecteur d'outliers θ qui contamine l'ensemble des valeurs de $\mathbf{x}$ tel que $\mathbf{x}^* = \mathbf{x} + \theta$, alors l'estimateur contaminé de β_G est :

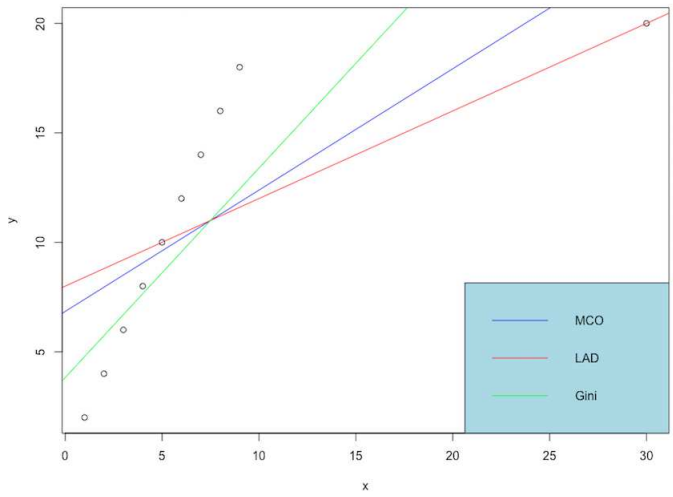

FIGURE 3 –
Illustration graphique de la présence d'outlier dans $\mathbf{x}$

$$\begin{aligned}
\widehat{\beta^{\theta}}_G &= \frac{Cov(y, \mathbf{r_x})}{Cov(\mathbf{x}^*, \mathbf{r_x})} \\
\widehat{\beta^{\theta}}_G &= \frac{Cov(y, \mathbf{r_x})}{Cov(\mathbf{x} + \theta, \mathbf{r_x})} \\
\widehat{\beta^{\theta}}_G &= \frac{Cov(y, \mathbf{r_x})}{Cov(\mathbf{x}, \mathbf{r_x}) + Cov(\theta, \mathbf{r_x})}
\end{aligned}$$

Puisque,

$$\begin{aligned}
\text{plim} \left[\frac{1}{n} \, y' \mathbf{r_x} \right] &= Cov(y, \mathbf{r_x}) \\
\text{plim} \left[\frac{1}{n} \, \mathbf{x}' \mathbf{r_x} \right] &= Cov(\mathbf{x}, \mathbf{r_x}) \\
\text{plim} \left[\frac{1}{n} \, \theta' \mathbf{r_x} \right] &= Cov(\theta, \mathbf{r_x})
\end{aligned}$$

alors :

$$\text{plim} \; \widehat{\beta^{\theta}}_G = \frac{Cov(y, \mathbf{r_x})}{Cov(\mathbf{x}, \mathbf{r_x}) + Cov(\theta, \mathbf{r_x})}$$

Par conséquent :
Si $Cov(\theta, \mathbf{r_x}) > 0 \Rightarrow \widehat{\beta^{\theta}}_G < \widehat{\beta}_G$.
Si $Cov(\theta, \mathbf{r_x}) < 0 \Rightarrow \widehat{\beta^{\theta}}_G > \widehat{\beta}_G$.

L'erreur d'observation dans la variable explicative $\mathbf{x}$ entraîne une contamination de l'estimateur du Gini caractérisé par le terme $Cov(\theta, \mathbf{r_x})$. Dans le cadre de la regression par MCO, la contamination provient du terme $Cov(\theta, \mathbf{x})$. Par conséquent, si $Cov(\theta, \mathbf{x}) > Cov(\theta, \mathbf{r_x})$, l'impact de l'erreur de mesure dans $\mathbf{x}$ est plus faible avec la regression Gini.
Pour identifier ce problème, un test d'exogénéité de type Hausman (1978) peut permettre de détecter un terme d'erreur attaché à une (ou plusieurs) variables explicatives. En cas de violation de l'hypothèse **H5**, le recourt à la méthode par variables instrumentales s'impose.

4.3 La méthode des variables instrumentales

Deux approches sont développées dans cette scetion, celle de la méthode des variables instrumentales par MCO (VI-MCO) et celle des variables instrumentales avec le rang des instruments (Gin-VI). L'avantage d'utiliser le rang est la possibilité d'obtenir de plus amples informations sur la relation qui lie les données, comme celle de savoir si la relation entre la variable dépendante et les variables explicatives ou celle entre les variables explicatives elles-mêmes est monotone ou non. Soit le modèle général :

$$\mathbf{y} = \mathbf{X}\,\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

tel que plim$[\mathbf{X}'\,\boldsymbol{\epsilon}/N] \neq 0$. La technique des variables instrumentales par MCO bien connue consiste à trouver des instruments à la fois corrélées aux variables explicatives et non corrélées au terme d'erreur. Nous considérons dans ce qui suit que le nombre de variables instrumentales est égal à celui des variables explicatives. Les principes de base de cette régression sont les suivants :
1) Asymptotiquement, il n'y a pas de corrélation entre les instruments et le terme d'erreur : pLim$(\mathbf{Z}'\,\boldsymbol{\epsilon}/n) = 0$
2) Il existe une forte corrélation entre les variables instrumentales $\mathbf{Z}$ et les variables explicatives $\mathbf{X}$: plim$(\mathbf{Z'X}/n) = \tau$.
3) Les variables instrumentales $\mathbf{Z}$ ne sont pas colinéaires : plim$(\mathbf{Z'Z}/n) = \tau^*$.

Rappelons que l'on peut déduire les estimateurs VI-MCO de deux manières différentes : une application directe d'une matrice d'instruments ; ou l'utilisation des moindres carrés en deux étapes. Pour la méthodologie Gini-VI, les deux méthodes peuvent donner des estimateurs totalement différents. Nous présentons dans un premier temps, les deux méthodes relatives aux MCO puis celles relatives au Gini.

L'estimateur VI-MCO est celui de l'estimateur MCO du modèle transformé $\mathbf{Z'y} = \mathbf{Z'X}\,\boldsymbol{\beta} + \mathbf{Z'}\boldsymbol{\varepsilon}$:

$$\hat{\beta}_{VI-MCO1} = (\mathbf{X'Z}\;\mathbf{Z'X})^{-1}\;\mathbf{X'Z}\;\mathbf{Z'y}$$

or $\mathbf{y} = \mathbf{X}\,\boldsymbol{\beta} + \boldsymbol{\epsilon}$, alors :

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{VI-MCO1} &= (\mathbf{X}'\mathbf{Z}\ \mathbf{Z}'\mathbf{X})^{-1}\ \mathbf{X}'\mathbf{Z}\ \mathbf{Z}'(\mathbf{X}\,\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\
&= (\mathbf{X}'\mathbf{Z}\ \mathbf{Z}'\mathbf{X})^{-1}\ (\mathbf{X}'\mathbf{Z}\ \mathbf{Z}'\mathbf{X})\ \boldsymbol{\beta} + (\mathbf{X}'\mathbf{Z}\ \mathbf{Z}'\mathbf{X})^{-1}\ \mathbf{X}'\mathbf{Z}\ \mathbf{Z}'\boldsymbol{\epsilon} \\
\hat{\boldsymbol{\beta}}_{VI-MCO1} &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{Z}\ \mathbf{Z}'\mathbf{X})^{-1}\ \mathbf{X}'\mathbf{Z}\ \mathbf{Z}'\boldsymbol{\epsilon}
\end{aligned}$$

La limite en probabilité de l'estimateur $\hat{\boldsymbol{\beta}}_{VI-MCO1}$ est $\boldsymbol{\beta}$ car $\text{plim}(\mathbf{Z}'\boldsymbol{\epsilon}/n) = 0$. L'estimateur est sans biais. L'expression simplifiée de l'estimateur VI-MCO est, lorsque le nombre d'instruments est égal au nombre de variables explicatives ($\mathbf{Z}'\mathbf{X}$ est une matrice carrée) :

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{VI-MCO1} &= (\mathbf{X}'\mathbf{Z}\ \mathbf{Z}'\mathbf{X})^{-1}\ \mathbf{X}'\mathbf{Z}\ \mathbf{Z}'\mathbf{y} \\
&= (\mathbf{Z}'\mathbf{X})^{-1}\ \overbrace{(\mathbf{X}'\mathbf{Z})^{-1}\ \mathbf{X}'\mathbf{Z}}^{=Id}\ \mathbf{Z}'\mathbf{y} \\
\hat{\boldsymbol{\beta}}_{VI-MCO1} &= (\mathbf{Z}'\mathbf{X})^{-1}\ \mathbf{Z}'\mathbf{y}
\end{aligned} \tag{40}$$

L'estimateur VI en deux étapes consiste d'abord à faire un modèle de type $\mathbf{X} = \mathbf{Z}\pi + \mathbf{r}$ où on estime π afin d'estimer $\hat{\mathbf{X}}$ que l'on utilise dans la deuxième étape qui consiste à régresser $\mathbf{y}$ sur $\widehat{\mathbf{X}}$. On retrouve ainsi $\hat{\beta}_{VI-MCO1} = (\mathbf{Z}'\mathbf{X})^{-1}\ \mathbf{Z}'\mathbf{y}$.

En 2004, Yitzhaki et Schechtman proposent une régression Gini avec variables instrumentales communément appelée Gini-VI. Comme dans le cas des MCO, deux méthodes permettent d'obtenir deux estimateurs, mais nous verrons qu'ils ne sont pas nécessairement identiques. La méthode directe Gini-VI est :

$$\hat{\beta}_{Gini-VI1} = (\mathbf{R}'_{\mathbf{Z}}\ \mathbf{X})^{-1}\ \mathbf{R}'_{\mathbf{Z}}\ \mathbf{y} \tag{41}$$

oú $\mathbf{R}_{\mathbf{Z}}$ est la matrice des rangs des variables instrumentales. La procédure en deux étapes se fait de la manière suivante. Premièrement :

$$\hat{\mathbf{X}} = \mathbf{Z}(\mathbf{R}'_{\mathbf{Z}})^{-1}\mathbf{R}'_{\mathbf{Z}}\mathbf{X} \tag{42}$$

Dans la seconde étape :

$$\begin{aligned}
\hat{\beta}_{Gini-VI2} &= (\mathbf{R}'_{\hat{\mathbf{X}}}\ \hat{\mathbf{X}})^{-1}\ \mathbf{R}'_{\hat{\mathbf{X}}}\mathbf{y} \\
&= [\ (\mathbf{R}'_{\hat{\mathbf{X}}}\mathbf{Z})\ (\mathbf{R}'_{\mathbf{Z}}\mathbf{Z})^{-1}\ \mathbf{R}'_{\mathbf{Z}}\mathbf{X}\ ]^{-1}\ \mathbf{R}'_{\hat{X}}\mathbf{y} \\
\hat{\beta}_{Gini-VI2} &= (\mathbf{R}'_{\mathbf{Z}}\mathbf{X})^{-1}\ \mathbf{R}'_{\mathbf{Z}}\mathbf{Z}\ (\mathbf{R}'_{\hat{\mathbf{X}}}\mathbf{Z})^{-1}\ \mathbf{R}'_{\hat{\mathbf{X}}}\mathbf{y}
\end{aligned}$$

Nous constatons que dans la méthode directe, nous avons la matrice $\mathbf{R}$ qui est la matrice des rangs des variables instrumentales ($\mathbf{R}_{\mathbf{Z}}$), tandis que dans la méthode en deux étapes, deux matrices $\mathbf{R}$ sont utilisées, celle des rangs de $\hat{\mathbf{X}}$ et celle de $\mathbf{Z}$. Par conséquent, à moins que les rangs de tous les instruments et de la variable explicative initiale soient identiques, nous devrions nous

attendre à des résultats différents. Cela peut s'expliquer par le fait que la fonction de répartition n'est pas en général une transformation linéaire de la variable explicative. Et donc, tout se passe comme si nous espérions avoir une relation linéaire malgré que nous utilisons une transformation non linéaire.

Il est comparé dans l'article de Yitzhaki et Schechtman (2004), les propriétés de la méthode d'estimation des variables instrumentales sous deux paramètres alternatifs (VI-MCO et Gini-VI). Le premier paramètre, dit paramètre standard (VI-MCO) est basé sur la minimisation d'une fonction quadratique de l'erreur, comme dans le cas de la régression par les MCO. Le second estimateur (Gini-VI) quant à lui, est une méthode d'estimation des variables instrumentales par le bias du GMD comme méthode de régression. Ainsi les coefficients de régression de ces différentes approches de la méthode des variables instrumentales peuvent être interprétés comme des moyennes pondérées des pentes entre les observations adjacentes. Par conséquent, face aux limites de la variance et de la régression par les MCO, il apparait que l'estimateur de la méthode des variables instrumentales issu de la régression Gini est moins sensible aux outliers et à la violation de l'hypothèse de linéarité que l'estimateur standard de la méthode des variables instrumentales.
Il est à noter que si le modèle utilisé est linéaire et que les hypothèses communément utilisées sont respectées, alors les deux méthodes produisent les mêmes estimateurs.

5 Inférence statistique

Le problème posé est de savoir quelle est la loi suivie par les estimateurs issus de la regression Gini $\widehat{\beta}_G$ et de déterminer les propriétés de ces derniers, notamment la convergence. Yitzhaki et Schechtman (2013) montrent que les estimateurs issus de la régression Gini sont des U-statistiques. La théorie des U-statistiques est une théorie d'échantillonnage qui permet de déterminer, sous des conditions peu exigeantes, que l'estimateur est sans biais et convergent (Hoeffding, 1948). Cette théorie permet par ailleurs de déterminer la variance des estimateurs et d'effectuer des tests statistiques de signification asymptotiques sur ces derniers.

Si la statistique dont on cherche l'estimateur s'écrit :

$$\theta(F) = \mathbb{E}[\phi(X_1, \cdots, X_m)] = \int \dots \int \phi(x_1, \dots, x_m) dF(x_1) \cdots dF(x_m),$$

où $\phi(x_1, \dots, x_m)$ est appelé le noyau de $\theta(F)$ de degré m. Alors l'estimateur U (sans biais) de θ existe tel que $U \sim \mathcal{N}$:

$$U_n := \binom{n}{m}^{-1} \sum_c \phi(X_{i_1}, X_{i_2}, \dots, X_{i_m})$$

où $\sum_c$ indique la somme pour toutes les combinaisons de m éléments $\{i_1, \cdots, i_m\}$ de $\{1, \cdots, n\}$.

Les conditions qui garantissent la convergence de U et l'absence de biais sont les suivantes :

- $\phi(x_1, \ldots, x_m)$ doit être une fonction symétrique ;
- les variables aléatoires X_i doivent être *i.i.d.* ;
- la taille de l'échantillon doit être suffisamment large.

Prenons quelques exemples usuels.

Exemple 5.1 *Soit $\theta(F_X) = \mathbb{E}(X)$. Le noyau est $\phi(X) = X$, il est symétrique de degré $m = 1$. Alors la U-statistique est :*

$$U_1 = \frac{1}{\binom{n}{1}} \sum_i x_i = \frac{1}{\frac{n!}{1!\,(n-1)!}} \sum_i x_i = \frac{1}{\frac{n\,(n-1)!}{1\,(n-1)!}} \sum_i x_i = \frac{1}{n} \sum_i x_i = \bar{x}$$

Exemple 5.2 *Soit $\theta(F_X) = \mathbb{E}(X_1 - X_2)^2$. Le noyau est $\phi(X) = \frac{1}{2}(X_1 - X_2)^2$, il est symétrique de degré $m = 2$. La U-statistique est donc :*

$$\begin{aligned} U &= \frac{1}{\binom{n}{2}} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{(x_i - x_j)^2}{2} \\ &= \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \end{aligned}$$

Revenons à la régression Gini linéaire simple. Nous avons vu que :

$$\begin{aligned} \beta_G &= \frac{Cov(\, y, F_{\mathbf{x}}\,)}{Cov(\, \mathbf{x}, F_{\mathbf{x}}\,)} \\ &= \frac{GCov(\, y, \mathbf{x}\,)}{GCov(\, \mathbf{x}, \mathbf{x}\,)} \end{aligned}$$

Proposition 5.3 – Yitzhaki et Schechtman (2013, Chapitre 9) – *: Soient (X_1, Y_1) et (X_2, Y_2) de dimension 2 d'une distribution bi-variée continue dont les deux premiers moments sont finis. Soit $h((X_1, Y_1), (X_2, Y_2)) = (X_1, Y_1)\mathbb{1}_{(Y_1 > Y_2)} + (X_2, Y_2)\mathbb{1}_{(Y_1 > Y_2)})$ où $\mathbb{1}_{a>b}$ est défini comme :*

$$\mathbb{1}_{a>b} = \begin{cases} 1 & si\ a > b \\ 0 & sinon. \end{cases}$$

Alors $h((X_1, Y_1), (X_2, Y_2))$ est un noyau symétrique de degré (2,2) pour $\Delta_{X,Y} = 4Cov(X, F(Y)) = 4GCov(X, Y)$. [4]

4. Un degré (2,2) signifie qu'on a besoin de deux X indépendants et deux Y indépendants afin d'obtenir un estimateur non biaisé.

Preuve:
Voir Yitzhaki et Schechtman (2013, Chapitre 9, p. 203-204). ■

En utilisant le noyau ci-dessus, la U-statistique est

$$\begin{aligned} U(\Delta_{X,Y}) &= \frac{1}{\binom{n}{2}} \sum_{i<j} \sum h((x_i,y_i),(x_j,y_j)) \\ &= \frac{1}{\binom{n}{2}} \sum_{i<j} \sum [(x_i - x_j)\mathbb{1}_{y_i>y_j} + (x_j - x_j)\mathbb{1}_{y_i>y_j}] \end{aligned}$$

Il s'agit d'une U-statistique pour $4Cov(X, F_Y)$ et par conséquent un estimateur sans biais et convergent. Une définition alternative, basée sur une combinaison linéaire des éléments concomitants des statistiques d'ordre, est donnée par :

$$U(\Delta_{X,Y}) = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n} (2i - 1 - n)\, x_{y(i)}$$

où $x_{y(i)}$ *et* $x_{(i)}$ sont des statistiques d'ordre avec $x_{y(i)}$ la valeur de x qui correspond à la ième statistique d'ordre de $y_1, \cdots, y_n$.

Nous avons vu que :

$$\beta_G = \frac{Cov(\, y, F_{\mathbf{x}}\,)}{Cov(\, \mathbf{x}, F_{\mathbf{x}}\,)}$$

Étant donné que $Cov(\mathbf{y}, F_{\mathbf{x}})$ et $Cov(\mathbf{x}, F_{\mathbf{x}})$ sont des fonctions symétriques, alors il existe une U-statistique, telle que les estimateurs de $Cov(\mathbf{y}, F_{\mathbf{x}})$ et $Cov(\mathbf{x}, F_{\mathbf{x}})$ sont sans biais et respectivement donnés par :

$$U_1 = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n} (2i - 1 - n)\, x_{y(i)} \tag{43}$$

$$U_2 = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n} (2i - 1 - n)\, x_{(i)}, \tag{44}$$

Théorème 5.1 *Pour U_1 et U_2 les U-statistiques des estimateurs respectifs de la $Cov(\mathbf{y}, F_{\mathbf{x}})$ et de la $Cov(\mathbf{x}, F_{\mathbf{x}})$, l'estimateur $\widehat{\beta}_G$ du paramètre β_G est une U-statistique tel que $\widehat{\beta}_G = \frac{U_1}{U_2} \overset{a}{\sim} \mathcal{N}$.*

Preuve:
Soit :

$$\widehat{\beta}_G = \frac{Cov(y, \mathbf{r}_{\mathbf{x}})}{Cov(\mathbf{x}, \mathbf{r}_{\mathbf{x}})}$$

Des équations (43) et (44), l'estimateur $\widehat{\beta}_G$ du paramètre β_G s'écrit :

$$\widehat{\beta}_G = \frac{U_1}{U_2}$$

L'estimateur $\widehat{\beta}_G$ étant un ratio de deux U-statistiques, alors $\widehat{\beta}_G$ est aussi une U-statistique. Du théorème 10.4 de Yitzhaki et Schechtman (2013), nous avons : Soit $(U') = U_1, \cdots, U_t$ t U-statistiques basées sur un échantillon $x_1, \cdots, x_n$ de taille n avec U_i correspondant à θ_i (avec un noyau h_i), $i = 1, \cdots, t$. Si la fonction $g(y) = g(y_1, \cdots, y_t)$ qui ne comporte pas n et est continue avec ses dérivées partielles à certains voisinages du point $(y) = (\theta) = (\ \theta_1, \cdots, \theta_t)$ et sous la condition que $\mathbb{E}[h_i^2(X_1, X_2, \cdots, X_{m_i})] < \infty$ alors l'estimateur $\widehat{\beta}_G$ tend vers une loi normale lorsque $n \longrightarrow \infty$. ■

L'expression de la variance d'une U-statistique est très complexe et d'utilisation pratique assez délicate dès lors que m est supérieur à 1. Une alternative est l'utilisation de la méthode de Jackknife qui permet de réduire le biais. De manière générale, l'estimateur de la variance par Jackknife d'un estimateur U est donné par :

$$\hat{V}(\hat{\beta}_G) = \frac{n-1}{n} \sum_{i=1}^{n} \left[U_{-i} - \frac{1}{n} \sum_{i=1}^{n} U_{-i} \right]^2$$

avec U_{-i} l'estimateur d'une U-statistique sur un échantillon de taille n sans la ième observation. De cette manière, la statistique de test pour le coefficient de la pente d'une régression Gini est déduit du fait que $\hat{\beta}_G \overset{a}{\sim} \mathcal{N}(\ \beta_G, \widehat{V}(\hat{\beta}_G)\ )$.

6 Conclusion

Cet article a pour objectif l'enrichissement de la littérature sur le Gini's Mean Difference (GMD). Il met d'abord en lumière les limites de la régression des moindres carrés ordinaires (MCO) en présence d'outliers dans les observations. De même en cas d'endogénéité de l'une des variables explicatives ou d'erreur de mesure, la régression par les MCO s'avère ne pas être efficace. Ensuite, une nouvelle approche est développée, la regression Gini (Yitzhaki et Schechtman - 2013) en vue de remédier à cette sensibilité de la méthode des MCO et surtout de s'affranchir de certaines hypothèses de base en économétrie. Cette approche avec en son coeur l'opérateur Cogini peut être perçue comme un mélange de la méthode de pur rang de Spearman (métrique ℓ_1) et la méthode de la variance (métrique ℓ_2).
En présence des points atypiques qui limitent l'utilisation des MCO, la régression Gini produit des estimateurs plus robustes que les estimateurs des MCO. De plus, les estimateurs issus de la régression Gini sont des U-statistiques qui autorisent l'inférence de différents estimateurs, permettant ainsi de tester la qualité des modèles étudiés.

Références bibliographiques

Alan Stuart. 1954. “ Limit Distribution For Total Rank ", *The British Psychological Society*, vol. 7, Issue 1, pp. 50 - 51

Corrado Gini. 1912. – “ Variabilità e Mutuabilità. Contributo allo Studio delle Distribuzioni e delle Relazioni Statistiche ", C. Cuppini, Bologna.

Haim Shalit and Sholmo Yitzhaki. 1984. “ Mean-Gini, portfolio Theory, and the Pricing of Risy Assets ", *The Journal of Finance*. vol. XXXIX, n. 5.

Hideo Tanaka and Satoru Uejima. 1982. " Linear regression analysis with fuzzy Model ", *IEEE Transaction on systems, man, and cybenetics*, vol. SMC-12 n. 6.

Kannan Senthamarai and Kuppusamy Manoj. 2015. “ Outlier Detection in Multivariate Data ", *Applied Mathematical Sciences*, vol. 9, n. 47, pp. 2317 - 2324.

Marcel Carcea and Robert Serfling. 2015. “ A Gini autocovariance function for time series modeling ", *Journal of Time Series Analysis*, 36, 817-838.

Marie Davidian and Raymond Carroll. 1987. "Variance function estimation", *Journal of the American Statistical Association*, 82, 400, (December), 1079-1091.

Ndéné Ka and Stéphane Mussard. 2015. “ l1 regressions : Gini estimators for fixed effects panel data ", *Journal of Applied Statistics*, , vol. 46, n.8, pp. 1436 - 1446.

Roger Koenker and Gilbert Bassett. 1978. “ Regression quantiles ", *Econometrica*, vol. 46, pp. 33 - 50.

Sholmo Yitzhaki. 2003. " Gini's Mean difference : a superior measure of variability for non-normal distributions ", *METRON - International Journal of Statistics*, vol. LXI, n. 2, pp. 285 - 316.

Shlomo Yitzhaki & **Edna Schechtman.** 2004. “ The Gini Instrumental Variable, or the "double instrumental variable estimator ", *METRON - International Journal of Statistics*, vol. LXII, n. 3, pp. 287 - 313.

Shlomo Yitzhaki & **Edna Schechtman.** 2013. “ The Gini Methodology : A Primer on a Statistical Methodology ", New York : Springer.

Wassily Hoeffding. 1984. " A class of Statistics with Asymptotically Normal Distribution ", *The Annals of Mathematical Statistics*, vol. 19 n. 3, pp. 293 - 325.