GREDI
Groupe de Recherche en Économie
et Développement International

Cahier de Recherche / Working Paper
15-02

ℓ_1 Regressions:
Gini Estimators for Fixed Effects Panel Data

Ndéné Ka
&
Stéphane Mussard

UNIVERSITÉ DE
SHERBROOKE

ℓ_1 Regressions: Gini Estimators for Fixed Effects Panel Data*

Ndéné Ka†
Lameta
Université Montpellier I

Stéphane Mussard‡
Lameta
Université Montpellier I

October 7, 2014

Abstract

Panel data, frequently employed in empirical investigations, provide estimators being strongly biased in the presence of atypical observations. The aim of this work is to propose a ℓ_1 Gini regression for panel data. It is shown that the fixed effects within-group Gini estimator is more robust than the OLS one when the data are contaminated by outliers. This semi-parametric Gini estimator is proven to be an U-statistics, consequently, it is asymptotically normal.

Keywords: Gini, Panel, Regression, U-statistics.

*The authors are greatly indebted to Shlomo Yitzhaki for very helpful comments and advices. They also acknowledge Benoît Mulkay for stimulating discussions. The usual disclaimer applies.

† Université Montpellier 1, UMR5474 LAMETA, F-34000 Montpellier, France, Faculté d'Economie, Av. Raymond Dugrand, Site de Richter C.S. 79606, 34960 Montpellier Cedex 2. E-mail: ka@lameta.univ-montp1.fr.

‡Université Montpellier 1, UMR5474 LAMETA, F-34000 Montpellier, France, Faculté d'Economie, Av. Raymond Dugrand, Site de Richter C.S. 79606, 34960 Montpellier Cedex 2,. Tel: 33 (0)4 67 15 83 82 / Fax : 33 (0)4 67 15 84 67 - e-mail: smussard@adm.usherbrooke.ca, Research Fellow at Grédi, Université de Sherbrooke and Ceps Luxembourg.

1 Introduction

Econometrics has devoted an important line of research to ℓ_1 regressions. The seminal work of Olkin and Yitzhaki (1992) has paved the way on a general ℓ_1 regression, the so-called Gini regression, embracing many other common and well-known target functions such as the Least Absolute Deviation (LAD) and the absolute deviation from a quantile, see Koenker and Bassett (1978).[1] LAD is actually regarded as a partial regression technique since it represents only one component of the Gini variability to be minimized: the between-group variability of the Gini index of the residuals (see Yitzhaki and Lambert, 2013).The Gini regression has been initiated with respect to two non-exclusive approaches. The first one, the parametric Gini regression, aims at determining (numerically) the coefficient estimates by minimization of the Gini index of the residuals. The second one, the semi-parametric Gini regression, offers estimates on the basis of averaging slope coefficients. Those Gini regressions are coincident if the linearity of the model is assessed. They also share the common property of being robust to outliers, that is, when data are contaminated by extreme values or more generally when the underlying distribution deviates from the multivariate normal – see Yitzhaki and Schechtman (2013) for an overview of the Gini methodology.[2]

Most of empirical findings are nowadays based on the use of panel or longitudinal data sets. Panel data benefits are: a much larger variability, less collinearity among the covariates (compared with cross-sectional data or time series), more degrees of freedom, more efficiency, and the ability to control for individual heterogeneity.

From our knowledge, Gini regressions are only available either for cross-sectional data or time series. In this note, a Gini regression for panel data is proposed. We pursue the idea that the employ of one particular variability is crucial to derive robust estimators. In panel data, the decomposition of the moment matrices into within- and between-group variability is known to produce within- and between-group fixed effects estimators. The fixed effects OLS estimators are very popular and convenient for empirical investigations, however outliers can drastically affect the estimates. Huber (1981) shows that only 3% of outliers in a set of observations are sufficient to change significantly the estimates (strongly biased in the presence of atypical observations). If outliers are removed, some part of the information in the sample is definitely lost.

[1] The Gini regression includes other regressions criteria based on the "city block" metric such as the mean absolute deviation (MAD). These different target functions also rely on the between-group Gini variability.

[2] It is also important to note that the OLS regression coefficients are very sensitive to monotonic transformation of the variables (Yitzhaki and Schechtman, 2013, Chapter 5). If the covariates are multinormal, the Gini estimates are close than those of OLS.

The aim of this note is to decompose the variability of the moment matrices into within- and between-group Gini variabilities in order to deduce a fixed effects semi-parametric Gini regression for panel data. We show that the within-group Gini estimator derived from this decomposition is a semi-parametric estimator. It is also an U-statistics, consequently, it is asymptotically normal.

The outline of the note is as follows. We begin with the standard Gini regression approaches for cross-sectional data (Section 2.1) before investigating the within-group Gini estimator for fixed effects panel data (Section 2.2) and its asymptotic properties (Section 2.3). A simple simulation illustrates the robustness of the within-group Gini estimator in the presence of outliers (Section 2.4). Section 3 closes the note.

2 Gini Regressions

2.1 The standard approaches for cross-sectional data

Consider a model $\mathbf{y} = a + b\mathbf{x}$ with $\mathbf{x}, \mathbf{y}$ some $N \times 1$ vectors. The semi-parametric Gini (simple) regression introduced by Olkin and Yitzhaki (1992), consists in averaging tangents b_{ij} (between observations i and j) with weights v_{ij}. Let the values of $\mathbf{x}$ and $\mathbf{y}$ be ranked by ascending order ($x_1 \leqslant \cdots \leqslant x_N$ and $y_1 \leqslant \cdots \leqslant y_N$), then the semi-parametric Gini estimator of the slope coefficient is given by:

$$\hat{b}^G = \sum_{i<j} v_{ij} b_{ij}, \text{ with } v_{ij} = \frac{(x_i - x_j)}{\sum_{i<j}(x_i - x_j)} \text{ and } b_{ij} = \frac{(y_i - y_j)}{(x_i - x_j)} \; \forall i < j \; ; \; i = 1, \ldots, N.$$

The authors also demonstrate that if the weights v_{ij} are replaced by quadratic ones such as $w_{ij} = \frac{(x_i - x_j)^2}{\sum_{i<j}(x_i - x_j)^2}$, then the standard OLS estimator of the slope coefficient is obtained: $\hat{b}^{OLS} = \sum_{i<j} w_{ij} b_{ij}$. Since it depends on quadratic weights, the OLS slope coefficient is shown to be heavily sensitive to outliers.

The parametric Gini regression (Olkin and Yitzhaki, 1992) solves the minimization of Gini index of the residuals ($e_i = y_i - \hat{y}_i$) and provides the following estimator (only numerically in the multiple regression case):

$$\hat{b}^{PG} = \arg\min_b G(\mathbf{e}) = \arg\min_b \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} |e_i - e_j| \, .$$

Based on all pairwise "city-block" distances, the parametric and non-parametric Gini regressions are equivalent ($\hat{b}^{PG} = \hat{b}^G$) if, and only if, the linearity of the model $\mathbf{y} = a\mathbf{x} + b$ is assessed. The semi-parametric Gini regression may be defined according to the cogini op-

erator[3], *i.e.* $\text{cog}(\mathbf{y},\mathbf{x}) := \text{cov}(\mathbf{y},\mathbf{r}(\mathbf{x}))$ and $\text{cog}(\mathbf{x},\mathbf{x}) := \text{cov}(\mathbf{x},\mathbf{r}(\mathbf{x}))$ where $\mathbf{r}(\mathbf{x})$ is the rank vector of $\mathbf{x}$:[4]

$$\hat{b}^G = \frac{\text{cog}(\mathbf{y},\mathbf{x})}{\text{cog}(\mathbf{x},\mathbf{x})} \ , \ \text{whereas} \ \ \hat{b}^{OLS} = \frac{\text{cov}(\mathbf{y},\mathbf{x})}{\text{cov}(\mathbf{x},\mathbf{x})}.$$

The semi-parametric Gini multiple regression depends on the rank matrix of the regressors. Let $\mathbf{X}$ be the $N \times K$ matrix of the regressors and $\mathbf{R_x}$ its rank matrix, which contains in columns the rank vectors $\mathbf{r}(\mathbf{x}_k)$ of the regressors $\mathbf{x}_k$ for all $k = 1, \ldots, K$. The semi-parametric Gini multiple regression yields the following estimator (a $K \times 1$ vector):

$$\hat{\mathbf{b}}^G = (\mathbf{R}'_\mathbf{x}\mathbf{X})^{-1}\mathbf{R}'_\mathbf{x}\mathbf{y} \ . \tag{0}$$

The Gini semi-parametric approach has the advantage of relying on a few assumptions, no linearity hypothesis is needed. The estimator $\hat{\mathbf{b}}^G$ is less sensitive to extreme values since it is built on the cogini matrices $\mathbf{R}'_\mathbf{x}\mathbf{X} =: \mathbf{G}^{total}_{xx}$ and $\mathbf{R}'_\mathbf{x}\mathbf{y} =: \mathbf{G}^{total}_{xy}$ rather than the moment matrices $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}'\mathbf{y}$ of the OLS estimator:

$$\hat{\mathbf{b}}^G = \mathbf{G}^{total}_{xx}\mathbf{G}^{total}_{xy} \ . \tag{1}$$

In the remainder, this technique is extended to fixed effects panel data.

2.2 Fixed effects panel data Gini estimators

Consider the simple formulation of the fixed effects linear panel data model:

$$y_{nt} = \beta_0 + \beta_n + \boldsymbol{\beta}'\mathbf{x}_{nt} + \varepsilon_{nt} \ , \tag{2}$$

where subscript n denotes the cross-section dimension ($n = 1, \ldots, N$) and where t denotes the time series dimension ($t = 1, \ldots, T$). The element y_{nt} of the $NT \times 1$ vector $\mathbf{y}$ represents the n-th observation at time t of the dependent variable, $\mathbf{x}_{nt}$ is the $K \times 1$ regressor vector of the n-th observation at time t, $\boldsymbol{\beta}' \in \mathbb{R}^K$ is a $1 \times K$ vector of the regression parameters, β_n the unobservable time-invariant individual fixed effect and β_0 the intercept. Finally, ε_{nt} denotes the disturbance term which is assumed to be uncorrelated through time and cross-sections. Averaging (2) over time and subtracting from (2) yields:

$$y_{nt} - y_{n.} = \boldsymbol{\beta}'(\mathbf{x}_{nt} - \mathbf{x}_{n.}) + \varepsilon_{nt} - \varepsilon_{n.} \ . \tag{3}$$

[3] Actually, it exists two coginis: $\text{cov}(\mathbf{y},\mathbf{r}(\mathbf{x}))$ and $\text{cov}(\mathbf{x},\mathbf{r}(\mathbf{y}))$. The cogini enables a new correlation statistics to be characterized, quite close to Pearson's coefficient, the G-correlation index $\Gamma = \text{cog}(\mathbf{y},\mathbf{x})/\text{cog}(\mathbf{y},\mathbf{y})$. It is bounded between $[-1, 1]$, it is insensitive to monotonic transformation of $\mathbf{x}$ and to linear transformation of $\mathbf{y}$, and it is nil if and only if $\mathbf{x}$ and $\mathbf{y}$ are independent, see Yitzhaki (2003).

[4] The rank vector of $\mathbf{x}$ (of size $N \times 1$) is obtained by replacing the elements of $\mathbf{x}$ by their rank (the smallest value of $\mathbf{x}$ being 1 and the highest being N).

A well-know result in panel data literature is that the within-group estimator of $\boldsymbol{\beta}$ (or Least Squares Dummy Variable) issued from (2) is equivalent to the OLS estimator issued from (3):

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{WOLS} &= \left[\sum_{n=1}^{N}\sum_{t=1}^{T}(\mathbf{x}_{nt}-\mathbf{x}_{n.})(\mathbf{x}_{nt}-\mathbf{x}_{n.})'\right]^{-1}\left[\sum_{n=1}^{N}\sum_{t=1}^{T}(\mathbf{x}_{nt}-\mathbf{x}_{n.})(y_{nt}-y_{n.})\right] \\ &=: (\mathbf{X}^{c\prime}\mathbf{X}^{c})^{-1}\mathbf{X}^{c\prime}\mathbf{y}^{c},\end{aligned} \tag{4}$$

where $\mathbf{X}^c$ is the $NT \times K$ matrix of the centered regressors and $\mathbf{y}^c$ the centered dependent variable. Mimicking the OLS estimator (4), one could think that the within-group semi-parametric Gini estimator for fixed effects panel data is simply given by, using (0),

$$\hat{\boldsymbol{\beta}}^{WGini} = (\mathbf{R}'_{\mathbf{x}^c}\mathbf{X}^c)^{-1}\mathbf{R}'_{\mathbf{x}^c}\mathbf{y}^c,$$

where $\mathbf{R}_{\mathbf{x}^c}$ is the rank matrix of $\mathbf{X}^c$. This result is misleading. Actually, in the OLS case, the estimator (4) is deduced from the decomposition of the moment matrices $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}'\mathbf{y}$ into within-group and between-group variabilities. In the following lines, the cogini matrices $\mathbf{R}'_{\mathbf{x}}\mathbf{X}$ and $\mathbf{R}'_{\mathbf{x}}\mathbf{y}$ are decomposed in order to assess the accurate within-group Gini estimator – we shall demonstrate in Section 2.3 that this estimator is a semi-parametric one.

Let the $K \times 1$ vector $\mathbf{x}_{..}$ be the average over time and individuals of $\mathbf{X}$ and let the rank matrix of $\mathbf{X}$ of size $NT \times K$ be $\mathbf{R}_{\mathbf{x}} =: (\mathbf{r}'_{11}(\mathbf{X}), \ldots, \mathbf{r}'_{nt}(\mathbf{X}), \ldots, \mathbf{r}'_{NT}(\mathbf{X}))$ where $\mathbf{r}'_{nt}(\mathbf{X})$ is the nt-th line of $\mathbf{R}_{\mathbf{x}}$, that is, a $1 \times K$ vector. Let $\mathbf{r}_{n.}(\mathbf{X})$ be the $K \times 1$ average rank vector of individual n over time, and $\mathbf{r}_{..}(\mathbf{X})$ the $K \times 1$ average rank vector over time and individuals.[5] Then, the decomposition of the cogini matrix $\mathbf{R}'_{\mathbf{x}}\mathbf{X}$ is:

$$\begin{aligned}\mathbf{G}_{xx}^{total} &= \sum_{n=1}^{N}\sum_{t=1}^{T}(\mathbf{x}_{nt}-\mathbf{x}_{..})(\mathbf{r}_{nt}(\mathbf{X})-\mathbf{r}_{..}(\mathbf{X}))' \\ &= \sum_{n=1}^{N}\sum_{t=1}^{T}(\mathbf{x}_{nt}+\mathbf{x}_{n.}-\mathbf{x}_{n.}-\mathbf{x}_{..})\mathbf{r}'_{nt}(\mathbf{X}) \\ &= \sum_{n=1}^{N}\sum_{t=1}^{T}(\mathbf{x}_{nt}-\mathbf{x}_{n.})\mathbf{r}'_{nt}(\mathbf{X}) + \sum_{n=1}^{N}\sum_{t=1}^{T}(\mathbf{x}_{n.}-\mathbf{x}_{..})\mathbf{r}'_{nt}(\mathbf{X}) \\ &= \underbrace{\sum_{n=1}^{N}\sum_{t=1}^{T}(\mathbf{x}_{nt}-\mathbf{x}_{n.})[\mathbf{r}_{nt}(\mathbf{X})-\mathbf{r}_{n.}(\mathbf{X})]'}_{\text{within-group variability: } \mathbf{G}_{xx}^{within}} + \underbrace{\sum_{n=1}^{N}\sum_{t=1}^{T}(\mathbf{x}_{n.}-\mathbf{x}_{..})[\mathbf{r}_{nt}(\mathbf{X})-\mathbf{r}_{..}(\mathbf{X})]'}_{\text{between-group variability: } \mathbf{G}_{xx}^{between}}.\end{aligned} \tag{5}$$

[5] $\mathbf{r}_{n.}(\mathbf{X}) = \frac{1}{T}\sum_{t=1}^{T}\mathbf{r}_{nt}(\mathbf{X})$ and $\mathbf{r}_{..}(\mathbf{X}) = \frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T}\mathbf{r}_{nt}(\mathbf{X})$.

The breakdown of the cogini matrix $\mathbf{R}'\mathbf{y}$ into within-group and between-group variabilities is derived in the same manner as before:

$$\mathbf{G}_{xy}^{total} = \sum_{n=1}^{N}\sum_{t=1}^{T}(y_{nt} - y_{n.})[\mathbf{r}_{nt}(\mathbf{X}) - \mathbf{r}_{n.}(\mathbf{X})] \qquad (\mathbf{G}_{xy}^{within})$$

$$+ \sum_{n=1}^{N}\sum_{t=1}^{T}(y_{n.} - y_{..})[\mathbf{r}_{nt}(\mathbf{X}) - \mathbf{r}_{..}(\mathbf{X})]. \qquad (\mathbf{G}_{xy}^{between})$$

In sum, the total variabilities (in the cogini sense) are given by:

$$\mathbf{G}_{xx}^{total} = \mathbf{G}_{xx}^{within} + \mathbf{G}_{xx}^{between} \quad \text{and} \quad \mathbf{G}_{xy}^{total} = \mathbf{G}_{xy}^{within} + \mathbf{G}_{xy}^{between}. \tag{6}$$

Following (1), the within-group Gini variabilities ($\mathbf{G}_{xx}^{within}$ and $\mathbf{G}_{xy}^{within}$) yield the within-group Gini estimator:

$$\hat{\boldsymbol{\beta}}^{WG} = [\mathbf{G}_{xx}^{within}]^{-1}\,[\mathbf{G}_{xy}^{within}]. \tag{7}$$

Let $\mathbf{R}^c$ be the $NT \times K$ rank matrix such that

$$\mathbf{R}^c := ((\mathbf{r}_{11}(\mathbf{X}) - \mathbf{r}_{1.}(\mathbf{X}))', \ldots, (\mathbf{r}_{nt}(\mathbf{X}) - \mathbf{r}_{n.}(\mathbf{X}))', \ldots, (\mathbf{r}_{NT}(\mathbf{X}) - \mathbf{r}_{N.}(\mathbf{X}))'),$$

then the within-group Gini estimator is also expressed as:

$$\hat{\boldsymbol{\beta}}^{WG} = (\mathbf{R}^{c\prime}\mathbf{X}^c)^{-1}\mathbf{R}^{c\prime}\mathbf{y}^c.$$

The between-group Gini estimator is:

$$\hat{\boldsymbol{\beta}}^{BG} = \left[\mathbf{G}_{xx}^{between}\right]^{-1}\left[\mathbf{G}_{xy}^{between}\right]. \tag{8}$$

Let us introduce the following matrices:

$$\mathbf{F}^{within} := \left[\mathbf{G}_{xx}^{within} + \mathbf{G}_{xx}^{between}\right]^{-1}\mathbf{G}_{xx}^{within} \tag{9}$$

$$\mathbf{F}^{between} := \left[\mathbf{G}_{xx}^{within} + \mathbf{G}_{xx}^{between}\right]^{-1}\mathbf{G}_{xx}^{between}. \tag{10}$$

From (6)-(10), the overall Gini estimator of the parameter $\boldsymbol{\beta}$ of (2) is decomposable as follows:

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{G} = [\mathbf{G}_{xx}^{total}]^{-1}[\mathbf{G}_{xy}^{total}] &= \left[\mathbf{G}_{xx}^{within} + \mathbf{G}_{xx}^{between}\right]^{-1}\left[\mathbf{G}_{xy}^{within} + \mathbf{G}_{xy}^{between}\right] \\ &= \left[\mathbf{G}_{xx}^{within} + \mathbf{G}_{xx}^{between}\right]^{-1}\left[\mathbf{G}_{xx}^{within}\hat{\boldsymbol{\beta}}^{WG} + \mathbf{G}_{xx}^{between}\hat{\boldsymbol{\beta}}^{BG}\right] \\ &= \mathbf{F}^{within}\hat{\boldsymbol{\beta}}^{WG} + \mathbf{F}^{between}\hat{\boldsymbol{\beta}}^{BG}. \end{aligned} \tag{11}$$

2.3 Inference on the within-group estimator

Yitzhaki and Schechtman (2013) show that all the estimators used in Gini regressions are U-statistics (first introduced by Heoffding, 1948), which possess desirable asymptotic properties. We prove in the sequel that the within-group Gini estimator $\hat{\boldsymbol{\beta}}^{WG}$ is a semi-parametric estimator and it can be estimated as a function of U-statistics.

We first recall the basic notions of U-statistics. Let X_1, $X_2, \ldots, X_N$ be N i.i.d. variables, and $\phi(X_1, X_2, \ldots, X_N)$ a symmetric function (the kernel) such that:

$$\phi^*(X_1, X_2, \ldots, X_N) = (m!)^{-1} \sum_{i_1, i_2, \ldots, i_m} \ldots \sum \phi(X_{i_1}, X_{i_2}, \ldots, X_{i_m}),$$

where m is the smallest number of observations needed to estimate ϕ^*. The U-statistic for the parameter ϕ^*, which is an unbiased estimate of ϕ^*, is written in the following form:

$$U(X_1, X_2, \ldots, X_N) = \binom{N}{m}^{-1} \sum_{i_1, i_2, \ldots, i_m} \ldots \sum \phi(X_{i_1}, X_{i_2}, \ldots, X_{i_m}).$$

The variance of an U-statistic, $Var(U)$, for the parameter ϕ^* of degree m (degree of the kernel) is giving by:

$$Var(U) = \binom{N}{m}^{-1} \sum_{i=1}^{m} \binom{m}{i} \binom{N-m}{m-i} \xi_i,$$

where,

$$\xi_i = Var[\phi_i^*(X_1, X_2, \ldots, X_N)] = \mathbb{E}(\phi_i^{*2}(X_1, X_2, \ldots, X_N)) - \mathbb{E}(\phi_i^*(X_1, X_2, \ldots, X_N))^2.$$

Another option to estimate the variance of U is the jackknife method:

$$Var(U) = \frac{N-1}{N} \sum_{i=1}^{N} \left[U_{-i} - \frac{1}{N} \sum_{i=1}^{N} U_{-i} \right]^2,$$

where U_{-i} is the estimator based on a sample of size N, without the ith observation.

In order to prove that $\hat{\boldsymbol{\beta}}^{WG}$ is a semi-parametric estimator, $\hat{\boldsymbol{\beta}}^{WG}$ is shown to be a function of slope coefficients stemming from simple semi-parametric Gini regressions. Let $\mathbf{r}_k^c$ be the kth column of $\mathbf{R}^c$ and $\mathbf{x}_k^c$ the kth column of $\mathbf{X}^c$, $k = 1, \ldots, K$. Since the within-group Gini estimator $\hat{\boldsymbol{\beta}}^{WG} = (\hat{\beta}_1^{WG}, \ldots, \hat{\beta}_K^{WG})$ yields,

$$\mathbf{y}^c = \hat{\beta}_1^{WG} \mathbf{x}_1^c + \cdots + \hat{\beta}_K^{WG} \mathbf{x}_K^c + \boldsymbol{\varepsilon},$$

then the following identities hold:[6]

$$\begin{aligned}
\text{cov}(\mathbf{y}^c, \mathbf{r}_1^c) &= \hat{\beta}_1^{WG}\text{cov}(\mathbf{x}_1^c, \mathbf{r}_1^c) + \cdots + \hat{\beta}_K^{WG}\text{cov}(\mathbf{x}_K^c, \mathbf{r}_1^c) + \text{cov}(\boldsymbol{\varepsilon}, \mathbf{r}_1^c) \\
\text{cov}(\mathbf{y}^c, \mathbf{r}_k^c) &= \hat{\beta}_1^{WG}\text{cov}(\mathbf{x}_1^c, \mathbf{r}_k^c) + \cdots + \hat{\beta}_K^{WG}\text{cov}(\mathbf{x}_K^c, \mathbf{r}_k^c) + \text{cov}(\boldsymbol{\varepsilon}, \mathbf{r}_k^c) \\
\text{cov}(\mathbf{y}^c, \mathbf{r}_K^c) &= \hat{\beta}_1^{WG}\text{cov}(\mathbf{x}_1^c, \mathbf{r}_K^c) + \cdots + \hat{\beta}_K^{WG}\text{cov}(\mathbf{x}_K^c, \mathbf{r}_K^c) + \text{cov}(\boldsymbol{\varepsilon}, \mathbf{r}_K^c).
\end{aligned}$$

Setting $\hat{\beta}_{\varepsilon j} := \frac{\text{cov}(\boldsymbol{\varepsilon}, \mathbf{r}_j^c)}{\text{cov}(\mathbf{x}_j^c, \mathbf{r}_j^c)}$, $\hat{\beta}_{0j} := \frac{\text{cov}(\mathbf{y}^c, \mathbf{r}_j^c)}{\text{cov}(\mathbf{x}_j^c, \mathbf{r}_j^c)}$ and $\hat{\beta}_{kj} := \frac{\text{cov}(\mathbf{x}_k^c, \mathbf{r}_j^c)}{\text{cov}(\mathbf{x}_j^c, \mathbf{r}_j^c)}$, dividing the three last equations by, respectively, $\text{cov}(\mathbf{x}_1^c, \mathbf{r}_1^c)$, $\text{cov}(\mathbf{x}_k^c, \mathbf{r}_k^c)$ and $\text{cov}(\mathbf{x}_K^c, \mathbf{r}_K^c)$ yields:

$$\begin{aligned}
\hat{\beta}_{01} &= \hat{\beta}_1^{WG} + \cdots + \hat{\beta}_K^{WG}\hat{\beta}_{K1} + \hat{\beta}_{\varepsilon 1} \\
\hat{\beta}_{0k} &= \hat{\beta}_1^{WG}\hat{\beta}_{1k} + \cdots + \hat{\beta}_K^{WG}\hat{\beta}_{Kk} + \hat{\beta}_{\varepsilon k} \\
\hat{\beta}_{0K} &= \hat{\beta}_1^{WG}\hat{\beta}_{1K} + \cdots + \hat{\beta}_K^{WG} + \hat{\beta}_{\varepsilon K}.
\end{aligned}$$

Setting the following column vectors $\hat{\mathbf{b}}_0 := (\hat{\beta}_{01}, \ldots, \hat{\beta}_{0K})$ and $\hat{\mathbf{b}}_\varepsilon := (\hat{\beta}_{\varepsilon 1}, \ldots, \hat{\beta}_{\varepsilon K})$, then it comes:

$$\begin{pmatrix} \hat{\beta}_1^{WG} \\ \vdots \\ \hat{\beta}_K^{WG} \end{pmatrix} = \begin{pmatrix} 1 & \hat{\beta}_{21} & \ldots & \hat{\beta}_{K1} \\ \vdots & \vdots & \ldots & \vdots \\ \hat{\beta}_{1K} & \hat{\beta}_{2K} & \ldots & 1 \end{pmatrix}^{-1} \begin{pmatrix} \hat{\beta}_{01} - \hat{\beta}_{\varepsilon 1} \\ \vdots \\ \hat{\beta}_{0K} - \hat{\beta}_{\varepsilon K} \end{pmatrix} =: \hat{\mathbf{B}}^{-1}\left[\hat{\mathbf{b}}_0 - \hat{\mathbf{b}}_\varepsilon\right].$$

The within-group Gini estimator is a function of slope coefficients of semi-parametric simple Gini regressions, and as such it is referred to as a semi-parametric Gini estimator. Yitzhaki and Schechtman (2013, Chapter 9) have proven that $\hat{\beta}_{0k}$, $\hat{\beta}_{\varepsilon k}$ and $\hat{\beta}_{kh}$ are function of U-statistics. If $\hat{\mathbf{B}}$ is a full rank matrix, then $\hat{\boldsymbol{\beta}}^{WG}$ is a function of U-statistics. By Slutzky's theorem, $\hat{\boldsymbol{\beta}}^{WG}$ is a consistent estimator of $\boldsymbol{\beta}^{WG}$, it is asymptotically normal.

2.4 An illustration with a simple simulation

In this Section, it is shown that the semi-parametric within-group Gini estimator is more robust than the OLS one when the data are contaminated by outliers. For that purpose, simple Monte Carlo simulations are performed. The contamination is concerned with only 1% of each sample. For all simulation samples, the same locations are chosen for the contamination. The 1% observations are contaminated by replacing their values by two times the maximum value of the regressor vector they belong to.[7]

[6] This technique has been introduced by Yitzhaki and Schechtman (2013, Chapter 8) for $\boldsymbol{\beta}^G$.

[7] Other contaminations have been made: changing the value of just one point or changing the values of 1% of the simulation samples by adding (subtracting) some values representing either 20, 30 or 50 times the value of the standard error of ε (valued to be 1). In each case, the OLS estimates highly deviate from their true values, not the within-group Gini estimates.

The steps are the following:

- Loop to $b = 1, \ldots, B = 10,000$;
  - ↪ The regressors are generated from a multivariate normal distribution $\mathbf{x} \sim \mathcal{N}(\mu, \Sigma)$ such that $\mu = (0, 10, 4)$, $\Sigma = \begin{pmatrix} 1 & 0.5 & 0.2 \\ & 1 & 0.15 \\ & & 1 \end{pmatrix}$, $\varepsilon \sim \mathcal{N}(0, 1)$, $\boldsymbol{\beta} = (0.7, 1.23, 0.13)$ and the fixed effects $\beta_n \sim \mathcal{N}(0, 0.02)$, $\beta_0 = 0.75$;
  - ↪ The dependent variable $\mathbf{y}$ is deduced from (1) ;
  - ↪ $\mathbf{y}$ is regressed on $\mathbf{x}$ [OLS (3) and Gini (7)] ;
  - ↪ Outliers are introduced into the regressors ($\mathbf{x}^o$), then $\mathbf{y}$ is regressed on $\mathbf{x}^o$ [OLS (3) and Gini (7)]: estimates $\hat{\boldsymbol{\beta}}_b^{OLS}$ and $\hat{\boldsymbol{\beta}}_b^{G}$ are deduced for each $b = 1, \ldots, B$;
- End b ;
- The mean of the estimates ($\bar{\hat{\boldsymbol{\beta}}}^{WG}$ and $\bar{\hat{\boldsymbol{\beta}}}^{OLS}$) and the mean squared error (MSE) are computed over B.

Table 1. Simulations

Estimates →	without outliers		with outliers	
$\boldsymbol{\beta} =$	$\bar{\hat{\boldsymbol{\beta}}}^{WOLS}$ (MSE)	$\bar{\hat{\boldsymbol{\beta}}}^{WG}$ (MSE)	$\bar{\hat{\boldsymbol{\beta}}}^{WOLS}$ (MSE)	$\bar{\hat{\boldsymbol{\beta}}}^{WG}$ (MSE)
0.7	0.70022 (0.000388)	0.70013 (0.000432)	0.96120 (0.018314)	0.79715 (0.001041)
1.23	1.22991 (0.000062)	1.22989 (0.000067)	1.35940 (0.206812)	1.29836 (0.004652)
0.13	0.13018 (0.000119)	0.13023 (0.000124)	0.24691 (0.195059)	0.15651(0.002849)

3 Conclusion

Whenever the distribution of the covariates is multivariate normal, both OLS and Gini estimates are very close. Using the Gini approach implies that the efficiency of the OLS is lost. This is the case for instance when extreme values or measurement errors alter the regressors.

We have shown that the fixed effects Gini regression does not consist in mimicking the application of the OLS on the centered model (3). It is based on a proper decomposition of the within-group and between-group cogini variabilities of the moment matrices. The semi-parametric within-group Gini estimator avoids to treat the multiple solutions that would arise in the minimization of a target function such as the Gini index of the residuals of the centered model. Future researches could be done to compare the minimization approach and the semi-parametric one in order to assess the linearity of the model.

References

[1] Hoeffding, W. (1948), A class of statistics with asymptotically normal distributions, *Annals of Statistics*, 19, 293-325.

[2] Huber, P. (1981), *Robust Statistics*, Chichester, John Wiley.

[3] Koenker, R. and G. Bassett (1978), Regression Quantiles, *Econometrica*, 46(1), 33-50.

[4] Olkin, I. and S. Yitzhaki (1992), Gini Regression Analysis, *International Statistical Review*, 60(2), 185-196.

[5] Yitzhaki, S. (2003), Gini's Mean difference: a superior measure of variability for non-normal distributions, *Metron*, LXI(2), 285-316.

[6] Yitzhaki, S. and P. Lambert (2013), The Relationship Between the Absolute Deviation from a Quantile and Gini's Mean Difference, *Metron*, 71, 97-104.

[7] Yitzhaki, S. and E. Schechtman (2013), *The Gini Methodology. A Primer on a Statistical Methodology*, Springer.