INFERENCE FOR MONOTONE SINGLE-INDEX CONDITIONAL MEANS: A LORENZ REGRESSION APPROACH

Cédric Heuchenne, Alexandre Jacquemain

LIDAM Discussion Paper ISBA
2021 / 42

UCLouvain

LIDAM
ISBA | Louvain Institute of Data Analysis and Modeling in economics and statistics

ISBA

Voie du Roman Pays 20 - L1.04.01

B-1348 Louvain-la-Neuve

Email : lidam-library@uclouvain.be

https://uclouvain.be/en/research-institutes/lidam/isba/publication.html

Inference for monotone single-index conditional means: a Lorenz regression approach*

Cedric Heuchenne[a], Alexandre Jacquemain[b,*]

[a] *HEC Liege, University of Liege, Rue Louvrex 14, Liege, B-4000, Belgium*
[b] *Universite catholique de Louvain, ISBA, Voie du Roman Pays 20, Louvain-la-Neuve, B-1348, Belgium*

Abstract

The Lorenz regression procedure quantifies the inequality of a response explained by a set of covariates. Formally, it gives a weight to each covariate to maximize the concentration index between the response and a weighted average of the covariates. The obtained index is called the explained Gini coefficient. Unlike methods based on decompositions of inequality measures, the procedure does not assume a linear relationship between the response and the covariates. Inference can be performed by noticing a similarity with the monotone rank estimator, introduced in the context of the single-index model. A continuity correction is presented in the presence of discrete covariates. The Lorenz-R^2 is a goodness-of-fit measure evaluating the proportion of explained inequality and is used to build a test of joint significance of several covariates. Monte-Carlo simulations and a real-data example are presented.

Keywords: Single-index model, Monotone rank estimator, Lorenz curve, Income inequality

1. Introduction

In the eighth of his *Ten Short Reflections on the Future of Income Inequality*, Milanovic (2016) expresses the idea that economics has methodologically shifted from a concern with representative agents and averages to a concern

*An Online Supplement providing technical details of proofs and extra simulation results is provided in a separate document.

*Corresponding author
Email address: alexandre.jacquemain@uclouvain.be (Alexandre Jacquemain)

with heterogeneity and inequality. This academic focus is paralleled with rising voices among social observers concerning increasing inequalities all over the world. In this context, it is of prime importance to understand what factors can explain the inequality pattern observed in, say, an income or wealth distribution. Practically speaking, we have in mind a microeconomist facing a cross-sectional dataset, with information on incomes, along with many other variables. She would like to determine to what extent income inequality is attributable to, say, gender, age or education.

Available tools to address such a challenge are mainly decomposition methods, where the observed income inequality is divided into the contributions of each of the explanatory variables. However, these procedures require to assume a linear model between income and the explanatory variables. To bypass this restrictive assumption, we set the problem differently. We propose the Lorenz regression, in which each explanatory variable is given a weight in order to maximize a measure of explained inequality. On a more statistical aspect, we show that the obtained optimization programme consists in a special case of the monotone rank estimator developed by Cavanagh and Sherman (1998) in the context of single-index models. This link will enable us to use existing results concerning this estimator to perform inference on the weight vector mentioned above.

In this paper, we view inequality as a statistical measure of relative dispersion of a random variable, where relative means that it is independent of the scale of the variable. It is interesting to see that a similar relative dispersion exists in risk analysis in finance, see for example Shalit and Yitzhaki (1984). Consequently, while we build interpretations in a context of inequality measurement, these could easily be translated to the risk of a financial asset. We close this introduction by a brief review of the existing literature.

1.1. Inference about inequality

The measurement of inequality has long been related to the well-known Lorenz curve and the Gini coefficient. Considering an income distribution, the Lorenz curve evaluated at p provides the share of total income owned by the $p\times100\%$-poorest individuals. The Gini coefficient is a summary measure of it. Both objects are precisely defined in Section 2, see (3) and (4). Several papers have used these tools to determine how the inequality of a variable of interest can be explained by some other variables. In what follows, we will call response the variable of interest, typically income, and covariates the explanatory variables.

Some decomposition ideas proved to be useful when the response can be exactly decomposed as a linear index of different sources. Lerman and Yitzhaki (1985) introduced a decomposition of the Gini coefficient of income in the contributions of various income sources. Using such methodology, Garner (1993) assessed the role of different household budget components in total expenditure inequality. Yitzhaki (1994) analysed the distributional effects of commodity taxations.

Another stream of research generalized the Lorenz curve with the concept of concentration curve. Introduced by Blitz and Brittain (1964), the concentration curve evaluated at p of, say, income with respect to education gives the share of total income owned by the $p\times100\%$-least educated individuals, see also Fei et al. (1978). A decomposition of its summary measure, the concentration index, can be found in health economics. The idea has been introduced in Wagstaff et al. (2003), and then used in Jones and Nicolás (2004) and van Doorslaer and Koolman (2004). Importantly, all these procedures require to assume a linear regression model between the response and the explanatory variables. It is also interesting to mention that decompositions of the concentration curve have been used outside the realm of inequality measurement. Shalit and Yitzhaki (2003) and Denuit et al. (2014) used a decomposition of the concentration curve itself to check the efficiency of a portfolio.

In order to analyze the impact of several covariates, Aaberge et al. (2005) introduced the pseudo-Lorenz regression curve. This curve is obtained by replacing the expected value in the numerator of (3) by a conditional expectation. The impact of the covariates is obtained by marginal changes in the pseudo-Lorenz curves or in pseudo-Gini coefficients. However, this approach requires to estimate curves in a nonparametric setup and, hence, will suffer from the curse of dimensionality.

In Section 2, we will illustrate the interest of our procedure in the measurement of inequality of opportunity (IOP). The concept of equality of opportunity, developped by Roemer (1998), considers that inequalities in some economic advantage are unjustified when they are the product of circumstances, i.e. variables over which individuals have no control. IOP will thus measure the extent of inequality which is related to circumstances. One difficulty arises when balancing the interpretability of the inequality measure with the statistical modelling of the advantage variable. As we have already stated, decompositions of inequality measures call for linear decompositions. Another strand of the literature uses log-linear regressions of the

advantage on the circumstances, see for example Bourguignon et al. (2007). Because a restrictive statistical model is imposed, these approaches "clip the wings of the econometricians" to borrow the expression used in Fleurbaey and Schokkaert (2009).

1.2. The semiparametric single-index model

In many applications, parametric models seem not robust enough to capture the relationship between variables. However, while offering more flexibility, nonparametric methods have also their own drawbacks such as the curse of dimensionality. In that sense, semiparametric models stand as appropriate compromises. A famous example of such method lies in the single-index model proposed by Ichimura (1993).

Let (X, Y) be an observation point in $\mathbb{R}^p \times \mathbb{R}$ with joint cumulative distribution function (CDF) $F_{X,Y}$, where Y is the response and $X = (X_1, \ldots, X_p)^\intercal$ a vector gathering the p covariates. Besides, let $E[Y|X = x]$ denote the conditional expectation of Y given that X takes the value x. Horowitz (2009) defines the single-index model as

$$E[Y|X = x] = H(x^\intercal \theta_0), \tag{1}$$

where H is left unspecified and $\theta_0 = (\theta_{0,1}, \ldots, \theta_{0,p})^\intercal$ is a vector of parameters. In the context of this paper, we will furthermore assume that H is strictly increasing. Some conditions on θ_0 are needed in order to ensure the identifiability of the model. Here, we will set $||\theta_0|| = 1$, where $||\cdot||$ denotes the L1-norm.

At first sight, the estimation of θ_0 seems to be contingent to that of H. Most estimation methods, such as those proposed by Ichimura (1993) and Powell et al. (1989) require to replace H, or its first derivative, by a suitable kernel estimator. In practice, the price to pay translates into the choice of some smoothing parameter. Cavanagh and Sherman (1998) exploit the monotonicity of H to escape this problem. Facing an i.i.d sample (X_i, Y_i), $i = 1, \ldots, n$ from $F_{X,Y}$, they propose the monotone rank estimator (MRE) of θ_0, which solves the following maximization programme

$$\max_{\theta} \sum_{i=1}^{n} M(Y_i) R_n(X_i^\intercal \theta) \qquad \text{s.t. } ||\theta|| = 1, \tag{2}$$

where M is an increasing function, $R_n(X_i^\intercal \theta) \coloneqq \sum_{j=1}^{n} \mathbb{1}\{X_j^\intercal \theta < X_i^\intercal \theta\}$ is the rank of $X_i^\intercal \theta$ in the vector $X^\intercal \theta$, and $\mathbb{1}\{\cdot\}$ is the indicator function. Under

some regularity conditions, Theorem 1 in Cavanagh and Sherman (1998) proves the consistency of the MRE when there is at least one continuous covariate. The situation where all covariates are discrete is discussed in Section 3. Given two supplementary regularity conditions, Theorem 2 from the same source proves the asymptotic normality of the estimator and proposes methods for the estimation of the covariance matrix.

The rest of this paper is organized as follows. In Section 2, we introduce our procedure in the continuous case. The central notion is the explained Gini coefficient, for which we show the linkage with the MRE and provide interpretations in the context of IOP. We then notice that an estimator for this coefficient can be easily built on that basis. In Section 3, we highlight issues arising when all covariates are discrete and propose a continuity correction to address them. We discuss the asymptotic properties of this solution as well as its practical relevance. Section 4 discusses the inference about the procedure. Among others, we present a bootstrap test of joint significance of multiple covariates. Section 5 addresses the actual performance of our procedure through a series of Monte-Carlo simulations. The estimation procedure is compared to the semiparametric least squares estimator defined in Ichimura (1993). We also display power curves of the testing procedure. Finally, Section 6 presents an application on wages data.

2. The Lorenz regression methodology

We start with some notations and definitions. Let $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$ with continuous joint CDF $F_{X,Y}$ and $0 < E[Y] < \infty$, where $E[\cdot]$ is the expected value. The case $-\infty < E[Y] < 0$ can be dealt with by analysing the random variable $-Y$ instead. We define the Lorenz curve of Y at p as

$$\mathrm{LC}_Y(p) \coloneqq \frac{E[Y\,\mathbb{1}\{F_Y(Y) \leq p\}]}{E[Y]}, \tag{3}$$

where $\mathbb{1}\{\cdot\}$ denotes the indicator function and F_Y the marginal CDF of Y. The LC passes through $(0, 0)$ and $(1, 1)$, and is always convex. Perfect equality is pictured by the 45° line. If $Y \in \mathbb{R}^+$, perfect inequality is displayed by the right-angle situation. For more intuitions underlying the use of Lorenz curves, we refer to Yitzhaki and Schechtman (2013).

While the Lorenz curve pictures inequality in a disaggregated way, the Gini coefficient summarizes this information in an index. The lower bound

of 0 is attained in situation of perfect equality. The upper bound is attained when perfect inequality occurs, at a value of 1 if $Y \in \mathbb{R}^+$. It is formally defined as

$$\text{Gi}_Y \coloneqq 2\int_0^1 [p - \text{LC}_Y(p)]dp = \frac{2C[Y, F_Y(Y)]}{E[Y]},$$

where $C[\cdot, \cdot]$ denotes the covariance operator. Introducing a second variable X_k with marginal CDF F_k, we define the concentration curve of Y with respect to X_k as

$$\text{CC}_{Y,X_k}(p) \coloneqq \frac{E[Y\, \mathbb{1}\{F_k(X_k) \leq p\}]}{E[Y]}. \tag{4}$$

The CC goes through $(0, 0)$ and $(1, 1)$ but is no longer necessarily convex. Intuitively, it pictures the inequality of Y that we can reproduce if we rank individuals in terms of X_k instead of ordering them with respect to Y. For illustration purposes, consider that X_k represent education and $Y \in \mathbb{R}^+$ is income. The 45° line still pictures a situation of perfect equality in the sense that, for all p, the $p\times100\%$ least-educated accumulate $p\times100\%$ of the total income. However, we may now face two different right-angle shaped situations, each corresponding to a situation of extreme inequality. It may now occur because all the income rests in the hands of the most educated individual (bottom right), or because it is concentrated at the least educated (top left). For more information, the reader is again referred to Yitzhaki and Schechtman (2013).

Similarly to the Gini coefficient, the concentration index summarizes the information of the curve in an index. If $Y \in \mathbb{R}^+$, it ranges from -1 to 1. A value of 1 or -1 indicates a situation of extreme inequality. It reaches 1 when the most-educated individual owns all the income and -1 if this occurs for the least-educated one. As before, a value of 0 indicates perfect equality. Formally, it is defined as

$$\text{Ci}_{Y,X_k} \coloneqq 2\int_0^1 [p - \text{CC}_{Y,X_k}(p)]dp = \frac{2C[Y, F_k(X_k)]}{E[Y]}.$$

In the Lorenz regression procedure, we will focus on a linear index $X^\intercal\theta$ of the covariates, where $\theta = (\theta_1, \ldots, \theta_p)^\intercal$, $||\theta|| = 1$ is a vector of weights. The

concentration index of Y with respect to $X^\intercal\theta$ is given by

$$\text{Ci}_{Y,X^\intercal\theta} = \frac{2C[Y, F_\theta(X^\intercal\theta)]}{E[Y]}, \tag{5}$$

where F_θ denotes the CDF of $X^\intercal\theta$. It represents the inequality of Y obtained when individuals are ranked in terms of $X^\intercal\theta$. Our objective consists in finding the vector of weights θ^* which maximizes (5). We call the obtained concentration index the explained Gini coefficient. Formally, it is given by

$$\text{Gi}_{Y,X} \coloneqq \max_\theta \frac{2C[Y, F_\theta(X^\intercal\theta)]}{E[Y]} = \frac{2C[Y, F_{\theta^*}(X^\intercal\theta^*)]}{E[Y]}.$$

We also define the proportion of explained inequality (PEI) as the ratio between the explained Gini coefficient and the acual one, i.e.

$$\text{PEI}_{Y,X} \coloneqq \frac{\text{Gi}_{Y,X}}{\text{Gi}_Y}.$$

In the interpretations, it will be useful to have $\text{PEI}_{Y,X} \in [0, 1]$. This result is a direct consequence of Theorem 3.1 and Remark 6.2 in Das Gupta (1999), which shows that, at any point, the concentration curve of Y with respect to X_k lies above the Lorenz curve of Y.

Proposition 1 establishes the connection between θ^*, the weight vector characterizing the explained Gini coefficient, and θ_0, the true parameter vector of the single-index model. For simplicity, we assume that all components of X are continuous. The situation where discrete covariates are introduced is discussed in Section 3.

Proposition 1. *Let $(Y, X) \in \mathbb{R} \times \mathbb{R}^p$, with $0 < E[Y] < \infty$, be continuous random variables satifsfying* (1) *with H strictly increasing. Then (i) $\theta^* = \theta_0$ is unique and (ii) the explained Gini coefficient is given by*

$$Gi_{Y,X} = \frac{2C[Y, F_{\theta_0}(X^\intercal\theta_0)]}{E[Y]} = \frac{2C[H(X^\intercal\theta_0), F_{\theta_0}(X^\intercal\theta_0)]}{E[H(X^\intercal\theta_0)]}.$$

(i) can be easily derived from the proof of Theorem 1 in Cavanagh and Sherman (1998). The connection arises because finding θ^* boils down to maximizing $G(\theta) \coloneqq E[YF_\theta(X^\intercal\theta)]$, which is the population version of (2) when $M(Y_i) = Y_i$. (ii) is a direct consequence of the first result and of the

single-index model. It indicates that the explained Gini coefficient is the Gini coefficient of $H(X^\intercal\theta_0)$, the explained part of the single-index model.

In the following paragraphs, we turn to the intepretation of our procedure in the field of socioeconomic inequality. As an illustration, we consider the case of equality of opportunity. First, the explained Gini coefficient is an appropriate measure of IOP. In this application, Y is an economic advantage and X a vector of circumstances. As shown in Proposition 1, the explained Gini coefficient is the Gini coefficient of $H(X^\intercal\theta_0)$ in the single-index model (1). It measures the inequality of the economic advantage predicted by the circumstances. Hence, we benefit from a natural measure of IOP while, at the same time, enjoying the statistical flexibility of a semiparametric model. As we will observe at the end of this section, the estimation of this measure does not entail an estimation of the link function H. Since the PEI always lies between 0 and 1, it can be interpreted as the proportion of inequality which corresponds to inequality of opportunity. We close this discussion with a fictitious example. Consider two countries A and B. We run a Lorenz regression on the same set of circumstances for each country. Assume that we obtain the results displayed in Table 1. Even though the Gini coefficient is

Table 1: (Explained) Gini coefficients in A and B

	A	B
$\widehat{\mathrm{Gi}}_Y$	0.5	0.4
$\widehat{\mathrm{Gi}}_{Y,X}$	0.3	0.35

larger in country A, the IOP captured by the explained Gini coefficient is higher in country B. To put this differently, while there is more inequality in the distribution of the advantage variable in country A, there is more unjustified inequality (in the IOP sense) in country B.

The sign of each weight $\theta_{0,k}$ is easy to interpret. A value of zero indicates that the circumstance does not contribute to the inequality of the advantage, i.e. it does not create inequality of opportunity. A positive (negative) weight rather refers to a notion of concordance (discordance). In order to understand what it entails, consider that the associated covariate represents the number of siblings. A positive weight corresponds to a situation where a

bigger share of advantage is amassed by the individuals with bigger families. If, instead, more advantage is gathered in the hands of individuals with less siblings, the weight would then exhibit a negative value.

Assuming (1), the ratio between two weights compares the relative impact of two circumstances on the advantage variable. Formally, we define the marginal rate of substitution (MRS) of X_k for X_l, all other things being equal, by

$$\mathrm{MRS}_{k,l} \coloneqq \frac{\partial E[Y|X=x]}{\partial x_k} \Big/ \frac{\partial E[Y|X=x]}{\partial x_l} = \frac{\theta_{0,k}}{\theta_{0,l}} \,.$$

Again, this quantity can be consistently estimated without the need to estimate the link function H.

We close this section by presenting an estimator of the explained Gini coefficient and by introducing the Lorenz-R^2. Concerning the estimation, the difficulty lies in estimating θ^*. An estimator $\hat{\theta}$ for this weight vector is obtained by maximizing $G_n(\theta) \coloneqq \sum_i Y_i R_n(X_i^\intercal \theta)$. Again by Theorem 1 from Cavanagh and Sherman (1998), $\hat{\theta}$ is a consistent estimator for θ^*, for which asymptotic properties have been derived. We estimate the explained Gini coefficient $\mathrm{Gi}_{Y,X}$ with

$$\widehat{\mathrm{Gi}}_{Y,X} \coloneqq \widehat{\mathrm{Ci}}_{Y,X^\intercal\hat{\theta}} = \frac{2}{n^2} \sum_{i=1}^{n} \frac{Y_i}{\overline{Y}} R_n(X_i^\intercal \hat{\theta}) - \frac{n+1}{n}. \tag{6}$$

The consistency of this estimator follows trivially from Theorem 1 from Cavanagh and Sherman (1998). These results open the door to inference exercises, both on θ^* and on $\mathrm{Gi}_{Y,X}$. In linear regression, the R^2 measures the proportion of variance which is explained by our covariates. Here, we are rather interested in comparing the inequality reproduced by our covariates with the total inequality, as pictured by the Gini coefficient of the outcome. The PEI, introduced previously precisely does that in the population. Accordingly, we build the Lorenz-R^2 as a simple translation in the sample. Formally, it is defined as

$$\mathrm{LR}^2 \coloneqq \frac{\widehat{\mathrm{Gi}}_{Y,X}}{\widehat{\mathrm{Gi}}_{Y}} = \frac{\frac{1}{n^2}\sum_{i=1}^{n} Y_i R_n(X_i^\intercal \hat{\theta}) - \frac{n+1}{n}\frac{\overline{Y}}{2}}{\frac{1}{n^2}\sum_{i=1}^{n} Y_i R_n(Y_i) - \frac{n+1}{n}\frac{\overline{Y}}{2}}.$$

Note that this measure always lies between 0 and 1.

3. Accommodating for discrete covariates

The issues brought about by the introduction of discrete covariates can be viewed from two different angles. First, when all covariates are discrete, the single-index model is not identified, see Horowitz (2009) for a thorough discussion. However, when the link function is increasing, some information can still be extracted. In our context, we will be able to identify θ_0 up to a pre-specified region, see Theorem 1 for the resulting specific consistency. The main message is that the more covariates we introduce, or similarly, the more categories the covariates have, the closer we will get to identifiability. When the total number of categories is low, other methods could be used, e.g. the pseudo-Lorenz approach of Aaberge et al. (2005). Second, the computation of the explained Gini coefficient requires that we are able to rank observations. But the presence of discrete covariates creates ties in the index, which raises the question of how ranks are defined. In essence, the CDF deals with this issue by giving the highest rank to all the ties. However, other options could be proposed. From Schechtman and Zitikis (2006), we know that keeping the CDF in the definition of the Gini coefficient leads to an inconsistency between its interpretation as a covariance on one side and a surface between the egalitarian line and the Lorenz curve on the other side. In order to restore the concordance between these two perspectives, it is enough to replace the CDF by a suitable alternative. One of such alternatives, presented in Lerman and Yitzhaki (1989), amounts to giving the average rank in every situation where ties occur. In this section, we propose another solution. It consists in a random assignment of the ranks affected by a situation of ties. We provide some arguments in favor of this choice.

We start by noticing that the concordance between deriving θ^* and maximizing $E[YF_\theta(X^\intercal\theta)]$ is compromised in the discrete case. Essentially, this is due to the fact that $F_\theta(X^\intercal\theta) \sim U[0,1]$ no longer holds. The following continuity correction restores this property.

Proposition 2. *Let Z be a non-continuous random variable with CDF F having a finite number of discontinuities, and define $F(z^-) \coloneqq P(Z < z)$. Also, let $V \sim U[0,1]$ independent of Z. Finally, $\tilde{F}(Z) \coloneqq F(Z^-)+V(F(Z)-F(Z^-))$. Then, $\tilde{F}(Z) \sim U[0,1]$.*

Proof: We need to show that $P(\tilde{F}(Z) \leq q) = q$ for $q \in [0,1]$. Let $F^{-1}(q) \coloneqq \inf\{z \in \mathbb{R} : F(z) \geq q\}$. In absence of discontinuity, it holds $F(F^{-1}(q)) = q$.

When a discontinuity occurs, $F(z)$ experiences a jump. Hence, it might also be that $F(F^{-1}(q)^-) = q$ when q is located at the start of a jump, or $F(F^{-1}(q)^-) < q < F(F^{-1}(q))$ when q is located in the middle of a jump. We focus here on this last case and treat the remaining two in Section 1 of the Online Supplement. We have

$$P(\tilde{F}(Z) \leq q) = P\Big(\tilde{F}(Z) \leq F(F^{-1}(q)^-)\Big) + P\Big(F(F^{-1}(q)^-) < \tilde{F}(Z) \leq q\Big).$$

Regardless of the value of V, $\tilde{F}(Z) \leq F(F^{-1}(q)^-)$ occurs when $Z < F^{-1}(q)$. Hence, the first piece boils down to $F(F^{-1}(q)^-)$. After conditioning on $Z = F^{-1}(q)$ and using the definition of $\tilde{F}(\cdot)$, the second piece becomes

$$\begin{aligned}
&P\Big(F(F^{-1}(q)^-) < F(F^{-1}(q)^-) + V\big[F(F^{-1}(q)) - F(F^{-1}(q)^-)\big] \leq q \Big| Z = F^{-1}(q)\Big) \\
&\qquad \times P(Z = F^{-1}(q)) \\
&= P\left(0 < V \leq \frac{q - F(F^{-1}(q)^-)}{F(F^{-1}(q)) - F(F^{-1}(q)^-)} \middle| Z = F^{-1}(q)\right) \\
&\qquad \times (F(F^{-1}(q)) - F(F^{-1}(q)^-)).
\end{aligned}$$

Using the independence and uniformity of V, we obtain the desired result. □

The idea is then to replace the discontinuous F_θ by $\tilde{F}_\theta$, where the latter is defined as in Proposition 2. We are now maximizing $\tilde{G}(\theta) \coloneqq E[Y\tilde{F}_\theta(X^\intercal\theta)]$, which, by Proposition 2, is equivalent to maximizing the following corrected concentration index

$$\widetilde{\mathrm{Ci}}_{Y,X^\intercal\theta} \coloneqq \frac{2C[Y, \tilde{F}_\theta(X^\intercal\theta)]}{E[Y]}.$$

In the sample, the corrected optimization programme becomes

$$\max_\theta \tilde{G}_n(\theta) \coloneqq \sum_{i=1}^n Y_i \tilde{R}_n^V(X_i^\intercal\theta) \qquad \text{s.t. } ||\theta|| = 1, \tag{7}$$

where the corrected rank vector $\tilde{R}_n^V$ is obtained as

$$\tilde{R}_n^V(X_i^\intercal\theta) = \sum_{j=1}^n \mathbb{1}[X_j^\intercal\theta < X_i^\intercal\theta] + \mathbb{1}[X_j^\intercal\theta = X_i^\intercal\theta, V_j < V_i],$$

and $V = (V_1, \ldots, V_n)^\intercal$ where $V_i \overset{\text{i.i.d}}{\sim} U[0,1]$. Equipped of our continuity correction, the modified MRE is obtained as a solution to (7). As in the continuous case, we denote it by $\hat{\theta}$. The explained Gini coefficient is estimated using (6) by replacing $R_n(X_i^\intercal\hat{\theta})$ with $\tilde{R}_n^V(X_i^\intercal\hat{\theta})$.

In what follows, we examine the asymptotic properties of the modified MRE. We start from a construction inspired by that developped in Section 5 of Cavanagh and Sherman (1998). However, we adapt it and argue that it best characterizes the extent of the identifiability of θ_0. Suppose X has possible values $x_1, \ldots, x_N$. Let $B_1, \ldots, B_l$ be open regions where $x_j^\intercal\theta \neq x_k^\intercal\theta$ for any $k \neq j$ and such that any pair of θ share the same ordering of $\{x_1^\intercal\theta, \ldots, x_N^\intercal\theta\}$ if and only if they live in the same region. These regions are bounded by hyperplanes $H_1, \ldots, H_m$ where $x_j^\intercal\theta = x_k^\intercal\theta$ for at least one pair $k \neq j$. Denote by $I(\theta)$ the subspace defined by the strict inequalities in the ordering of $\{x_1^\intercal\theta, \ldots, x_N^\intercal\theta\}$ implied by θ. Our conceptualization starts from the region where θ_0 falls, whether it's an open region or an hyperplane. The important step is the following. We merge all the regions characterized by $I(\theta) \subseteq I(\theta_0)$. As an illustration, take $N = 3$ and consider the following situation

$$\begin{aligned} B_1 &= \{x_1^\intercal\theta > x_2^\intercal\theta > x_3^\intercal\theta\}, \\ H_1 &= \{x_1^\intercal\theta > x_2^\intercal\theta = x_3^\intercal\theta\}, \\ B_2 &= \{x_1^\intercal\theta > x_3^\intercal\theta > x_2^\intercal\theta\}. \end{aligned}$$

If $\theta_0 \in B_1$, or if $\theta_0 \in B_2$, the set of strict inequalities fully describes the region and no grouping needs to be made. If $\theta_0 \in H_1$, the set of strict inequalities is $I(\theta_0) = \{x_1^\intercal\theta > x_2^\intercal\theta \text{ and } x_1^\intercal\theta > x_3^\intercal\theta\}$. In this case, for any θ in B_1 or B_2, we have $I(\theta) \subseteq I(\theta_0)$. Hence, the three regions need to be merged. This construction formalizes the idea that θ_0 can be identified up to the region defined by the strict inequalities in the ordering of the index. The resulting regions, as well as the remaining hyperplanes and open regions, are then summarized by a vector of representatives. The new parameter space is denoted by Θ^*. Theorem 1 formalizes the consistency result concerning the estimated parameter vector and implies the consistency of the estimated explained Gini coefficient.

Theorem 1. *Let X be defined as above and Y be a continuous random variable. Also, let $V \sim U[0,1]$ independent of (X, Y). Assume $E[Y|X = x] =$*

$H(x^\intercal\theta_0)$ with H strictly increasing on its points of definition, and $E[Y] < \infty$. Besides, let $\tilde{G}(\theta)$ and $\tilde{G}_n(\theta)$ be defined as above. Then (i) $\tilde{G}_n(\theta) \to \tilde{G}(\theta)$ almost surely for all $\theta \in \Theta^*$ and (ii) $P(\hat{\theta}^* \neq \theta_0^*) = O(1/n)$, where $\hat{\theta}^*$ is the solution to (7) maximized on Θ^* and θ_0^* denotes the representative of the region where θ_0 lies.

Proof: The strategy of the proof is exactly similar to that of Theorem 3 in Cavanagh and Sherman (1998). (i) follows from standard results on U-statistics. (ii) is guaranteed by the almost sure convergence of $\tilde{G}_n(\theta)$, the continuity of $\tilde{G}(\theta)$, and the fact that $\tilde{G}(\theta)$ is uniquely maximized in θ_0. Since all functions are continuous in the discrete topology, we are left with proving the last point. In order to rewrite $\tilde{G}(\theta)$, we show that

$$E\Big[YV\big[F_\theta(X^\intercal\theta) - F_\theta(X^\intercal\theta^-)\big]\Big] = E\big[Y_1\,\mathbb{1}[V_1 \geq V_2, X_1^\intercal\theta = X_2^\intercal\theta]\big] \tag{8}$$

$$= \frac{1}{2}E\big[Y_1\,\mathbb{1}[X_1^\intercal\theta = X_2^\intercal\theta]\big], \tag{9}$$

where (X_1, Y_1, V_1) and (X_2, Y_2, V_2) are i.i.d copies of (X, Y, V). Let $Q(t, v) \coloneqq P(X^\intercal\theta = t, V \leq v) = v[F_\theta(t) - F_\theta(t^-)]$, where the last equality is due to the independence and uniformity of V. (8) is obtained by noticing that $E\Big[YV\big[F_\theta(X^\intercal\theta) - F_\theta(X^\intercal\theta^-)\big]\Big] = E[YQ(X^\intercal\theta, V)]$. (9) is again a consequence of the independence and uniformity of V. Hence, we can rewrite $\tilde{G}(\theta)$ as

$$\tilde{G}(\theta) = E[Y_1\,\mathbb{1}[X_1^\intercal\theta > X_2^\intercal\theta]] + \frac{1}{2}E[Y_1\,\mathbb{1}[X_1^\intercal\theta = X_2^\intercal\theta]], \tag{10}$$

Using the law of total expectation, we have

$$\begin{aligned} 2\tilde{G}(\theta) = {} & E[H(X_1^\intercal\theta_0)\,\mathbb{1}[X_1^\intercal\theta > X_2^\intercal\theta]] + E[H(X_2^\intercal\theta_0)\,\mathbb{1}[X_2^\intercal\theta > X_1^\intercal\theta]] \\ & + E\left[\frac{H(X_1^\intercal\theta_0) + H(X_2^\intercal\theta_0)}{2}\,\mathbb{1}[X_1^\intercal\theta = X_2^\intercal\theta]\right]. \end{aligned}$$

If $p_j \coloneqq P(X = x_j)$ and $p_k \coloneqq P(X = x_k)$, we have

$$\begin{aligned} 2\tilde{G}(\theta) = \sum_{j,k} \Bigg[& H(x_j^\intercal\theta_0)\,\mathbb{1}[x_j^\intercal\theta > x_k^\intercal\theta] + H(x_k^\intercal\theta_0)\,\mathbb{1}[x_k^\intercal\theta > x_j^\intercal\theta] \\ & + \frac{H(x_j^\intercal\theta_0) + H(x_k^\intercal\theta_0)}{2}\,\mathbb{1}[x_j^\intercal\theta = x_k^\intercal\theta]\Bigg] p_j p_k. \end{aligned} \tag{11}$$

From this equation and by noticing that $H(x_j^\intercal\theta_0) > H(x_k^\intercal\theta_0)$ whenever $x_j^\intercal\theta_0 > x_k^\intercal\theta_0$, it is clear that $\tilde{G}(\theta)$ is maximized in θ_0^*. In order to show that the maximum is unique, suppose per contra that there exists $\theta_1^* \neq \theta_0^*$ which also maximizes $\tilde{G}(\theta)$. Then, there must exist at least one pair (x_j, x_k) such that

$$x_j^\intercal\theta_0^* > x_k^\intercal\theta_0^* \text{ and } x_j^\intercal\theta_1^* < x_k^\intercal\theta_1^* \qquad \text{if } \theta_1^* \in \{B_1, \ldots, B_l\}, \tag{12}$$

$$\text{or } x_j^\intercal\theta_1^* = x_k^\intercal\theta_1^* \qquad \text{if } \theta_1^* \in \{H_1, \ldots, H_m\}. \tag{13}$$

Since $x_j^\intercal\theta_0^* > x_k^\intercal\theta_0^*$, we have $H(x_j^\intercal\theta_0^*) > H(x_k^\intercal\theta_0^*)$. Let us now examine the contribution of (x_j, x_k) to (11) for θ_0^* and for θ_1^*. It consists of $H(x_j^\intercal\theta_0^*)$ for θ_0^*. For θ_1^*, it amounts either to $H(x_k^\intercal\theta_1^*)$ in the situation described by (12), or to $\frac{H(x_j^\intercal\theta_1^*)+H(x_k^\intercal\theta_1^*)}{2}$ in (13). In any case, this contradicts the fact that $\tilde{G}(\theta)$ is maximized in θ_1. □

As we have already stated, there exist other solutions to deal with the ties issue. It is then important to judge the relevance of our proposal compared to others. From (10), we observe that our random solution and the average rank method lead to the same objective function in the population. In practice, however, the two methods are not equivalent. A first argument in favor of the random solution is illustrated in Appendix B and briefly explained here. In order to specify the extent of identifiability of our model, we restricted the parameter space to a vector of representatives. Some of these are representatives of the open regions $B_1, \ldots, B_l$ while the others come from the hyperplanes $H_1, \ldots, H_m$. Working with the new parameter space Θ^* and $\theta_0 \in \{B_1, \ldots, B_l\}$ leads to identifiability up to a specific region while $\theta_0 \in \{H_1, \ldots, H_m\}$ needs a further merging operation to reach such an identifiability. Even though the situation where θ_0 lies in a hyperplane cannot be identified in the usual sense, we argue in Appendix B that the random solution has more chance to tend to the exact θ_0 in that latter situation than the average rank solution. A second interest of the random solution consists in its ability to judge the importance of the ties issue on the value taken by the explained Gini coefficient. The rest of this section is dedicated to this question.

Beyond statistical considerations, each procedure to solve the ties issue induces a different definition of the explained Gini coefficient and, hence, a different value for this quantity. Facing one dataset, it would be interesting to have an idea of how small or large the coefficient can get. In this respect,

our random assignment solution offers some help. We express this variability issue differently. In our case, variability is induced by our further ranking in terms of a vector of uniform random variables. If we repeat the generation of V a large number of times in the estimation of the explained Gini coefficient, the random solution will sweep through the different possibilities of attaching ranks to the ties. In this way, we internalize the variability related to the ties issue and are able to assess its influence. We formalize this procedure in the remaining of this section. Interestingly, we will see that, under mild assumptions, this variability is independent from the estimation of θ_0. As such, we repeat the generation of V when we are estimating the explained Gini coefficient, but using a fixed $\hat{\theta}$.

Fix a dataset $(x_i, y_i)_{i=1,\dots,n}$ and suppose that ties in the index only happen if $x_i = x_j$. We can write the corrected rank vector as

$$\tilde{R}_n^V(x_i^\intercal\theta) = R_i^M(\theta) + R_i^D(V), \tag{14}$$

where $R_i^M(\theta)$ is the rank of observation i obtained with the average-rank method and is defined as

$$R_i^M(\theta) \coloneqq \sum_{j=1}^{n} \left(\mathbb{1}[x_j^\intercal\theta < x_i^\intercal\theta] + \frac{1}{2}\,\mathbb{1}[x_j^\intercal\theta = x_i^\intercal\theta] \right) - \frac{1}{2}.$$

This part depends on θ but not on V and is hence deterministic. On the other hand, $R_i^D(V)$ is the random deviation from the average-rank for observation i. This part is a random variable with expected value 0 and variance $(n_i^2 - 1)/12$, where n_i is the number of observations with the same vector of covariates as x_i. However, this part does not depend on θ thanks to our simplifying assumption that ties occur independently of θ. We propose the following procedure. Choose M large. For $m = 1, \dots, M$, generate $V^m = (V_1^m, \dots, V_n^m)^\intercal$, with $V_i^m \overset{\text{i.i.d}}{\sim} U[0,1]$. For each iteration m, estimate the explained Gini coefficient as

$$\widehat{\text{Gi}}_{Y,X}^m = \frac{2}{n^2} \sum_{i=1}^{n} \frac{y_i}{\overline{y}} \tilde{R}_n^{V^m}(x_i^\intercal\hat{\theta}) - \frac{n+1}{n}, \tag{15}$$

where $\hat{\theta}$ does not depend on m, meaning that θ^* is estimated only once, sparing a lot of computation time. This choice is justified by (14), which entails that we can separate out the impact of V from the impact of θ in the computation of the rank vector. We illustrate this procedure in Section 5.2.

4. Inference about the Lorenz regression

In terms of inference, our main objective is twofold. First, we wish to construct confidence intervals and tests for the explained Gini coefficient. Second, we wish to test the significance of one or several covariates in explaining the inequality of the response. Since the explained Gini coefficient involves θ_0, we start by discussing the inference on the latter. Cavanagh and Sherman (1998) derive the asymptotic normality of the MRE in Theorem 2 and discuss the estimation of the asymptotic covariance matrix. However, as thoroughly discussed in Subbotin (2007), the proposed solutions have several drawbacks. To name one, they all involve crucial choices of smoothing parameters. This issue undermines the construction of confidence intervals on that basis and the development of a Wald test. An alternative lies in the use of bootstrap. The advantage of such method lies in the absence of any smoothing parameter. Interestingly, Subbotin (2007) thoroughly discusses the consistency of bootstrap procedures for the MRE. Confidence intervals and tests may then be constructed either by bootstrapping the asymptotic covariance matrix only (hybrid bootstrap), or rather by bootstrapping the distribution of $\hat{\theta}$ directly (basic bootstrap). Theorem 3 in Subbotin (2007) implies the consistency of the basic bootstrap while Theorem 4 from the same source guarantees the consistency of the hybrid bootstrap. A third method to construct confidence intervals consists in directly plugging the quantiles of the bootstrap distribution of $\hat{\theta}$ (percentile bootstrap). More precisely, $(1-\alpha)$-level confidence intervals for $\theta_{0,k}$ are given by

$$\begin{aligned}
\text{CI}_{\text{Basic}} &= \left[2\hat{\theta}_k - q_{\hat{\theta}^*_k;1-\frac{\alpha}{2}}; 2\hat{\theta}_k - q_{\hat{\theta}^*_k;\frac{\alpha}{2}}\right], \\
\text{CI}_{\text{Percentile}} &= \left[q_{\hat{\theta}^*_k;\frac{\alpha}{2}}; q_{\hat{\theta}^*_k;1-\frac{\alpha}{2}}\right], \\
\text{CI}_{\text{Hybrid}} &= \left[\hat{\theta}_k \pm z_{1-\frac{\alpha}{2}} \frac{\sqrt{\hat{\Sigma}^*_{kk}}}{\sqrt{n}}\right],
\end{aligned}$$

where $\hat{\theta}^*$ is the MRE in the bootstrap sample. Besides, $q_{\hat{\theta}^*_k;a}$ is the bootstrap estimator of the a-quantile of the distribution of $\hat{\theta}^*_k$ and $\hat{\Sigma}^*$ is the bootstrap estimator of the asymptotic variance-covariance matrix of $\hat{\theta}^*$. Finally, z_a is the a-quantile of the standard normal distribution. With this in mind, it is easy to obtain bootstrap confidence intervals and to build tests concern-

ing $\text{Gi}_{Y,X}$. For example, we could test the equality of the explained Gini coefficients between two countries.

Central to the motivation of this paper is the idea of assessing whether a set of covariates explains significantly the inequality of the response. Accordingly, we develop a bootstrap test for the joint significance of d variables. Assume $Y_i = H(X_i^\intercal \theta_0) + \epsilon_i$, with H increasing, and $E[\epsilon_i | X_i] = 0$. Formally, we wish to test $H_0 : \theta_{0,k}, \ldots, \theta_{0,k+d-1} = 0$ vs. $H_1 : \exists j \in [0, d-1]$ such that $\theta_{0,k+j} \neq 0$. The idea underlying the following testing procedure lies in comparing the Lorenz-R^2 under the null hypothesis with the unconstrained one. Accordingly, we will reject H_0 if dropping the d variables leads to a significant decrease in explained inequality. Specifically, we use

$$U \coloneqq \frac{\text{LR}^2}{\text{LR}^2_{H_0}} = \frac{\frac{1}{n^2}\sum_{i=1}^n Y_i R_n(X_i^\intercal \hat{\theta}) - \frac{\overline{Y}}{2}}{\frac{1}{n^2}\sum_{i=1}^n Y_i R_n(X_i^\intercal \hat{\theta}^{(0)}) - \frac{\overline{Y}}{2}},$$

where

$$\hat{\theta}^{(0)} \coloneqq \underset{\theta \in \Theta_{-\{k,\ldots,k+d-1\}}}{\arg\max} \sum_{i=1}^n Y_i R_n \left(X_i^{(0)\intercal} \theta \right),$$

and $X_i^{(0)}$ is obtained by dropping columns k to $k+d-1$ from X_i. Intuitively, we will reject H_0 if the observed value of U is sufficiently larger than 1. The residuals under H_0 are obtained as

$$\hat{\epsilon}_i^{(0)} = Y_i - \hat{H}^{(0)} \left(X_i^{(0)\intercal} \hat{\theta}^{(0)} \right),$$

where $\hat{H}^{(0)}$ is an estimator of H under H_0. Then, $\epsilon_i^{(0)*}$ is randomly drawn with replacement from $\hat{\epsilon}_1^{(0)}, \ldots, \hat{\epsilon}_n^{(0)}$. The bootstrap sample is obtained as $(X_i, Y_i^*)_{i=1}^n$, where

$$Y_i^* = \hat{H}^{(0)} \left(X_i^{(0)\intercal} \hat{\theta}^{(0)} \right) + \epsilon_i^{(0)*},$$

We can then approximate the p-value of our test using $p^* = P(U^* \geq U_{\text{obs}})$, where U_{obs} is the observed value of U in the original sample and U^* is obtained as

$$U^* = \frac{\frac{1}{n^2}\sum_{i=1}^n Y_i^* R_n(X_i^\intercal \hat{\theta}^*) - \frac{\overline{Y^*}}{2}}{\frac{1}{n^2}\sum_{i=1}^n Y_i^* R_n(X_i^\intercal \hat{\theta}^{(0)*}) - \frac{\overline{Y^*}}{2}},$$

where $\hat{\theta}^*$ is the unconstrained MRE in the bootstrap sample and $\hat{\theta}^{(0)*}$ the constrained one. Finally, p^* can be estimated via Monte-Carlo simulations.

The described procedure opens the door to the estimation of the link function H. Upon estimation of θ_0 with the MRE, we have at our disposal an estimated index $X^\intercal\hat{\theta}$. In a second stage, we can estimate H as a nonparametric smoother of Y given $X^\intercal\hat{\theta}$, subject to the constraint that H is increasing. To achieve this, we can use the procedure laid down by Chernozhukov et al. (2009), hereafter called rearrangement method, which starts from an initial estimator of H and makes it increasing by taking its quantile function. Interestingly, this rearrangement operation weakly reduces the estimation error of the initial estimator, see Proposition 1 in Chernozhukov et al. (2009).

5. Monte-Carlo simulations

This section is divided into three Monte-Carlo exercises. In the first, we implement the MRE with a genetic algorithm and compare its performance with the semiparametric least squares (SLS) estimator defined in Ichimura (1993). In the second, we illustrate the procedure laid down in Section 3 to assess the impact of ties on the range of values taken by the explained Gini coefficient. Finally, we examine the performance of the bootstrap.

5.1. Performance of the estimation

The computation of the MRE requires to solve (7), a discrete and hence non-convex maximization programme. In this respect, our contribution is to propose a genetic algorithm with the double advantage of being fast and reducing the risk of local minima. The detailed functioning of the algorithm as well as a pseudo code are provided in Appendix C. We judge the quality of the estimation via three criteria : the mean L2-distance between the true and the estimated θ, the mean integrated squared error (MISE) of the regression curve and, most importantly, the mean squared error (MSE) of the explained Gini coefficient. These last two quantities are given by

$$\text{MISE.Curve} \coloneqq E\left[\int \left(H(x^\intercal\hat{\theta}) - H(x^\intercal\theta_0)\right)^2 dF_X(x)\right],$$
$$\text{MSE.Gini} \coloneqq E\left[\left(\widehat{\text{Gi}}_{Y,X} - \text{Gi}_{Y,X}\right)^2\right],$$

where F_X is the joint CDF of the X vector. These criteria are estimated with 1000 Monte-Carlo replications. We use the following data generating process (DGP)

$$Y_i = H\left(\theta_1 X_i^1 + \ldots + \theta_c X_i^c + \theta_{c+1} Z_i^1 + \ldots + \theta_{c+d} Z_i^d\right) \epsilon_i, \tag{16}$$

where $i = 1, \ldots, n$ and $||\theta|| = 1$. The X_i's are c standard normal variables while the Z_i's are d Bernoulli variables with $1/2$ probability of success. Finally, ϵ_i is a lognormal noise with mean 1 and a variance set to ensure a signal-to-noise ratio of 3. Throughout this section, we consider the following link function

$$H(t) = 1000 \exp\left[1 + \frac{1}{2}(t-1)^3\right].$$

Ichimura's method is available in the `R` package `np`. Importantly, this package uses a different normalization for θ. There, the first element of the vector is fixed to one, while we rather impose a unit norm. Both setups are comparable if we properly adjust the vector of parameters and the link function. To see this, consider a dataset generated with the DGP presented at (16). This exact dataset can be generated with the single-index model using the other normalization, with vector of parameters $\theta^* = \theta/\theta_1$ and link function $H^*(t) = H(\theta_1 t)$.

In what follows, we use 4 continuous and 2 discrete covariates. The parameter vector is $\theta = (0.0923, -0.0889, 0.0506, 0.2838, -0.2074, 0.2270)^\intercal$. We focus first on the explained Gini coefficient. In that respect, it is important to mention that under the SLS estimator, the link function does not need to be increasing. This implies that the ordering given by the index $X^\intercal\hat{\theta}$ does not necessarily coincide with the ordering implied by the regression curve $\hat{H}(X^\intercal\hat{\theta})$. Let us illustrate this with a simple example. Assume that the first covariate X_1 has a negative marginal impact on the response. Recalling that the SLS estimator forces $\theta_1 = 1$, we expect the estimated link function to be decreasing. In this context, we should order individuals in terms of $-X^\intercal\hat{\theta}$. To bypass this issue, we rank the observations in terms of $\hat{H}(X^\intercal\hat{\theta})$. Of course, this issue does not appear in a Lorenz regression since the estimated link function will always preserve the ranking provided by the index. The boxplot of the explained Gini coefficient estimated with each method is displayed in Figure 1. The horizontal line corresponds to the actual value

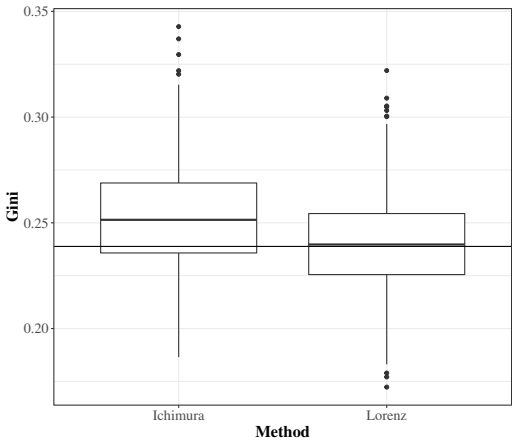

Figure 1: Distribution of $\widehat{\text{Gi}}_{Y,X}$

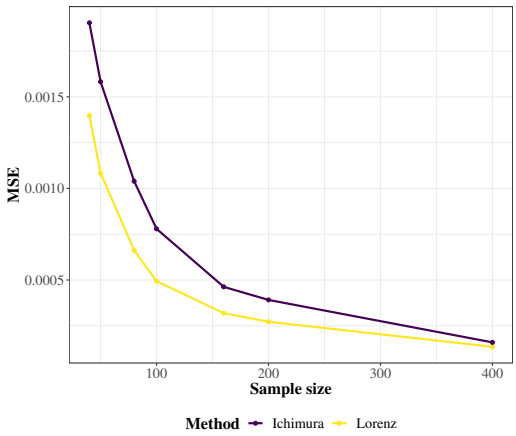

Figure 2: MSE of $\widehat{\text{Gi}}_{Y,X}$

of the parameter, i.e. around 23.88%. While the distribution of Lorenz estimates is centered on the true value, this is not the case for the SLS estimator. Since the link function is not restricted to be increasing, the SLS estimator tends to overfit the data and, hence, to overestimate the explained Gini coefficient. Figure 2 reinforces our conclusion that Lorenz regression provides a better estimator, even though the gap closes down with sample size.

We now compare the two estimators in terms of the other two criteria. Figure 3 and 4 show their evolution with sample size. They confirm the better performance of Lorenz regression since the L2-distance is always lower in comparison with the SLS estimator. A similar pattern emerges for the estimation of the regression curve.

5.2. Range of the Explained Gini coefficient in presence of ties

With this example, we strive first to highlight the definition problem of the explained Gini coefficient inherent to the discrete case. Second, we wish to illustrate the procedure explained at the end of Section 3. Note that the point of the exercise is not to assess the estimation quality of the explained Gini coefficient. We work on one dataset, generated with the same DGP and link function as before and concentrate on a pure discrete scenario with $d = 4$. We generate $M = 1000$ uniform random vectors $V^1, \ldots, V^M$ and use these to estimate a vector of explained Gini coefficient, following (15). This computation is undertaken using θ_0, i.e. the value we used to generate the

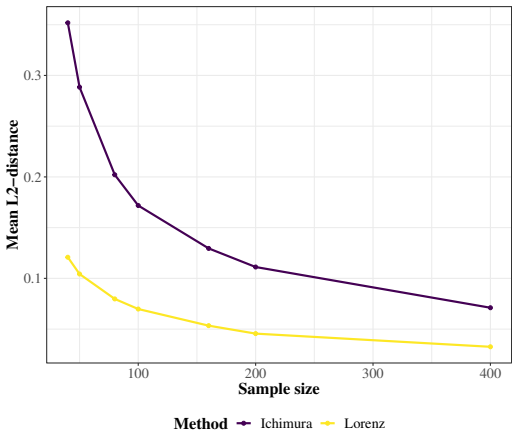

Figure 3: Mean L2-distance between $\hat{\theta}$ and θ_0

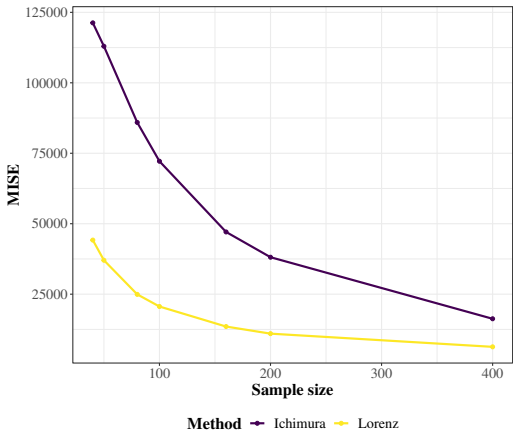

Figure 4: MISE of $\hat{H}(X^\intercal\hat{\theta})$

data, as well as using $\hat{\theta}$. Figure 5 summarizes our results. Both boxplots display the distribution of the explained Gini coefficient estimated with the random solution. In the first, we plug the estimated vector of parameters $\hat{\theta}$. In the second, we rather use θ_0. The dashed and dotted lines represent the explained Gini coefficients obtained with the average-rank method for θ_0 and $\hat{\theta}$ respectively. As expected, the two boxplots are centered on these lines. On average, the rank-method provides the same explained Gini coefficient as the average-rank solution. We also observe that the two boxplots display the same variability. This stems from the fact that the variable part of the explained Gini coefficient is independent from the value taken by the parameter vector θ. Finally, the total size of the boxplots indicate that the ties issue is not too severe. Indeed, their range is only slightly larger than 0.003.

5.3. Performance of the bootstrap

We assess the performance of the bootstrap test of joint significance using the same DGP as in Section 5.1, again with 4 continuous and 2 discrete covariates. The bootstrap procedure described in Section 4 can still be used, by transforming the response Y into $\log(Y)$ before estimation. We test the joint significance of θ_4 and θ_5, with a significance level of $\alpha = 0.05$ and 500 Monte-Carlo replications for the bootstrap. Specifically, we use the following vector of parameters.

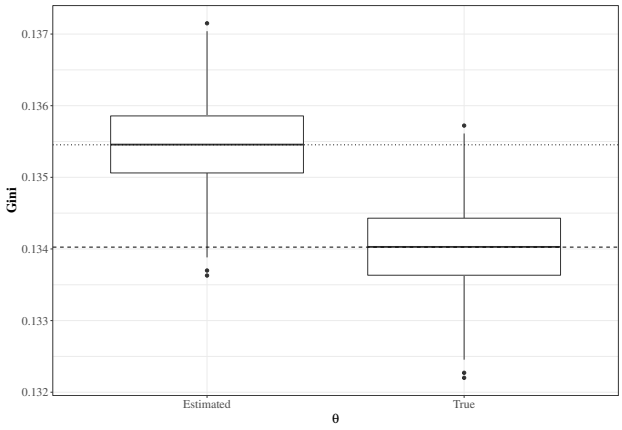

Figure 5: Distribution of the Explained Gini coefficient for different generations of V

	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6
Under H_0	0.5504	0.1790	-0.1724	0	0	0.0982
Under H_1	0.44032	0.14320	-0.13792	0.1	0.1	0.07856

In what follows, we quantify the power of our testing procedure. It is interesting to examine the power if we were to know the distribution of the test statistic, hereafter mentioned as theoretical power, isolating the power loss due to bootstrapping. Figure 6 displays the power curves estimated on 600 Monte-Carlo replications. As expected, both power curves converge to 1 as the sample size increases. The vertical distance between the two curves illustrate the power loss related to bootstrapping. Though this difference is not negligible for small and medium sample sizes, it is never larger than 22% and it closes down relatively quickly. With a sample of 100 observations, our procedure already attains a power of 96.5%. In Section 2 of the Online Supplement, we appraise the performance of the bootstrap confidence intervals, both for one element of θ and for the explained Gini coefficient.

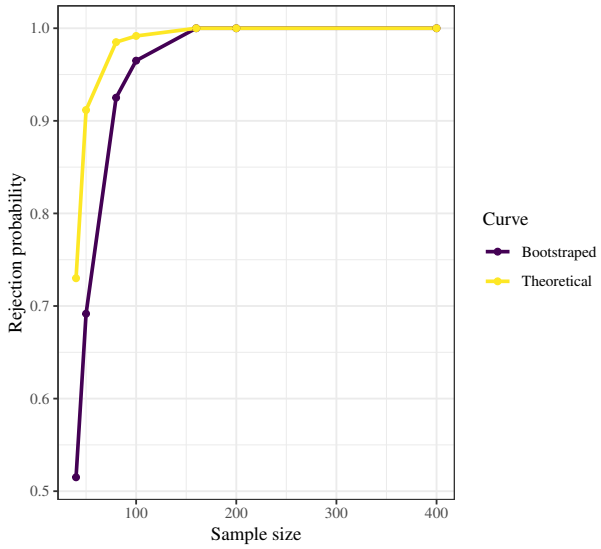

Figure 6: Theoretical and bootstrapped power curves

6. Empirical illustration

In economics, attention has often been directed to identify the main determinants of wages. A classical starting point is the wage equation proposed by Mincer and Polachek (1974) of the form

$$\log(W) = \beta_0 + \beta_1 S + \beta_2 E + \beta_3 E^2 + \epsilon, \tag{17}$$

where W is the wage, S is schooling, E is professional experience and ϵ is an unobservable error term. Focusing on these two covariates is of course restrictive and many papers went beyond this basic setup. For example, Griliches (1976) drew attention on a bias which stems from ignoring ability while it is at the same time expected to be influenced by schooling and to have an effect on wages. In this section, we apply the Lorenz regression methodology on such determinants with the objective of visualizing the inequality that they help to reproduce.

Our discussion is based on data resulting from the Young Men's Cohort of the National Longitudinal Survey (NLS-Y), a survey started in 1966 on individuals of ages 14-24. The excerpt we use is available in the dataset `Griliches` contained in the `R` package `Ecdat`. Besides wage, schooling and

experience, we include the following variables in our analysis: age (A), ability (IQ), marital status (MS) and degree of urbanisation (DU). Ability was computed as IQ scores collected in a school survey conducted in 1968. All the remaining variables were observed in 1980. The degree of urbanisation is a dummy determining whether the individual lives in a metropolitan area. Before delving into modelling, we note that the considered wage distribution exhibits a mean of 1000, a median of 948 and a Gini coefficient of 0.222.

We first restrict the attention to schooling and experience, allowing for a quadratic term on experience. The underlying models are presented in (18) and (19) and are more general than Mincer and Polachek (1974) equation since they do not impose a log link function.

$$E[W|S,E] = H_{[1]}\left(\theta_S S + \theta_E E + \theta_{E^2} E^2\right), \tag{18}$$

$$E[W|S,E] = H_{[2]}\left(\theta_S S + \theta_E E\right). \tag{19}$$

Table 2 displays the estimated parameters, bootstrap standard deviations

Table 2: Lorenz regressions for (18) and (19)

	(18)			(19)		
	Estimate	Std dev	CI	Estimate	Std dev	CI
θ_S	0.552	0.133	[0.398;0.912]	0.786	0.055	[0.733;0.908]
θ_E	0.437	0.237	[-0.279;0.584]	0.214	0.055	[0.092;0.267]
θ_{E^2}	-0.011	0.011	[-0.018;0.020]			

and 95% percentile bootstrap confidence intervals corresponding to (18) and (19). We observe that only schooling is significant when a quadratic experience term is included. However, experience becomes significant when the quadratic term is dropped. As expected, the Lorenz-R^2 decreases from (18) to (19) very slightly, from 0.391 to 0.389.

As a second step, we augment our model of several control variables. In (20), we introduce marital status, degree of urbanisation, age and ability, as

measured by IQ.

$$E[W|S, MS, DU, A, E, IQ] = H_{[4]}(\theta_S S + \theta_{MS} MS + \theta_{DU} DU + \theta_A A + \theta_E E + \theta_{IQ} IQ). \quad (20)$$

Table 3 displays the output of the Lorenz regression corresponding to (20),

Table 3: Lorenz and log-linear regressions for (20)

	Lorenz regression			Log-linear regression		
	Estimate	Std dev	CI	Estimate	Std dev	CI
Intercept	/	/	/	4.689	0.182	[4.332;5.047]
θ_S	0.123	0.030	[0.086;0.206]	0.054	0.008	[0.038;0.069]
θ_{MS}	0.372	0.117	[0.034;0.490]	0.169	0.044	[0.082;0.255]
θ_{DU}	0.429	0.097	[0.344;0.686]	0.202	0.029	[0.145;0.260]
θ_A	0.029	0.021	[0.002;0.085]	0.014	0.005	[0.004;0.024]
θ_E	0.037	0.016	[0.008;0.067]	0.016	0.004	[0.008;0.024]
θ_{IQ}	0.01	0.005	[0.005;0.024]	0.004	0.001	[0.002;0.007]

as well as the results of a linear regression on log-wages, using the same covariates. Both methods offer similar conclusions. However, the Lorenz regression provides us with further interpretations. The Lorenz-R^2 is of 0.48, indicating that we can reproduce 48% of the observed inequality with our covariates. Figure 7 displays the observed and explained Lorenz curves. By computing marginal rates of substitution, we can also compare the magnitude of two weights. For example, the MRS of DU with respect to MS is of 1.153 meaning that, all other things being equal, the degree of urbanisation has a marginal impact on wages 1.153 times more important than marital status. Also, the MRS of S with respect to E is of 3.324, meaning that, all other things being equal, spending one more year of schooling has three-times more impact than a further year of professional experience.

Interestingly, our dataset contains a variable recording the individual's residency. Namely, we know if she lives in a Southern or Northern state. Dividing

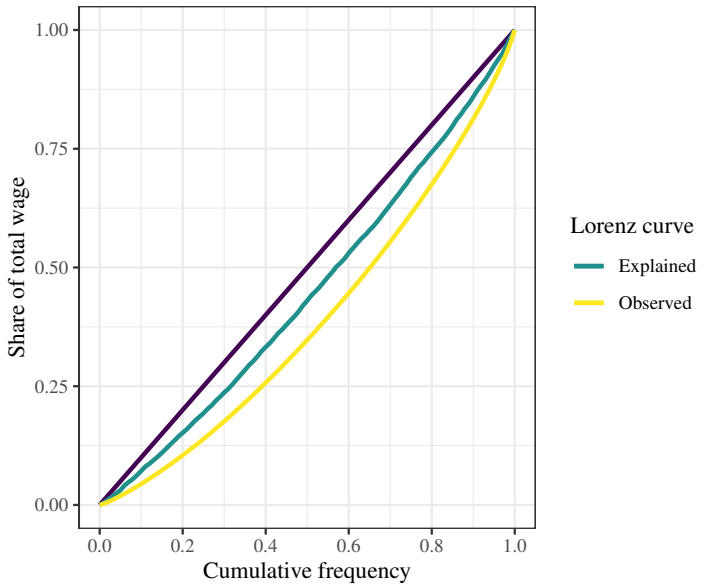

Figure 7: Observed and explained Lorenz curve

our dataset on the basis of this information and using model (20), we fit two distinct Lorenz regressions. This allows us to compare the explained Gini coefficient in Northern and Southern states. Results are gathered in Table 4. The Gini coefficient obtained when ranking individuals in terms of explained income is higher in Southern states (13.47% against 9.96% in Northern states. In some sense, this was expected since the Gini coefficient of wages is higher in Southern states (25.17% against 20.96% in Northern states). Still, it is worth mentionning that the unexplained part of inequality, i.e. the difference between the actual and the explained Gini coefficients, is rather similar between the two situations (11.17% in Southern states against 11% in Northern states). This stems from the fact that we explain a larger proportion of the observed inequality in the Northern states. Indeed, the Lorenz-R^2 there is of 53.5% against 47.53% in Southern states.

Having estimated the index using the MRE on (20), we may estimate the link function using the rearrangement method and compare it with an exponential fit. The latter is obtained from a classical linear regression of log-wages on the estimated index. More specifically, we assume $\log(W) = \alpha + \beta T + \varepsilon$, where T is the estimated index, $E[\varepsilon|T] = 0$ and ε is independent from T.

Table 4: Lorenz regressions distinguishing between Northern and Southern states

	South	North		South	North
θ_S	0.212	0.097	Lorenz-R^2	0.535	0.4753
θ_{MS}	0.255	0.253	$\widehat{\text{Gi}}_Y$	0.2517	0.2096
θ_{DU}	0.431	0.570	$\widehat{\text{Gi}}_{Y,X}$	0.1347	0.0996
θ_A	0.003	0.056	$\widehat{\text{Gi}}_Y - \widehat{\text{Gi}}_{Y,X}$	0.117	0.11
θ_E	0.094	0.011			
θ_{IQ}	0.006	0.013			

Denote by $\hat{\alpha}$ and $\hat{\beta}$ the OLS estimators of α and β. We have $E[W|T] = \exp(\alpha + \beta T)E[\exp(\varepsilon)]$. The first piece is estimated using $\exp(\hat{\alpha} + \hat{\beta}T)$ while the second is estimated by the empirical mean of the $W/\exp(\hat{\alpha} + \hat{\beta}T)$ vector. As we can observe on Figure 8, both curves exhibit the same behaviour.

The purpose of this empirical example was to provide a first illustration of the Lorenz regression and the interpretations that can be build upon. In this case, this methodology yields quite similar results to a linear regression on log-wages. This provides some evidence supporting the logarithm as link function in the wage equation proposed by Mincer and Polachek (1974). This closeness is also attributable to the low inequality present in the wage distribution. Facing a distribution with more inequality, typically containing a few very rich individuals, we can expect linear and Lorenz regressions to yield quite different results. Because of its sensitivity to outliers, linear regression would be greatly driven by the richest individuals. With its part reliance on ranks, we can expect Lorenz regression to be less affected by such outliers.

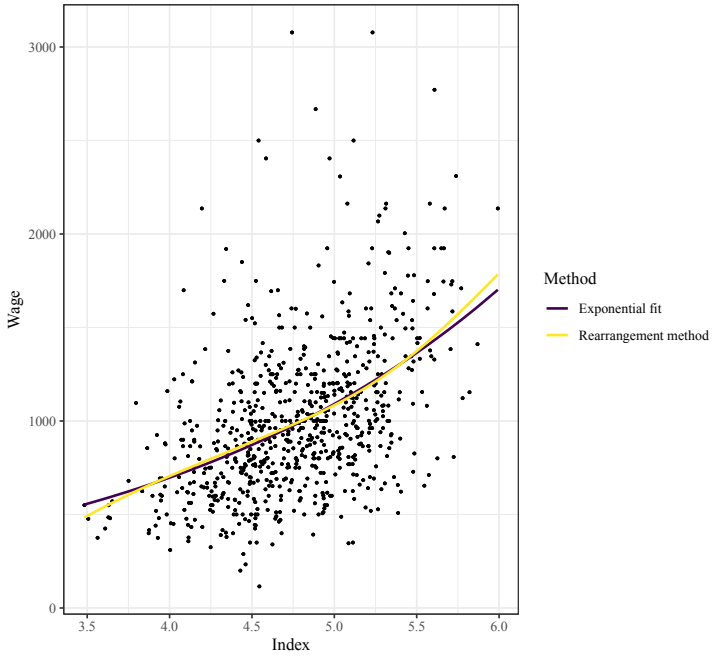

Figure 8: Nonparametric and parametric fit of the wage equation

Acknowledgments

Computational resources have been provided by the supercomputing facilities of the Université catholique de Louvain (CISM/UCL) and the Consortium des Équipements de Calcul Intensif en Fédération Wallonie Bruxelles (CÉCI) funded by the Fond de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under convention 2.5020.11.

Appendix A. The genetic algorithm

The genetic algorithm iteratively proposes solution candidates and assesses their score through a fitness function. Convergence is met when the fitness has not improved for a sufficiently large number of generations. The `R` library `GA` (Scrucca, 2013) allows us to build such algorithm. The only technical difficulty lies in the incorporation of the unit-norm constraint. To address this issue, we propose the following principles.

(a) We ensure that the initial population of candidates satisfies the unit-norm constraint.

(b) Candidates γ are vectors of size $p-1$ defining two possible solutions $\tilde{\theta}_1$ and $\tilde{\theta}_2$, where the last element is either $-(1-||\gamma||)$ or $(1-||\gamma||)$.

(c) The fitness function exhibits a penalty to ensure that candidates unable to satisfy the unit-norm constraint are discarded. Formally, the fitness function is defined by

$$f(\theta) \coloneqq \sum_{i=1}^{n} Y_i R_n(X_i^\intercal \theta) - K|\ ||\theta|| - 1\ |,$$

where K is a sufficiently large constant. Figure A.9 presents the pseudo-code of the algorithm. We use the notation α for a scalar, $\alpha[]$ for a vector, $\alpha[,]$ for a matrix and $\alpha[,,]$ for a three-dimensional array. The function `NextGeneration` describes the passage from one generation to another using the default choices of the package, i.e. arithmetical crossover and uniform mutation. Finally, `run` is the number of consecutive run without improvement triggering the algorithm to stop and ϵ is a tolerance level based on machine precision.

Appendix B. An advantage of the random-rank method

In this appendix, we provide an example illustrating the better performance of the random solution compared to the average-rank method.

Consider data generated from the DGP used in Section 5.1, with only two discrete covariates and using parameter vector $\theta_0 = (\frac{1}{2}, \frac{1}{2})^\intercal$. In such a setup, the number of bounding regions is quite low and a discrete algorithm could be used to estimate the parameter vector. Besides θ_0, we consider two alternative vectors of parameters

$$\theta_1 \coloneqq \left(\frac{3}{4}, \frac{1}{4}\right)^\intercal,$$
$$\theta_2 \coloneqq \left(\frac{1}{4}, \frac{3}{4}\right)^\intercal,$$

which, without loss of generality, are the representatives of their regions. Finally, we suppose that our sample is made of $n = 20$ observations and consists in a balanced design, i.e. 5 observations in each of the 4 possible values of X. In what follows, we focus on observations for which $X = (1,0)^\intercal =: x_1$

Input : A vector $Y[]$ of size n and a matrix X of size $n \times p$
Output: A solution vector $\hat{\theta}^*[]$ of size p

Initial population of S candidates, gathered in $\gamma[0,,]$;
for $s \leftarrow 1$ **to** S **do**
  for $j \leftarrow 1$ **to** p **do**
    $u[s,j] \sim U[-1,1]$;
    $v[s,j] \leftarrow \frac{u[s,j]}{||u[s,.]||}$;
  end
  $\gamma[0,s,] \leftarrow (v[s,1], v[s,2], \ldots, v[s,p-1])$;
end
The algorithm goes on until convergence is met;
for $k \geq 1$ **do**
  $\gamma[k,,] \leftarrow$ `NextGeneration(`$\gamma[k-1,,]$`)`;
  $\tilde{\theta}_1[k,,] \leftarrow (\gamma[k,,1], \gamma[k,,2], \ldots, \gamma[k,,p-1], 1 - ||\gamma[k,,j]||)$;
  $\tilde{\theta}_2[k,,] \leftarrow (\gamma[k,,1], \gamma[k,,2], \ldots, \gamma[k,,p-1], -(1 - ||\gamma[k,,j]||))$;
  The fitness of the candidate is given by;
  $\text{Fit}[k,] \leftarrow \max(f(\tilde{\theta}_1[k,,]), f(\tilde{\theta}_2[k,,]))$;
  The algorithm stops if there are enough consecutive runs without improvement;
  $\text{Fit.max}[k] \leftarrow \max_s(\text{Fit}[k,s])$;
  $\text{run.attained} \leftarrow \sum_{l \leq k} \mathbb{1}\{\text{Fit.max}[l] \geq \max_m(\text{Fit.max}[m]) - \epsilon\}$;
  if *run.attained* $\geq$ `run` **then**
    break;
  end
end
Denote by $\gamma^*[]$ the solution given by the algorithm;
$\theta_1^*[] \leftarrow (\gamma^*[1], \gamma^*[2], \ldots, \gamma^*[p-1], 1 - ||\gamma^*[]||)$;
$\theta_2^*[] \leftarrow (\gamma^*[1], \gamma^*[2], \ldots, \gamma^*[p-1], -(1 - ||\gamma^*[]||))$;
$\hat{\theta}^*[] \leftarrow \arg\max(\theta_1^*[], \theta_2^*[])$;

Figure A.9: Pseudo code of the genetic algorithm used to find the MRE

and those for which $X = (0,1)^\intercal =: x_2$. In this way, $x_1^\intercal\theta_0 = x_2^\intercal\theta_0$, while $x_1^\intercal\theta_1 > x_2^\intercal\theta_1$ and $x_1^\intercal\theta_2 < x_2^\intercal\theta_2$. Table B.5 shows the ranks that each method attaches to those observations if the index is built upon either θ_0, θ_1 or θ_2. Note that $U\{.,.\}$ denotes the discrete uniform distribution. For example, each observation with value x_1 is given a rank of 10.5 if we are considering θ_0 and the average-rank solution. On the other hand, they have a rank randomly drawn between 6 and 15 if we are using the random solution.

Table B.5: Rank granted to each group of observations

	Average rank		Random rank	
	x_1	x_2	x_1	x_2
θ_0	10.5	10.5	$U\{6,15\}$	$U\{6,15\}$
θ_1	13	8	$U\{11,15\}$	$U\{6,10\}$
θ_2	8	13	$U\{6,10\}$	$U\{11,15\}$

The derivation of the MRE requires to maximize the scalar product between the response and the vector of index ranks. Denote by $\overline{y}_1$ ($\overline{y}_2$) the mean response for observations with covariate vector x_1 (x_2). From the table, it is clear that θ_0 will never be chosen by the average-rank solution, unless $\overline{y}_1 = \overline{y}_2$. Indeed, if $\overline{y}_1 > \overline{y}_2$, we would pick θ_1. Conversely, if $\overline{y}_1 < \overline{y}_2$, we would pick θ_2. This is not the case with the random rank solution since it depends on the generation of the uniform random variables. We performed 1000 Monte-Carlo simulations to confirm this intuition, using the aforementioned DGP but relaxing the assumption of balanced design. In each iteration, we recorded the maximizer(s) of the objective function. Table B.6 counts the number of times θ_0, θ_1 and θ_2 maximize the objective function.

As we can observe, θ_0 is almost never chosen with the average-rank solution. The 8 instances where θ_0 maximizes the objective function correspond to situations where there is no observation either in x_1 or x_2. In this case, the three values of θ provide the same value for the objective function. This is not the case for the random solution. In total, θ_0 maximizes the objective

Table B.6: Number of times each parameter vector maximizes the objective function

	θ_0	θ_1	θ_2
Average rank	8	505	503
Random rank	177	435	420

function in 177 instances. Beyond the same 8 instances where θ_0, θ_1 and θ_2 all reach the optimum, there are 16 instances where θ_0 and one of either θ_1 or θ_2 maximize the objective function. This situation occurs when the generation of the uniform random variables ends up yielding the same rank vectors for θ_0 and for either θ_1 or θ_2. Hence, the 153 remaining instances correspond to situations where θ_0 is the sole maximizer of the objective function. Recall that it never occurs with the average rank method. The overall conclusion of this exercise runs as follows: the average-rank solution discriminates against certain types of solutions. These solutions are characterized by equal values of the index but made up with different values of the covariates. While the random solution remains neutral about this, the average-rank solution tries to disaggregate the index as much as possible.

References

Aaberge, R., Bjerve, S., Doksum, K., 2005. Decomposition of rank-dependent measures of inequality by subgroups. Metron - International Journal of Statistics LXIII, 493–503.

Blitz, R., Brittain, J., 1964. An extension of the Lorenz diagram to the correlation of two variables. Metron XXIII, 137–143.

Bourguignon, F., Ferreira, F.H.G., Menéndez, M., 2007. Inequality of Opportunity in Brazil. Review of Income and Wealth 53, 585–618.

Cavanagh, C., Sherman, R.P., 1998. Rank estimators for monotonic index models. Journal of Econometrics 84, 351–381.

Chernozhukov, V., Fernández-Val, I., Galichon, A., 2009. Improving point and interval estimators of monotone functions by rearrangement. Biometrika 96, 559–575.

Das Gupta, S., 1999. Gini association and pseudo lorenz curve. Communications in Statistics - Theory and Methods 28, 2181–2199.

Denuit, M., Huang, R.J., Wang, C., 2014. Almost marginal conditional stochastic dominance. Journal of Banking & Finance 41, 57.

van Doorslaer, E., Koolman, X., 2004. Explaining the differences in income-related health inequalities across European countries. Health Economics 13, 609–628.

Fei, J.C.H., Ranis, G., Kuo, S.W.Y., 1978. Growth and the Family Distribution of Income by Factor Components. The Quarterly Journal of Economics 92, 17–53.

Fleurbaey, M., Schokkaert, E., 2009. Unfair inequalities in health and health care. Journal of Health Economics 28, 73–90.

Garner, T.I., 1993. Consumer Expenditures and Inequality: An Analysis Based on Decomposition of the Gini Coefficient. The Review of Economics and Statistics 75, 134–138.

Griliches, Z., 1976. Wages of Very Young Men. Journal of Political Economy 84, S69–S85.

Horowitz, J.L., 2009. Single-Index Models, in: Horowitz, J.L. (Ed.), Semiparametric and Nonparametric Methods in Econometrics. Springer-Verlag, New York. Springer Series in Statistics, pp. 7–51.

Ichimura, H., 1993. Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. Journal of Econometrics 58, 71–120.

Jones, A.M., Nicolás, A.L., 2004. Measurement and explanation of socioeconomic inequality in health with longitudinal data. Health Economics 13, 1015–1030.

Lerman, R.I., Yitzhaki, S., 1985. Income Inequality Effects by Income Source: A New Approach and Applications to the United States. The Review of Economics and Statistics 67, 151–156.

Lerman, R.I., Yitzhaki, S., 1989. Improving the accuracy of estimates of Gini coefficients. Journal of Econometrics 42, 43–47.

Milanovic, B., 2016. What Next?, in: Global Inequality: A New Approach for the Age of Globalization. Harvard University Press, Cambridge, Massachusetts, pp. 212–239.

Mincer, J., Polachek, S., 1974. Family Investment in Human Capital: Earnings of Women. Journal of Political Economy 82, S76–S108.

Powell, J.L., Stock, J.H., Stoker, T.M., 1989. Semiparametric Estimation of Index Coefficients. Econometrica 57, 1403–1430.

Roemer, J.E., 1998. Equality of Opportunity. Harvard University Press.

Schechtman, E., Zitikis, R., 2006. Gini indices as areas and covariances: What is the difference between the two representations? Metron - International Journal of Statistics LXIV, 385–397.

Scrucca, L., 2013. GA: A Package for Genetic Algorithms in R. Journal of Statistical Software 53, 1–37.

Shalit, H., Yitzhaki, S., 1984. Mean-Gini, Portfolio Theory, and the Pricing of Risky Assets. The Journal of Finance 39, 1449–1468.

Shalit, H., Yitzhaki, S., 2003. An Asset Allocation Puzzle: Comment. American Economic Review 93, 1002–1008.

Subbotin, V., 2007. Asymptotic and Bootstrap Properties of Rank Regressions. SSRN Scholarly Paper ID 1028548. Social Science Research Network. Rochester, NY.

Wagstaff, A., van Doorslaer, E., Watanabe, N., 2003. On decomposing the causes of health sector inequalities with an application to malnutrition inequalities in Vietnam. Journal of Econometrics 112, 207–223.

Yitzhaki, S., 1994. On The Progressivity of Commodity Taxation, in: Eichhorn, W. (Ed.), Models and Measurement of Welfare and Inequality. Springer Berlin Heidelberg, pp. 448–466.

Yitzhaki, S., Schechtman, E., 2013. The Lorenz Curve and the Concentration Curve, in: Yitzhaki, S., Schechtman, E. (Eds.), The Gini Methodology: A Primer on a Statistical Methodology. Springer New York, New York, NY. Springer Series in Statistics, pp. 75–98.