DECLARATORIA SULLA TESI DI DOTTORATO

La sottoscritta

COGNOME	Raffinetti
NOME	Emanuela
Matr.	1194984

Titolo della tesi:

Multivariate dependence measures through Lorenz curves and their generalization

Dottorato di ricerca in	Statistica
Ciclo	XXII
Tutor del dottorando	Prof. Paolo Giudici
Anno di discussione	2011

DICHIARA

sotto la sua responsabilità di essere a conoscenza:

1) che, ai sensi del D.P.R. 28.12.2000, N. 445, le dichiarazioni mendaci, la falsità negli atti e l'uso di atti falsi sono puniti ai sensi del codice penale e delle Leggi speciali in materia, e che nel caso ricorressero dette ipotesi, decade fin dall'inizio e senza necessità di nessuna formalità dai benefici previsti dalla presente declaratoria e da quella sull'embargo;

2) che l'Università ha l'obbligo, ai sensi dell'art. 6, comma 11, del Decreto Ministeriale 30 aprile 1999 prot. n. 224/1999, di curare il deposito di copia della tesi finale presso le Biblioteche Nazionali Centrali di Roma e Firenze, dove sarà consentita la consultabilità, fatto salvo l'eventuale embargo legato alla necessità di tutelare i diritti di enti esterni terzi e di sfruttamento industriale/commerciale dei contenuti della tesi;

3) che il Servizio Biblioteca Bocconi archivierà la tesi nel proprio Archivio istituzionale ad Accesso Aperto e che consentirà unicamente la consultabilità on-line del testo completo (fatto salvo l'eventuale embargo);

4) che per l'archiviazione presso la Biblioteca Bocconi, l'Università richiede che la tesi sia consegnata dal dottorando alla Società NORMADEC (operante in nome e per conto dell'Università) tramite procedura on-line con contenuto non modificabile e che la Società Normadec indicherà in ogni piè di pagina le seguenti informazioni:

- tesi di dottorato "*Multivariate dependence measures through Lorenz curves and their generalization*";

- di *Raffinetti Emanuela*;
- discussa presso l'Università commerciale Luigi Bocconi – Milano nell'anno *2011;*
- La tesi è tutelata dalla normativa sul diritto d'autore (legge 22 aprile 1941, n.633 e successive integrazioni e modifiche). Sono comunque fatti salvi i diritti dell'Università Commerciale Luigi Bocconi di riproduzione per scopi di ricerca e didattici, con citazione della fonte;
- **solo nel caso sia stata sottoscritta apposita altra dichiarazione con richiesta di embargo**: La tesi è soggetta ad embargo della durata di ……. mesi (indicare durata embargo);

5) che la copia della tesi depositata presso la NORMADEC tramite procedura on-line è del tutto identica a quelle consegnate/inviate ai Commissari e a qualsiasi altra copia depositata negli Uffici dell'Ateneo in forma cartacea o digitale e che di conseguenza va esclusa qualsiasi responsabilità dell'Ateneo stesso per quanto riguarda eventuali errori, imprecisioni o omissioni nei contenuti della tesi;

6) che il contenuto e l'organizzazione della tesi è opera originale realizzata dal sottoscritto e non compromette in alcun modo i diritti di terzi (legge 22 aprile 1941, n.633 e successive integrazioni e modifiche), ivi compresi quelli relativi alla sicurezza dei dati personali; che pertanto l'Università è in ogni caso esente da responsabilità di qualsivoglia natura, civile, amministrativa o penale e sarà dal sottoscritto tenuta indenne da qualsiasi richiesta o rivendicazione da parte di terzi;

7) **scegliere l'ipotesi 7a o 7b indicate di seguito**:

~~7a)~~ che la tesi di dottorato non è il risultato di attività rientranti nella normativa sulla proprietà industriale, non è stata prodotta nell'ambito di progetti finanziati da soggetti pubblici o privati con vincoli alla divulgazione dei risultati; non è oggetto di eventuali registrazioni di tipo brevettale o di tutela, e quindi non è soggetta a embargo;

Oppure

7b) che la tesi di Dottorato rientra in una delle ipotesi di embargo previste nell'apposita dichiarazione "**RICHIESTA DI EMBARGO DELLA TESI DI DOTTORATO**" sottoscritta a parte.

Data 30 Gennaio 2011

F.to Emanuela Raffinetti

UNIVERSITÀ COMMERCIALE “LUIGI BOCCONI”
Milano

PhD in Statistics

Multivariate dependence measures through Lorenz curves and their generalization

Tutor: Prof. Paolo Giudici

Author: Emanuela Raffinetti

UNIVERSITÀ COMMERCIALE “LUIGI BOCCONI”

DECISION SCIENCE DEPARTMENT

PhD IN STATISTICS

Coordinator: **Pr. Pietro Muliere**

The thesis Committee of the candidate **Emanuela Raffinetti** is composed by:

Tutor:	**Pr. Paolo Giudici**
Internal Examiner 1:	**Pr. Marco Bonetti**
Internal Examiner 2:	**Pr. Sonia Petrone**
External Examiner:	**Dr. Chiara Gigliarano**

Ringraziamenti

Desidero rivolgere innanzitutto un doveroso ringraziamento al mio relatore Prof. Paolo Giudici che, con grande professionalità, serietà e costante disponibilità, integrate da una lodevole capacità di ascolto e comprensione, mi ha seguita nello svolgimento di questa tesi.

Ringrazio i docenti del corso di Dottorato in Statistica dell'Università Bocconi, il Dott. Valsecchi e il Dott. Grillo per la gentilezza e la partecipazione dimostrate nei miei confronti e i dottorandi che ho conosciuto in questi anni di studio, in particolare Massimo, amico leale e sincero.

Voglio ringraziare Paola, mia collega di ufficio, perché é il contatto quotidiano con lei che mi ha fatto capire di avere accanto un'amica dai profondi sentimenti.

Ed ora i ringraziamenti rivolti alle persone più importanti della mia vita senza le quali nulla sarebbe mai stato possibile. Sono immensamente grata a mia madre, a mio padre e a mia sorella Alessandra che, con infinita pazienza e amorevole disponibilità, mi hanno sempre incoraggiata spronandomi a credere in me stessa, nelle mie capacità e nei miei obiettivi. Non può di certo mancare un affettuoso grazie a Fabio per il prezioso ed incrollabile supporto morale che ha saputo darmi e per l'entusiasmo che mi ha trasmesso nel mio percorso di ricerca.

Ai miei genitori, a mia sorella Alessandra e a Fabio

Colui che desidera deve perseguire il suo desiderio

anche se il mondo intero lo ostacola.

Colui che persegue il desiderio

trova il proprio cammino disseminato di ostacoli.

(Frederick Rolfe, “Il desiderio e la ricerca del tutto”.)

x

Preface

This PhD Thesis is the result of a collection of three papers[1] concerning the study of dependence based on the application of particular statistical tools such as the Lorenz curve, its dual and the Lorenz zonoids. The research work consists in exploiting the theoretical topics present in literature, related to the dependence study in order to develop further relevant approaches able to provide additional statistical methods. Each chapter corresponds to a specific paper: even if the thesis is represented by three distinguished articles, these are strictly related because they are based on the same central issue represented by the Lorenz curve which can assume different characterizations.

The first part is represented by the introduction whose role consists in focusing on the general dependence study and briefly providing the summary of the three chapters, the central part is represented by the three different papers devoted to the dependence and concordance problem analysis and the last part is represented by the conclusions about all proposed contributions in this research context.

All the presented themes can be useful as starting points for further research proposals.

Milano, January 2011

Emanuela Raffinetti

[1] Submitted to international journals. In particular, the second and third chapters contents have been published in:

- Volume of selected papers “Statistical Methods for the analysis of large data-sets”. Springer (2011);
- Statistics and Probability Letters 81, pp. 133-139 (January, 2011).

Contents

Chapter 1

Introduction

This PhD Thesis is devoted to the dependence analysis in a multivariate context, an area of probability and statistics with increasing activity and applications in several fields. It also contributes to the employment of particular statistical tools as alternative to the classic dependence indices or measures usually used in order to capture information about the existence and the nature of dependence relations.

Original contributions in the thesis include:

- the definition of a novel ranks-based index useful to study the dependence between a quantitative response variable and a predictor, linear function of a number of categorical explanatory variables;
- the introduction, in the above context, of a new theoretical Gini measure decomposition providing a new kind of dependence, in terms of concordance and discordance, named *"Gini Rank Dependence"*;
- the formalization of a new measure able to substitute the partial correlation coefficient in the selection between different nested predictors. This is particularly useful when the most "relevant" explanatory variables assume categorical nature. This new measure may work also as a stopping rule when considering a *stepwise regression* procedure.

All the approaches discussed in this thesis lead to the development of measures which have the same role of the traditional dependence indices (R^2, residual analysis and partial correlation coefficient) but they overcome their disadvantages when the analysis involves mainly categorical variables.

The main purpose of our personal contribution consists in providing a k-dimensional (with $k > 2$) extension of dependence measures through specific statistical tools based on the *Lorenz curve* and the *dual Lorenz curve* as described in [12] and [13]: more precisely, the most interesting issue is provided by the direct relation between the area between the mentioned Lorenz curves and the Gini measure. In fact, when considering the unidimensional context, the Gini measures exactly correspond to the so called *Lorenz zonoid* (see e.g. [6]).

The present section is a review of concepts, contained in the literature, that are needed for the subsequent chapters. There is no attempt to be exhaustive in mentioning all dependence concepts that have been proposed in the literature. A summary of the used dependence concepts is the following:

1. the concept of the ranks-based approach through the application of the Lorenz concentration and the dual Lorenz concentration curves built on discrete data. The idea of employing the concentration curves, in order to describe the dependence relations (concordance and discordance relations), has its source in the taxation problem and is finalized to the "horizontal equity" satisfaction. The ranks-based approach allows to define a dependence measure more sensitive than the classical ones existing in literature (such as the Spearman's ρ, the ranking Gini index and the Kendall's τ);

2. the concepts of the Lorenz curve and monotone dependence orderings. By considering the Lorenz curve of a regression function $E(Y|X)$, one can define a partial order of dependence on the class of non-negative

bivariate random variables with given marginals (see e.g [13]);

3. the concepts of the lift zonoid and the Lorenz zonoid, the latter corresponding to the lift zonoid of relative data. More precisely, the Lorenz zonoid represents the Lorenz curve generalization in the multivariate context (see e.g. [6] and [7]). Subsequently, we will introduce also a proposal attempt of building dependence measures by considering a general regression function characterized by only a covariate.

In order to provide a framework to the subsequent three original chapters, which represents the collection of three papers concerning the applications of the Lorenz curves and the Gini index to the study of multivariate dependence, a deeper review of the mentioned concepts will be now discussed. Each subsequent paragraph content is based on the description of the existing literature topics that define a basic support for each chapter research contribution.

1.1 The concordance problem

The purpose of introducing a novel multivariate index devoted to capture a kind of dependence expressed in terms of concordance and discordance relations (Chapter 2) is focused on the employment of the Lorenz curves. The reason of considering these statistical tools is given by the fact that the existing ranking indices present some disadvantages, whose details will be discussed in Chapter 2.

1.1.1 A brief overview of the ranking indices

The role of the ranking indexes consists in providing a dissimilarity measure between the modalities assumed by the $j-th$ unit and the $i-th$ unit with reference to a specific character: this measure is denoted with d_{ij}. Obviously

d_{ij} has value zero if $i = j$ but also if $i \neq j$ and the $i-th$ and $j-th$ units assume the same character modality: the last property satisfies the relation $d_{ij} = -d_{ji}$ (for more details see [9]).

Now, let us define with d the difference between the unit 1 and the unit 2 modalities, with regard to character 1, and with δ the same difference referred to character 2. The concordance condition is verified if to $d > 0$ corresponds $\delta > 0$ or if to $d < 0$ corresponds $\delta < 0$. All the previous topics lead to provide an index of concordance/discordance degree, for every units pair, defined by the following measure $\sum_i \sum_j d_{ij}\delta_{ij}$, with $i \neq j$. By dividing this term by its admissible maximum value, provided by the Cauchy inequality, and taking into account that $d_{ij} = -d_{ji}$, one can define the relative index

$$\Omega = \frac{\sum_{ij} d_{ij}\delta_{ij}}{\sqrt{\sum_{ij} d_{ij}^2 \sum_{ij} \delta_{ij}^2}}, \quad \text{with } i \neq j, \tag{1.1}$$

where $-1 \leq \Omega \leq +1$. When $-1 \leq \Omega \leq 0$ one can speak about discordance whereas when $0 \leq \Omega \leq +1$ one can speak about concordance. In particular, if the property $d_{ij} = -d_{ji}$ is satisfied, then the expression (1.1) can be defined as

$$\Omega = \frac{2\sum_{i<j} d_{ij}\delta_{ij}}{\sqrt{(2\sum_{i<j} d_{ij}^2)(2\sum_{i<j} \delta_{ij}^2)}} = \frac{\sum_{i<j} d_{ij}\delta_{ij}}{\sqrt{\sum_{i<j} d_{ij}^2 \sum_{i<j} \delta_{ij}^2}}. \tag{1.2}$$

If the collective of units are arranged in a ranking with respect to two characters, every character assumes the position in ranking as modality. This is shown in the table below

Ranking	$i-th$ unit position					
I	p_1	p_2	...	p_i	...	p_n
II	π_1	π_2	...	π_i	...	π_n

where p_i and π_i are ordinal numbers. In order to establish the concordance or the discordance degree between the two rankings we refer to rel-

ative indices called *ranking indices*. In literature there are specific ranking indices each of them has specific characteristics. In the following we recall the Greiner-Kendall's τ, the Spearman's ρ and the ranking Gini index properties.

The Greiner-Kendall's τ index

Starting from the original assumption that two units don't fill the same position (so that $\pi_i \neq \pi_j$, $p_i \neq p_j$, for every $i \neq j$), by placing in (1.2)

$$d_{ij} = \begin{cases} +1 & \text{if } p_i < p_j \\ -1 & \text{if } p_i > p_j \end{cases} \text{ and } \delta_{ij} = \begin{cases} +1 & \text{if } \pi_i < \pi_j \\ -1 & \text{if } \pi_i > \pi_j \end{cases},$$

one defines the Greiner-Kendall's τ index. Every unit is compared with the subsequent one obtaining $\frac{n(n-1)}{2}$ matches. Being $d_{ij}^2 = \delta_{ij}^2 = 1$ (with $i \neq j$), then $\sum_{i<j} d_{ij}^2 = \frac{n(n-1)}{2} = \sum_{i<j} 1 = \sum_{i<j} \delta_{ij}^2$ and the denominator of (1.2) becomes $\frac{n(n-1)}{2}$: in fact

$$\tau = \frac{\sum_{i<j} d_{ij}\delta_{ij}}{\frac{n(n-1)}{2}}. \tag{1.3}$$

In order to simplify the τ index definition, one can consider one of the two rankings in an increasing order, so that $d_{ij} = 1$, for every $i \neq j$ and $\sum_{i<j} d_{ij}\delta_{ij} = \sum_{i<j} \delta_{ij} = \sum^{+} \delta_{ij} - \sum^{-} |\delta_{ij}|$ (where $\sum^{+} \delta_{ij}$ represents the sum of positive dissimilarities and $\sum^{-} |\delta_{ij}|$ identifies the sum in absolute value of negative dissimilarities). The τ-index expression corresponds to

$$\tau = \frac{\sum^{+} \delta_{ij} - \sum^{-} |\delta_{ij}|}{\frac{n(n-1)}{2}}. \tag{1.4}$$

The Spearman's ρ index

Let $d_{ij} = p_j - p_i$ and $\delta_{ij} = \pi_j - \pi_i$, so that the relative concordance index (1.1) becomes

$$\Omega = \frac{\sum_i \sum_j (p_j - p_i)(\pi_j - \pi_i)}{\sqrt{\sum_i \sum_j (p_j - p_i)^2 \sum_i \sum_j (\pi_j - \pi_i)^2}}. \tag{1.5}$$

Through some computations one can show that the Spearman index is equivalent to

$$\rho = 1 - \frac{6\sum_i (p_i - \pi_i)^2}{n(n^2-1)}. \tag{1.6}$$

***Remark* 1.1.1** *The Spearman's index differs from the Greiner-Kendall's one because it provides also information about the distance between the position filled by a unit with respect to the position assumed by another one.*

The ranking Gini index

The ranking Gini index employs the opposite ranking of one of the two considered rankings. Let us denote with p the ranking whose positions are represented by $p_1, p_2, \ldots, p_n$ and with π the other ranking whose positions are represented $\pi_1, \pi_2, \ldots, \pi_n$. Finally, let us point out with π' the ranking complementary to π. Furthermore $\pi_1^{'}, \pi_2^{'}, \ldots, \pi_n^{'}$ define the position filled by every unit. The following difference

$$\sum |p_i - \pi_i^{'}| - \sum |p_i - \pi_i| \tag{1.7}$$

assumes positive value if there is concordance and negative value if there is discordance between the rankings. In order to obtain a relative index which ranges between -1 and $+1$, the expression (1.7) has to be divided by its maximum value, if it is positive, or by its minimum value (in absolute value), if it is negative. These values are equivalent in absolute value and correspond to $\frac{n^2}{2}$, if n is even, and to $\frac{n^2-1}{2}$, if n is odd. In conclusion, the ranking Gini index expression is given by

$$G^* = \begin{cases} \frac{\sum |p_i - \pi_i^{'}| - \sum |p_i - \pi_i|}{\frac{n^2}{2}} & \text{if } n \text{ is even} \\ \frac{\sum |p_i - \pi_i^{'}| - \sum |p_i - \pi_i|}{\frac{n^2-1}{2}} & \text{if } n \text{ is odd} \end{cases}.$$

These indices present some restrictions because, even if they are invariant to every monotone transformation, they are based only on the units ranks without considering the extent variation of the considered character. For this reason new measures, taking into account the previous condition, have to be defined.

1.1.2 The ranks-based approach in the bivariate case

The ranks-based approach employment is strictly related to the taxation context.

The theoretical contributions to the condition of "perfect" taxation have focused the attention on the relation between efficiency and vertical equity without considering the issue of horizontal equity. Only recently some measures, devoted to the taxation equity topic satisfaction, have been introduced in literature.

According to the traditional definition of the "horizontal equity" notion "[...] people in equal positions should be treated equally", (see e.g. [15]).

For this reason, the concordance problem appears strictly related to the taxation context (the subsequent description of the concordance problem concerning the taxation field has been developed in [12]). However, as it will be discussed in Chapter 2, one can provide an extension of the concordance problem with regard to a more general point of view, in order to develop new dependence measures in a multivariate context.

Let us consider an ordered set of n real numbers (x_i, y_i), $i = 1, \ldots, n$[1], whose components denote the measure of two quantitative characters. Let F and G be the two distribution functions of the variables X (describing the income level before taxation) and the variable Y (describing the income level after taxation): let $r(x_i)$ and $r(y_i)$ be the rank of the $i-th$ individual with respect to the char-

[1] We intend all the pairs defined as $(x_1, y_1), \ldots, (x_n, y_n)$.

acter X and to the character Y. Supposing that $x_i \neq x_j$, $y_i \neq y_j$, with $i \neq j$, the horizontal equity definition implies that

$$r(x_i) = r(y_i), \quad i = 1, \ldots, n, \tag{1.8}$$

in a condition of perfect horizontal equity.

In the opposite case

$$r(y_i) = n + 1 - r(y_i), \quad i = 1, \ldots, n \tag{1.9}$$

corresponds to a perfect horizontal iniquity situation.

The previous definition implies the knowledge of the units ordering before the taxation process and the income level of each individual after the redistribution process.

According to these conditions, the proposal of identifying a new index able to stress the existence of functional monotone relations between the involved two characters is required. An appropriate solution to this problem has been detecting in the employment of the Lorenz concentration curves tool.

In 1905, Max Lorenz proposed a simple graphical means in order to summarize the inequality of wealth in a finite population of individuals (see [10]): known subsequently as the Lorenz concentration curve, it has survived well and indeed still occupies a preeminent place in discussion of the quantification of inequality. Lorenz curve birth is due to a data set which provides the proportion of the total population earning less than the given value together with the proportion of the total wealth of the population accruing to those individuals. The percentage of the population is plotted against the y-axis and the proportion of the total wealth is plotted against the x-axis. Then, one joins the points through a smooth curve without giving hint about how this interpolating curve is selected: we would likely interchange the axes in the diagram and

interpolate linearly to obtain what we call the Lorenz concentration curve.

Let us describe the Lorenz curves application in the taxation field so that let us consider the single entities of the character X describing the income amount before the taxation process. Let sort them in an increasing sense such that $x_1 < x_2 < \ldots < x_n = x_{(i)}$ with respect to the following restriction $\sum_{i=1}^{n} x_i = nM_X$, with n and M_X fixed (where M_X represents the X variable mean value). From well known means inequalities one gets $S_i \leq iM_X$, $i = 1, \ldots, n$, being $S_i = \sum_{j=1}^{i} x_{(j)}$. The set of all ordered pairs $(i/n, S_i/(nM_X))$, $i = 1, \ldots, n$ defines the Lorenz concentration function L_X with regard to the discrete context: the sequence of differences $(iM_X - S_i)$ represents a valuable tool in order to appreciate the X variable concentration. A concentration measure can be obtained through the ratio between the sum $C_n = \sum_{i=1}^{n}(i/n - S_i/nM_X)$ and its value, assumed in a maximum concentration situation, is given by $S_n = nM_X$. The obtained indicator represents the classical Gini concentration ratio $R_X = 2C_n/n - 1$ (as denoted in [12]).

The Y variable concentration function (representing the income amount after taxation and denoted with L_Y) is obtained in an analogous way. In order to evaluate the concordance degree between the two involved variables, a new curve, based on the Y values ordered according to the X variable ranks $(y_i|r(x_i))$, is needed. For sake of clarity in the following we denote these values with y_i^*. The function obtained by the set of pairs $(i/n, (1/(nM_X))\sum_{j=1}^{i} y_j^*)$ represents the concordance curve, denoted with $C(Y|r(x_i))$. The concordance curve construction in the bivariate case is proposed in [12]. The Lorenz curves obtained from Y and $Y|r(x_i)$ are then compared through the ratio of their areas (i.e. the sum of the distances from each curve's points to the egalitarian line given by the points $(i/n, i/n)$, $i = 1, \ldots, n)$). This index, which assumes always values between -1 and $+1$, satisfies several properties which will be discussed in Chapter 2, whose content is based on an extension of this mea-

sure to a multivariate context.

Since this Phd thesis is focused on the dependence analysis through the application of the Lorenz curves, in the following sections all the main properties concerning these specific statistical tools will be discussed starting from the notion of monotone dependence and monotone dependence orderings.

1.2 Lorenz curve and monotone dependence ordering approach

In this paragraph, a partial order of monotone dependence on the class of non-negative bivariate random variables with given marginals is defined, based on the notion of the Lorenz curve of the regression function. In the developments, illustrated and discussed in Chapters 2,3 and 4, the considered regression function will be linear.

Statistical literature provides a large number of families of bivariate distributions with a natural interpretation of monotone dependence: the intuitive meaning of monotone dependence for a bivariate random vector (X,Y) is that large values of Y stochastically correspond to large values of X (positive dependence) or, in the opposite case, large values of Y correspond to small values of X (negative dependence). This concept of monotone dependence has played a fundamental role in many recent new ideas in statistics leading to the development of some important concepts such as quadrant dependence (see [8]), association (see [2]) and concordance (see [16]). Literature concerning monotone dependence, can be found, for instance, in [2].

In order to compare two bivariate distributions having the same pairs of marginals to determine whether one distribution is more positively dependent than the other, several partial orderings on the class of bivariate distributions

with fixed marginals have been introduced. In this field of research a new characterization of monotone dependence is achieved in drawing a comparison between the Lorenz curve of the regression function $E(Y|X)$ and the Lorenz curve of Y will be introduced. This notion is appropriate when asking the relation to be invariant under increasing transformation of X but sensible to increasing transformation of Y. As a consequence, a partial ordering of monotone dependence on the class of nonnegative bivariate random vectors with given marginals is then defined (for more details see e.g. [13]): the motivation of the proposed partial ordering arises in the study of economic problems (e.g. taxation, see [12]) and in the study of several applications such as selection discriminant problems and statistical quality control.

1.2.1 Lorenz curve: definition and properties

Let X be a non-negative random variable with distribution function F_X and finite expectation $E(X) > 0$: let us provide the following definition of the Lorenz curve of X

$$L_X(p) = \frac{1}{E(X)} \int_0^p F_X^{-1}(z)dz, \qquad 0 \leq p \leq 1, \tag{1.10}$$

where $F_X^{-1}(z) = \inf\{x : F_X(x) \geq z\}$, $0 \leq z \leq 1$.

Let us now consider a bivariate non-negative random vector (X,Y) with marginals F_X and F_Y respectively and suppose that $E(X)$ and $E(Y)$ are positive and finite: denoting with $m(X)$ the general regression function $E(Y|X)$, the aim is in examining some properties of the Lorenz function of $m(X) = E(Y|X)$, as illustrated in [13].

For sake of simplicity, the attention is restricted to the class $\pi(F_X, F_Y)$ of bivariate non-negative random vectors with continuous marginal distribution functions F_X and F_Y: furthermore one assumes that for all $(X,Y) \in \pi(F_X, F_Y)$ the regression function $m(x) = E(Y|X)$ is continuous with finite first derivative

$m^{'}(x)$ (all the following details are contained and discussed in [13]).

The Lorenz curve of the regression function $E(Y|X)$ assumes the following expression (see e.g. [3])

$$L_{E(Y|X)}(p) = \frac{1}{E(E(Y|X))} \int_0^{x_p} m(t) dF_X(t) = \frac{1}{E(Y)} \int_0^p m(F_X^{-1}(z)) dz, \quad (1.11)$$

where $0 \leq p \leq 1$ and $x_p = F_X^{-1}(p)$.

Let us recall some of its important properties.

Property 1.

1. $L_{E(Y|X)}(p)$ passes through the points $(0,0)$ and $(1,1)$;

2. $L_{E(Y|X)}(p)$ is increasing if and only if $m(x) > 0$ for all x;

3. $L_{E(Y|X)}(p)$ is concave if and only if $m(x)$ is nondecreasing for all x.

Proof.

(1.) follows from definition. (2.) follows immediately observing that: $\frac{\partial L_{E(Y|X)}(p)}{\partial p} = \frac{m(x_p)}{E(Y)}$. (3.) $\frac{\partial^2 L_{E(Y|X)}(p)}{\partial p^2} = \frac{m(x_p)^{'}}{E(Y) f_X(x_p)}$, where f_X is the density function of F_X. Since $f_X(x_p) > 0$, the sign of the second derivative is that of $m^{'}(x)$. ∎

A second criterion states that $L_{E(Y|X)}(p)$ is above (below) the egalitarian line if the elasticity $\eta(x) = \frac{x m^{'}(x)}{m(x)}$ is less (grater) than zero for all $x \geq 0$.

Property 2.

$L_Y(p) \leq L_{E(Y|X)}(p) \leq L_Y^{'}(p)$, where

$$L_Y^{'}(p) = \frac{1}{E(Y)} \int_{1-p}^1 F_Y^{-1}(z) dz, \quad 0 \leq p \leq 1.$$

The usefulness and the proof of Property 2 will be discussed in Chapter 4.

Property 3.

$L_{E(Y|X)}(p) = p$ for all $p \in (0,1)$ if and only if $E(Y|X) =_{st} E(Y)$.

Proof.

Recall that the notation $=_{st}$ means equivalence of random variables, defined as $X =_{st} Y$, if the probability of the event $(X = Y)$ is equal to one.

$L_{E(Y|X)}(p) = p$ for all $p \in (0,1)$ can be reexpress as

$$\int_0^{x_p} E(Y|X = x)dF_X(x) = \int_0^{x_p} E(Y)dF_X(x) \text{ for all } p \in (0,1),$$

proving that $E(Y|X) =_{st} E(Y)$. ■

Property 4.

$L_{E(Y|X)}(p) = L_Y(p)$ for all $p \in (0,1)$ if and only if $m(x)$ is increasing and such that $E(Y|X) =_{st} Y$ and $L_{E(Y|X)}(p) = L_Y^{'}(p)$ for all $p \in (0,1)$ if and only if $m(x)$ is decreasing and such that $E(Y|X) =_{st} Y$.

Proof.

About proof details see [13].

Property 5.

If $E(Y|X = x) = \alpha + \beta x$, then

$$L_{E(Y|X)}(p) = p - \beta \frac{E(X)}{E(Y)}(p - L_X(p)). \tag{1.12}$$

Proof.

In general we have that $L_{E(Y|X)}(p) = \frac{1}{E(Y)} \int_0^p m(F_X^{-1}(z))dz$, where $x = F_X^{-1}(z)$,

then

$$\begin{aligned} L_{E(Y|X)}(p) &= \frac{1}{E(Y)}\int_0^p (\alpha+\beta F_X^{-1}(z))dz \\ &= \frac{1}{E(Y)}\left[\int_0^p \alpha dz+\beta\int_0^p F_X^{-1}(z)\right] = \frac{1}{E(Y)}\left[\alpha p+\beta\int_0^p F_X^{-1}(z)\right] \\ &= \frac{1}{E(Y)}[\alpha p+\beta E(X)L_X(p)] = \frac{\alpha}{E(Y)}p+\frac{\beta}{E(Y)}E(X)L_X(p); \end{aligned}$$

since in the linear regression model $\alpha = E(Y)-\beta E(X)$, then

$$\begin{aligned} L_{E(Y|X)}(p) =& \frac{p}{E(Y)}[E(Y)-\beta E(X)]+\frac{\beta}{E(Y)}E(X)L_X(p) \\ =& p\frac{E(Y)}{E(Y)}-\beta\frac{E(X)}{E(Y)}p+\beta\frac{E(X)}{E(Y)}L_X(p) \\ =& p-\beta\frac{E(X)}{E(Y)}[p-L_X(p)]. \quad \blacksquare \end{aligned}$$

Property 6.

$L_{E(Y|X)}(p)$ is invariant under F_X-increasing[2] transformation of X.

Property 7.

Let (X,Y) and $(X',Y') \in \pi(F_X,F_Y)$. Then $L_{E(Y|X)}(p) = L_{E(Y'|X')}(p)$ for all $p \in (0,1)$ if and only if $E(Y|X)$ and $E(Y'|X')$ have the same distribution.

Proof.

About proof details see [13].

1.2.2 Orderings of positive and negative dependence

The Lorenz curve represents a widely used tool for ordering distributions: given two non-negative random variables, X and Y, with finite positive ex-

[2] A function $h:\mathbb{R}\to\mathbb{R}$ is F_X-increasing if for all $s,t\in\mathbb{R}$: $F_X(s)<F_Y(s)$ implies $h(s)<h(t)$.

pectations such that $X \sim F_X$ and $Y \sim F_Y$, one can say that the distributions F_X and F_Y (or equivalently the random variables X and Y) are ordered by Lorenz ordering if the Lorenz curve of X is nowhere below the Lorenz curve of Y, that is (see e.g. [11])

$$X \preceq_L Y \quad \text{if} \quad L_X(p) \geq L_Y(p) \text{ for every } p \in (0,1). \tag{1.13}$$

The aim is measuring monotone dependence by drawing a comparison between the Lorenz curve of $E(Y|X)$ and the Lorenz curve of Y. Intuitively the more is the Lorenz curve of $E(Y|X)$ similar, or far from, the Lorenz curve of Y, the stronger is the positive or negative dependence of Y on X. The idea, described in [13], is based on the introduction of a monotone dependence structure based on $L_{E(Y|X)}$: let us denote with $\mathbf{L}$ the family of all bivariate random vector (X,Y) with monotonic dependence distinguished in $\mathbf{L}^+$ and $\mathbf{L}^-$ in case of positive or negative dependence.

***Definition* 1.2.1** *The set of all ordered pairs (X,Y) satisfying*

$$L_{E(Y|X)} \leq p \text{ for all } p \in (0,1)$$

is denoted with $\mathbf{L}^+$: if the inequality is reverse it is denoted with $\mathbf{L}^-$.

The dependence structure $\mathbf{L}$, when considering a bivariate context, presents some important relations with the covariance measure: in fact, if two variables tend to vary together (that is, when one of them is above its expected value, then the other variable tends to be above its expected value too), then the covariance between the two variables will be positive. On the other hand, if one of them tends to be above its expected value when the other variable is below its expected value, then the covariance between the two variables will be negative. So the covariance definition perfectly reflects the monotone dependence definition as one can show through the following proposition:

***Proposition* 1.2.2** *If* $(X,Y) \in \mathbf{L}^+$*, then* $Cov(X,Y) \geq 0$*. If* $(X,Y) \in \mathbf{L}^-$*, then* $Cov(X,Y) \leq 0$*. Finally,* $Cov(X,Y) = 0$ *if and only if* $E(Y|X) \doteq E(Y)$*.*

Proof.

About proof details see [13].

Let us consider the following propositions to provide an ordering in the family $\boldsymbol{L}^+$ $[\boldsymbol{L}^-]$.

***Definition* 1.2.3** *For each* (X,Y) *and* (X',Y') *belonging to* $\mathbf{L}^+$ $[\mathbf{L}^-]$*:*

$$(X,Y) \preceq_{\mathbf{L}^+} (X',Y') \quad [\preceq_{\mathbf{L}^-}]$$

if $F_X = F_{X'}$*,* $F_Y = F_{Y'}$ *and* $L_{E(Y|X)}(p) \geq L_{E(Y'|X')}(p)$ $[\leq]$ *for all* $p \in (0,1)$*.*

Of particular interest is the relation between the ordering based on $\mathbf{L}^+$ $[\mathbf{L}^-]$ and the covariance measure expressed by the following proposition:

***Proposition* 1.2.4** *Let* (X,Y) *and* $(X',Y') \in \mathbf{L}^+$*. If*

$$(X,Y) \preceq_{\mathbf{L}^+} (X',Y') \text{ then } Cov(X,Y) \leq Cov(X',Y').$$

The case of negative dependence can be easily derived by reversing the inequalities. The ordering $\preceq^+_{\mathbf{L}}$ is only partial: two bivariate random vectors are comparable, under $\preceq^+_{\mathbf{L}}$ whenever their Lorenz curves intersect. In order to compare pairs of bivariate random vectors that are not ordered by $\preceq^+_{\mathbf{L}}$ ordering, it is wise to choose (partial or total) orders that are finer than $\preceq^+_{\mathbf{L}}$. An order $\preceq_A$ is finer than another partial order $\preceq_B$ if $(X,Y) \preceq_B (X',Y')$ implies $(X,Y) \preceq_A (X',Y')$, i.e., $\preceq_A$ orders all bivariate distributions that $\preceq_B$ orders. An order is total if it orders every pair of bivariate distributions.

A measure I of $\mathbf{L}$-dependence is a functional of the bivariate random vector that induces a total order $\preceq_I$ defined by

$$(X,Y) \preceq_I (X',Y') \text{ if } I(X,Y) \leq I(X',Y')$$

which is finer than $\preceq_{\mathbf{L}^+}$ and $\succeq_{\mathbf{L}^-}$.

Through all these theoretical notions, one can propose a measure of $\mathbf{L}$-dependence given by the area between the egalitarian line and the Lorenz curve $L_{E(Y|X)}$,

$$A_{YX} = \frac{1}{2} - \int_0^1 L_{E(Y|X)}(p)dp. \tag{1.14}$$

By normalizing the area A_{YX}, a relative measure of $\mathbf{L}$-dependence which assumes values in $[-1,+1]$ is provided.

***Proposition* 1.2.5** *Let $A_Y = \int_0^1 [p - L_Y(p)]dp$, then*

1. *$C_{YX} = A_{YX}/A_Y$ is a relative measure of $\mathbf{L}-$ dependence, such that $-1 \leq C_{YX} \leq +1$;*

2. *$|C_{YX}| = 1$ if and only if $E(Y|X) \doteq Y$: more specifically, $C_{YX} = +1$ if $m(x) = E(Y|X = x)$ is a nondecreasing function for all x, and $C_{YX} = -1$ if $m(x)$ is a non increasing function for all x;*

3. *$C_{YX} = 0$ if $E(Y|X) \doteq E(Y)$*

4. *If $E(Y|X = x) = \alpha + \beta x$, then*

$$C_{YX} = \beta \frac{E(X)}{E(Y)} \frac{A_X}{A_Y} = \rho \frac{\sigma_Y}{\sigma_X} \frac{E(X)}{E(Y)} \frac{A_X}{A_Y},$$

where ρ is the linear correlation coefficient.

Our research contribution takes advantages from these described notions existing in literature (all the details can be found in [13]).

In order to be exhaustive, the last topic to be illustrated, concerns the Lorenz curve extension to the multivariate context through the so called Lorenz zonoid approach. In Chapter 4, the Lorenz zonoid tool, has been employed in order to develop partial dependence measures. For this reason, we devote

the following section to a brief discussion about the Lorenz zonoid definition and its relevant properties.

1.3 Lorenz and lift zonoids

In the previous paragraphs the Lorenz curve definition, both in the discrete and continuous context, has been introduced when considering dimension one: however, in dimensions greater than one, the disparity can be capture by employing a new tool represented by the Lorenz zonoid. The Lorenz zonoid has been introduced by Koshevoy (see [4], [5]) and Mosler (see [14]): the Lorenz zonoid represents the Lorenz curve generalization for relative multivariate data. When considering non relative data, one can apply the so called lift zonoid.

The lift zonoid application proves to be very useful: first, the lift zonoid is related to random convex sets and to convex hull of a multivariate random sample and second, the set inclusion of lift zonoids defines an ordering of random vectors that reflects their variability. In fact, our interest is a comprehensive investigation of the ordering among random vectors that is induced by the inclusion of lift and Lorenz zonoids.

Starting from the multivariate Lorenz curve generalization, one can define the multivariate Lorenz dominance via inclusion of the Lorenz zonoids: one of several equivalent characterization of this ordering states that a d-dimensional random vector $\mathbf{X} = (X_1, \ldots, X_d)$ is Lorenz dominated by another random vector $\mathbf{Y} = (Y_1, \ldots, Y_d)$ if for all $p_1, \ldots, p_d \in \mathbb{R}$, the random variable $\sum_{j=1}^{d} p_j X_j$ is Lorenz dominated by the random variable $\sum_{j=1}^{d} p_j Y_j$. One can compare the Lorenz dominance with the scaled convex order, usually used in order to get information about the variability of random vectors.

All these topics will be discussed in Chapter 4, since they define the basic

support for the extensions that will be presented and discussed.

1.3.1 Lift zonoids and Lorenz zonoids of multivariate data

Here we provide the definitions of the lift zonoid and Lorenz zonoid when considering multivariate data: all the details are provided in [7].

Consider a population of n economic units, say households, which are endowed with quantities of d commodities or other attributes of well-being. Let $\mathbf{x}_i = (x_{1i}, \ldots, x_{di})' \in \mathbb{R}^d$ denote the endowment of household $i = 1, \ldots, n$ and the matrix

$$\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)' = \begin{pmatrix} x_{11} & \ldots & x_{d1} \\ \vdots & & \vdots \\ x_{1n} & \ldots & x_{dn} \end{pmatrix}$$

describe an empirical distribution $F_{\mathbf{X}}$ of endowments among households[3]. Assume that in every attribute j, the mean endowment is positive: for each j we can consider the classical Lorenz curve. It refers to relative data (i.e. data over their mean)

$$\tilde{x}_{ji} = \frac{n x_{ji}}{\sum_{k=1}^{n} x_{kj}}, \quad \tilde{\mathbf{x}}_i = (\tilde{x}_{1i}, \ldots, \tilde{x}_{di}), \quad \tilde{\mathbf{X}} = (\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n)'.$$

In case of only one attribute, say income, the endowments are real numbers $x_1, \ldots, x_n$ and can be ordered. Let $\tilde{x}_{(1)}, \ldots, \tilde{x}_{(n)}$ be the $\tilde{x}_i$'s in ascending order.

The Lorenz curve is the piecewise linear connection of points

$$\left(\frac{k}{n}, \frac{1}{n} \sum_{i=1}^{k} \tilde{x}_{(i)} \right), \quad k = 0, \ldots, n, \tag{1.15}$$

in the unit square. The generalized Lorenz curve is the Lorenz curve with $\tilde{x}_{(i)}$ replaced by $x_{(i)}$.

[3] Note that in this case we use the same symbol $\mathbf{X}$ for a random variable in $\mathbb{R}^d$ and an $n \times d$ data matrix.

***Example* 1.3.1** *Consider two households which receive* $x_1 = 2400$ *and* $x_2 = 5600$ *dollars of income, respectively. The bold lines in Figure 1.1 show the Lorenz curve (a.) and the generalized Lorenz curve (b.) of this two-point distribution. Each point* z_0, z_1 *on the Lorenz curve indicates that the poorer* $z_0 \cdot 100$ *per cent of the population receives* $z_1 \cdot 100$ *per cent of total income.*

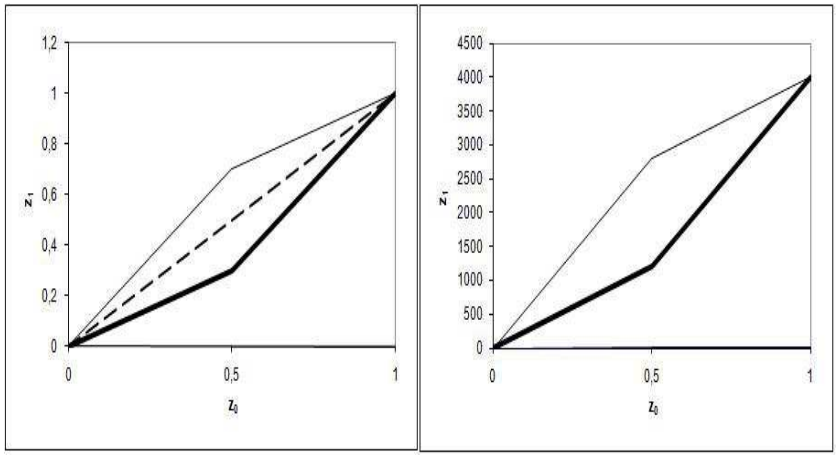

Figure 1.1: a. Lorenz curve and Lorenz zonoid $LZ(2400, 5600)$, b. generalized Lorenz curve and lift zonoid $\hat{Z}(2400, 5600)$ of the univariate two-point distribution in Example 1.3.1.

The data $X = (x_1, \ldots, x_n)'$ of a single attribute is *Lorenz dominated* by some other data $Y = (y_1, \ldots, y_m)'$, $X \preceq_L Y$, if the Lorenz curve of X lies below the Lorenz curve of Y. this is usually interpreted that X contains less inequality than Y: if $n = m$ Lorenz dominance is equivalent to *majorization* of n-vectors. $X \preceq_L Y$ if and only if

$$\sum_{i=1}^{k} \tilde{x}_{(i)} \geq \sum_{i=1}^{k} \tilde{y}_{(i)} \quad \text{for} \quad k = 1, \ldots, n-1. \tag{1.16}$$

Definition (1.15) and restriction (1.16) do not easily carry over to more than one attribute, since there is no natural complete order of d-dimensional space when $d > 1$. To avoid the ordering of vectors, the principal idea is regarding each Lorenz curve as the boundary of a properly chosen set and order these sets by inclusion. In order to construct such sets, a natural choice can be

employing a centrally symmetric set. In case $d = 1$, the set is south-east bordered by the Lorenz curve and north-west bordered by the curve symmetric to it, that is the piecewise linear connection of points

$$\left(\frac{n-k}{n}, 1 - \frac{1}{n} \sum_{i=1}^{k} \tilde{x}_{(i)} \right), \quad k = 0, \ldots, n.$$

This curve is the dual Lorenz curve, which has been denoted with $L_X^{'}$. In Figure 1.1 (a.), it corresponds to the upper border of the hatched area.

For general $d \geq 1$, the lift zonoid of data $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^{'}$ is a set in $(d+1)$-space, defined by

$$\hat{Z}(\mathbf{X}) = \frac{1}{n} \sum_{i=1}^{n} [(0, \mathbf{0}), (1, \mathbf{x}_i)]. \tag{1.17}$$

Here $[(0, \mathbf{0})], (1, \mathbf{x}_i)]$ denotes the line segment that extends from the origin $(0, \mathbf{0})$ to the point $(1, \mathbf{x}_i)$ in $\mathbb{R}^{d+1}$. The sum is the Minkowski sum, that is, the sets are added point by point. For example the Minkowski sum of the two segments $[(0,0), (1, 0.6)]$ and $[(0,0), (1, 1.4)]$ in $\mathbb{R}^2$ amounts to the parallelogram spanned by the vectors (1,0.6) and (1,1.4).

We define the Lorenz zonoid of $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^{'}$ as the lift zonoid of the relative data

$$LZ(\mathbf{X}) = \hat{Z}(\tilde{\mathbf{X}}). \tag{1.18}$$

In Example 1.3.1 of two households, with $d = 1$, one obtains the Lorenz zonoid

$$LZ(2400, 5600) = \hat{Z}(0.6, 1.4) = \frac{1}{2}[(0,0), (1, 0.6)] + \frac{1}{2}[(0,0), (1, 1.4)], \tag{1.19}$$

which is depicted in Figure 1.1 (a.). Here, the lower border (in boldface) describes the Lorenz curve, while the upper border, the dual Lorenz curve, is a rotation symmetric to it. In order to get (1.19) one has to apply the Minkowski sum (see [7]), so that one computes the sum of all the pairs of vertices points.

Then,

- $(0,0)+(0,0)=(0,0)$;
- $(0,0)+(1,1.4)=(1,1.4)$;
- $(1,0.6)+(0,0)=(1,0.6)$;
- $(1,0.6)+(1,1.4)=(2,2)$.

Multiplying by $1/2$, on the basis of (1.17) and (1.18), one gets the set of points $[(0,0),(0.5,0.3),(0.5,0.7),(1,1)]$ which represents the vertices corresponding to the parallelogram defined as the set of points between the Lorenz curve and the dual Lorenz curve.

1.3.2 Properties of lift and Lorenz zonoids

In many applications it is convenient to employ random variables and general probability distributions rather than empirical distributions only.

The lift zonoid of a general d-variate random vector is defined as follows. Consider the set $\mathscr{X}^d$ of random vectors in $\mathbb{R}^d$ that have finite expectation, the subset $\mathscr{X}^{d+} \subset \mathscr{X}^d$ of those vectors that have positive (in each component) expectation, and the subset $\mathscr{X}_+^{d+} \subset \mathscr{X}^{d+}$ of those that have, in addition, support in $\mathbb{R}_+^d$. Define, for $\mathbf{X} \in \mathscr{X}^d$, the lift zonoid

$$\hat{Z}(\mathbf{X}) = \left\{ E[h(\mathbf{X}), h(\mathbf{X})\mathbf{X}] : h : \mathbb{R}^d \to [0,1] \text{ measurable} \right\}. \tag{1.20}$$

The lift zonoid is a convex compact set in $\mathbb{R}^{d+1}$ that includes the expectations of all d-variate random vectors $(h(\mathbf{X}), h(\mathbf{X})\mathbf{X})$. With data $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$, the definition of the lift zonoid specializes to the previous one (1.17): by setting

$$\lambda_i = \frac{1}{n} h(\mathbf{x}_i)$$

one can define

$$\hat{Z}(\mathbf{X}) = \left\{ \left(\sum_{i=1}^{n} \lambda_i, \sum_{i=1}^{n} \lambda_i \mathbf{x}_i \right) : 0 \leq \lambda_i \leq \frac{1}{n}, i = 1, \ldots, n \right\}$$
$$= \frac{1}{n} \sum_{i=1}^{n} [(0, \mathbf{0}), (1, \mathbf{x}_i)].$$

The lift zonoid has many attractive properties, which make it useful for a broad range of applications. The first important property is uniqueness. Like the generalized Lorenz curve in dimension one, the lift zonoid contains full information about the underlying distribution, that is, given the lift zonoid, the data can be completely regained.

Proposition* 1.3.2 *(Uniqueness) *The lift zonoid $\hat{Z}(\mathbf{X})$ uniquely determines the distribution $F_{\mathbf{X}}$, for $\mathbf{X} \in \mathscr{X}^d$.*

The second property regards marginalization. If we restrict to one or several dimensions of the data, the lift zonoid of the marginal distribution is a simple projection of the lift zonoid of the joint distribution.

Proposition* 1.3.3 *(Marginalization) *For any $J \subset \{1, 2, \ldots, d\}$ consider the projection $pr_J : \mathbf{z} \to (z_0, \mathbf{z}_J)$, $\mathbf{z} \in \mathbb{R}^{d+1}$, where $\mathbf{z}_J$ is the subvector containing components z_j, $j \in J$. Then*

$$pr_J(\hat{Z}(\mathbf{X})) = \hat{Z}(\mathbf{X}_j).$$

Thirdly, we state that the lift zonoid is additive in the distribution.

Proposition* 1.3.4 *(Additivity) *For two random vectors $\mathbf{X}$ and $\mathbf{Z}$ in $\mathscr{X}^d$ with distributions $F_{\mathbf{X}}$ and $F_{\mathbf{Z}}$ and $\alpha \in [0, 1]$,*

$$\hat{Z}(\alpha F_{\mathbf{X}} + (1 - \alpha) F_{\mathbf{Z}}) = \alpha \hat{Z}(F_{\mathbf{X}}) + (1 - \alpha) \hat{Z}(F_{\mathbf{Z}}).$$

For data $\mathbf{X}$ and $\mathbf{Y}$, the last equation reads

$$\tilde{Z}(\mathbf{x}_1,\ldots,\mathbf{x}_n,\mathbf{y}_1,\ldots,\mathbf{y}_m) = \frac{n}{n+m}\hat{Z}(\mathbf{x}_1,\ldots,\mathbf{x}_n) + \frac{m}{n+m}\hat{Z}(\mathbf{y}_1,\ldots,\mathbf{y}_m). \quad (1.21)$$

The additivity property is useful with indices based on the lift zonoid: it allows to decompose an index into subgroups of the population.

The general definition of the Lorenz zonoid of a random vector and its properties are straightforward (see e.g. [7]). The Lorenz zonoid of some $\mathbf{X} \in \mathscr{X}_+^{d+}$ is the lift zonoid of the relative vector, i.e. the vector componentwise divided by its expectation. With

$$\tilde{\mathbf{X}} = \left(\frac{X_1}{E[X_1]},\ldots,\frac{X_d}{E[X_d]}\right)$$

define

$$\begin{aligned} LZ(\mathbf{X}) = \hat{Z}(\tilde{\mathbf{X}}) &= \left\{E[h(\tilde{\mathbf{X}}), h(\tilde{\mathbf{X}})\tilde{\mathbf{X}}] : h : \mathbb{R}_+^d \to [0,1] \text{ measurable}\right\} \quad (1.22)\\ &= \{\mathbf{z} \in \mathbb{R}^{d+1} : \mathbf{z} = (z_0, z_1, \ldots, z_d) = \zeta(h),\\ &\quad h : \mathbb{R}_+^d \to [0,1] \text{ measurable}\}, \end{aligned}$$

where

$$\zeta(h) = \left(\int_{\mathbb{R}_+^d} h(\mathbf{x})dF(\mathbf{x}), \int_{\mathbb{R}_+^d} \psi_F(\mathbf{x})h(\mathbf{x})dF(\mathbf{x})\right),$$

and $\psi_F(\mathbf{x}) = \tilde{\mathbf{x}} = (\tilde{x}_1,\ldots,\tilde{x}_d)$, $\tilde{x}_j = \frac{x_j}{\mu_j}$, for $j = 1,\ldots,d$. The Lorenz zonoid of a random vector $\mathbf{X} \in \mathscr{X}_+^{d+}$ is a convex compact set in $\mathbb{R}^{d+1}$.

From an economic view point the function $h : \mathbb{R}_+^d \to [0,1]$ can be seen as a selection of some part of the population. Of all households, which have relative endowment vector $\tilde{\mathbf{X}}$, the percentage $h(\tilde{\mathbf{X}})$ is selected. $E[\tilde{\mathbf{X}}h(\tilde{\mathbf{X}})]$ amounts to the total portion vector held by this subpopulation and $E[h(\tilde{\mathbf{X}})]$ is the size of the subpopulation selected by h.

The following three principal properties of the Lorenz zonoid correspond to those of the lift zonoid.

(Uniqueness): the Lorenz zonoid of a random vector determines its distribution uniquely, up to d scale factors.

(Marginalization): the Lorenz zonoid of a marginal $\mathbf{X}_J$ is equal to the projection of the Lorenz zonoid of $\mathbf{X}$, for any $J \subset \{1,2,\ldots,d\}$.

(Additivity): given two random vectors $\mathbf{X}$ and $\mathbf{Z}$ in $\mathscr{X}^{d+}$ with distributions $F_{\mathbf{X}}$ and $F_{\mathbf{Z}}$ and some $\alpha \in [0,1]$ consider a random vector $\mathbf{Y}$ that has distribution $F_{\mathbf{Y}} = \alpha F_{\mathbf{X}} + (1-\alpha)F_{\mathbf{Z}}$. Then $LZ(\mathbf{Y})$ corresponds to

$$\alpha LZ(\mathbf{X}) diag\left(1, \frac{E[X_1]}{E[Y_1]}, \ldots, \frac{E[X_d]}{E[Y_d]}\right) + (1-\alpha) LZ(\mathbf{Z}) diag\left(1, \frac{E[Z_1]}{E[Y_1]}, \ldots, \frac{E[Z_d]}{E[Y_d]}\right).$$

By exploiting the marginalization property and the main features related to the lift zonoids tool, dependence measures in the bivariate context can be obtained. All the details concerning the construction, that will be discussed here, have been presented in an unpublished draft of Koshevoy and Muliere (2002).

Let (X,Y) be a bivariate random vector with distribution $F(x,y)$. The lift zonoid $\hat{Z}(F)$ is a convex compact in $\mathbb{R}^3$ by definition: the projection $pr_J(\hat{Z}(F))$ is the lift zonoid $\hat{Z}(F^1)$, which is the area between the X variable generalized Lorenz curve and its generalized dual Lorenz curve, denoted with $GL'_X(t) = E(X) - GL_X(1-t)$, where $GL_X(t)$ represents the generalized Lorenz curve which corresponds to the graph of points $\left(\int_{-\infty}^{a} dF^1(x), \int_{-\infty}^{a} x dF^1(x)\right)$, with $a \in$

$(-\infty,+\infty)$, in $\mathbb{R}^2$. Let us compute the inverse image $pr_1^{-1}(GL_X(t))$:

$$pr_1^{-1}(GL_X(t)) = \left(\int_{-\infty}^{a} dF^1(x), \int_{(x,y):x\leq a} xF(dx,dy), \int_{(x,y):x\leq a} yF(dx,dy) \right).$$

The projection of this point under pr_2 corresponds to the following collection of points

$$\left(\int_{-\infty}^{a} dF^1(x), \int_{(x,y):x\leq a} yF(dx,dy) \right),$$

which can be expressed as

$$\left(\int_{-\infty}^{a} dF^1(x), \int_{-\infty}^{a} E(Y|X=x)F^1(dx) \right).$$

In fact, $F(dx,dy) = F(dy|dx)F^1(dx)$: by substituting it in $\int_{(x,y):x\leq a} yF(dx,dy)$, we get $\int_{-\infty}^{a} yF(dy|dx)F^1(dx) = \int_{-\infty}^{a} E(Y|X=x)F^1(dx)$. All these points define the $E(Y|X)$ generalized Lorenz curve, denoted with $GL_{E(Y|X)}$, when a ranges over $(-\infty,+\infty)$. Similarly, the inverse image of the $E(Y|X)$ generalized dual Lorenz curve, denoted with $GL'_{E(Y|X)}$, is the collection of the form

$$\left(\int_{a}^{+\infty} F^1(dx), \int_{a}^{+\infty} E(Y|X=x)F^1(dx) \right),$$

when $a \in (-\infty,+\infty)$. Let us consider the convex hull of the collection of points that characterizes $GL_{E(Y|X)}$ and $GL'_{E(Y|X)}$. Let us denote it with $\hat{Z}(E(Y|X))$. This set is centrally symmetric, therefore, since we are interested to the bidimensional context of analysis, there exists a distribution, indicated with $\mu_{E(Y|X)}$, on the real line, such that $\hat{Z}(\mu_{E(Y|X)}) = \hat{Z}(E(Y|X))$. The measure $\mu_{E(Y|X)}$ represents the *dependency measure* of Y on X (Koshevoy and Muliere (2002)).

All these notions will be very important in order to establish the partial contribution, in terms of dependence measures, realted to the introduction of a

new covariate into a linear regression model, as it will be discussed in Chapter 4.

1.4 A brief summary of the thesis contents

This section is devoted to provide a brief summary of all the original issues developed in the following three chapters.

More precisely:

- the second chapter is focused on the definition of a *concordance index*, intended as a dependence measure, in a multivariate context: for this reason a k-variate ($k > 2$) concordance index is provided by the means of a multiple linear regression function combined with the Lorenz concentration curve and its dual. Besides defining the novel index, we introduce and prove its main properties and finally we show its behavior in a practical application;

- the third chapter has as purpose the introduction of a new approach concerning the decomposition of the Gini measure in terms of concordance and discordance shares. A relevant information that assures us to derive the aforementioned decomposition is based on the fact that the Gini measure, in the univariate context, corresponds to the area between the Lorenz curve and its dual. Through this approach, a new kind of dependence, the *Gini Rank Dependence* (GRD), will be illustrated and discussed;

- in the last chapter the idea consists in focusing the attention on the Lorenz zonoid tool: when considering only the univariate case, the Lorenz zonoid corresponds to the Gini measure. Our aim is extending the Lorenz zonoids application to the multivariate context of analysis. In particular,

we consider the Lorenz zonoid of a multiple linear regression function characterized by k explanatory variables and we define the partial contribution due to the introduction of a $(k+1)$ explanatory variable in terms of dependence measures. The evident effect of a new explanatory variable introduction into the model is translated into an increase of the "explained" model variance. The final result is characterized by the definition of a new dependence measure that we call the "*Relative Gini Index*".

The final section of this PhD thesis regards the conclusions and the main research developments currently in progress.

Bibliography

[1] Block, H. W., Sampson, A. R., Savits, T. H.: *Topics in Statistics Dependence*. Institute of Mathematical Statistics, Lecture Notes, Vol16, Harward (1990)

[2] Esary, J. D., Prochan, F., Walkup, D.: *Association of random variables with applications*. Ann. Math. Stat. 38, 1466–1474 (1967)

[3] Kakwani, N. C.: *Applications of Lorenz curves in Economic Analysis*. Econometrica, No. 45 (1977)

[4] Koshevoy, G.: *Multivariate inequality indices and orderings on a product of symmetric groups*. Economica i Matematicheskie Methody, 29 (1993)

[5] Koshevoy, G.: *Multivariate Lorenz majorization*. Social Choice and Welfare, 12 (1995)

[6] Koshevoy, G., Mosler, K.: *The Lorenz Zonoids of a Multivariate Distribution*. Journal of the American Statistical Association, 91, No. 434, Theory and Methods (1996)

[7] Koshevoy, G., Mosler, K.: *Multivariate Lorenz dominance based on zonoids*. AStA Advances in Statistical Analysis, Vol. 91, No. 1 (2007)

[8] Lehmann, E. L.: *Some concpts of dependence*. The Annals of Mathematical Statistics No 37, pp 1137-1153 (1966)

[9] Leti, G.: *Descriptive Statistics* (in Italian). Il Mulino (1983)

[10] Lorenz, M. O.: *Methods of measuring the concentration of wealth Publications*. Journal of the American Statistical Association No 9, pp 209-219 (1905)

[11] Marshall, A., Olkin, I.: *Inequalities: Theory of Majorization and its Applications*. Accademic Press. New York (1979)

[12] Muliere, P.: *Some remarks about the horizontal equity of a taxation* (in Italian). Ed. by Bocconi Comunicazione **2**, (Milano, 1986)

[13] Muliere, P., Petrone, S.: *Generalized Lorenz curve and monotone dependence orderings*. Metron, Vol. L, No. 3–4 (1992)

[14] Mosler, K.: *Majorization in economic disparity measures*. Linear Algebra and its applications, 220 (1994)

[15] Musgrave, R.A.: *The Theory of Public Finance*. New York, Mc Graw Hill (1959)

[16] Tchen A. H.: *Inequalities for Distributions with Given Marginals*. Annals of Probability, Vol. 8, No. 4, pp 814-827 (1980)

Chapter 2

A multivariate Ranks-based concordance index

Abstract[1]***.*** The aim of this paper consists in defining concordance indices, as dependence measures, in a multivariate context. For this reason a k-variate $(k > 2)$ concordance index is provided through the employment of a multiple linear regression model. By considering the response variable Lorenz concentration curve and its dual, one builds the concordance curve defined as a set of points characterized by the response variable values ordered according to the ranks assigned to the corresponding linear estimates. Besides defining the novel index, we introduce and prove its main properties. We show its behaviour in a practical application.

Keywords: concordance and discordance indices, Lorenz curve, multiple linear regression model.

2.1 An introduction to concordance index problem

The issue of defining a concordance index often recurs in the statistical and economical literature. Although this presentation is general we will refer, for sake of clarity, to the taxation example throughout: in particular, the concordance index is strictly connected to the "horizontal equity" topic according to which people who own the same income level have to be taxed for the same

[1]Paper published in the volume of selected papers "Statistical Methods for the analysis of large data-sets". Springer (2011)

amount (see e.g. [6]).

The analysis is focused on considering n ordered pairs of real values, (x_i, y_i), $i = 1, 2, \ldots, n$[2], whose components describe measures of two quantitative variables referred to each element of a statistical population: let us denote with X and Y the income amount before taxation and the income amount after taxation. Our interest is in defining the i-th individual rank with respect to variable X (denoted with $r(x_i)$) and to variable Y (denoted with $r(y_i)$). Furthermore, suppose that $x_i \neq x_j$, $y_i \neq y_j$, $i \neq j$.

In a situation of perfect horizontal equity one can show that

$$r(x_i) = r(y_i), \qquad i = 1, 2, \ldots, n$$

whereas, in a situation of perfect horizontal iniquity, one gets

$$r(y_i) = n + 1 - r(x_i), \qquad i = 1, 2, \ldots, n.$$

Obviously the definition of the "horizontal equity" requires the existence of an *ordering* among individuals before taxation and the knowledge of each individual income amount after taxation. Furthermore, getting an equity index requires that the available data are referred to the single considered units and not to grouped data because the latter do not allow the identification of individuals reordering after the taxation process. The purpose is then identifying an index able to stress potential *functional monotone relations* between variables leading to study the degree of concordance or discordance among the involved variables.

The statistical literature provides a wide set of association indicators such as the Kendall-τ, the Spearman-ρ and the ranking Gini index: as well known, these indices assume values between -1 and $+1$ and, in particular one can

[2] The set of n ordered pairs of real values has to be intended as the set of pairs $(x_1, y_i), (x_2, y_2), \ldots, (x_n, y_n)$, characterized by an ordering on the i index and not on the x and y values.

show that they are equal to

$$-1 \text{ if } r(y_i) = n+1-r(x_i)(\forall i)$$
$$+1 \text{ if } r(x_i) = r(y_i)(\forall i).$$

We remark that these indices, even if invariant with respect to monotone transformations, are unfortunately based on observations ranks and the same ranks can remain unchanged also after the redistribution process in spite of each individual income extent is substantially changed.

For this reason one has to define equity measures based also on the variability of the considered extent character: a possible solution to this problem can be identified in the employment of the Lorenz curve and the dual Lorenz curve. In the taxation context the literature is limited to the bivariate case (see e.g. [4]): in the following sections we consider the extension of this problem in a more general case when one considers more then two variables.

We finally remark that R^2 is not suited in this context as it looks for linear relationships: this remark will be discussed in Sections 2.2 and 2.3.

2.2 A multivariate concordance index

The objective of this analysis concerns the definition of concordance measures in a multidimensional context: the study is then oriented to the achievement of a concordance index in presence of a random vector $(Y, X_1, X_2, \ldots, X_{k-1})$.

The procedure consists in applying a model able to describe the relation among the target variable Y and the explanatory variables $X_1, X_2, \ldots, X_{k-1}$: more precisely, let us suppose that the response variable Y assumes non-negative values. Furthermore, since this approach will be applied when the most relevant explanatory variables have categorical nature, they are always characterized by non-negative values representing the corresponding assigned

label values. In order to define a concordance index in the hypothesis that the dependent variable Y is conditioned by more then one explanatory variable, one can apply the multiple linear regression model (see e.g. [5]). Thus the estimated variable Y values can be obtained by employing the following relation

$$\hat{E}(Y_i|X_1,X_2,\ldots,X_{k-1}) = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \ldots + \hat{\beta}_{k-1} x_{(k-1)i} = \hat{y}_i.$$

The gap between the observed value y_i and the estimated value $\hat{y}_i$ (where $i = 1,2,\ldots,n$) represents the residual deviance of the model that has to be minimized (see e.g. [3]). Furthermore, let us recall that our index can be applied when $y_i \neq y_j$ and $\hat{y}_i \neq \hat{y}_j$, for every $i \neq j$.

Our purpose, is introducing and describing an index assuming the role of an alternative measure of goodness of fit for the estimated linear regression function: in particular, its properties lead it to be very useful in specific contexts of analysis, as it will be discussed in the following.

The starting point of our proposal is building the response variable Lorenz curve, L_Y (characterized by the set of ordered pairs $(i/n, (1/(nM_Y))\sum_{j=1}^{i} y_{(j)})$, where $y_{(j)}$ denotes the y_j ordered in an increasing sense and M_Y is the mean of Y) and the so called dual Lorenz curve of the variable Y, $L_Y^{'}$, (characterized by the set of ordered pairs $(i/n, (1/(nM_Y))\sum_{j=1}^{i} y_{(n+1-j)})$, where $y_{(n+1-j)}$ denotes the y_j ordered in a decreasing sense) (see e.g. [8]). The analysis proceeds in estimating the variable Y values according to the multiple linear model application. First of all we estimate the regression coefficients using the usual ordinary least square method: the purpose is getting the estimated Y values, $\hat{y}_i$, for each $i = 1,2,\ldots,n$.

Once computed the $\hat{y}_i$, one can proceed with the construction of the concordance function based on ordering the Y values with respect to the ranks assigned to the $\hat{y}_i$. Let us denote this ordering with $(y_i|r(\hat{y}_i))$ and, more specifically, with y_i^*: the set of pairs $\left(i/n, (1/(nM_Y))\sum_{j=1}^{i} y_j^*\right)$ defines the concordance

curve denoted with $C(Y|r(\hat{y}_i))$.

Through a direct comparison between the set of points that represent the Lorenz curve, L_Y, and the set of points that represent the concordance curve, $C(Y|r(\hat{y}_i))$, one can show that a perfect "overlap" occurs only if

$$\sum_{j=1}^{i} y_{(j)} = \sum_{j=1}^{i} y_j^* \text{ for every } i = 1,2,\ldots,n, \tag{2.1}$$

that is, if and only if $r(y_i) = r(\hat{y}_i)$. Note that this implies that if the residual deviance of the model decreases, the concordance index assumes higher values (meaning that the concordance degree increases), due to the fact that the y_i preserve their original ordering also with respect to $r(\hat{y}_i)$.

The result (2.1) is trivial to prove. By noticing that the set of points defining the concordance curve $C(Y|r(\hat{y}_i))$ and the Y Lorenz curve are characterized by the same x-axis values, they differ only in their y-axis values. In particular, recalling that the y-axis values related to set of Lorenz curve points correspond to $(1/(nM_Y))\sum_{j=1}^{i} y_{(i)}$ and those of the concordance curve are given by $(1/(nM_Y))\sum_{j=1}^{i} y_j^*$, then a perfect overlaps is verified if and only if $(1/(nM_Y))\sum_{j=1}^{i} y_{(i)} = (1/(nM_Y))\sum_{j=1}^{i} y_j^*$, that is if and only if $\sum_{j=1}^{i} y_{(i)} = \sum_{j=1}^{i} y_j^*$.

The comparison between the set of points that represent the Y dual Lorenz curve, $L_Y^{'}$, and the set of points that represent the concordance curve, $C(Y|r(\hat{y}_i))$, allows to conclude that there is a perfect "overlap" if and only if

$$\sum_{j=1}^{i} y_{(n+1-j)} = \sum_{j=1}^{i} y_j^* \text{ for every } i = 1,2,\ldots,n, \tag{2.2}$$

that is if and only if $r(y_i) = n+1-r(\hat{y}_i)$.

The proof of (2.2) can be obtained analogously to (2.1): in fact, recalling that the Y dual Lorenz curve is defined by the set of points of coordinates

$(i/n, (1/(nM_Y))\sum_{j=1}^{i} y_{(n+1-j)})$, then a perfect overlap between it and the concordance curve $C(y|r(\hat{y}_i)$ is provided if and only if $(1/(nM_Y))\sum_{j=1}^{i} y_{(n+1-j)} = (1/(nM_Y))\sum_{j=1}^{i} y_j^*$, that is if and only if $\sum_{j=1}^{i} y_{(n+1-j)} = \sum_{j=1}^{i} y_j^*$.

Recalling the following inequalities

$$\begin{cases} \sum_{j=1}^{i} y_j^* \geq \sum_{j=1}^{i} y_{(j)} \\ \sum_{j=1}^{n} y_j^* = \sum_{j=1}^{n} y_{(j)} \end{cases} \tag{2.3}$$

and

$$\begin{cases} \sum_{j=1}^{i} y_j^* \leq \sum_{j=1}^{i} y_{(n+1-j)} \\ \sum_{j=1}^{n} y_j^* = \sum_{j=1}^{n} y_{(n+1-j)} \end{cases} \tag{2.4}$$

one can show that $L_Y^{'} \leq C(Y|r(\hat{y}_i)) \leq L_Y$.

In order to prove that $L_Y^{'} \leq C(Y|r(\hat{y}_i)) \leq L_Y$, let us notice that all the involved three curves are characterized by the same x-axis values but different y-axis values. The concordance curve differs from the Y Lorenz curve and the Y dual Lorenz curve with regard to the measure $\sum_{j=1}^{i} y_j^*$, where $i = 1, \ldots, n$. In fact, the response variable Lorenz curve is described by the sum of its first i values ordered in an increasing sense (denoted with $\sum_{j=1}^{i} y_{(j)}$) and the response variable dual Lorenz curve is described by the sum of its first i values ordered in a decreasing sense (denoted with $\sum_{j=1}^{n} y_{(n+1-j)}$).

Since, by the employment of inequality (2.3), it results that

$$\sum_{j=1}^{i} y_j^* \geq \sum_{j=1}^{i} y_{(j)},$$

implying that $C(Y|r(\hat{y}_i)) \leq L_Y$.

In the same manner, by the employment of inequality (2.4), one obtains that

$$\sum_{j=1}^{i} y_j^* \leq \sum_{j=1}^{i} y_{(n+1-j)},$$

implying that $C(Y|r(\hat{y}_i)) \leq L_Y^{'}$. ■

***Definition* 2.2.1** *A multivariate concordance index*[3] *can then be provided: its expression is the following*

$$C_{Y,X_1,X_2,\ldots,X_{k-1}} = \frac{\sum_{i=1}^{n-1}\left\{i/n - (1/(nM_Y))\sum_{j=1}^{i} y_j^*\right\}}{\sum_{i=1}^{n-1}\left\{i/n - (1/(nM_Y))\sum_{j=1}^{i} y_{(j)}\right\}}. \tag{2.5}$$

Note that the index represents the ratio between the Y and $(Y|r(\hat{y}_i))$ concentration areas (equivalent to the ratio between the correspondent Gini indices): it enables to express the contribution of the k explanatory variables to the variable concentration. In particular the numerator of (2.5) describes the "gap" between the ordinates of points that lie on the egalitarian line and the ordinates of points that lie on the concordance curve, provided that these points have the same x-axis values: in the same manner the denominator of (2.5) defines the "gap" between the ordinates of points that lie on the egalitarian line and the ordinates of points that lie on the Y Lorenz curve.

Even if the proposed index appears as a bivariate concordance index between the y_i values and the estimated $\hat{y}_i$ values, through the employment of the Y variable linear estimated values, one defines a relation between the Y variable and the considered $k-1$ covariates implying the multivariate nature of the introduced index.

We now introduce and prove some properties of the proposed multivariate concordance index.

Property 1

Through some mathematical steps one can provide an alternative concor-

[3]The proposed index has been called "concordance index" in order to highlight that the linear dependence strength, between Y and the covariates, is based on the concordance or discordance relation between the response variable values and their corresponding linear regression estimated values. For this reason it can be considered as a *dependence index*.

dance index expression, easier to calculate

$$C_{Y,X_1,X_2,\ldots,X_{k-1}} = \frac{2\sum_{i=1}^{n} i y_i^* - n(n+1)M_Y}{2\sum_{i=1}^{n} i y_{(i)} - n(n+1)M_Y}. \tag{2.6}$$

Proof.

Let us try to simplify (2.5) by operating both in the numerator and in the denominator in the same manner.

For the numerator we get

$$C_{Y,X_1,X_2,\ldots,X_{k-1}} = \frac{M_Y\left(\sum_{i=1}^{n} i - n\right) - \sum_{i=1}^{n}\sum_{j=1}^{i} y_j^* + \sum_{j=1}^{n} y_j^*}{M_Y\left(\sum_{i=1}^{n} i - n\right) - \sum_{i=1}^{n}\sum_{j=1}^{i} y_{(j)} + \sum_{j=1}^{n} y_{(j)}}.$$

Since $\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$, one obtains

$$\begin{aligned} C_{Y,X_1,X_2,\ldots,X_{k-1}} &= \frac{M_Y[(n(n+1)/2) - n] - \sum_{i=1}^{n}\sum_{j=1}^{i} y_j^* + \sum_{j=1}^{n} y_j^*}{M_Y[(n(n+1)/2) - n] - \sum_{i=1}^{n}\sum_{j=1}^{i} y_{(j)} + \sum_{j=1}^{n} y_{(j)}} \\ &= \frac{M_Y(n(n+1)/2) - nM_Y - \sum_{i=1}^{n}\sum_{j=1}^{i} y_j^* + \sum_{j=1}^{n} y_j^*}{M_Y(n(n+1)/2) - nM_Y - \sum_{i=1}^{n}\sum_{j=1}^{i} y_{(j)} + \sum_{j=1}^{n} y_{(j)}}, \end{aligned}$$

being $nM_Y = \sum_{j=1}^{n} y_j^* = \sum_{j=1}^{n} y_{(j)}$. One can show that

$$C_{Y,X_1,X_2,\ldots,X_{k-1}} = \frac{n(n+1)M_Y - 2\sum_{i=1}^{n}\sum_{j=1}^{i} y_j^*}{n(n+1)M_Y - 2\sum_{i=1}^{n}\sum_{j=1}^{i} y_{(j)}}; \tag{2.7}$$

finally, verified that $\sum_{i=1}^{n}\sum_{j=1}^{i} y_j^* = \sum_{i=1}^{n}(n+1-i)y_j^*$ and $\sum_{i=1}^{n}\sum_{j=1}^{i} y_{(j)} = \sum_{i=1}^{n}(n+1-i)y_{(j)}$ are jointly true, we have

$$\sum_{i=1}^{n}\sum_{j=1}^{i} y_j^* = n(n+1)M_Y - \sum_{i=1}^{n} i y_i^* \tag{2.8}$$

which substituted in (2.7) gives directly

$$C_{Y,X_1,X_2,\ldots,X_{k-1}} = \frac{2\sum_{i=1}^{n} iy_i^* - n(n+1)M_Y}{2\sum_{i=1}^{n} iy_{(i)} - n(n+1)M_Y}. \quad \blacksquare \tag{2.9}$$

Note that $\sum_{i=1}^{n} iy_i$ is an arrangement increasing function. By arrangement we mean a real valued function of a vector arguments in $\mathbb{R}^n$ that increases in value if the components of the vector arguments become more similarly arranged (see e.g. [1]).
We can thus conclude that:

Property 2

$$-1 \leq C_{Y,X_1,X_2,\ldots,X_{k-1}} \leq +1. \tag{2.10}$$

Proof.
To prove (2.10) it is sufficient to prove that $\sum_{i=1}^{n} iy_{(i)} \geq \sum_{i=1}^{n} iy_i^*$. This can be proved, for instance, directly by looking at the inequalities (2.3) and (2.4): since

$$\sum_{j=1}^{i} y_j^* \geq \sum_{j=1}^{i} y_{(j)}$$

is true for all i, we also have that

$$\sum_{i=1}^{n}\sum_{j=1}^{i} y_j^* \geq \sum_{i=1}^{n}\sum_{j=1}^{i} y_{(j)};$$

now, because of the aforementioned relationship (2.8) we have

$$n(n+1)M_Y - \sum_{i=1}^{n} iy_i^* \geq n(n+1)M_Y - \sum_{i=1}^{n} iy_{(i)},$$

which gives $\sum_{i=1}^{n} iy_{(i)} \geq \sum_{i=1}^{n} iy_i^*$. $\blacksquare$

Property 3 $C_{Y,X_1,X_2,\ldots,X_{k-1}} = +1$ if and only if the concordance function over-

laps with the Lorenz curve.

Proof.

The concordance function overlaps with the Lorenz curve if and only if $\sum_{j=1}^{i} y_{(j)} = \sum_{j=1}^{i} y_j^*$. This implies that $r(y_i) = r(y_i^*)$ for every $i = 1, 2, \ldots, n$. ∎

Property 4

$C_{Y,X_1,X_2,\ldots,X_{k-1}} = -1$ if and only if concordance function overlaps with the dual Lorenz curve.

Proof.

This property can be proved, similarly to Property 3, from the inequality (2.4) noticing that:

$$\sum_{i=1}^{n} (n+1-i) y_{(i)} = \sum_{i=1}^{n} y_{(n+1-i)} i,$$

so

$$\sum_{i=1}^{n} i y_{(i)} = n(n+1) M_Y - \sum_{i=1}^{n} y_{(n+1-i)} i$$

and therefore, by applying this equivalence in the denominator of (2.9), we get an equivalent formulation of the concordance index based on $L_Y^{'}$:

$$C_{Y,X_1,\ldots,X_k} = \frac{2\sum_{i=1}^{n} i y_i^* - n(n+1)M_Y}{n(n+1)M_Y - 2\sum_{i=1}^{n} i y_{(n+1-i)}} = -\frac{2\sum_{i=1}^{n} i y_i^* - n(n+1)M_Y}{2\sum_{i=1}^{n} i y_{(n+1-i)} - n(n+1)M_Y}.$$

Finally, since from the inequality (2.4) we have $\sum_{j=1}^{i} y_j^* \leq \sum_{j=1}^{i} y_{(n+1-i)}$, $\forall i$, the result follows as in the proof of Property 3. ∎

These results have been obtained also in [4] with regard to the bivariate context: our analysis represents an extension to the multivariate case (when $k > 2$).

An alternative concordance measure, which provides a measure of distance between the concordance function and the Y Lorenz curve, is the Plot-

nick indicator (see e.g. [7]) expressed by

$$I^*_{Y,X_1,X_2,\ldots,X_{k-1}} = \frac{\sum_{i=1}^n iy_{(i)} - \sum_{i=1}^n iy_i^*}{2\sum_{i=1}^n iy_{(i)} - (n+1)\sum_{i=1}^n y_{(i)}}. \tag{2.11}$$

We remark that, the construction that uses the Y values ordered according to the ranks of the corresponding linear estimated values, is a novel construction with respect to the one proposed by Plotnick (1981). In fact, the Plotnick index has been introduced as an iniquity indicator able to measure the distance between the concordance curve and the Lorenz curve when considering the taxation context, as described in [4]. In our analysis, the Plotnick indicator has been adapted to the multivariate setting by exploiting the already discussed ranks-based approach.

The Plotnick index assumes values in the range $[0,+1]$: in particular,

- $I^*_{Y,X_1,X_2,\ldots,X_{k-1}} = 0$, in a perfect concordance situation;
- $I^*_{Y,X_1,X_2,\ldots,X_{k-1}} = +1$, in a perfect discordance situation.

A direct comparison between this indicator and the multivariate concordance index (2.6) is needed in order to discuss about the corresponding lower and upper bounds. The upper bound of the Plotnick indicator conceptually coincides with the lower bound of the multivariate concordance index whereas, its lower bound conceptually corresponds to the multivariate concordance index upper bound.

Let us consider the upper bound of the concordance index.

If $C_{Y,X_1,X_2,\ldots,X_{k-1}} = +1$, by exploiting (2.6), one can conclude that $2\sum_{i=1}^n iy_i^* = 2\sum_{i=1}^n iy_{(i)} \Rightarrow \sum_{i=1}^n iy_i^* = \sum_{i=1}^n iy_{(i)} \Rightarrow r(y_i) = r(\hat{y}_i)$. Since a concordance index assuming value $+1$ defines a perfect concordance situation, in the same manner, a Plotnick index with value 0 represents the same situation. In fact, if $I^*_{Y,X_1,X_2,\ldots,X_{k-1}} = 0$, by exploiting the (2.11) one obtains that $\sum_{i=1}^n iy_{(i)} = \sum_{i=1}^n iy_i^* \Rightarrow r(y_i) = r(\hat{y}_i)$.

Let us now focus the attention on the lower bound of the concordance index. If $C_{Y,X_1,X_2,\ldots,X_{k-1}} = -1$, it follows that $\sum_{i=1}^{n} i y_i^* + \sum_{i=1}^{n} i y_{(i)} = n(n+1)M_Y$; being $M_Y = (1/n)\sum_{i=1}^{n} y_{(i)}$[4], it results that $\sum_{i=1}^{n} i y_i^* + \sum_{i=1}^{n} i y_{(i)} = (n+1)\sum_{i=1}^{n} y_{(i)} \Rightarrow r(y_i) = n+1-r(\hat{y}_i)$. In the same manner, if $I^*_{Y,X_1,X_2,\ldots,X_{k-1}} = +1$ then $\sum_{i=1}^{n} i y_i^* + \sum_{i=1}^{n} i y_{(i)} = (n+1)\sum_{i=1}^{n} y_{(i)}$, meaning exactly that $r(y_i) = n+1-r(\hat{y}_i)$.

2.2.1 Final remarks

Properties 2 - 4 allow us to give the interpretation of the introduced index. First of all it can be usefully employed in order to study dependence between a dependent quantitative character (continuous or discrete) and a number of explanatory ones: the kind of dependence, which is captured by the employment of the aforementioned index, is expressed in terms of concordance between the y_i values and their corresponding linear estimated values $\hat{y}_i$. In this context, the term "concordance" is used in order to highlight that the ranks assigned to the linear estimated values are unchanged with respect to those assigned to the corresponding original Y values.

Furthermore, the reason that allows to conclude that the concordance between the Y values and its linear estimated values holds can be motivated in the following manner. By the means of a linear regression function, we define the estimated values vector $\hat{Y}$ as a function of the explanatory variables $X_1, \ldots, X_{k-1}$: when using a multiple linear regression function, one obtains information about the variable mean value variation as a consequence of the explanatory variables variation. The linear regression model satisfies the property of variance decomposition which allows to define the R^2 index intended as a goodness of fit measure. Obviously, the greater the residual deviance (that summarizes the residual values), the worst the model. Therefore,

[4] The Y variable mean value is unchanged even if one considers its values ordered in an increasing sense ($y_{(i)}$, where $i = 1, \ldots, n$).

the proposed index could represent a useful measure in evaluating the model performance: in fact, if the residuals associated to each observation assume small values (providing an high R^2 value) the ranks assigned to the linear Y variable estimated values tend to be the same of those related to the original ones. This result provides the satisfaction of a concordant relation between the response variable Y and its estimated values vector $\hat{Y}$: since the $\hat{Y}$ is obtained through a multiple linear regression function with $k-1$ covariates, one can conclude that the selected model has a good performance. On the other hand, if the the residuals assume high values and, in particular, lead to a reversal in the original Y variable ordered values, one can conclude the relation between Y and $\hat{Y}$ is discordant: more precisely, negative values imply that the kind of dependence among the response variable Y and $X_1,\ldots,X_{k-1}$ is not described by a linear regression function, but it could be well fitted by an alternative regression function. To summarize, the concordance relation between Y and $\hat{Y}$ implies the existence of a linear dependence relation among Y and the $k-1$ covariates and to suggest which regression function best explains the response variable.

On the basis of the above remarks, when the proposed index assumes positive values, there is evidence of (non linear) dependence[5] between Y and $\hat{Y}$ and, therefore, the fitted model well explains the observed values. Of course, the higher the value, the better. On the other hand, when the index assumes negative values, there is evidence of (non linear) dependence between Y and $\hat{Y}$ which implies weak linear dependence between Y and the explanatory variables. The higher the (negative) value, the worse is such linear dependence among Y and $X_1,\ldots,X_{k-1}$.

Finally, when index values are around zero, we are in an intermediate situation, because of ranks compensation between different fitted values.

[5]Linear dependence is satisfied indeed among the response variable Y and the $k-1$ explanatory variables through the adopted linear estimated regression model.

Our proposal, concerning the construction of a concordance index through the application of the Lorenz curves, is more appropriate than simply comparing the ranks of the response variable values with those of the corresponding linear estimated values. In fact, by building the concordance curve (which lies between the Y Lorenz curve and its dual) one can define the single contribution due to each observation in terms of concordance and discordance between Y and $\hat{Y}$ and not only between their ranks: thereby, one can establish each observation contribution to the fitting of data provided by the estimated multiple linear regression model.

We also remark that the proposed index assumes a relevant role when evaluating model performance, especially when the analysis context is characterized by categorical variables: this topic represents the key contribution of our approach.

All these conclusions have been supported by the results obtained through a simulation study that we now briefly report. For sake of simplicity, let us consider only an explanatory variable: obviously one can extend this computations with regard to linear regression models defined by more than one covariate. The aim consists in evaluating the values assumed by our proposed index in the following different cases:

- the response variable Y and the covariate X are positively correlated;
- the response variable Y and the covariate X are negatively correlated;
- the response variable Y and the covariate X are not correlated.

In our simulation study, since our response variable assumes only non-negative values, we generate random numbers from a Chi-Square distribution instead of a Normal distribution. Furthermore, since we are dealing with categorical explanatory variables, we generate random numbers from a Poisson distribution (each number represents a label assigned to the categorical covariate).

Let us now introduce and comment the obtained results.

Suppose first that Y and X_1 are positive correlated. The proposed concordance index results to be $C_{Y,X_1} \cong 0.166$ against an $R^2 = 0.2856$: the smaller value of the concordance index means that, the linear dependence among Y and the covariate X is not strong meaning that the ranks assigned to the original response variable have a different position with respect to the ranks of its corresponding estimated values.

The second simulation study, characterized by a negative correlation between Y and X, provides a concordance index $C_{Y,X_1} \cong 0.573$, whereas the $R^2 = 0.1821$ means that our index captures better the linear dependence relation between Y and X.

The last step is based on computing the concordance index when there is not a linear correlation between Y and the covariate X: thereby, the R^2 assumes a value very close to 0 (more precisely $R^2 = 0.00003947$). With regard to our index it assumes an high negative value ($C_{Y,X_1} \cong -0.754$) meaning that the adopted estimated linear regression model does not fit the data as it leads to ranks reversal: in fact, between the response variable Y and the covariate X does not exists a linear dependence relation.

2.2.2 Application

In order to illustrate our proposed index we introduce a simple data example. Suppose to have data concerning 18 business companies on three characters: Sales Revenues (Y) (expressed in thousands of Euros), Selling price (X_1) (expressed in Euros) and Advertising investments (X_2) (expressed in thousand of Euros). These data are shown in Table 2.1.

The model used to describe relations among the involved variables is based on linear regression. The application of ordinary least square method leads to the following estimated linear regression coefficients $\hat{\beta}_0 \equiv 98.48$, $\hat{\beta}_1 \equiv 0.63$,

$\hat{\beta}_2 \equiv 4.57$ so the regression line is

$$\hat{y}_i = 98.48 + 0.63x_{1i} + 4.57x_{2i}$$

Once obtained the estimated Y values, in order to build our index we need

ID Business company	Sales revenues	Selling price	Advertising investments
01	350	84	45
02	202	73	19
03	404	64	53
04	263	68	31
05	451	76	58
06	304	67	23
07	275	62	25
08	385	72	36
09	244	63	29
10	302	54	39
11	274	83	35
12	346	65	49
13	253	56	22
14	395	58	61
15	430	69	48
16	216	60	34
17	374	79	51
18	308	74	50

Table 2.1: *Data describing Sales revenues, Selling price and Advertising investments expressed in Euros*

to assign their ranks and finally order the Y values according to $\hat{y}_i$ ranks. All the results are summarized in Table 2.2. From Table 2.2 we can compute the multivariate concordance index using (2.5), recalling that y_i^* represent the Y variable values ordered with respect to the $\hat{y}_i$ ranks. The concordance index assumes the value 0.801 proving that there is a strong concordance relation among the response variable Y and the explanatory variables X_1, X_2: this conclusion is clear in Figure 2.1, where the concordance curve (denoted with the continuous black line), is very close to the Y variable Lorenz curve (denoted

y_i	$r(y_i)$	$\widehat{y}_i$	ordered $\widehat{y}_i$	$r($ ordered $\widehat{y}_i)$	y_i ordered by $r($ ordered $\widehat{y}_i)$
202	1	231.07	231.07	1	202
216	2	291.41	234.10	2	253
244	3	270.46	245.57	3	304
253	4	234.10	251.56	4	275
263	5	282.73	270.46	5	244
274	6	310.41	282.73	6	263
275	7	251.56	291.41	7	216
302	8	310.47	308.07	8	385
304	9	245.57	310.41	9	274
308	10	373.26	310.47	10	302
346	11	363.04	357.71	11	350
350	12	356.71	360.99	12	430
374	13	380.97	363.04	13	346
385	14	308.07	373.26	14	308
395	15	413.45	380.68	15	404
404	16	380.68	380.97	16	374
430	17	360.99	411.05	17	451
451	18	411.05	413.45	18	395

Table 2.2: *Results*

with the dashed dot line).

A further verification of this result is provided by the Plotnick indicator (2.11), whose numerical value is very close to 0, meaning that the distance between concordance function and the Lorenz curve is close to the minimum.

2.3 Conclusions

In this chapter we have introduced a novel multivariate concordance index that can be usefully employed to study dependence between quantitative characters and a number of explanatory ones. This index is of simple calculation, has some interesting properties and can be compared to known alternatives, such as the Plotnick index. Furthermore, it can be very useful as a measure of fit when the relevant explanatory variables have categorical nature because it is based on the response variable values ordered according to the ranks as-

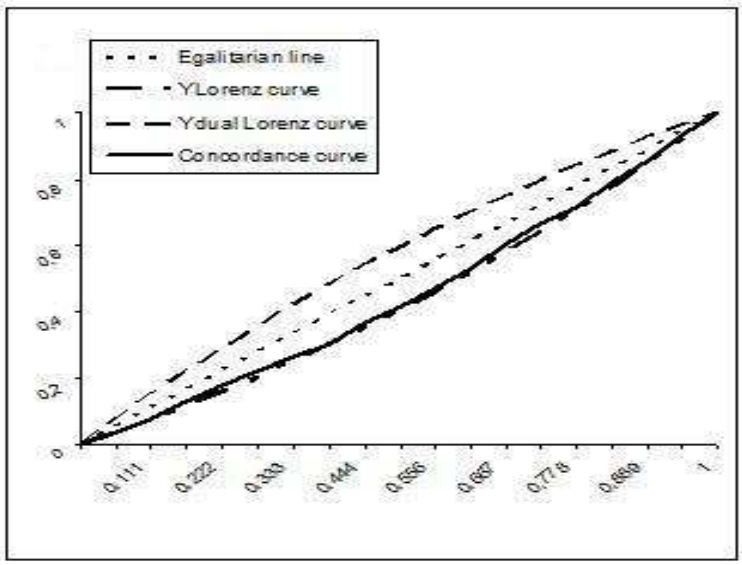

Figure 2.1: Y Lorenz curve, Y dual Lorenz curve and concordance function

signed to their corresponding estimated values. A positive value of the concordance index means a dependent relation between Y and $\hat{Y} = f(X_1, \ldots, X_{k-1})$, whereas if the index value is negative such dependence relation is weak.

Note that the index can be applied beyond the linear regression context as long as the response variable is quantitative, discrete or continuous and the adopted model is a regression tree or a polynomial regression (see e.g. [2]). The extension to ordinal and categorical further topic could represent a possible future and interesting development context.

Bibliography

[1] Boland, P.J., Proschan, F.: *Multivariate arrangement increasing functions with applications in probability and statistics*. Journal of Multivariate Analysis, Vol. 25, Issue 2, 286–298 (1988)

[2] Giudici, P.: *Applied Data Mining*. Wiley (2003)

[3] Leti, G.: *Descriptive Statistics* (in Italian). Il Mulino (1983)

[4] Muliere, P.: *Some remarks about the horizontal equity of a taxation* (in Italian). Ed. by Bocconi Comunicazione, Vol. 2, (Milano, 1986)

[5] Muliere, P., Petrone, S.: *Generalized Lorenz curve and monotone dependence orderings*. Metron, Vol. L, No. 3–4 (1992)

[6] Musgrave, R.A.: *The Theory of Public Finance*. New York, Mc Graw Hill (1959)

[7] Plotnick, R.: *A Measure of Horizontal Inequity*. The review of Economics and Statistics, **2**, 283–288 (1981)

Chapter 3

On the Gini measure decomposition

Abstract[1]***.*** The purpose of this paper is to introduce a new approach to the decomposition of the Gini measure in terms of concordance and discordance shares: a new kind of dependence, the *Gini Rank Dependence* (*GRD*), and its formal definition are provided.

Keywords: Gini measure, concordance curve, Gini Rank Dependence.

3.1 Introduction

The Gini measure decomposition always assumed the role of partitioning the total inequality of a population into two components, concerning the inequality between and within subpopulations[2]. Theil (1967), partitioning the total population into *h* subpopulations, decomposed the total inequality T into the inequality *within* (T_w) and *between* (T_b) the h subpopulations, such that $T = T_w + T_b$ (where T_b is Theil inequality between the income means of the h subpopulations weighted by the subpopulations sizes). This decomposition approach stimulated further research: for example, researchers concentrated on the Gini ratio, deriving important transformations to capture the idea of decomposability (see, for example, [9] and [11]). More recently Dagum (1997)

[1] Paper published in Statistics and Probability Letters 81, pp. 133-139 (January, 2011)

[2] This condition is true only in case of not overlapping subpopulations.

suggested to decompose the Gini ratio into three components: the Gini inequality *within* the subpopulations (G_w), the Gini inequality *between* subpopulations (G_b) and the intensity of transvariation between subpopulations.

Our research aim consists in proposing a new approach devoted to employ the Gini measure decomposition in terms of concordance and discordance. Section 3.2 describes the main statistical tools useful in characterizing this new approach. Section 3.3 is focused on defining a new kind of dependence, the *Gini rank dependence* (GRD), whose formal definition is provided. Finally, Section 3.4 is devoted to the conclusion and further research developments.

3.2 Background

The aim of this section is introducing the main topics concerning the statistical tools needed in obtaining the Gini measure decomposition in terms of multivariate concordance and discordance.

These topics have been illustrated in order to define a new multidimensional concordance index through the employment of the Lorenz curves (see e.g. [8]): in this context of analysis, we now recall all the background elements that allow to establish the decomposition of the Gini measure through the so called concordance curve.

Let us suppose to consider a k-variate random vector $(Y, X_1, X_2, \ldots, X_{k-1})$, on which one can apply a model able to describe the relation among a response variable Y and the explanatory variables $X_1, X_2, \ldots, X_{k-1}$ (for more details see [4]). For this purpose one can employ, for example, the estimated multiple linear regression function

$$\hat{E}(Y|X_1, X_2, \ldots, X_{k-1}) = \hat{Y}. \tag{3.1}$$

Given the response variable Y, our starting point is based on building its Lorenz curve, denoted with L_Y, and its dual, denoted with $L_Y^{'}$: the former is characterized by the set of ordered pairs $(i/n, (1/(nM_Y))S_i)$, $i = 1, \ldots, n$, where $S_i = \sum_{j=1}^{i} y_{(j)}$, denoting the sum of the y_i ordered in an increasing sense and M_Y is the Y variable mean. The latter is characterized by the set of ordered pairs $(i/n, (1/(nM_Y))S_i^{'})$, where $S_i^{'} = \sum_{j=1}^{i} y_{(n+1-j)}$ denotes the sum of the y_i ordered in a decreasing sense. Once computed the estimated Y values, $\hat{y}_i$, through (3.1), one can proceed by the construction of the concordance curve based on ordering the Y values with respect to the ranks assigned by its estimated values $\hat{y}_i$. Let us denote this ordering with $(y_i|r(\hat{y}_i))$ and, more specifically, by y_i^*. The set of pairs $(i/n, (1/(nM_Y))\sum_{j=1}^{i} y_j^*)$ defines the concordance curve, denoted with $C(Y|r(\hat{y}_i))$.

Through a direct comparison between the set of points that represent the Y Lorenz curve, L_Y, and the set of points that represent the concordance curve, $C(Y|r(\hat{y}_i))$, one can show that a perfect "overlap" is provided only if

$$\sum_{j=1}^{i} y_{(j)} = \sum_{j=1}^{i} y_j^* \text{ for every } i = 1, 2, \ldots, n, \tag{3.2}$$

that is, if and only if $r(y_i) = r(\hat{y}_i)$. The further comparison between the set of points that represent the Y dual Lorenz curve, $L_Y^{'}$, and the set of points that represent the concordance curve, $C(Y|r(\hat{y}_i))$, allows to conclude that there is a perfect "overlap" if and only if

$$\sum_{j=1}^{i} y_{(n+1-j)} = \sum_{j=1}^{i} y_j^* \text{ for every } i = 1, 2, \ldots, n. \tag{3.3}$$

that is, if and only if $r(y_{(n+1-i)}) = r(\hat{y}_i)$.

Note that the egalitarian line, which represents the bisector of the unit side square in which the Lorenz and the dual Lorenz curves lie, splits the Gini

measure in two equal parts: we call the upper area, located between the egalitarian line and the Y dual Lorenz curve, the *discordance area* while the lower area, located between the egalitarian line and the Y Lorenz curve, the *concordance area*.

3.3 The Gini measure decomposition: a proposal

This section is devoted to the decomposition of the Gini measure. The role of the Gini measure consists in describing the statistical dispersion and, therefore, one can consider it as a variability measure.

3.3.1 The concentration curve, its dual and the Gini measure

Let us start our analysis providing the definition of the Gini measure. To help our illustration, Figure 3.1 represents an example of a Lorenz and a dual Lorenz curve.

The Gini measure is defined as the ratio of the areas on the Lorenz curve

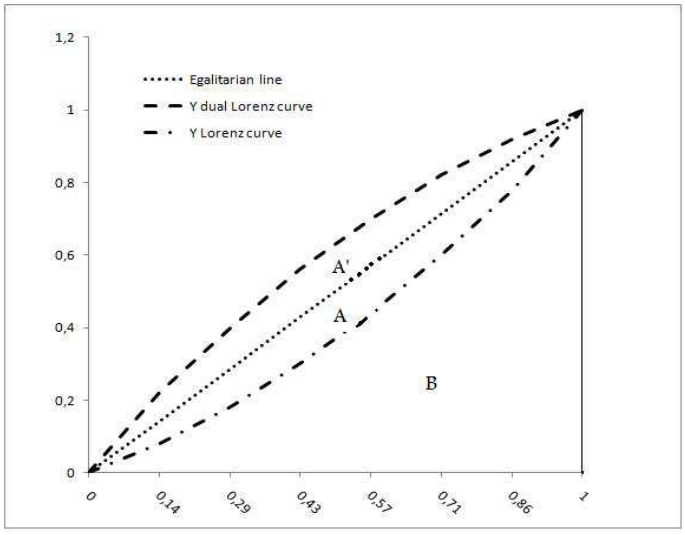

Figure 3.1: The Gini measure diagram: Gini measure=$A^{'}+A$.

diagram. If the area between the egalitarian line and the Lorenz curve is A

(see Figure 3.1), and the area under the Lorenz curve is B, then the Gini measure is $A/(A+B)$. Since $A+B=0.5$, the Gini measure can be defined as $G=A/(0.5)=2A=1-2B$. If the Lorenz curve of the Y variable is denoted with $L_Y(t)$, the value of B can be found by integration and

$$G=1-2\int_0^1 L_Y(t)dt, \text{ with } 0\leq t\leq 1. \tag{3.4}$$

It is trivial to establish that if the concordance curve $C(Y|r(\hat{y}_i))$ is close to the Y dual Lorenz curve, the relationship between the response variable Y and the estimated response $\hat{Y}$ based on the explanatory variables, points towards discordance, in the other case towards concordance. For this reason we call the area between the Y dual Lorenz curve and the egalitarian line "*discordance area*" and the area between the Y Lorenz curve and the egalitarian line "*concordance area*".

Let us now provide the measures associated to these two different areas. The concordance area (CA) can be defined as follows:

$$CA=\frac{1}{2}-\int_0^1 L_Y(t)dt \text{ with } 0\leq t\leq 1; \tag{3.5}$$

on the other hand, being the Y dual Lorenz curve, denoted with $L_Y^{'}(t)$, equivalent to $1-L_Y(1-t)$ (see e.g. [6]), with $0\leq t\leq 1$, the discordance area (DA) can be computed as follows:

$$\begin{aligned} DA &= \int_0^1 L_Y^{'}(t)dt-\frac{1}{2}=\int_0^1 [1-L_Y(1-t)]dt-\frac{1}{2} \\ &= 1-\int_0^1 L_Y(1-t)dt-\frac{1}{2}=\frac{1}{2}-\int_0^1 L_Y(1-t)dt. \end{aligned} \tag{3.6}$$

The relationship between the Gini measure and the discordance and concordance areas can thus be expressed as

$$G = CA + DA = \frac{1}{2} - \int_0^1 L_Y(t)dt + \frac{1}{2} - \int_0^1 L_Y(1-t)dt \qquad (3.7)$$
$$= 1 - \int_0^1 L_Y(t)dt - \int_0^1 L_Y(1-t)dt = 1 - 2\int_0^1 L_Y(t)dt,$$

since $\int_0^1 L_Y(1-t)dt = \int_0^1 L_Y(t)dt$, for all $0 \leq t \leq 1$. Note that, in particular, the maximum concordance and discordance areas are equal to each other and assume value $G/2$.

A direct implication of the previous definition is that the Gini measure corresponds to the area between the Y variable Lorenz curve and the Y variable dual Lorenz curve, that is the area given by the sum of A and A' in Figure 3.1.

This chapter proposes a decomposition of the mutual variability, analogous to the residual analysis, in terms of the mutual variability "explained" by the Y rank dependence with respect to the explanatory variables[3].

3.3.2 The Gini Rank Dependence

Our main purpose is now decomposing the Gini measure in terms of concordance and discordance. More precisely, we define the share of the Gini measure which corresponds to a concordance or to a discordance situation between Y and $\hat{Y}$. This kind of proceeding implies the study of a new form of dependence that we call *rank dependence*: the measure associated to this dependence form will be called *Gini Rank Dependence* and denoted with GRD.

Our proposal can be explained considering three different cases illustrated in Figure 3.2, 3.3 and 3.4.

[3] Let us recall that the Gini measure can be interpreted as a measure of mutual variability, based on differences $|y_i - y_j|$, as opposed to variability, based on differences $|y_i - \mu|$.

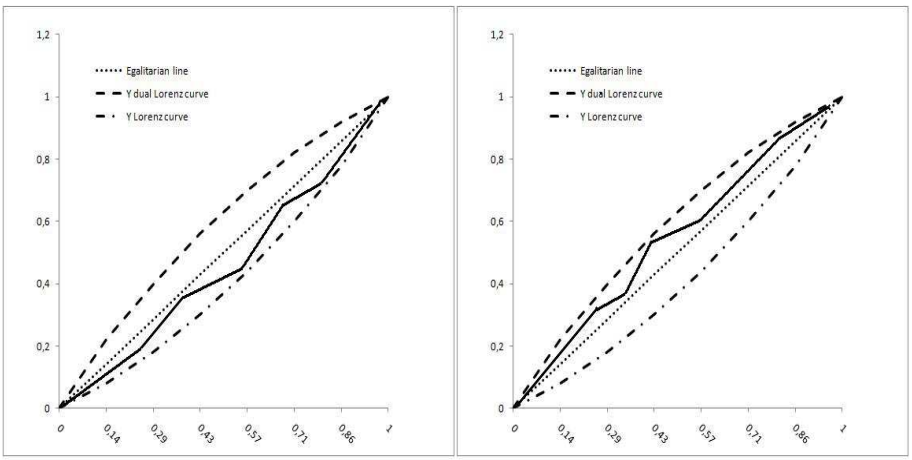

Figure 3.2: a) Case 1 and b) Case 2. The concordance curve is the continuous curve.

Case 1: the concordance curve $C(Y|r(\hat{y}_i))$ completely lies in the concordance area.

Our aim is to measure the "concordance" degree: in other terms, to define the measure of the Y variable concentration "explained" by the Y rank dependence with respect to the explanatory variables.

The Gini measure, explained by the rank dependence (GRD), assumes the following expression

$$GRD = \frac{1}{2} - \int_0^1 C(Y|r(\hat{y}_i))dy : \tag{3.8}$$

which can be normalized by dividing it by $G/2$.

Case 2: the concordance curve $C(Y|r(\hat{y}_i))$ completely lies in the discordance area.

In this case the aim is the mirror image of the previous one, in fact one has to define the measure of the Y variable concentration "explained" by the Y rank dependence with respect to the explanatory variables, so

$$GRD = \int_0^1 C(Y|r(\hat{y}_i))dy - \frac{1}{2}, \tag{3.9}$$

which can be normalized by dividing it by $G/2$.

Case 3: the concordance curve $C(Y|r(\hat{y}_i))$ partially lies in the concordance area and partially in the discordance area meaning that between the concordance curve and the egalitarian line there are one or more intersection points. This case is a little more complex than the first two and deserves a further development: we need to describe the main conditions that can occur in relation to the number of intersection points, between the concordance curve and the egalitarian line, and the concordance curve initial position with respect to the concordance or discordance area. More precisely, in order to measure the concordance and discordance areas, one has to consider the following steps:

1. identifying the number of intersection points between the concordance curve and the egalitarian line;

2. defining the concordance and discordance "extent" by the employment of a series of integrals whose integration extremes are represented by the intersection points x-axis values.

Through some examples we can get the "size" of Gini measure explained by the rank dependence in terms of concordance and discordance, distinguishing between two main situations, as follows.

Even number of intersection points. Let us start considering an even number of intersection points: for instance let us suppose that the concordance curve intersects the egalitarian line in two points, A and B, respectively of x-axis values a_1 and a_2.

The subsequent step concerns the identification of the the first segment position of the concordance curve with respect to the concordance or discordance area: let us denote with Con the concordance area share and with Dis

the discordance area share, whose bounds are defined by the concordance and the egalitarian curves. Two subcases have to be taken into account.

Subcase 1: the first segment of the concordance curve lies in the discor-

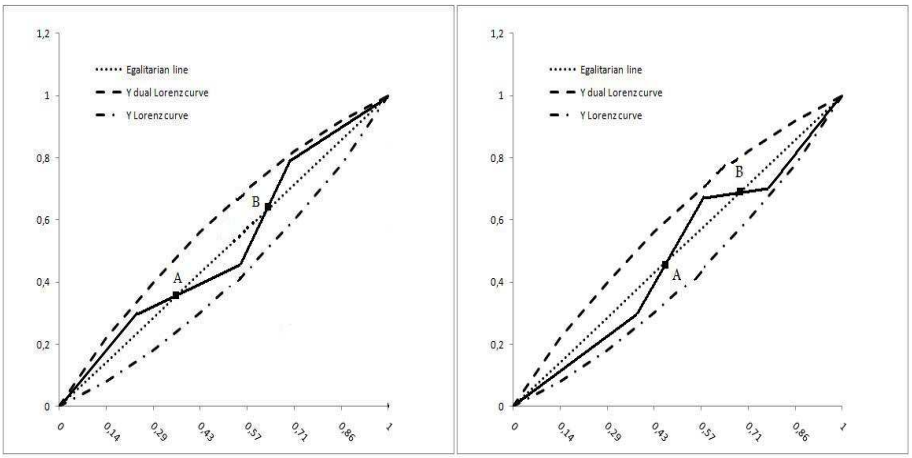

Figure 3.3: a) Subcase 1: even number of intersection points and first segment of the concordance curve located in the discordance area and b) Subcase 2: even number of intersection points and first segment of the concordance curve located in the concordance area.

dance area.

Since $Con = \int_{a_1}^{a_2} t\,dt - \int_{a_1}^{a_2} C(Y|r(\hat{y}_i))dy$, with $0 \leq t \leq 1$, and $Dis = \left[\int_0^{a_1} C(Y|r(\hat{y}_i))\right.$ $\left. dy - \int_0^{a_1} t\,dt\right] + \left[\int_{a_2}^{1} C(Y|r(\hat{y}_i))dy - \int_{a_2}^{1} t\,dt\right]$ the measure of the Y variable concentration "explained" by the Y rank dependence with respect to the explanatory variables is obtained by the following expression:

$$GRD = Con + Dis = \int_0^{a_1} C(Y|r(\hat{y}_i))dy - \int_{a_1}^{a_2} C(Y|r(\hat{y}_i))dy + \\ + \int_{a_2}^{1} C(Y|r(\hat{y}_i))dy + \left[-a_1^2 + a_2^2 - \frac{1}{2}\right].$$

Subcase 2: the first segment of the concordance curve lies in the concordance area.

The GRD is obtained by the following expression:

$$GRD = Con + Dis = -\int_0^{a_1} C(Y|r(\hat{y}_i))dy + \int_{a_1}^{a_2} C(Y|r(\hat{y}_i))dy + \\ -\int_{a_2}^{1} C(Y|r(\hat{y}_i))dy + \left[a_1^2 - a_2^2 + \frac{1}{2}\right],$$

being $Con = \left[\int_0^{a_1} t dt - \int_0^{a_1} C(Y|r(\hat{y}_i))dy\right] + \left[\int_{a_2}^{1} t dt - \int_{a_2}^{1} C(Y|r(\hat{y}_i))dy\right]$ and $Dis = \int_{a_1}^{a_2} C(Y|r(\hat{y}_i)dy - \int_{a_1}^{a_2} t dt$.

Odd number of intersection points. Let us continue considering an odd number of intersection points: for instance, let us suppose that the concordance curve intersects the egalitarian line in three points, A, B and C respectively of x-axis values a_1, a_2 and a_3. Then, as described previously, we have to proceed to the identification of the first segment position of the concordance curve with respect to the concordance or discordance area. We have to take into account two subcases, as before.

Subcase 1: the first segment of the concordance curve lies in the discor-

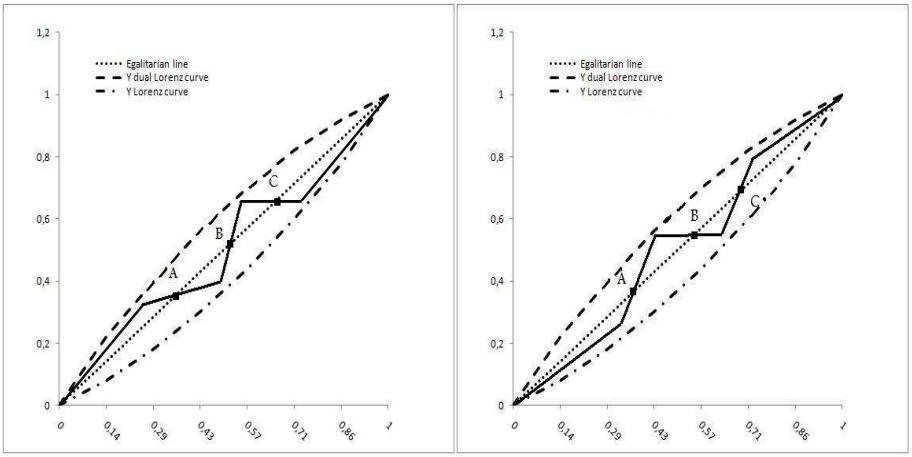

Figure 3.4: a) Subcase 1: odd number of intersection points and first segment of the concordance curve located in the discordance area, b) Subcase 2: odd number of intersection points and first segment of the concordance curve located in the concordance area.

dance area.

The measure of the Y variable concentration "explained" by the Y rank dependence with respect to the explanatory variables is obtained by the following expression:

$$GRD = Con + Dis = \int_0^{a_1} C(Y|r(\hat{y}_i))dy - \int_{a_1}^{a_2} C(Y|r(\hat{y}_i))dy + \\ + \int_{a_2}^{a_3} C(Y|r(\hat{y}_i))dy - \int_{a_3}^{1} C(Y|r(\hat{y}_i))dy + \left[-a_1^2 + a_2^2 - a_3^2 + \frac{1}{2} \right],$$

where $Con = \left[\int_{a_1}^{a_2} t dt - \int_{a_1}^{a_2} C(Y|r(\hat{y}_i))dy \right] + \left[\int_{a_3}^{1} t dt - \int_{a_3}^{1} C(Y|r(\hat{y}_i))dy \right]$ and $Dis = \left[\int_0^{a_1} C(Y|r(\hat{y}_i))dy - \int_0^{a_1} t dt \right] + \left[\int_{a_2}^{a_3} C(Y|r(\hat{y}_i))dy - \int_{a_2}^{a_3} t dt \right]$.

Subcase 2: the first segment of the concordance curve lies in the concordance area.

The GRD is obtained by the following expression:

$$GRD = Conc + Dis = -\int_0^{a_1} C(Y|r(\hat{y}_i))dy + \int_{a_1}^{a_2} C(Y|r(\hat{y}_i))dy + \\ - \int_{a_2}^{a_3} C(Y|r(\hat{y}_i))dy + \int_{a_3}^{1} C(Y|r(\hat{y}_i))dy + \left[a_1^2 - a_2^2 + a_3^2 - \frac{1}{2} \right],$$

being $Con = \left[\int_0^{a_1} t dt - \int_0^{a_1} C(Y|r(\hat{y}_i))dy \right] + \left[\int_{a_2}^{a_3} t dt - \int_{a_2}^{a_3} C(Y|r(\hat{y}_i))dy \right]$ and $Dis = \left[\int_{a_1}^{a_2} C(Y|r(\hat{y}_i))dy - \int_{a_1}^{a_2} t dt \right] + \left[\int_{a_3}^{1} C(Y|r(\hat{y}_i))dy - \int_{a_3}^{1} t dt \right]$.

3.3.3 The GRD generalization formula

The previous construction inductively suggests to find a general formulation of GRD. Through the computations related to the GRD definition, one can obtain a recursive form consisting in alternate signs in integral terms and in constant terms, represented by the intersection points x-axis values: this recursive for-

mula depends either on the nature of the intersection points number as well as on the first segment position of the concordance curve with respect to the concordance or discordance area.

Let us suppose to have p intersection points (whose x-axis values can be denoted with $a_1, a_2, \ldots, a_p$).

Before proceeding we have to define some conditions:

- $a_j = 1$ with $j = p+1$;
- $a_{j-1} = 0$ with $j = 1$.

In order to achieve a general formulation, we need to take into account the difference between the even or odd number of the intersection points but also the first segment concordance curve position with respect to the concordance and discordance area. We thus need to consider a further term in the expression characterizing GRD. This term is a multiplicative factor equivalent to $(-1)^s$, where s can assume only two values 0 or 1: in particular

- if $s = 0$, then the first segment of the concordance curve is located in the discordance area;
- if $s = 1$, then the first segment of the concordance curve is located in the concordance area.

In conclusion, GRD can be defined as:

$$GRD = (-1)^s \left\{ \sum_{j=1}^{p+1} (-1)^{j+1} \left[\int_{a_{j-1}}^{a_j} C(Y|r(\hat{y}_i))dy - a_j^2 \right] + \frac{1}{2} \right\}, \qquad (3.10)$$

for p even and

$$GRD = (-1)^{s+p} \left[\sum_{j=1}^{p+1} (-1)^{j+1} \left(a_j^2 - \int_{a_{j-1}}^{a_j} C(Y|r(\hat{y}_i))dy \right) \right] + (-1)^{s+1} \frac{1}{2}, \quad (3.11)$$

for p odd.

Our proposed index satisfies the following property:

Property

The *Gini Rank Dependence* assumes values in the range $[0, +1]$.

Proof.

In order to establish the range of values that define the *Gini Rank Dependence index*, let us consider Case 1 and Case 2, which represent the extreme cases.

If the concordance curve completely lies in the concordance area, then $GRD = \frac{1}{2} - \int_0^1 C(Y|r(\hat{y}_i))dy$: by normalizing, one obtains the following GRD expression:

$$GRD = \frac{\frac{1}{2} - \int_0^1 C(Y|r(\hat{y}_i))dy}{G/2} = \frac{1 - 2\int_0^1 C(Y|r(\hat{y}_i))dy}{G}. \tag{3.12}$$

In particular, if the concordance curve perfectly overlaps with the response variable Y Lorenz curve, then (3.12) becomes $GRD = \frac{1-2\int_0^1 L_Y(t)dt}{G}$, so

$$GRD = \frac{1 - 2\int_0^1 L_Y(t)dt}{G} = \frac{G}{G} = 1. \tag{3.13}$$

On the other hand, if the concordance curve completely lies in the discordance area, then $GRD = \int_0^1 C(Y|r(\hat{y}_i))dy - \frac{1}{2}$: by normalizing, one obtains the following GRD expression:

$$GRD = \frac{\int_0^1 C(Y|r(\hat{y}_i))dy - \frac{1}{2}}{G/2} = \frac{2\int_0^1 C(Y|r(\hat{y}_i))dy - 1}{G}. \tag{3.14}$$

In particular, if the concordance curve perfectly overlaps with the response variable Y dual Lorenz curve, then (3.14) becomes $GRD = \frac{2\int_0^1[1-L_Y(1-t)]dt-1}{G}$,so

$$GRD = \frac{2\int_0^1[1-L_Y(1-t)]dt-1}{G} = \frac{1-2\int_0^1 L_Y(1-t)dt}{G} \qquad (3.15)$$
$$= \frac{1-2\int_0^1 L_Y(t)dt}{G} = \frac{G}{G} = 1.$$

By considering the extreme cases, one can show that the upper bound of GRD is equivalent to $+1$.

The last step consists in defining the GRD lower bound: if the concordance curve perfectly overlaps with the egalitarian line, then $C(Y|r(\hat{y}_i)) = t$. For this reason

$$GRD = \frac{\frac{1}{2}-\int_0^1 tdt}{G/2} = \frac{\frac{1}{2}-\frac{1}{2}}{G/2} = 0, \qquad (3.16)$$

implying that $0 < GRD < +1$. ■

Note that the extreme cases, characterized by a null number of intersection points, occur when the concordance curve completely lies in the concordance area (Case 1) or in the discordance area (Case 2). In order to conduct these cases under the general formulation, we have to determine the nature of the 0 number. Pitaghora considers this number neither even nor odd even if the number 0 satisfies some typical even numbers conditions. Indeed to support considering the number 0 as even, in our context, note that the existence of an even number of intersection points implies that both the first and the last segments of the concordance curve lie in the same concordance or discordance area. Since our current discussion regards the cases where the concordance curve does not present any position reversal (in particular any reversal in the first or in the last segment) we will thus consider the generalized expression concerning p even.

Now if we let:

- $a_{j-1} = 0$ if $j = 1$, then $a_0 = 0$;
- $a_j = 1$ if $j = p+1$, then $a_1 = 1$,

we can show that the two extreme cases are special cases of the general formulation:

Case 1: the concordance curve $C(Y|r(\hat{y}_i))$ completely lies in the concordance area $\Rightarrow s = 1$.

$$\begin{aligned} GRD &= (-1)^1 \left\{ \sum_{j=1}^{1} (-1)^{j+1} \left(\int_0^1 C(Y|r(\hat{y}_i))dy - a_1^2 \right) + \frac{1}{2} \right\} \\ &= (-1)^1 \left\{ \int_0^1 C(Y|r(\hat{y}_i))dy - 1 + \frac{1}{2} \right\} \\ &= \frac{1}{2} - \int_0^1 C(Y|r(\hat{y}_i))dy. \quad \blacksquare \end{aligned}$$

Case 2: the concordance curve $C(Y|r(\hat{y}_i))$ completely lies in the discordance area $\Rightarrow s = 0$.

$$\begin{aligned} GRD &= (-1)^0 \left\{ \sum_{j=1}^{1} (-1)^{j+1} \left(\int_0^1 C(Y|r(\hat{y}_i))dy - a_1^2 \right) + \frac{1}{2} \right\} \\ &= (-1)^2 \left(\int_0^1 C(Y|r(\hat{y}_i))dy - 1 \right) + \frac{1}{2} = \int_0^1 C(Y|r(\hat{y}_i))dy - \frac{1}{2}. \quad \blacksquare \end{aligned}$$

The GRD measure provides a new approach to residual analysis when the “relevant” involved explanatory variables assume mostly categorical nature. Through the concordance curve construction one is able to detect the statistical units which contribute to the concordance and discordance shares:

let us recall that the discordance shares correspond to a relevant change of the original Y response variable ranks with respect to the ranks assigned to the corresponding estimated values. If the concordance curve is completely lying in the discordance area, there is a situation of no linear dependence among the response variable Y and the involved covariates, meaning that our adopted linear estimated regression model is not appropriate to fit the data. In the opposite case, the linear dependence between Y and the $k-1$ considered covariates is satisfied. Obviously, one can obtained intermediate cases characterized by a concordance curve lying partially in the discordance area and partially in the concordance area: in this context, one can detect which statistical units well fit the data and which ones do not fit well the data. Since our index assumes value $+1$, either when the concordance curve perfectly overlaps with the Y Lorenz curve and when the concordance curve perfectly overlaps with the Y dual Lorenz curve, the evaluation of the existence of a linear dependence can be deduced by considering the concordance and discordance shares size from a graphical point of view, as it occurs with residual analysis.

3.4 Conclusions

This study has introduced a new approach to the decomposition of the Gini measure into a concordance share and a discordance share. This in order to obtain the "quota" of the the response variable Y concentration explained by $\hat{Y}$, function of the explanatory variables $X_1, \ldots, X_{k-1}$, decomposed into the contribution of each statistical unit.

Our approach can be seen as a useful measure of fit when the "relevant" dependent variables are of categorical nature (nominal or ordinal): in the classical literature one can use the measure of fit represented by R^2 and residual

analysis, but when qualitative variables are involved and prevalent in the explanation of the response variable, R^2, based on the euclidean distance, may not be appropriate. Instead one can employ the Gini measure in place of R^2, and to its decomposition, seen in this chapter, in place of residual analysis. Furthermore, our proposed index, intended as a measure of goodness of fit, is based on a ratio between the concentration areas of $\hat{Y}$ and Y rather than on the euclidean distance between $\hat{Y}$ and Y.

We remark that our proposal can be applied to the model assessment context, particularly when one wants to compare alternative classifiers (e.g. polynomial regression, tree regression). In this context, a frequent model performance measure is the area under the ROC curve (AUC). It can be shown that the ROC curve is the equivalent to the Lorenz dual curve (see e.g. [5]). The AUC then corresponds to the Gini measure defined above (see e.g. [1], [3], [5] and [7]).

In this respect, in this chapter we have provided a way to decompose the contribution to the AUC measure in terms of the contributions of the different statistical units: the idea is similar to the ROC curve criterion. In fact, any classifier that appears in the lower right triangle performs worse than random guessing: this triangle is therefore usually empty in ROC graphs. The same result is obtained when employing our approach: in this case, every point which is located in the upper triangle, and in particular, lying in the discordance area, performs much worse than random.

In particular, one can split the response variable values in deciles (obtained by ordering these values in an increasing sense), one can establish the position of each value, with respect to the ranks assigned by its corresponding regression estimates, in order to define the single contribution to the concordance or discordance. If the position of the observed variable values are different with respect to the fitted ones, one can conclude that the predictive

selected model does not allow a good fitting.

Finally we would like to remark that our approach is nonparametric and can thus be applied, generally, for all predictive models such as regression tress and polynomial regression. A relevant further research topic could concern the definition of the concordance curve analytical expression, characterized as a functional of order statistics, under alternative parametric assumptions.

Bibliography

[1] Bradley, A. P.: *The use of the area under the ROC curve in the evaluation of machine learning algorithms*. Pattern Recognition, 30, 1145–1159 (1997)

[2] Dagum, C.: *A new Approach to the Decomposition of the Gini Income Inequality Ratio*. Empirical Economics 22, 515–531 (1997)

[3] Fawcett, T. : *An introduction to ROC analysis*. Pattern Recognition, 27, 861–874 (2006)

[4] Giudici, P., Raffinetti, E.: *Multivariate Ranks-based concordance indexes*. Published in the volume of selected papers “Statistical Methods for the analysis of large data-sets”. Springer (2011)

[5] Hand, D. J.: *Measuring classifier performance: a coherent alternative to the area under the ROC curve*. Mach Learn, 77, 103–123 (2009)

[6] Koshevoy, G., Mosler, K.: *The Lorenz Zonoids of a Multivariate Distribution*. Journal of the American Statistical Association, 91, No. 434, Theory and Methods (1996)

[7] Krzanowski, W. J., Hand, D. J.: *ROC curves for continuous data*. London: Chapman and Hall (2009)

[8] Muliere, P., Petrone, S.: *Generalized Lorenz curve and monotone dependence orderings*. Metron, Vol. L, n. 3–4 (1992)

[9] Rao, V. M.: *Two decompositions of concentration ratio*. Journal of the Royal Statistical Society, CXXXII, A, 418–425 (1969)

[10] Theil, H.: *Economics and information Theory*. North-Holland Publishing Company, Amsterdam (1967)

[11] Yitzhaki, S.: *Economic distance and overlapping of distributions*. Journal of Econometrics, 61, 147–159 (1994)

Chapter 4

Multivariate dependence through Lorenz zonoids

Abstract. During the last years, the dependence analysis context has assumed a relevant role both in economical and statistical applications: the literature provides a wide set of statistical tools focused in obtaining information about the dependence problem. In this paper we focus the attention on the Lorenz zonoid tool: when considering only the univariate case, the Lorenz zonoid corresponds to the Gini measure. Our aim is extending the Lorenz zonoid application to a multivariate dimension. In particular we consider the Lorenz zonoid of a linear regression function characterized by k explanatory variables and we define the partial contribution due to the introduction of a $(k+1)$ explanatory variable. This leads to the definition of a new dependence measure that we call *"Relative Gini Index" (RGI)*.

Keywords: Lorenz zonoid, Relative Gini Index, multiple linear regression models.

4.1 Introduction

Statistical dependence is a type of relation between any two features of units under study. These units may, for instance, be individuals, or objects, or various aspects of the environment.

The current statistical literature describes different approaches finalized to the study of dependence between random variables. Many of the notions of

positive dependence are defined by means of some comparison of the joint distribution of two random variables, X and Y, with their distribution under the theoretical assumption that X and Y are independent (see e.g. [14]). Often such a comparison can be extended to general pairs of bivariate distributions with given marginals: this fact led reaserchers to introduce various notions of positive dependence orderings (all the definitions have been introduced in Chapter 1). The role of these orders consists in comparing the strength of the positive dependence of the two underlying bivariate distributions (see e.g. [4]).

Since dependence analysis is strictly related to concordance studies, a useful tool focused in getting information about the dependence degree among the involved variables is the so called ranks-based approach (see e.g. [17] for the bivariate case and [6] for the multivariate case).

Recalling that the existent literature provides a large number of families of bivariate distributions with a natural interpretation of monotone dependence (whose definition has been already discussed in Chapter 1), the intuitive meaning of monotone dependence for a bivariate random vector (X,Y) is that large values of Y stochastically correspond to large values of X (positive dependence) or, in the opposite case, large values of Y correspond to small values of X (negative dependence) (see e.g. [18]).

Different notions of monotone dependence are defined through the conditional distribution of $Y|X=x$ or through the regression function $E(Y|X=x)$: examples can be found in regression dependence (see e.g. [14] and [24]), total positive of order two (see e.g. [8]) and in monotone regression (see e.g. [22]).

General literature concerning monotone dependence can be found, for instance, in [2] and in [18] (as already mentioned in Chapter 1).

In order to compare two bivariate distributions having the same pairs of

marginals to determine whether one distribution is positively dependent on the other, several partial orderings on the class of bivariate distributions with fixed marginals have been introduced: the existing literature in this research field can be found in [23] and [20]. Furthermore, a general concept of a positive dependence ordering can be found in [9], [21] and more recently in [18]. The motivation of using a partial ordering arises in the study of economic problems, such as the effects of taxation, and in the study of several applications such as discriminant problems and statistical quality control. In this field of research a new characterization of monotone dependence is in drawing a comparison between the Lorenz curve of the regression function $E(Y|X)$ and the Lorenz curve of Y (see e.g. [18]): this notion is appropriate when asking the relation to be invariant under increasing transformation of X but sensible to increasing transformation of Y.

Here we consider a partial ordering based on a particular statistical tool named Lorenz zonoid: when considering multivariate data, the Lorenz zonoid represents a specific characterization of the Lorenz curve. The Lorenz zonoid has been introduced by Koshevoy (see e.g. [10] and [11]) for empirical distributions and by Mosler (see e.g. [16]) for general probability distributions. The Lorenz zonoid of a d-dimensional random vector corresponds to a convex set in $\mathbb{R}^{d+1}$ whose role is in analyzing and comparing random vectors. Through the Lorenz zonoid representation one can establish an ordering of random vectors that reflects their variability: in fact, of our interest in this chapter is a comprehensive investigation of the ordering among random vectors that is induced by the inclusion between Lorenz zonoids. All the details about this contents are provided in [13].

In Section 4.2 the Lorenz zonoid will be defined and its inclusion property will be applied to the context of a multiple linear regression model.

4.2 The Lorenz zonoid approach to the dependence study

The Lorenz curve of a continuous random variable X, having finite expectation $\mu > 0$ and support in $\mathbb{R}_+$, is the graph of the function (see e.g. [18])

$$L_X(t) = \frac{1}{\mu}\int_0^p F_X^{-1}(z)dz, \quad 0 \le p \le 1,$$

where F_X^{-1} is the quantile function of X that is $F_X^{-1}(z) = \min\{x : F_X(x) \ge z\}$, $0 \le z \le 1$.

The Lorenz zonoid of a general d-variate random vector is defined as follows (see e.g. [13] and [12]). Consider the set $\mathscr{X}^d$ of random vectors in $\mathbb{R}^d$ that have finite expectation, the subset $\mathscr{X}^{d+} \subset \mathscr{X}^d$ of those vectors that have positive (in each component) expectation, and the subset $\mathscr{X}_+^{d+} \subset \mathscr{X}^{d+}$ of those that have, in addition, support in $\mathbb{R}_+^d$.

For $\mathbf{X} \in \mathscr{X}_+^{d+}$, we introduce the notation

$$\psi(\mathbf{x}) = \tilde{\mathbf{x}} = (\tilde{x}_1, \ldots, \tilde{x}_d), \quad \tilde{x}_j = \frac{x_j}{\mu_j}, \quad j = 1, \ldots, d,$$

where $\mu_j = \int_{\mathbb{R}_+^d} x_j dF(\mathbf{x}) > 0$.

The set

$$LZ(\mathbf{X}) = \{\mathbf{z} \in \mathbb{R}^{d+1} : \mathbf{z} = (z_0, \ldots, z_d) = \zeta(h), h : \mathbb{R}_+^d \to [0,1] \text{ measurable } \},$$

where $\zeta(h) = \left(\int_{\mathbb{R}_+^d} h(\mathbf{x})dF(\mathbf{x}), \int_{\mathbb{R}_+^d} h(\mathbf{x})\psi(\mathbf{x})dF(\mathbf{x})\right)$, is called *Lorenz zonoid*.

Furthermore, when $d = 1$, the Lorenz zonoid[1] is the area between the Lorenz curve and the dual Lorenz curve[2]: furthermore, it corresponds to the Gini measure (see e.g. [12]).

[1] What is called "Lorenz zonoid" is actually the "area of the Lorenz zonoid": however, it can be simply called the Lorenz zonoid as denoted in [12].

[2] The dual Lorenz curve corresponds to the Lorenz curve built by ordering all the interested variable values in a decreasing sense.

The Lorenz zonoid has many attractive properties, which makes it useful for a broad range of applications: in order to introduce our results in terms of Lorenz zonoids based dependence measures, the following notions are needed. First we show that the Lorenz curve of a general linear regression function lies always between the response variable Lorenz curve and its dual: this condition has been already introduced in Chapter 1 and more precisely it partially corresponds to Property 2 (see [18]).

***Proposition* 4.2.1** $L_Y(p) \leq L_{E(Y|X)}(p) \leq L_Y^{'}(p)$, *where* $L_Y^{'}(p) = \frac{1}{E(Y)} \int_{1-p}^{1} F_Y^{-1}(z)dz$, $0 \leq p \leq 1$. *Furthermore,* $L_{E(Y|X)}^{'}(p) \leq L_Y^{'}(p)$.

Proof.

Let us denote with $x_p = F_X^{-1}(p)$, where $0 \leq p \leq 1$. In order to show that $L_Y(p) \leq L_{E(Y|X)}(p)$, note that

$$L_{E(Y|X)}(p) = \frac{1}{E(Y)} E(Y|X \leq x_p) F_X(x_p)$$

and

$$L_Y(p) = \frac{1}{E(Y)} E(Y|Y \leq y_p) F_Y(y_p).$$

Now, $(Y|X \leq x_p)$ is stochastically larger than $(Y|Y \leq y_p)$, that is

$$P(Y \leq y|X \leq x_p) \leq P(Y \leq y|Y \leq y_p) \text{ for all } y \in \mathbb{R},$$

and the result follows.

In the same manner, one can prove that the dual Lorenz curve of the general regression function always lies below the dual Lorenz curve of the response variable Y. In order to show that $L_{E(Y|X)}^{'}(p) \leq L_Y^{'}(p)$ let us define

$$L_Y^{'}(p) = \frac{1}{E(Y)} E(Y|Y > y_{1-p})[1 - F_Y(y_{1-p})]$$

and

$$L'_{E(Y|X)}(p) = \frac{1}{E(Y)} E(Y|X > x_{1-p})[1 - F_X(x_{1-p})].$$

Since $E(Y|Y > y_{1-p})$ is stochastically larger than $E(Y|X > x_{1-p})$, the result follows. ■

Proposition 4.2.1 is basic for avoiding the intersection between the response variable and the general regression function Lorenz curves: let us note that the Lorenz curves are special cases of Lorenz zonoids in the univariate case. In fact, when one consider the unidimensional context, the area between the Lorenz curve and its dual of a generic variable corresponds to the so called Lorenz zonoid (see e.g. [12]). In other words, one can conclude that the Lorenz zonoid of the general linear regression function is always included in the Lorenz zonoid of the response variable: the Lorenz zonoids inclusion property assumes a relevant role as it will be illustrated in the following. Figure 4.1 shows this result in depicted way. The aforementioned graphical representation can be obtained supposing to consider the data matrix whose columns are represented by the values of the random variables (Y, X). Let us suppose to apply a simple linear regression function in order to establish a linear relation between the response variable Y and the explanatory variable X. The choice of resorting to a linear regression function is motivated by the fact that our aim consists in discussing partial linear dependence measures able to stress the existence and the strength of dependence relations, much better than the classical ones, when the context of analysis is characterized by “relevant” categorical covariates. For sake of simplicity, here only a covariate has been introduced.

The first step consists in building the response variable Lorenz zonoid; once computed its corresponding estimated values ($E(Y|X)$), one can proceed to the construction of the $E(Y|X)$ Lorenz zonoid. Depending on Proposition 4.2.1, the Lorenz zonoid of the response variable contains the Lorenz zonoid

of the corresponding estimated values, as one can deduce in Figure 4.1: we denote this condition with the following notation $LZ(E(Y|X)) \subset LZ(Y)$.

Furthermore, by recalling that the Gini measure (which is equivalent to the Lorenz zonoid in the univariate case, as already mentioned in [12]) is often used in order to measure the variability of a random variable, with regard to the linear model variance analysis, one can conclude that the Lorenz zonoid of $E(Y|X)$ can be considered as a useful tool in establishing the total model variance "explained" by the response variable estimated values. In fact, in or-

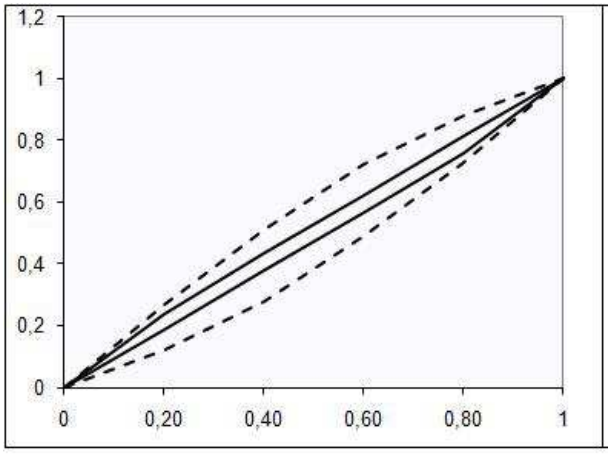

Figure 4.1: Y Lorenz zonoid (area between dashed lines) and $E(Y|X)$ Lorenz zonoid (area between continuous lines).

der to analyze the variability of random vectors in $\mathscr{X}_{+}^{d+}$, one can then make use of the Lorenz zonoids and consider the order between random vectors that is induced by their inclusion: in the following subsections we propose an overview of all the literature tools (see e.g. [4]) that have been taken into account in order to provide our proposed contributions in terms of dependence measures.

At first one focuses the attention on the multivariate context and, subsequently, one reduces the application with regard to the dimension one since our approach is based on the employment of a multiple linear regression function characterized by only one response variable end several covariates of

different nature.

4.2.1 Lorenz dominance

In this subsection the multivariate Lorenz dominance, via set inclusion of Lorenz zonoids, is defined: we do this for general random vectors, including the case of multivariate data (see e.g. [13]).

For two random vectors $\mathbf{X}$ and $\mathbf{Y}$ in $\mathscr{X}_+^{d+}$ define the Lorenz dominance $\preceq_L$ by

$$\mathbf{X} \preceq_L \mathbf{Y} \text{ if } LZ(\mathbf{X}) \subset LZ(\mathbf{Y}). \tag{4.1}$$

First consequences of the definition are that the Lorenz dominance $\preceq_L$ is a preorder[3] on $\mathscr{X}_+^{d+}$.

As the Lorenz zonoid of a random vector depends on its distribution only, we can also write $F_{\mathbf{X}} \preceq_L F_{\mathbf{Y}}$ in place of $\mathbf{X} \preceq_L \mathbf{Y}$. The preorder is scale invariant, meaning that $(X_1,\ldots,X_d) \preceq_L (Y_1,\ldots,Y_d)$ implies $(\lambda_1 X_1,\ldots,\lambda_d X_d) \preceq_L (\lambda_1 Y_1,\ldots,\lambda_d X_d)$ for any $\lambda_1,\ldots,\lambda_d > 0$.

The following proposition characterizes the multivariate Lorenz dominance.

Proposition* 4.2.2 *(Characterization of Lorenz dominance)
The following statements are equivalent:

1. $(X_1,\ldots,X_d) \preceq_L (Y_1,\ldots,Y_d)$.
2. *For every* $\mathbf{p} = (p_1,\ldots,p_d) \in \mathbb{R}^d$, *the generalized Lorenz curve of* $\sum_{j=1}^d p_j \tilde{Y}_j$ *lies above that of* $\sum_{j=1}^d p_j \tilde{X}_j$.
3. *For every convex function* $\psi : \mathbb{R} \to \mathbb{R}$ *and* $\mathbf{p} \in \mathbb{R}^d$,

$$E\left[\psi\left(\sum_{j=1}^d p_j \tilde{X}_j\right)\right] \leq E\left[\psi\left(\sum_{j=1}^d p_j \tilde{Y}_j\right)\right]. \tag{4.2}$$

[3] Preorders are binary relations that are reflexive and transitive.

4. *For every increasing*[4] *convex function* $\psi : \mathbb{R} \to \mathbb{R}$ *and* $\mathbf{p} \in \mathbb{R}^d$*, (4.2) holds.*

5. *For every* $\mathbf{p} \in \mathbb{R}^d$ *there exists a random variable* $U_{\mathbf{p}}$ *such that* $E[U_{\mathbf{p}} | \sum_{j=1}^d p_j X_j] = 0$ *and*

$$\sum_{j=1}^{d} p_j \tilde{Y}_j =_{st} \sum_{j=1}^{d} p_j \tilde{X}_j + U_{\mathbf{p}}, \tag{4.3}$$

where the notation $=_{st}$ *denotes equality in distribution.*

Part (2.) of the Proposition 4.2.2 is also called *directional majorization* of $\tilde{Y}$ over $\tilde{X}$. It can be interpreted in terms of ”prices” and ”expenditures”: when, with properly chosen units, the mean endowment of each commodity amounts to one and $\mathbf{X}$ and $\mathbf{Y}$ mean alternative distributions of the commodity vector, then $\sum_{j=1}^d p_j \tilde{X}_j$ and $\sum_{j=1}^d p_j \tilde{Y}_j$ stand for the distributions of expenditures given the price vector $\mathbf{p}$. The proposition says that $\mathbf{X}$ has less multivariate disparity than $\mathbf{Y}$ if and only if the first distribution of expenditures is less unequal than the second, in the sense of usual Lorenz dominance, for every price vector in d-space.

In terms of expected utility, part (4.) of Proposition 4.2.2 means that every risk seeking person which has to choose between the random expenditures $\sum_{j=1}^d p_j \tilde{X}_j$ and $\sum_{j=1}^d p_j \tilde{Y}_j$ will, for any prices, prefer the latter, and that every risk adverse person will do the opposite. Part (3.) says the same, with not necessarily increasing utilities.

In part (5.), $U_{\mathbf{p}}$ may be interpreted as a perturbation of expenditures or “noise”. The distribution of expenditures for $\tilde{\mathbf{Y}}$ is, for all prices, ”noisier” than all distribution of expenditures for $\tilde{\mathbf{X}}$.

[4] “Increasing” is always meant in the weak sense.

4.2.2 Lift zonoids and variability of random vectors

In order to analyze the variability of random vectors one can replace a probability distribution in $\mathbb{R}^d$ by its lift zonoid and consider the order between random vectors that is induced by the inclusion of their lift zonoids. This result is also achieved when one resorts to another extension of the Lorenz dominance, the so called multivariate *scaled convex order*. The aforementioned order will be considered and contrasted with the Lorenz dominance in the following subsection.

As already defined in Chapter 1, the lift zonoid[5] represents the Lorenz zonoid of multivariate non-relative data.

***Definition* 4.2.3** *For $\mathbf{X}$ and $\mathbf{Y} \in \mathscr{X}^d$, we introduce the lift zonoid order $\preceq_{LZ}$,*

$$\mathbf{X} \preceq_{LZ} \mathbf{Y} \text{ if } \hat{Z}(\mathbf{X}) \subset \hat{Z}(\mathbf{Y}),$$

where $\hat{Z}(\mathbf{X})$ and $\hat{Z}(\mathbf{Y})$ represent the lift zonoids of the random vectors $\mathbf{X}$ and $\mathbf{Y}$.
Let now $\mathbf{X}$ and $\mathbf{Y}$ be distributed as $F_{\mathbf{X}}$ and $F_{\mathbf{Y}}$, then

$$\mathbf{X} \preceq_{LZ} \mathbf{Y} \text{ if } F_{\mathbf{X}} \preceq_{LZ} F_{\mathbf{Y}}.$$

The lift zonoid order ranks random vectors by their variability.

In the remainder of this subsection the properties of $\preceq_{LZ}$ will be investigated starting with a well known Proposition and the definition of the dilation order.

***Proposition* 4.2.4** *Let $\mathbf{X}$ and $\mathbf{Y} \in \mathscr{X}^d$. The following three conditions are equivalent.*

[5] Let us recall the lift zonoid formal definition. For $\mathbf{X} \in \mathscr{X}^d$, the set

$$LZ(\mathbf{X}) = \{\mathbf{z} \in \mathbb{R}^{d+1} : \mathbf{z} = (z_0, \ldots, z_d) = \zeta(h), h : \mathbb{R}^d \to [0,1] \text{ measurable}\},$$

where $\zeta(h) = \left(\int_{\mathbb{R}^d} h(\mathbf{x}) dF(\mathbf{x}), \int_{\mathbb{R}^d} h(\mathbf{x}) \mathbf{x} dF(\mathbf{x}) \right)$, is called *lift zonoid*.

1. $E[\phi(\mathbf{X})] \leq E[\phi(\mathbf{Y})]$ *for all convex* $\phi : \mathbb{R}^d \to \mathbb{R}$ *for which the expectation exists.*

2. $E[\mathbf{X}] = E[\mathbf{Y}]$ *and* $E[\phi(\mathbf{X})] \leq E[\phi(\mathbf{Y})]$ *for all increasing convex* $\phi : \mathbb{R}^d \to \mathbb{R}$ *for which the expectations exist.*

3. $\mathbf{Y} =_{st} \mathbf{X} + \mathbf{U}$ *for some* $\mathbf{U}$ *for which* $E[\mathbf{U}|\mathbf{X}] = 0$.

The random vector $\mathbf{U}$ in Proposition 4.2.4 can be interpreted as "noise", so that $\mathbf{Y}$ is distributed as $\mathbf{X}$ plus some noise.

If one, and hence all, conditions of Proposition 4.2.4 are satisfied, $\mathbf{Y}$ is called a *dilation* of $\mathbf{X}$, $\mathbf{X} \preceq_{dil} \mathbf{Y}$.

***Corollary* 4.2.5** $\mathbf{X} \preceq_{dil} \mathbf{Y} \Rightarrow \mathbf{X} \preceq_{lZ} \mathbf{Y}$ *(see the details in [4]).*

If $d > 1$, the reverse implication, $\mathbf{X} \preceq_{lZ} \mathbf{Y} \Rightarrow \mathbf{X} \preceq_{dil} \mathbf{Y}$, does not hold in general (see e.g. [4]).

4.2.3 Scaled convex order

We have seen that the lift zonoid order is equivalent to the dilation order when $d = 1$: now, the aim is to detect similarities between the dilation order and the ordering based on Lorenz zonoids. The ordering based on Lorenz zonoids is generally speaking the Lorenz dominance already described. However, the classical Lorenz dominance can be extended defining new orderings such as the *scale convex order* which is now considered and contrasted with the Lorenz dominance (for more details see [4] and [13] for more details).

***Proposition* 4.2.6** *For* $\mathbf{X}, \mathbf{Y} \in \mathscr{X}^d$ *the following statements are equivalent:*

1. $E[\phi(\tilde{\mathbf{X}})] \geq E[\phi(\tilde{\mathbf{Y}})]$ *if* $\phi : \mathbb{R}^d \to \mathbb{R}$ *is concave and the expectations exist.*

2. $E[\phi(\tilde{\mathbf{X}})] \leq E[\phi(\tilde{\mathbf{Y}})]$ *if* $\phi : \mathbb{R}^d \to \mathbb{R}$ *is convex and the expectations exist.*

3. $E[\phi(\tilde{\mathbf{X}})] \leq E[\phi(\tilde{\mathbf{Y}})]$ *if* $\phi : \mathbb{R}^d \to \mathbb{R}$ *is increasing convex and the expectations exist.*

4. $\tilde{\mathbf{Y}} =_{st} \tilde{\mathbf{X}} + \mathbf{U}$ *with* $E[\mathbf{U}|\tilde{\mathbf{X}}] = 0$.

$\mathbf{X}$ is said to be not larger than $\mathbf{Y}$ in *scaled convex order*, $\mathbf{X} \preceq_{scx} \mathbf{Y}$, if one of these equivalent restriction is satisfied. Proposition 4.2.6 is similar to Proposition 4.2.4: in fact, Proposition 4.2.6 satisfies the same conditions of Proposition 4.2.4 when considering relative data. For this reason one can define the scaled convex preorder[6] as an order able to order random vectors by the inclusion of their Lorenz zonoids.

In view of Propositions 4.2.2 (3.) and 4.2.6 (2.) and, as every convex-linear function is convex, the scaled convex order implies the Lorenz dominance,

$$\mathbf{X} \preceq_{scx} \mathbf{Y} \Rightarrow \mathbf{X} \preceq_{L} \mathbf{Y}. \tag{4.4}$$

Thereby, in case $d = 1$ the scaled convex order is the same as the Lorenz dominance (see [4]): for this reason, $\mathbf{X} \preceq_{L} \mathbf{Y} \Rightarrow \mathbf{X} \preceq_{dil} \mathbf{Y}$ holds.

By summarizing, the Lorenz dominance, corresponds to the scaled convex order which is strictly related to the dilation order built on non-relative data. Moreover, since the Lorenz dominance allows to provide (in dimension one) a partial ordering of random variables according to their variability by the corresponding Lorenz zonoids inclusion, one can exploit this property in order to describe the linear regression variance "explained" by each considered covariate. In particular, through all the discussed properties, one can establish the partial contribution to the multiple linear regression model "explained" variance provided by the introduction of an additional explanatory variable into the model.

All these issues represent exactly our contribution in terms of dependence

[6] The scaled convex order is a preorder in $\mathscr{X}^d$.

study through the employment of the Lorenz zonoids.

4.3 Our proposal: partial Lorenz dependence measures

Lorenz zonoids have been employed in statistical applications only in the bivariate case (see Chapter 1).

The question is how extending the dependence analysis when considering a linear regression function characterized by more than one explanatory variable.

Our idea is to define a single partial dependence measure whose role consists in expressing the "share" of the Y Lorenz zonoid "explained" by each single considered explanatory variable. Let us define the response variable Y Lorenz zonoid, intended as the area between the Lorenz curve and its dual, and let us denote it with $A(Y)$: since this area corresponds to the Gini measure[7], the procedure will be simplified by referring all the analysis to the univariate context.

Assuming the application of a linear regression function and a bivariate vector (Y, X_1), one can build the Lorenz zonoid of $\hat{Y}_{X_1} = E(Y|X_1)$ and denote it with $A(\hat{Y}_{X_1})$. Now, supposing to operate with a trivariate random vector (Y, X_1, X_2), one can thus compute the Lorenz zonoid of the estimated Y variable values according to the second explanatory variable X_2, obtaining $A(\hat{Y}_{X_2})$. Furthermore, by exploiting the multiple linear regression tool, one can get the estimated Y variable values $\hat{Y}_{X_1,X_2} = E(Y|X_1, X_2)$, considering both the covariate X_1 and X_2. The area that ranges between the Lorenz curve of $\hat{Y}_{X_1,X_2}$ and its dual can be denoted with $A(\hat{Y}_{X_1,X_2})$. Forward selection[8] of other variables can be dealt in a similar way without loss of generality.

[7] In fact we are considering the dimension one.

[8] According to the *forward selection* procedure, at each stage the best unselected covariate is added until no further candidates remain. See Section 4.4.

These areas, corresponding to the Lorenz zonoids, are particularly useful in order to obtain partial dependence measures. In fact, through the Lorenz zonoids tool application, one can exploit the main property of the Gini measure in evaluating the variability of its underlying variable. Through the response variable Lorenz zonoid one gets information about its related variability: the greater is this area, the greater is the variability of the underlying data . The idea of extending the Lorenz zonoid approach to the multiple linear regression model context allows to introduce an alternative procedure, with respect to the classical ones, of measuring the strength of the existing linear dependence relations among the involved variables. In general, through the Lorenz zonoid of a general linear regression function $E(Y|\cdot)$ and by exploiting the already discussed inclusion property, one defines the share of the Y variability "explained" by the considered covariates. Thereby, a partial measure, representative of the share of the Y Lorenz zonoid "explained" by covariate X_1 alone, is derived through the following ratio

$$G_1 = \frac{A(\hat{Y}_{X_1})}{A(Y)}. \tag{4.5}$$

By the means of the (4.5) ratio, one can evaluate the Y variability amount which is "explained" by the explanatory variable X_1 alone. The Lorenz zonoid inclusion property assumes a basic role in order to measure the area of the response variable Lorenz zonoid filled by its corresponding estimated values Lorenz zonoid.

In the same manner, a partial measure able to describe the share of the Y variable Lorenz zonoid "explained" by covariate X_2 alone, is derived through the ratio

$$G_2 = \frac{A(\hat{Y}_{X_2})}{A(Y)}. \tag{4.6}$$

The partial dependence measures G_1 and G_2 always assume values in the range $[0, +1]$. The Gini measure achieves its maximum value $+1$ when the underlying data are characterized by a high variability: on the other hand, it assumes its minimum value 0 when the variability is null. For this reason the index G_1 and G_2, obtained as a ratio between two Gini measures, can take values only in the range $[0, +1]$.

In order to obtain partial dependence measures expressed in terms of Lorenz zonoids, since we are using a multiple linear regression model, our idea consists in determining the partial contribution of each explanatory variable to the overall linear regression model. To achieve this purpose we can exploit the existing relation between the Lorenz zonoids and the dilation measure[9] whose role consists, in general, in defining the degree of variability of all the involved variables.

Consider, now the general context characterized by k explanatory variables: our proposal is focused on establishing the effect expressed in terms of dependence measures, connected to the introduction of a new $(k+1)$ explanatory variable into the regression model. Through the Lorenz zonoids inclusion, one is able to compare the variability of the interested variables. For this reason, when one builds the Lorenz zonoid of the response variable Y, one describes its variability: the greater the Lorenz zonoid, the greater the variable dispersion. If one applies an estimated multiple linear regression function characterized by several covariates, (that is $E(Y|X_1, \ldots, X_k)$), the corresponding Lorenz zonoid is included in the response variable Lorenz zonoid defining the Y variable variability explained by the k covariates. Since, we are operating by the employment of an estimated linear regression model, the classical assumptions imply that the addition of a covariate defines an increase of the "explained" model variance. This result, in terms of Lorenz zonoids, translates

[9] The dilation measure notion is related to the dilation ordering issue. It allows to compare and order random variables according to their variability.

into an enlargement of the $E(Y|X_1,\ldots,X_k)$ Lorenz zonoid. In the classical linear regression context, the addition of a new explanatory variable into the model is related to the corresponding reduction of the residual standard error, implying that the new fitted values are more similar to the values assumed by the response variable Y. Thereby, the y-axis values characterizing the Lorenz curve of the estimated regression function $E(Y|X_1,\ldots,X_k,X_{k+1})$ (denoted with $L_{E(Y|X_1,\ldots,X_k,X_{k+1})}$) come closer to the ones characterizing the response variable Lorenz curve: for this reason, one obtains that $L_Y \leq L_{E(Y|X_1,\ldots,X_k,X_{k+1})} \leq L_{E(Y|X_1,\ldots,X_k)}$. On the other hands, when considering the dual context, it results that $L'_{E(Y|X_1,\ldots,X_k)} \leq L'_{E(Y|X_1,\ldots,X_k,X_{k+1})} \leq L'_Y$. The aforementioned Lorenz zonoid enlargement represents the further response variable Lorenz zonoid share "explained" by the added covariate.

In the well-known multiple linear regression model, properly, the contribution of a single variable to the regression plane is additive and, therefore, the addition of a new explanatory variable translates into an increment of the multiple determination coefficient (see e.g.[5]): this measure depends on the correlation degree between the current and the additional variable and on the degree between the current explanatory variables and the response variable. More precisely, suppose to build a linear regression model characterized by k explanatory variables. Let us introduce an additional $(k+1)$ explanatory variable: its contribution corresponds to an increase of the variance "explained" by the regression plane and can be defined as the difference between $Var(\hat{Y}_{k+1})$ and $Var(\hat{Y}_{X_k})$. The squared partial correlation coefficient[10] is expressed as the ratio between this same additional contribution, in terms of "explained variance", and the "unexplained" variability of $\hat{Y}_{X_k}$.

$$r^2_{Y,X_{k+1}|X_1,\ldots,X_k} = \frac{Var(\hat{Y}_{X_{k+1}}) - Var(\hat{Y}_{X_k})}{Var(Y) - Var(\hat{Y}_{X_k})}. \tag{4.7}$$

[10] The root of the squared partial correlation coefficient corresponds to the correlation of residuals.

Our aim is to build a partial dependence measure based on Lorenz zonoids that "parallels" the construction of the squared partial correlation coefficient. In our context the purpose is to obtain a ratio whose numerator is characterized by a term denoting the contribution due to the addition of the $(k+1)$ explanatory variable to the Y variable Lorenz zonoid, whereas the denominator is defined by a term describing the share of the Y Lorenz zonoid not "explained" by the $\hat{Y}_{X_k}$ Lorenz zonoid.

The additional contribution related to the $(k+1)$ variable introduction can be measured by the difference between the Lorenz zonoid of $\hat{Y}_{X_{k+1}}$ and that of $\hat{Y}_{X_k}$, that is

$$A(\hat{Y}_{X_{k+1}}) - A(\hat{Y}_{X_k}). \tag{4.8}$$

A relative measure, able to stress the additional contribution of the X_{k+1} variable to the regression model, in terms of "explained" Y variability, can be obtained in analogy with the squared partial correlation coefficient construction. The role of this measure consists in capturing the partial contribution due to the introduction of an additional covariate into the model, in terms of "explained" variance (represented by the response variable Lorenz zonoid).

***Definition* 4.3.1** *The "partial Lorenz dependence measure" related to the introduction of a new covariate into the linear regression model defined as*

$$RGI = \frac{A(\hat{Y}_{X_{k+1}}) - A(\hat{Y}_{X_k})}{A(Y) - A(\hat{Y}_{X_k})}, \tag{4.9}$$

will be called Relative Gini Index and denoted with RGI.

In other words the so called "partial Lorenz dependence measure" represents an alternative to the squared partial correlation coefficient which can be used as a stopping rule in the stepwise regression context. It is particularly useful in order to overcome the R^2 measure restrictions when the "relevant" explanatory variables assume mostly categorical nature.

***Proposition* 4.3.2** *The RGI measure can be decomposed additively as the squared partial correlation coefficient.*

Proof.

For sake of simplicity, let us prove this construction when only two covariates are considered.

In this case, the squared correlation coefficient assumes the following expression

$$r^2_{Y|X_1,X_2} = r^2_{Y,X_1} + r^2_{Y,X_2|X_1}(1 - R^2_{Y,X_1}). \tag{4.10}$$

In a linear regression model the multiple determination coefficient R^2 defines the linear dependence among all the involved variables: in terms of Lorenz zonoids the total dependence measure corresponds to $LZ(E(Y|X_1,X_2))$ which is equivalent to the area $A(\hat{Y}_{X_1,X_2})$ lying between the Lorenz curve of $\hat{Y}_{X_1,X_2} = E(Y|X_1,X_2)$ and its dual. In terms of Lorenz zonoids, the relation (4.10) becomes

$$LZ(E(Y|X_1,X_2)) = LZ(E(Y|X_1)) + LZ(E(Y|(X_2|X_1)))(LZ(Y) - LZ(E(Y|X_1))),$$

w**here** $LZ(E(Y|X_1)) = A(\hat{Y}_{X_1})$, $LZ(E(Y|(X_2|X_1))) = RGI$ and $LZ(Y) = A(Y)$. Now, substituting to RGI its expression in (4.9), one gets

$$\begin{aligned} LZ(E(Y|X_1,X_2)) &= A(\hat{Y}_{X_1}) + RGI(A(Y) - A(\hat{Y}_{X_1})) \\ &= A(\hat{Y}_{X_1}) + \frac{A(\hat{Y}_{X_1,X_2}) - A(\hat{Y}_{X_1})}{A(Y) - A(\hat{Y}_{X_1})}(A(Y) - A(\hat{Y}_{X_1})) \\ &= A(\hat{Y}_{X_1}) + A(\hat{Y}_{X_1,X_2}) - A(\hat{Y}_{X_1}) \\ &= A(\hat{Y}_{X_1,X_2}). \quad \blacksquare \end{aligned}$$

This proof simply highlights the similarity between the squared partial correlation coefficient and the RGI measure which is particularly useful when considering a quali-quantitative context, that is, when the covariates can have

both quantitative and categorical nature. A further advantage related to our approach application regards its invariance with respect to scale transformation concerning the response variable (as it will be discussed in Section 4.6).

The RGI measure role will be deeply discussed in Section 4.4.

The obtained partial Lorenz dependence measure, RGI, defines the partial contribution to the Y Lorenz zonoid due to the addition of a new covariate into the model. This index is always non-negative because both the numerator and the denominator are positive, due to the Lorenz zonoids inclusion property. However, when sample data are involved in the analysis, the following remark is needed in order to guarantee the validity of our proposed approach.

***Remark* 4.3.3** *The Lorenz zonoids inclusion property always holds for the population joint distribution and it is satisfied also when employing an estimated linear regression function under the restriction of linear dependence between Y and the explanatory variables $X_1, \ldots, X_k$. This is satisfied, for example, when the adjusted R-squared*[11] *of each considered covariate must assume non-negative values. In fact, a negative adjusted R-squared, associated to a covariate, implies that there is no relationship between the considered covariate and the involved response variable. More precisely, one can conclude that there is no linear relation and that the regressor is then useless. Since the Lorenz zonoid inclusion is satisfied if there is linear dependence between the covariate and the response variable, this result holds for each covariate non-negative Adjusted R-squared. An adjusted R-squared negative value very close to zero can generate intermediate situations, so it is better to consider only positive values.*

We now discuss the statistical interpretation to assign to the obtained partial Lorenz dependence measures: we will consider two examples that combines

[11] The adjusted R-squared is defined as $1 - \frac{(1-R^2)(n-1)}{(n-k-1)}$, where n represents the number of observations and k the number of predictors.

multiple linear regression and Lorenz zonoids theory.

Let us suppose to consider the data in Table 4.1, characterized by five observations and two covariates X_1 and X_2.

Y	X_1	X_2
350	84	45
202	73	19
404	64	53
261	68	31
450	76	58

Table 4.1: *First Dataset Example*

Using the multiple linear regression model and the Lorenz zonoid tool, we get $A(Y) = 0.1533293$, $A(\hat{Y}_{X_1}) = 0.01264018$, $A(\hat{Y}_{X_2}) = 0.1512618$. Through (4.5) and (4.6), single partial dependence measures can be derived so that $G_1 = 0.08243813$ and $G_2 = 0.9865155$: the relative interpretation is very simple. The Lorenz zonoid of $E(Y|X_1) = G_1$ represents the 8.24% of the Lorenz zonoid of Y and the Lorenz zonoid of $E(Y|X_2) = G_2$ represents the 98.65% of the Lorenz zonoid of Y. In order to define the partial contribution due to the introduction of a further variable into the linear regression model, one can follow the forward selection procedure by choosing as first covariate the one that better "explains" the Y variable variability. According to the obtained partial measures values, G_1 and G_2, the first variable that has to be included in the model is the covariate X_2. Let us evaluate the contribution related to the addition of the second covariate X_1 into the model in terms of the RGI measure. The first step consists in computing the $E(Y|X_1, X_2)$ Lorenz zonoid denoted as $A(\hat{Y}_{X_1,X_2})$. The $E(Y|X_1, X_2)$ Lorenz zonoid is characterized by an area whose amount is equivalent to 0.1512201, which is smaller than $A(\hat{Y}_{X_2}) = 0.1512618$, implying $LZ(E(Y|X_1)) \not\subset LZ(E(Y|X_2, X_1))$. In fact, if we compute the RGI_{X_2,X_1}[12]

[12] RGI_{X_2,X_1} denotes the partial contribution due to the introduction of the X_1 covariate into the model already characterized by the introduction of the covariate X_2.

measure it assumes a negative value being $A(\hat{Y}_{X_1,X_2}) < A(\hat{Y}_{X_2})$. The reason of such a result has to be found in the covariate X_1 adjusted R-squared value. In fact, when one computes the response variable Y estimated values according to the explanatory variable X_1 alone, one can notice that it is characterized by a negative adjusted R-squared (equals to -0.3242), meaning that between Y and X_1 there is not a linear relation. The linear relation lack between the response variable and the first covariate translates into a non inclusion of the the $E(Y|X_2)$ Lorenz zonoid into the $E(Y|X_2,X_1)$ Lorenz zonoid. For this reason, before adding a covariate into the model, one has always to check the related Adjusted R-squared values: in case of negative values, the associated covariate does not satisfy a linear relation with the response variable Y so it has to be deleted “a priori”.

Let us now take into account the data contained in Table 4.2 and let us retrace the same proposed approach steps. The results are the following:

Y	X_1	X_2
350	49	45
202	23	19
404	50	53
263	38	31
451	47	58
304	17	23
275	22	25
385	42	36
244	13	29
302	54	39
274	33	35
346	45	49
253	20	22
395	58	61
430	49	48
216	40	34
374	60	51
308	54	50

Table 4.2: *Second Dataset Example*

$A(Y) = 0.1288281$, $A(\hat{Y}_{X_1}) = 0,08276167$ (Adjusted R-squared: 0.3916), $A(\hat{Y}_{X_2})$ $= 0.1049543$ (Adjusted R-squared: 0.6414),
$A(\hat{Y}_{X_2,X_1}) = 0,1051045$ (Adjusted R-squared: 0.6298). The partial dependence measures are provided by $G_1 = 0.642419395$ and $G_2 = 0.81468484$.

The RGI_{X_2,X_1} index amounts to $0,006291416$ meaning that the addition of the covariate X_1 into the model implies a very low increase of the Y Lorenz zonoid "explained" share: thereby, the contribution of this new variable is almost null.

4.4 The selection of covariates according to the RGI measure

The aim of this section consists in introducing and discussing about a graphical test, based on the RGI scree plot, particularly useful in selecting the explanatory variables to be added into the model according to the issues of the forward selection method.

In the following subsections a general overview of the forward selection procedure, adapted to the proposed Lorenz zonoids approach, has been considered: furthermore, a test able to define the number of "relevant" covariates that mainly contribute to the response variable "explained" variability, intended as "explained" model variance, has been implemented.

4.4.1 The forward selection procedure and the Lorenz zonoids approach

In general, the problem of modeling reduces to finding one or more appropriate parsimonious sets of covariates corresponding to a model matrix $\mathbf{Z}$ of order $n \times k$. In [15] the justification for seeking a parsimonious model to represent a set of data has been discussed. Parsimony implies that covariates

having no detectable effects on the response should ordinarily be excluded from the linear predictor. However, the selection of a useful set of covariates from a large set of possible covariates to form a parsimonious model is then a non-trivial exercise.

On the statistical side, the problem is that of defining the balance to be struck between two opposing effects of including a new term in the model. The good effect may be a reduction in the discrepancy between the data and the fitted values. The bad effect is that, unless there is good prior knowledge that the covariate has a non-negligible influence on the response, inclusion of the covariate usually complicates the model and statements of conclusions derived from it. The usual F-statistic, whose definition can be found for instance in [5] and [19], represents the basis of most criteria for selection of covariates: it is based on the reduction of the sum of squares. In order to exclude irrelevant terms, the significance level for acceptance is set at low level, but it must not be set so low that important terms are thereby excluded (all these detailed topics can be found in [5]).

In literature, approximate methods for generating an “optimum” set of selected covariates include:

- *forward selection*, whereby at each stage the best unselected covariate satisfying the selection criterion is added until no further candidates remain;

- *backward elimination*, which begins with the full set and eliminates the worst covariates one by one until all remaining covariates are necessary;

- *stepwise regression* which combines the two previous procedures, following backward elimination by forward selection until both fail to change the model.

The central topic of the contribution presented in this subsection is focused on the implementation of the forward selection adapted to our context of analysis characterized by the employment of the Lorenz zonoids tool. The idea is based on using the RGI measure in order to establish, through the application of the multiple linear regression model and through the forward selection of each covariate into the model, which explanatory variables have to be considered "relevant" in "explaining" the model variance. The covariate that provides the greater contribution to the response variable "explained" variability is added into the model.

Let us suppose to have k explanatory variables. The first step consists in computing all the considered covariates partial dependence measures $G_1, \ldots, G_k$: the G_i, with $i = 1, \ldots, k$, with the highest value provides the corresponding covariate that has to be added into the model. Once selected the covariate with the highest G value, one has to introduce all the remaining covariates, one by one, and measuring their relative contribution in terms of the RGI measure: the highest RGI measure value implies the addition of the corresponding covariate into the model. This method is similar to the classical forward selection. However, whereas in the classical forward selection procedure, one can use the F-statistics value at a preselected significance level as a stopping rule, in this context for all the covariates one computes the related G measure by including them into the model. In the following subsection, we propose a test based on the RGI index, able to provide a stopping rule in the forward selection procedure.

4.4.2 The RGI scree-test

The issue consists in defining a stopping rule in the forward selection procedure: how many covariates have to be taken into account in order to study the linear dependence relationships among them and the interested response

variable? In other words, we may wish to test whether the addition of a further covariate significantly improves the fit in terms of Lorenz dependence measures.

The solution of this problem can be provided by the employment of a graphical tool widely used in factor analysis: we are speaking about the so called "scree test" (see e.g. [3]). Factor analysis can be intended as a data reduction method, that is, as a method for reducing the number of variables. The problem here is in establishing the number of factors that have to be extracted. Note that as we extract consecutive factors, they account for less and less variability. The decision of when to stop extracting factors basically depends on when there is only very little "random" variability left (see e.g. [7]). The variances extracted by the factors are the eigenvalues. Once having a measure of how much variance each successive factor extracts, one can return to the question of how many factors to retain: there are some guidelines that are commonly used, and that, in practice, seem to yield the best results. One of these guidelines consists in applying the scree test (for more details about its construction see e.g. [7]).

The factor analysis setting is similar to that of our Lorenz zonoids approach. Also in this context, the idea is specifying the covariates that better contribute to the model variance explanation: as a parallel of the factor analysis we want to define the total number of covariates that could be considered relevant in terms of dependence relations, by exploiting the information contained in the corresponding RGI values.

The adapted[13] scree-test suggests to set, on the graphical point of view, the number of covariates which deserve to be taken into account. One builds a plot characterized by a vertical axis on which one locates the number of covariates and an horizontal axis on which one denotes the RGI values cor-

[13]The term "adapted" is here used in order to highlight that the factor analysis scree-test has been adapted to our specific context of analysis based on Lorenz zonoids.

responding to each covariate. The RGI measures are represented in the plot as points joined through a line. According to this method, we suggest to find the place where the smooth decreases and the RGI values appears to level off to the right of the plot. To the right of this point, presumably, we find only "RGI scree": "scree" is the geological term referring to the debris which collects on the lower part of a rocky slope. In other words, we have to find where the graph seems to behave randomly, like rocks falling on a scree down a hill (see, for instance, [3]). The number of covariates corresponds to the number of RGI values preceding this scree. We call this test the RGI *scree-test* and the related plot the RGI *scree-plot*.

To exemplify let us try to interpret the RGI scree-plot results by focusing the attention on Figure 4.2, representing the RGI scree-plot with regard to a total number of six covariates. The purpose is evaluating the number of covariates

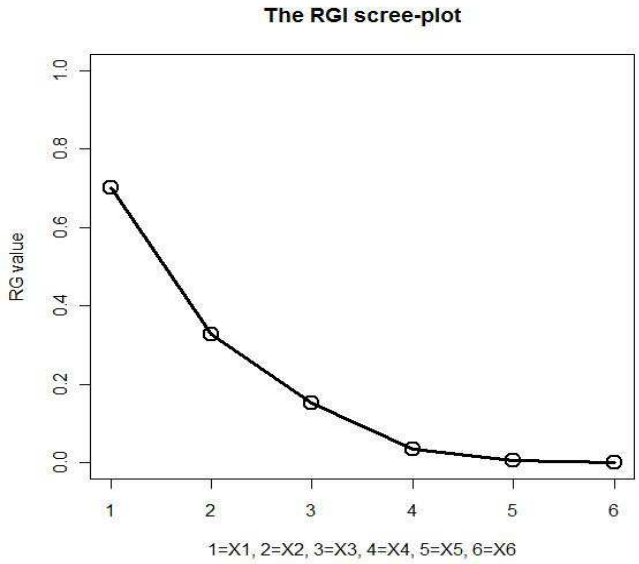

Figure 4.2: The RGI plot concerning the forward selection procedure.

that have truly to be added into the linear regression model: the choice is related to the partial contribution in terms of "explained" model variance due to each new introduced explanatory variable. By exploiting the forward selection

procedure, one begins by adding the explanatory variable that provides the greatest contribution in terms of the response variable “explained” variability. Once computed all the covariates partial dependence measures $G_1, \ldots, G_7$, one can conclude that the covariate associated to the highest G value is the covariate X_1: in fact, G_1 is greater than 0.70 meaning that the Lorenz zonoid of X_1 explains over the 70% of the response variable Lorenz zonoid. The second step, consists in including one by one the remaining covariates into the linear regression model already characterized by the covariate X_1. The second covariate that will be considered “relevant” will be the one whose RGI value will be the highest. From the Figure 4.2 it results that the covariate X_2 provides a partial contribution equivalent to 0.32. By iterating the procedure one can conclude that the addition of X_3 incurs a partial contribution equivalent to 0.15, the addition of X_4 incurs a partial contribution equivalent to 0.03 and the last two covariates X_5 and X_6 provide a low contribution very close to zero. On the graphical point of view the last added covariate that precedes the so called “scree” is the X_4 covariate. For this reason, according to the RGI scree-test, one can conclude that the number of covariates that have to be taken into account are four and precisely the covariates X_1, X_2, X_3 and X_4.

4.5 A case-study: “INVALSI” dataset

Our proposal can be applied in many real contexts: this last section will be devoted to the analysis of the so called “INVALSI” dataset, whose details will be discussed and developed in the following.

As already anticipated in the previous sections, Lorenz zonoids can be carefully employed to problems in which categorical explanatory variables are involved. In this case, in fact, the usage of classical euclidean measures of fit (such as R^2 or residual analysis) may lead to inappropriate conclusions bring-

ing, for instance, in deleting relevant covariates from the adopted linear regression model. However, through our approach one can detect the existence of linear dependence relations which could be excluded if only considering the R^2 measure.

Let us now focus the attention on the "INVALSI" dataset. Every year the Italian Department for Education builds, at the national level, a mathematical test to submit to a selected sample of primary school last classes in order to evaluate the education results. The dataset collects data concerning the scholastic year $2008-2009$ and is characterized by covariates of different nature. In our case we consider a total number of 8 variables: more precisely, four variables assume quantitative nature and four variables assume qualitative nature (in particular two of them are nominal and the other two are ordinal). More in detail, the involved variables are:

- the response variable Y describing the student score obtained in the mathematical test;
- the student citizenship, denoted with X_1;
- the student mark in Italian, denoted with X_2;
- the student mark in Maths, denoted with X_3;
- the student weekly school timetable, denoted with X_4;
- the student father education degree, denoted with X_5;
- the student mother education degree, denoted with X_6;
- the student geographic location, denoted with X_7.

The main purpose consists in analyzing the dependence relations among the score obtained in mathematical test and the parents degree, the school

timetable, the student marks, his citizenship and his geographic location.

The explanatory variable X_1 and X_7 have nominal nature. Below we denote the corresponding assumed values:

- X_1 assumes values $1 =$ "Italian citizenship", $2 =$ "No Italian citizenship";
- X_7 assumes values $1 =$ "North-East Italy", $2 =$ "North-West Italy", $3 =$ "Middle Italy", $4 =$ "South Italy", $5 =$ "South Italy and Islands";

The explanatory variables X_5 and X_6 have ordinal nature, and they assume values between 1 and 6 according to the parents education degree: the value 1 corresponds to the lowest education degree, a "primary school degree", and the value 6 corresponds to the highest education degree, a "university degree".

The covariates X_2, X_3 and X_4 are quantitative variables. The corresponding assumed values are:

- X_2 and X_3 assume values between 3 and 10, with $3 =$ "the lowest mark" and $10 =$ "the highest mark";
- X_4 assumes values $1 =$ "less than 27 hours in a week", $2 =$ "$[27;29]$ hours in a week", $3 =$ "30 hours in a week", $4 =$ "31 hours in a week", $5 =$ "$[32,35]$ hours in a week", $6 =$ "$[36,39]$ hours in a week", $7 =$ "40 hours in a week".

Our purpose is managing the available data in order to evaluate the number of covariates that have to be considered "relevant" and be added into the linear regression model, according to the corresponding contribution to the Y variability explanation. For this reason we compute all the partial measures which allow to define the highest G measure values related to each considered variable according to the usual forward selection procedure. The results highlight that the highest G measure value is provided by the covariate X_3 denoting the student mark in Math, in fact, in this case $G_3 = 0.401191204$.

The first explanatory variable that has to be added into the linear regression model is X_3: the second step consists in detecting the second covariate that gets the largest partial contribution to the model variance explanation. The measure to be computed is the $RGI_{X_1,X_?}$ measure: one can proceed by adding, one by one, the remaining explanatory variables and computing the corresponding RGI measures. The greater RGI measure is obtained by the introduction of covariate X_5, concerning the father education degree, in fact $RGI_{X_3,X_5} = 0.022728391$. One can conclude that the second covariate to be considered in the model is X_5. All the following obtained results are denoted below:

- the third covariate which is added is X_2, concerning the student mark in Italian ($RGI_{X_3,X_5,X_2} = 0.006148842$);
- the fourth covariate which is added is X_7, concerning the student geographic location ($RGI_{X_3,X_5,X_2,X_7} = 0.004200605$);
- the fifth covariate which is added is X_1, concerning the student citizenship ($RGI_{X_3,X_5,X_2,X_7,X_1} = 0.000785606$);
- the sixth covariate which is added is X_6, concerning the student mother education degree ($RGI_{X_3,X_5,X_2,X_7,X_1,X_6} = 0.000686691$);
- the last covariate which is added is the last one X_4, concerning the student weekly school timetable ($RGI_{X_3,X_5,X_2,X_7,X_1,X_6,X_4} = 0.000468083$).

Once obtained all the RGI measures, connected to the forward selection procedure, one has to provide the actual number of explanatory variables that are significantly relevant in the study of the dependence relations among the involved variables. Let us apply the RGI scree-test, whose results can be deduced in Figure 4.3. By following the same factor analysis decisional setting, on the graphical point of view, the last added covariate that precedes the so

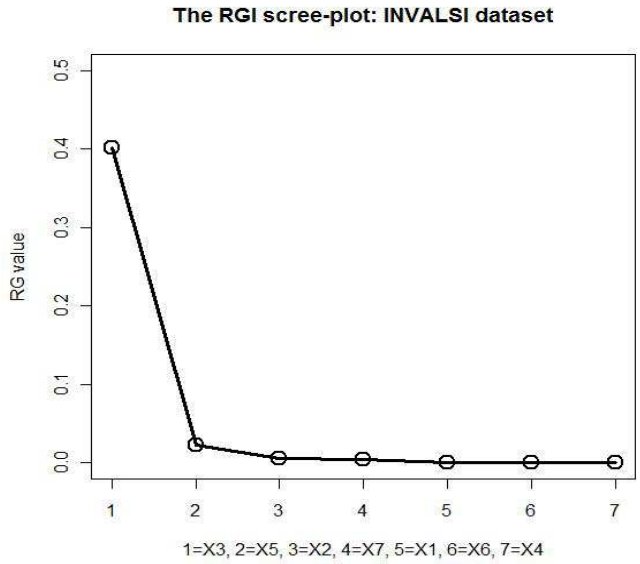

Figure 4.3: The *RGI* plot concerning the forward selection procedure. On the x-axis, the label 1 denotes the first introduced covariate into the model (X_3), the label 2 denotes the second covariate added into the model (X_5), the label 3 denotes the third covariate added into the model (X_2), the label 4 denotes the fourth covariate added into the model (X_7), the label 5 denotes the fifth covariate added into the model (X_1), the label 6 denotes the sixth covariate added into the model (X_6), the label 7 denotes the last covariate added into the model (X_4).

called “scree”, is the X_5 covariate: in conclusion, the two explanatory variables that have to be taken into account and that contribute to define the dependence relations with the student score obtained in the mathematical test, are the student mark in Maths and his father degree.

One of the main problem that can be generated by analyzing datasets characterized by related covariates is provided by a situation of multicollinearity. This situation verifies when in the model two or more predictors can be expressed one as function of the other: obviously it takes place when the correlation between two regressors is so high to make useless the contribution related to the introduction into the model of one of these two covariates. In our context the regressors that present a high correlation value are

- the covariate X_2 (the student mark in Italian) and the covariate X_3 (the student mark in Maths), with a correlation coefficient $\rho = 0.8227445$;

- the covariates X_5 (the father education degree) and X_6 (the mother education degree), with a correlation coefficient $\rho = 0.6154206$.

In order to avoid wrong conclusions about the "relevant" covariates to introduce into the considered model in presence of multicollinearity problems (in this case particularly emerging between the covariate X_2 and X_3) one can repeat the variables selection method through the employment of the backward elimination procedure. In fact, in a context characterized by the multicollinearity problem, the backward elimination could provide more reliable results (see e.g. [1]). According to our data, the backward elimination procedure, starting from a model defined by all the considered covariates, implies the covariates elimination in the following ordering: X_4, X_2, X_1, X_6, X_7 and the last two X_5, X_3. The last covariates that remain in the model, at step two, are the covariate X_3 and X_5 meaning that both the procedure (forward selection and backward elimination), combining together through the stepwise regression approach, come to the same results in terms of variables selection. In other words, the backward elimination procedure allows to validate the conclusions achieved by the forward selection procedure.

4.6 Conclusions

In this research contribution we have shown how Lorenz zonoids can be usefully employed to verify statistical dependence. When considering the multivariate case, one can obtain partial dependence measures based on Lorenz zonoids whose role consists in capturing information on the "explained" variability topic. The discussed Lorenz zonoids approach presents some similarities with the factor analysis: in fact, through the definition of the partial measures G and RGI, one can detect the main covariates that contribute to the explanation of the model variance. Furthermore, the RGI measure rep-

resents a useful stopping rule in the variable selection context: this role is supported by the graphical test, the RGI scree-test, which works based on the same procedure of the scree-test employed in factor analysis.

Another important advantage linked to the Lorenz zonoids approach is provided by the fact that the aforementioned procedure can be applied to explanatory variables expressed in different measure scales because it assures a consistent standardization: this topic could be very interesting in a model choice context. In fact, the classical criteria, such as the AIC and the BIC ones, suffer from the fact that they do not allow a good interpretation when considering different measure scales.

In conclusion, our proposed approach is able to provide a better measure of fit when the available data are characterized also by categorical covariates. The R^2 measure, typically used in order to capture information about the existence and the strength of linear dependence relations among the involved variables, could provide inappropriate results in a quali-quantitative context because, in some cases, it could imply the exclusion of relevant covariates.

Bibliography

[1] Agresti, A.:*Categorical Data Analysis*. Edited by John Wiley and Sons (2002)

[2] Block, H. W., Sampson, A. R., Savits, T. H.: *Topics in Statistics Dependence*. Institute of Mathematical Statistics, Lecture Notes, Vol16, Harward (1990)

[3] Cattell, R. B.: *The scree test for the number of factors*. Multivariate Behavioral Research, 1(2), 245-276 (1966)

[4] Colangelo, A., Scarsini, M., Shaked, M.: *Some Positive Dependence Stochastic Orders*. Journal of Multivariate Analysis, Volume 97, Issue 1, pp 46–78 (2006).

[5] Giudici, P.: *Applied Data Mining*. Wiley (2003)

[6] Giudici, P., Raffinetti, E.: *Multivariate Ranks-based concordance indexes*. Published in the volume of selected papers “Statistical Methods for the analysis of large data-sets”. Springer (2011)

[7] Johnson, R. A., Wichern, D. W.: *Applied Multivariate Statistical Analysis*. Prentice-Hall, Upper Saddle River, NJ (2002)

[8] Karlin, S.: *Total Positivity*. Vol. I, Stanford University Oress, Stanford (1968)

[9] Kimeldorf, G., Sampson, A.R.: *Positive dependemce orderings*. Ann. Inst. Statist. Math., 39 (1987)

[10] Koshevoy, G.: *Multivariate inequality indices and orderings on a product of symmetric groups*. Economica i Matematicheskie Methody, 29 (1993)

[11] Koshevoy, G.: *Multivariate Lorenz majorization*. Social Choice and Welfare, 12 (1995)

[12] Koshevoy, G., Mosler, K.:*The Lorenz Zonoids of a Multivariate Distribution*. Journal of the American Statistical Association, 91, No. 434, Theory and Methods (1996)

[13] Koshevoy, G., Mosler, K.: *Multivariate Lorenz dominance based on zonoids*. AStA Advances in Statistical Analysis, Vol 91, No. 1 (2007)

[14] Lehmann, E.L.: *Some concepts of dependence*. Ann. Math. Statist., 37 (1966)

[15] McCullagh, P., Nelder, J.A.: *Generalized Linear Models*. Chapman and Hall, Second Edition (1989)

[16] Mosler, K.: Majorization in economic disparity measures. Linear Algebra and its Applications, 220 (1994)

[17] Muliere, P.: *Some remarks about the horizontal equity of a taxation* (in Italian). Ed. by Bocconi Comunicazione **2**, (Milano, 1986)

[18] Muliere, P., Petrone, S.: *Generalized Lorenz curve and monotone dependence orderings*. Metron, Vol. L, No. 3–4 (1992)

[19] Rencher, A. C.: *Methods of multivariate analysis*. Edited by John Wiley and Sons (2002)

[20] Rinott, Y., Pollak, M.: *A stochastic ordering induced by a concept of positive dependence and monotonicity of asymptotic test sizes*. The Annals of Statistics, Vol. 8, No 1 (1980)

[21] Scarsini, M.: *On measures of concordance*. Stochastica, 8, (1984)

[22] Shea, G.: *Monotone regression and covariance structure*. Annals of Statistics, Vol. 7, (1979)

[23] Tchen A. H.: *Inequalities for Distributions with Given Marginals*. Annals of Probability, Vol. 8, No. 4, pp 814-827 (1980)

[24] Tukey, J.: *A problem of Berkson and minimum variance orderly estimation*. Ann. Math. Statist., 29, (1958)

Conclusions

The present PhD thesis focuses on the extension of multivariate dependence concepts: the current statistical literature provides different approaches devoted to the dependence analysis. As already discussed, the analysis of monotone dependence can be covered by using particular statistical tools, such as the Lorenz curves and their generalization in the multivariate context, named Lorenz zonoids. Some attempts, devoted to the definition of bivariate dependence measures, have enriched the literature through the introduction of innovative theoretical methodologies (see Chapter 1).

Starting from the Lorenz and dual Lorenz curves definition until defining the Lorenz zonoid concept, which is related to the Gini measure when $d = 1$ (as described in Chapters 3 and 4), one can obtain a complete and coherent overview of the different approaches that can be applied in order to define dependence measures in a multivariate context. The main result highlighted in this discussion concerns the link that we have been able to establish among all the three different ways of proceeding illustrated in the central chapters.

In Chapter 2, the proposal of a multivariate concordance index, has its source in the “horizontal equity” taxation problem resolution: the aim consists in considering the influence of $(k-1)$ explanatory variables on the response variable. In order to depict the relation among the response variable and the involved covariates, a multiple linear regression function has been applied. By the employment of the ranks-based approach, the observed values are

ordered with respect to the ranks assigned to the corresponding estimated values. Through this ordering one can build the *concordance curve* which lies in the area between the response variable Lorenz curve and its dual. By exploiting the ranks-based ordering a novel multivariate index, that we have called the *concordance index*, has been proposed. The aforementioned index, which is shown belonging to the range $[-1, 1]$, is more sensitive than the Gini's or Kendall's ones, since it is grounded on the real values of the response variable, rather than on their corresponding rank transformation. In particular, the concordance index assumes positive values if there is concordance among the response variable and the involved explanatory variables, whereas it assumes negative values if, among the response variable and the involved explanatory variables, there is discordance. The extreme values, respectively -1 and $+1$, are reached when the concordance curve perfectly overlaps with the response variable dual Lorenz curve and when the concordance curve perfectly overlaps with the response variable Lorenz curve. The original contribution provided in this chapter, is represented by the definition of an innovative model goodness of fit measure which overcomes the restrictions related to other performance measures, particularly useful when the most relevant explanatory variables are categorical. For this reason, a substantial support in models evaluation context, can be given by the application of the proposed index. Moreover, the ranks-based approach could represent a basic support for extending multivariate dependence measures involving ordinal or nominal response variables: this new research key is currently in progress.

In Chapter 3, the proposal of Chapter 2 is presented in terms of a *new decomposition of the Gini measure*. More precisely, this approach provides the decomposition of the mutual variability, almost analogous to the variance decomposition, in terms of the mutual variability explained by the regression of Y on the considered $(k-1)$ explanatory variables. Through this approach, a

new kind of dependence, called the *Gini Rank Dependence (GRD)*, has been formalized. A first aim of the proposed decomposition consists in defining the concordance and discordance degree. In particular, the GRD index provides a new approach to the model diagnostic approach known as residual analysis, especially useful when the relevant involved explanatory variables have categorical nature. Through the concordance curve construction one is able to identify the observation which contribute to the concordance share and to the discordance share: let us recall that the discordance shares correspond to a relevant change of the original Y response variable ranks with respect to the ranks assigned to the corresponding estimated values. If the concordance curve is completely lying in the discordance area, there is a situation of no linear dependence among the response variable Y and the involved covariates, meaning that our adopted linear estimated regression model is not appropriate to fit the data. In the opposite case, the linear dependence between Y and the $(k-1)$ considered covariates is satisfied. A possible further development could consist in defining the concordance function analytical expression by the employment of order statistics.

Chapter 4 moves from the notion of *Lorenz zonoid*, that extends the Lorenz curve to the multivariate context. In this chapter the aim is to define a *partial dependence measure* whose role consists in expressing the share of the Y Lorenz zonoid “explained” by each explanatory variable. The theoretical setting is characterized by considering a multiple linear regression model: the purpose is characterized by measuring the additional contribution related to the introduction of a new covariate (denoted with $k+1$) into the model in terms of Lorenz zonoids. The obtained partial dependence measures are defined through a ratio whose numerator describes the difference between the Lorenz zonoid of the response variable estimated values, related to the introduction of an additional explanatory variable, and the Lorenz zonoid built by consider-

ing only the k original explanatory variables. The denominator corresponds to the difference between the response variable Lorenz zonoid and the Lorenz zonoid based on the k original explanatory variables. Even if we are operating in a multivariate context, selecting a multiple linear regression model, we are able to bring back the analysis to the univariate dimension allowing to satisfy the correspondence between the Lorenz zonoid and the Gini measure. Since the role of the Gini measure consists in describing the statistical dispersion, one can exploit it as a variability measure.

By the means of the Lorenz zonoids inclusion, the partial Lorenz dependence measures assure to define the contribution related to the additional explanatory variable in terms of the "explained" model variance. A specific graphical test has been illustrated in order to establish the total number of covariates to be added into the model through the forward inclusion procedure. This test is called the RGI scree-test since it is based on a graphical support similar to the one used in factor analysis and aimed at defining the number of factor to be extracted. The proposed partial dependence measure parallels the partial correlation index: however, its employment is motivated when one considers a quali-quantitative context (that is when the most "relevant" explanatory variables assume categorical nature). Furthermore, the RGI presents as a good criterion in the variable selection context. A possible interesting future research topic, could consist in providing inferential results about the RGI measure: more precisely, studying its asymptotic distribution and then constructing a non-parametric test based on the related statistics.

Summarizing, the basic research contribution provided by this PhD thesis consists in defining new dependence measures in a multivariate context particularly useful for model goodness of fit and model selection when most explanatory variables are categorical.