Lorenz zonoid measures to compare predictive accuracy

Paolo Giudici and Emanuela Raffinetti

Abstract In this paper, a novel approach for model comparison is presented by means of the Gini approach tools. Specifically, the standard linear regression model framework is considered and extended through the employment of the Lorenz and concordance curves. The Lorenz curve, obtained by re-ordering the normalised values of the variable in non-decreasing sense, gives raise to the Gini coefficient. The more the Lorenz curve moves away from the bisector curve, corresponding to the case of degenerate variables, the greater is the variable variability measured by the Gini coefficient. Starting from these premises, a novel criterion for the evaluation of the contribution to the response variable explanation associated with each new covariate included into the model is introduced. This criterion, named concordance criterion, is defined in terms of the distance between the points lying on the concordance curve, obtained by re-ordering the normalised response variable values according to the related model fitted values, and those lying on the response variable Lorenz curve. In a model choice perspective, the concordance criterion fulfills interesting properties: it mimics the partial correlation coefficient and assures a consistent standardization, being invariant to the scale transformations of the involved variables.

Key words: Linear models, Lorenz and concordance curves, concordance criterion, partial correlation coefficient, market price analysis.

Paolo Giudici
Department of Economics and Management, University of Pavia, Via San Felice 5, 27100 Pavia (Italy), e-mail: paolo.giudici@unipv.it

Emanuela Raffinetti
Department of Economics, Management and Quantitative Methods, University of Milan, Via Conservatorio 7, 20122 Milan (Italy), e-mail: emanuela.raffinetti@unimi.it

1 Introduction

A very important problem in applied statistics is to compare alternative models on a given database, for example in terms of their predictive accuracy.

The traditional paradigm compares statistical models within the theory of statistical hypotheses testing, in which a model is chosen through a sequence of pairwise comparisons. These criteria are generally not applicable to machine learning models, which do not necessarily have an underlying probabilistic model and, therefore, do not allow the application of statistical hypotheses testing theory. In these cases models are compared in terms of information criteria such as AIC or BIC.

The last few years have witnessed the growing importance of model comparison methods based on the direct calculation of the predictive accuracy of a model, through cross-validation methods. In the cross-validation process, the data is split in two or more datasets, with training datasets used to fit a model and validation datasets used to compare the predictions made by the fitted model with the actual observed values. When the response variable is continuous, a typical cross-validation summary criterion is the root mean squared error (RMSE) which calculates the difference between the observed and the predicted value.

A problem with cross-validation measures, such as the RMSE, is that they are not normalised, similarly to what occurs with information based criteria but differently from what occurs with statistical tests. A second problem is that, when the number of explanatory variables increases the RMSE does not necessarily decrease.

We aim to overcome these drawbacks with a new cross-validation measure that is normalised and owns an "inclusion property" such that, when a new variable is added, the predictive accuracy improves. To achieve this aim we resort to statistical dependency tools.

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. In literature, several approaches addressed to dependence have been developed. In such a context, one of the main occurring problem concerns the application of standard dependence measures to capture information about the real existence of dependence relations among the involved variables. Let us suppose, for instance, to assume the employment of a multiple linear regression model built on a quantitative response variable. Typically, the existence of dependence relations among the response variable and the considered explanatory ones can be detected through the so called multiple linear determination coefficient. This measure allows also to assess the data goodness of fit: anyway, the multiple linear regression coefficient is affected by some relevant restrictions since based on the euclidean distance and then more appropriate in a quantitative context of analysis, as argued by [1]. For this reason, an interesting research field is represented by the definition of novel and more reliable dependence measures, as proved in the contribution provided in [4], where an alternative to residual analysis is proposed through a new Gini measure decomposition in terms of concordance and discordance. The recall to the Gini measure and thus to the underlying Lorenz curve is motivated by the last decade research proposals in dependence analysis. In [12], for instance, a partial ordering of monotone dependence on the class of non-negative bivariate random vectors with given marginals was defined.

The rest of the paper is organized as follows. Section 3 provides a background on the Lorenz zonoid tool, especially on its main features and properties. Section 3 focuses on the formalization of the Lorenz-zonoid based dependence measures in the linear model framework. Section 4 illustrates an application to crypto market price data. Finally, Section 5 briefly concludes the paper.

2 Background

In this paper we refer to a partial ordering based on a specific statistical tool, named Lorenz zonoid. When considering multivariate data, the Lorenz zonoid represents the multidimensional extension of the Lorenz curve. The Lorenz zonoid has been introduced by [6] for empirical distributions and [11] for general probability distributions. More precisely, the Lorenz zonoid of a d-dimensional random vector corresponds to a convex set in $\mathbb{R}^{d+1}$, whose role is analyzing and comparing random vectors. Through the Lorenz zonoid representation one can establish an ordering of random vectors that reflects their variability: the investigation of such ordering is induced by the inclusion between Lorenz zonoids. This aspect provides a helpful support for our proposed development.
Let us introduce the Lorenz curve definition of a non-negative variable Y, as reported in [8]. The Lorenz curve of a random variable Y having expectation μ is the graph of the function

$$t \mapsto \mu^{-1} \int_0^t F_Y^{-1}(s)ds, 0 \leq t \leq 1$$

where F_Y^{-1} is the quantile function of X, $F_Y^{-1} = \min\{y : R(y) \geq t\}, \quad 0 < t \leq 1$. Roughly speaking, given n observation, the Y variable L_Y Lorenz curve (see [10]) is given by the set of points $(i/n, \sum_{j=1}^{i} y_{(j)}/(n\bar{y}))$, for $i = 1,\dots,n$, where $y_{(i)}$ indicates the Y variable values ordered in a non-decreasing sense and $\bar{y}$ is the Y variable mean value. Analogously, the Y variable can be re-ordered in a non-increasing sense providing the $L_Y^{'}$ dual Lorenz curve, which is defined as the set of points $(i/n, \sum_{j=1}^{i} y_{(n+1-j)}/(n\bar{y}))$. The area lying between the L_Y and $L_Y^{'}$ Lorenz curves corresponds to the Gini coefficient, which is typically employed as an indicator of inequality, especially when dealing with income data. This holds in the univariate case. When considering more than one variable, the Lorenz curve generalization in d dimensions is the so-called Lorenz zonoid.

The Lorenz zonoid of a general d-variate random vector is defined as follows (see e.g. [7]). Consider the set $\mathscr{Y}^d$ of random vectors in $\mathbb{R}^d$ that have finite expectation, the subset $\mathscr{Y}^{d+} \subset \mathscr{Y}^d$ of those vectors that have positive (in each component) expectation, and the subset $\mathscr{Y}_+^{d+} \subset \mathscr{Y}^{d+}$ of those that have, in addition, support in $\mathbb{R}_+^d$.

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

$$\tilde{\mathbf{Y}} = \left(\frac{Y_1}{E(Y_1)}, \dots, \frac{Y_d}{E(Y_d)}\right),$$

in order to point out the relative vector[1] that is the vector componentwise divided by its expectation.

The Lorenz zonoid of a random vector $\mathbf{Y} \in \mathscr{Y}^{d+}$ is a convex compact set in $\mathbb{R}^{d+1}$, defined as follows:

$$LZ(\mathbf{Y}) = \left\{E[(g(\tilde{\mathbf{Y}}), g(\tilde{\mathbf{Y}})\tilde{\mathbf{Y}}] : g : \mathbb{R}^d \to [0,1] \text{ measurable}\right\}.$$

For the sake of clarity, a function $g : E \to \mathbb{R}$ is measurable if E is a measurable set and for each real number r, the set $\{y \in E : g(y) > r\}$ is measurable. It derives that continuous and monotone functions are measurable. We remark that if $\mathbf{X} \in \mathscr{Y}_{+}^{d+}$, i.e. has support in $\mathbb{R}_{+}^{d}$, the Lorenz zonoid is contained in the hypercube of $\mathbb{R}^{d+1}$. The Lorenz zonoid fulfills many attractive properties, some of which are basic for the contribution proposed here.

***Property* 1** *A linear dependence preorder $\preceq_{ld}$ on $\mathscr{Y}^{d+}$ is defined as follows (see, for instance, [2])*

$$\mathbf{Y} \preceq_{ld} \mathbf{X} \quad \textit{if} \quad LZ(\mathbf{X}) \subset LZ(\mathbf{Y}), \tag{1}$$

where $LZ(\mathbf{X})$ and $LZ(\mathbf{Y})$ are the Lorenz zonoids of the random vectors $\mathbf{X}$ and $\mathbf{Y}$.

***Property* 2** *For $\mathbf{X}$ and $\mathbf{Y} \in \mathscr{Y}^{d+}$, the Lorenz zonoid order (Lorenz dominance) $\preceq_L$, is defined as (see, for instance, [8]):*

$$\mathbf{Y} \preceq_L \mathbf{X} \quad \textit{if} \quad LZ(\mathbf{X}) \subset LZ(\mathbf{Y}).$$

Let us denote with $\preceq_{dil}$ the dilation order. A perfect equivalence between the dilation order and the Lorenz zonoid order is provided by the following corollary:

***Corollary* 1** $\mathbf{X} \preceq_{dil} \mathbf{Y} \Rightarrow \mathbf{X} \preceq_L \mathbf{Y}$.

Of our interest is the Lorenz zonoid in the univariate case, i.e., the Gini coefficient represented by the area between the variable Lorenz and dual Lorenz curves. Henceforth we denote the Gini coefficient with the notation $LZ_{d=1}(\cdot)$, indicating the Lorenz zonoid in the univariate case. The condition of linear dependence reported in Proposition 1, can be further on re-formalized to cover the case of variables whose linear relationships may be investigated through a linear regression model.

***Proposition* 1** *Consider the vector of variables (Y,X) and apply a simple linear regression model, such that $\hat{y} = \hat{\alpha} + \hat{\beta}x$. Assume that $\hat{Y}$ takes non-negative values. Denote respectively with $L_Y(t)$ and $L_Y'(t)$ the Y Lorenz curve and its dual, and with*

[1] Relative data are data divided by their mean value as in the classical definition of the Lorenz curve.

$L(\hat{Y})$ and $L'_{\hat{Y}}(t)$ the $\hat{Y}$ Lorenz curve and its dual. One can prove that $L_Y(t) \leq L_{\hat{Y}}(t) \leq L'_Y(t)$ (see e.g. [12]), where $L'_Y(t) = \frac{1}{E(Y)} \int_{1-t}^{1} F_Y^{-1}(s)ds$, $0 \leq t \leq 1$. Furthermore, $L'_{\hat{Y}}(t) \leq L'_Y(t)$.

We denote Proposition 1 as the Lorenz zonoid (when $d = 1$) "*inclusion property*". The existence of a linear dependence relation between Y and X translates into an inclusion between the response variable Y and the linear estimated variable $\hat{Y}$ Lorenz zonoids. Fig. 1 shows this outcome in a depicted way. By establishing a linear relation between the response variable Y and the X covariate, through a simple linear regression function, one computes the response variable estimated values $\hat{y}$ and then proceeds to the construction of the $\hat{Y}$ Lorenz zonoid. The Y variable Lorenz zonoid contains the corresponding linear estimated variable $\hat{Y}$ Lorenz zonoid, as one can deduce from Fig. 1.

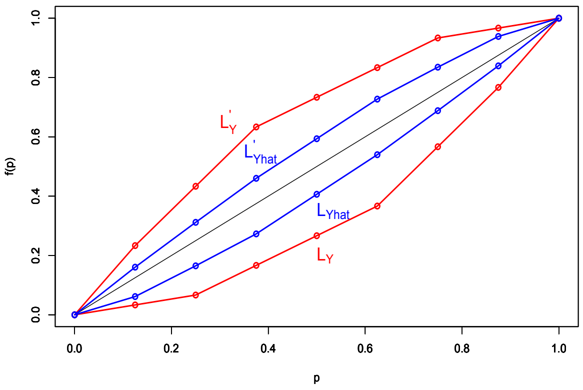

Fig. 1 Y Lorenz zonoid (area between red lines) and $\hat{Y}$ Lorenz zonoid (area between blue lines).

A direct implication of Proposition 1 is that the $\hat{Y}$ Lorenz zonoid appears as a useful tool to define the total variability explained by the response variable linear estimated values. Note that if $d = 1$, the reverse implication of $\hat{Y} \preceq_{dil} Y \Rightarrow \hat{Y} \preceq_L Y$ in Corollary 1 and equivalent to $\hat{Y} \preceq_L Y \Rightarrow \hat{Y} \preceq_{dil} Y$ holds, meaning that through the Lorenz dominance an ordering based on the variable variability degree can be specified.

***Remark* 1** *Note that the inclusion property is fulfilled if the $\hat{Y}$ estimated values are all non-negative. In the case of also some negative values, the $\hat{Y}$ Lorenz curve partially lies under the x-axis and the $\hat{Y}$ dual Lorenz curve partially lies over the y-axis upper bound 1. A direct implication follows from this: the $\hat{Y}$ Lorenz zonoid may take values greater than one.*

3 Proposal: measuring the contribution of explanatory variables in linear models

The employment of the Lorenz zonoids extended to the evaluation of the partial contribution provided by the single explanatory variable differs from the standard indices typically used in the classical linear modeling framework. In our perspective, we exploit the essence of the Lorenz zonoid ($LZ_{d=1}(\cdot)$) to be a measure assessing the variability characterizing the phenomenon of interest. This feature strictly relates wth the goals associated with the classical linear models aimed at assessing the contribution of the single independent variable in explaining the variability of the response variable. In this section we illustrate our proposal by: first introducing new marginal dependence measures, addressed to explain the response variable Lorenz zonoid share "explained" by each single considered explanatory variable; and subsequently by measuring the effect related to the introduction of a new explanatory variable into the linear regression model.

3.1 Marginal dependence measures

Let $LZ_{d=1}(Y)$ be the Lorenz zonoid of the response variable Y and X_1 be the independent variable, such that the $\hat{Y}_{X_1}$ is the vector of the linear estimated values computed by resorting to the simple linear regression model, such that $\hat{Y} = \hat{\alpha} + \hat{\beta} X_1$. Define $LZ_{d=1}(\hat{Y}_{X_1})$ as the Lorenz zonoid of $\hat{Y}_{X_1}$. Suppose to consider an additional independent variable X_2 and to apply a simple linear regression model such that $\hat{Y}_{X_2} = \hat{\alpha} + \hat{\beta} X_2$. We denote with $LZ_{d=1}(\hat{Y}_{X_2})$ the corresponding Lorenz zonoid. Following [9], in the univariate case, the Lorenz zonoid of a variable may be expressed by resorting to the covariance formula. Consider the response variable Y. In such a case,

$$LZ_{d=1}(Y) = \frac{2Cov(Y, F(Y))}{\mu}, \tag{2}$$

where μ is the response variable Y mean value and $F(Y)$ is uniformly distributed between the close range $[0,1]$. In the same manner, $LZ_{d=1}(\hat{Y}_{X_1})$ and $LZ_{d=1}(\hat{Y}_{X_2})$ can be expressed as

$$LZ_{d=1}(\hat{Y}_{X_1}) = \frac{2Cov(\hat{Y}_{X_1}, F(\hat{Y}_{X_1}))}{\mu} \quad \text{and} \quad LZ_{d=1}(\hat{Y}_{X_2}) = \frac{2Cov(\hat{Y}_{X_2}, F(\hat{Y}_{X_2}))}{\mu}, \tag{3}$$

where $E(\hat{Y}_{X_1}) = E(E(Y|\hat{Y}_{X_1})) = \mu$ and $E(\hat{Y}_{X_2}) = E(E(Y|\hat{Y}_{X_2})) = \mu$; $F(\hat{Y}_{X_1})$ and $F(\hat{Y}_{X_2})$ share the same feature of $F(Y)$ to be uniformly distributed in the close range $[0,1]$.

Let $R(Y)$, $R(\hat{Y}_{X_1})$ and $R(\hat{Y}_{X_2})$ be the ranks of each observation i characterising the Y, $\hat{Y}_{X_1}$ and $\hat{Y}_{X_2}$ variables. Since $R(\cdot)/n$ terms are the empirical representation of $F(\cdot)$, the covariance has to be divided by n, so that expressions in (2) and (3) become

$$LZ_{d=1}(Y) = \frac{2Cov(Y,R(Y))}{n\mu}, \quad LZ_{d=1}(\hat{Y}_{X_1}) = \frac{2Cov(\hat{Y}_{X_1},R(\hat{Y}_{X_1}))}{n\mu}$$
$$\text{and } LZ_{d=1}(\hat{Y}_{X_2}) = \frac{2Cov(\hat{Y}_{X_2},R(\hat{Y}_{X_2}))}{n\mu}. \tag{4}$$

Proof. Consider the response variable Y. We have to prove that

$$LZ_{d=1}(Y) = \frac{2Cov(Y,F(Y))}{\mu} = \frac{2Cov(Y,R(Y))}{n\mu}. \tag{5}$$

The term $Cov(Y,F(Y))$ is equivalent to $Cov\left(Y,\frac{R(Y)}{n}\right)$. Through some computations, we obtain that

$$Cov\left(Y,\frac{R(Y)}{n}\right) = \frac{1}{n}\sum_{i=1}^{n} Y_i \frac{R(Y_i)}{n} - \mu\frac{\bar{R}(Y)}{n} = \frac{1}{n}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i R(Y_i) - \mu\bar{R}(Y)\right]$$
$$= \frac{1}{n}Cov(Y,R(Y)),$$

and the equivalence in (5) follows. Trivially, this result holds also for variables $\hat{Y}_{X_1}$ and $\hat{Y}_{X_2}$.

Given a sample data of size n, formulas in (2) and (3) may be re-expressed as follows:

$$LZ_{d=1}(y) = \frac{2Cov(y,r(y))}{n\bar{y}}, \quad LZ_{d=1}(\hat{y}_{x_1}) = \frac{2Cov(\hat{y}_{x_1},r(\hat{y}_{x_1}))}{n\bar{y}}$$
$$\text{and } LZ_{d=1}(\hat{y}_{x_2}) = \frac{2Cov(\hat{y}_{x_2},r(\hat{y}_{x_2}))}{n\bar{y}}, \tag{6}$$

where y, $\hat{y}_{x_1}$ and $\hat{y}_{x_1}$ are the vectors of the observed and estimated values; $r(y)$, $r(\hat{y}_{x_1})$ and $r(\hat{y}_{x_2})$ are the ranks of the observed values; $\bar{y}$ is the sample mean.

When $d = 1$, the Lorenz zonoid can be also derived as a function of the distance between the y-axis values of the points lying on the Lorenz curve and those of the points lying on the bisector curve (the black curve in Fig. 1).

Proof. Consider the response variable Y, whose values arranged in non-decreasing sense are denoted with $y_{(i)}$, for $i = 1,\ldots,n$. Let d be the distance between the the y-axis values of the points lying on the Lorenz curve, characterised by coordinates $(i/n, \sum_{j=1}^{i} y_{(j)}/n\bar{y})$, and those of the points lying on the bisector curve, characterised by coordinates $(i/n, i/n)$. It follows that

$$d=\sum_{i=1}^{n}\left\{\frac{i}{n}-\frac{1}{n\bar{y}}\sum_{j=1}^{i}y_{(j)}\right\}=\sum_{i=1}^{n}\frac{i}{n}-\frac{1}{n\bar{y}}\sum_{i=1}^{n}\sum_{j=1}^{i}y_{(j)} \tag{7}$$

Because $\sum_{i=1}^{n}i=\frac{n(n+1)}{2}$ and $\sum_{i=1}^{n}\sum_{j=1}^{i}y_{(j)}=n(n+1)\bar{y}-\sum_{i=1}^{n}iy_{(i)}$, the term on the right side of equation (7) can be written as

$$\begin{aligned}d=\frac{n(n+1)}{2n}-\frac{1}{n\bar{y}}\left[n(n+1)\bar{y}-\sum_{i=1}^{n}iy_{(i)}\right]&=\frac{1}{\bar{y}}\left[\frac{1}{n}\sum_{i=1}^{n}iy_{(i)}-\frac{n(n+1)}{2n}\bar{y}\right]\\&=\frac{1}{\bar{y}}\left[\frac{1}{n}\sum_{i=1}^{n}iy_{(i)}-\frac{(n+1)}{2}\bar{y}\right].\end{aligned}$$

For the covariance approach, note that the mean of ranks $\bar{r}(y)$ equals to $\bar{i}=(n+1)/2$ providing that

$$d=\frac{1}{\bar{y}}cov(y_{(i)},r(y)). \tag{8}$$

From (8), it derives that that $LZ_{d=1}(y)$ is function of d, so that

$$LZ_{d=1}(y)=\frac{2}{n\bar{y}}\left[\frac{1}{n}\sum_{i=1}^{n}iy_{(i)}-\frac{n(n+1)}{2n}\bar{y}\right] \tag{9}$$

Let $\hat{y}_{(x_1i)}$ and $\hat{y}_{(x_2i)}$, for $i=1,\ldots,n$, be the $\hat{Y}_{X_1}$ and $\hat{Y}_{X_2}$ values arranged in a non-decreasing sense. Analogously, $LZ_{d=1}(\hat{y}_{x_1})$ and $LZ_{d=1}(\hat{y}_{x_2})$ may expressed in terms of distance between the the y-axis values of the points lying on the Lorenz curve and those of the points lying on the bisector curve, as follows:

$$LZ_{d=1}(\hat{y}_{x_1})=\frac{2}{n\bar{y}}\left[\frac{1}{n}\sum_{i=1}^{n}i\hat{y}_{(x_1i)}-\frac{n(n+1)}{2n}\bar{y}\right] \tag{10}$$

$$LZ_{d=1}(\hat{y}_{x_2})=\frac{2}{n\bar{y}}\left[\frac{1}{n}\sum_{i=1}^{n}i\hat{y}_{(x_2i)}-\frac{n(n+1)}{2n}\bar{y}\right]. \tag{11}$$

The term included into the squared brackets appearing in (9), (10) and (11) equations is the covariances between the y, $\hat{y}_{(x_1)}$ and $\hat{y}_{(x_2)}$ values and their corresponding ranks.

Through the Lorenz zonoids-based orderings, marginal dependence measures, we denote with *MGC* (acronym of "*Marginal Gini Coefficient*"), can be formalized in order to evaluate the Y Lorenz zonoid (Gini coefficient) share "explained" by covariate X_1 and covariate X_2 alone. More precisely, the following ratios are introduced:

$$MGC_{(Y|X_1)} = \frac{LZ_{d=1}(\hat{Y}_{X_1})}{LZ_{d=1}(Y)} = \frac{2Cov(\hat{Y}_{X_1}, R(\hat{Y}_{X_1}))/n\mu}{2Cov(Y, R(Y))/n\mu} = \frac{Cov(\hat{Y}_{X_1}, R(\hat{Y}_{X_1}))}{Cov(Y, R(Y))} \tag{12}$$

and

$$MGC_{(Y|X_2)} = \frac{LZ_{d=1}(\hat{Y}_{X_2})}{LZ_{d=1}(Y)} = \frac{2Cov(\hat{Y}_{X_2}, R(\hat{Y}_{X_2}))/n\mu}{2Cov(Y, R(Y))/n\mu} = \frac{Cov(\hat{Y}_{X_2}, R(\hat{Y}_{X_2}))}{Cov(Y, R(Y))}. \tag{13}$$

Generally, given k explanatory variable the marginal contribution associated with the *h-th* explanatory variable (with $h = 1, \dots, k$) is

$$MGC_{(Y|X_h)} = \frac{LZ_{d=1}(\hat{Y}_{X_h})}{LZ_{d=1}(Y)} = \frac{2Cov(\hat{Y}_{X_h}, R(\hat{Y}_{X_h}))/n\mu}{2Cov(Y, R(Y))/n\mu} = \frac{Cov(\hat{Y}_{X_h}, R(\hat{Y}_{X_h}))}{Cov(Y, R(Y))} \tag{14}$$

whose sample version is

$$MGC_{(y|x_h)} = \frac{\frac{2}{n\bar{y}}\left[\frac{1}{n}\sum_{i=1}^{n} i\hat{y}_{(x_h i)} - \frac{n(n+1)}{2n}\bar{y}\right]}{\frac{2}{n\bar{y}}\left[\frac{1}{n}\sum_{i=1}^{n} iy_{(i)} - \frac{n(n+1)}{2n}\bar{y}\right]} = \frac{Cov(\hat{y}_{x_h}, r(\hat{y}_{x_h}))}{Cov(y, r(y))}. \tag{15}$$

Due to their features, the *MGC* measures may be used in the forward stepwise regression procedure, i.e., the independent variable with the largest contribution in explaining the share of the response variable Lorenz zonoid (variability) measured by the associated *MGC*, is introduced as the first one into the linear regression model. Without loss of generality, the inclusion of the remaining independent variables occurs in the same manner, by resorting to the partial measures introduced in the following subsection.

3.2 *Partial dependence models*

Consider the general context characterized by k explanatory variables: our proposal is to determine the effect related to the introduction of a new $(k+1)$-*th* explanatory variable into the linear regression model. The inclusion of a new explanatory variable provides an enlargement of the $\hat{Y}$ Lorenz zonoid. The Lorenz zonoid of the Y linear estimated values, denoted with $LZ_{d=1}(\hat{Y}_{X_1,\dots,X_k})$, corresponds to the dilation measure of the Y response variable Lorenz zonoid $LZ_{d=1}(Y)$. Therefore, the introduction of an additional covariate in multiple linear regression models translates into an increase of the "explained" Y variability.

In the well-known 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 increase of the multiple determination coefficient (see e.g. [3]). More precisely, suppose to build a linear regression model characterized by k explanatory variables. Let us introduce an additional

$(k+1)$-*th* explanatory variable: its contribution determines an increase of the Y variable "explained" variability, defined as the difference between $Var(\hat{Y}_{X_1,\ldots,X_{k+1}})$ and $Var(\hat{Y}_{X_1,\ldots,X_k})$[2]. The squared partial correlation coefficient is expressed as

$$r^2_{Y,X_{k+1}|X_1,\ldots,X_k} = \frac{Var(\hat{Y}_{X_1,\ldots,X_{k+1}}) - Var(\hat{Y}_{X_1,\ldots,X_k})}{Var(Y) - Var(\hat{Y}_{X_1,\ldots,X_k})}, \tag{16}$$

where $Var(Y) - Var(\hat{Y}_{X_1,\ldots,X_k})$ identifies the Y variable variability not explained by the $X_1,\ldots,X_k$ ovariates.

We aim at building a partial dependence measure that "parallels" the partial correlation coefficient construction. Specifically, we propose as partial dependence measure the ratio between a numerator characterized by a term denoting the contribution generated by the $(k+1)$-*th* explanatory variable and a denominator including a term which describes the share of the Y Lorenz zonoid "not explained" by the $\hat{Y}_{X_k}$ Lorenz zonoid. The additional contribution related to the $(k+1)$-*th* explanatory variable inclusion can be measured through the difference between the $\hat{Y}_{X_1,\ldots,X_{k+1}}$ and $\hat{Y}_{X_1,\ldots,X_k}$ Lorenz zonoids, that is $LZ_{d=1}(\hat{Y}_{X_1,\ldots,X_{k+1}}) - LZ_{d=1}(\hat{Y}_{X_1,\ldots,X_k})$.

A relative index, measuring the additional contribution provided by the X_{k+1} independent variable is obtained in analogy with the partial correlation coefficient construction. Such a measure, which we call "*Relative Gini Index*", is expressed as:

$$RGI_{Y,X_{k+1}|X_1,\ldots,X_k} = \frac{LZ_{d=1}(\hat{Y}_{X_1,\ldots,X_{k+1}}) - LZ_{d=1}(\hat{Y}_{X_1,\ldots,X_k})}{LZ_{d=1}(Y) - LZ_{d=1}(\hat{Y}_{X_1,\ldots,X_k})}. \tag{17}$$

By resorting to the covariance formulas, equation in (17) becomes

$$\begin{aligned} RGI_{Y,X_{k+1}|X_1,\ldots,X_k} &= \frac{\frac{2}{n\mu}Cov(\hat{Y}_{X_1,\ldots,X_{k+1}}, R(\hat{Y}_{X_1,\ldots,X_{k+1}})) - \frac{2}{n\mu}Cov(\hat{Y}_{X_1,\ldots,X_k}, R(\hat{Y}_{X_1,\ldots,X_k}))}{\frac{2}{n\mu}Cov(Y,R(Y)) - \frac{2}{n\mu}Cov(\hat{Y}_{X_1,\ldots,X_k}, R(\hat{Y}_{X_1,\ldots,X_k}))} \\ &= \frac{Cov(\hat{Y}_{X_1,\ldots,X_{k+1}}, R(\hat{Y}_{X_1,\ldots,X_{k+1}})) - Cov(\hat{Y}_{X_1,\ldots,X_k}, R(\hat{Y}_{X_1,\ldots,X_k}))}{Cov(Y,R(Y)) - Cov(\hat{Y}_{X_1,\ldots,X_k}, R(\hat{Y}_{X_1,\ldots,X_k}))}. \end{aligned} \tag{18}$$

It is worth noting that for the first *h-th* variable included into the model, the equivalence $MGC_{(Y|X_h)} = RGI_{Y|X_h}$ holds.
Through some manipulations, the $RGI_{Y,X_{k+1}|X_1,\ldots,X_k}$ computed on sample data is expressed as:

$$RGI_{y,x_{k+1}|x_1,\ldots,x_k} = \frac{\sum_{i=1}^n i(\hat{y}_{(x_1,\ldots,x_{k+1}i)} - \hat{y}_{(x_1,\ldots,x_k i)})}{\sum_{i=1}^n i(y_{(i)} - \hat{y}_{(x_1,\ldots,x_k i)})}. \tag{19}$$

[2] $Var(\hat{Y}_{X_1,\ldots,X_k})$ denotes the Y variability "explained" by the $X_1,\ldots,X_k$ independent variables, whereas $Var(\hat{Y}_{X_1,\ldots,X_{k+1}})$ denotes the Y variability "explained" by the $X_1,\ldots,X_{k+1}$ independent variables.

The obtained partial Lorenz dependence measure, *RGI*, defines the possible partial contribution to the Y Lorenz zonoid, related to the addition of a new explanatory variable into the model. We now discuss about the statistical interpretation of our proposed dependence measures. We will consider an example that combines multiple linear regression with Lorenz zonoids theory. Suppose to consider data in Table 1.

Table 1 Data

Y	350	202	404	263	451	304	275	385	244	102	74	346	53	395	430
X_1	1	2	3	2	1	3	5	1	4	2	2	3	4	3	3
X_2	4	4	1	5	2	4	4	4	3	3	2	4	2	1	3

By resorting both to multiple linear regression model and Lorenz zonoid tools, one obtains $LZ_{d=1}(y) \simeq 0.245$, $LZ_{d=1}(\hat{y}_{x_1}) \simeq 0.051$, $LZ_{d=1}(\hat{y}_{x_2}) \simeq 0.008$ and $LZ_{d=1}$ $(\hat{y}_{x_1,x_2}) \simeq 0.054$. Through equation (15), the following marginal dependence measures can be derived: $MGC_{(y|x_1)} \simeq 0.209$ and $MGC_{(y|x_2)} \simeq 0.032$. The Lorenz zonoid of $\hat{Y}_{X_1}$ represents the 20.9% of the Y Lorenz zonoid and the Lorenz zonoid of $\hat{Y}_{X_2}$ represents the 3.2% of the Y Lorenz zonoid. The relative measure describing the possible additional contribution of covariate X_1 to the Y Lorenz zonoid is $RGI_{y,x_1|x_2} \simeq 0.194$, meaning that the introduction of covariate X_1 allows to increase the dilation of the $LZ_{d=1}(\hat{Y}_{X_2})$ in measure equivalent to 19.4%. Finally, the relative measure describing the possible additional contribution of covariate X_2 to the Y Lorenz zonoid is $RGI_{y,x_2|x_1} \simeq 0.013$, meaning that the introduction of covariate X_2 into the model allows to increase the dilation of the $LZ_{d=1}(\hat{Y}_{X_1})$ in measure equivalent to 1.3%. Thus, we can conclude that an increase of the dilation measure implies a reduction of the unexplained response variable variability.

4 Application to Crypto markets price

In this section we refer to an application on cryptocurrency prices data illustrated in a recent work of [5]. Similarly to [5], here we apply our proposal to assess if the the daily bitcoin prices in different crypto exchanges may be affected by the prices of classical assets.

Our data collect information on the daily bitcoin prices in eight different crypto exchanges from 18 May, 2016 to 30 April, 2016. Since the bitcoin price dynamics are very similar, for the sake of brevity we only focus on Coinbase Bitcoin and HitBtc Bitcoin, which represent the response variables of interest. The explanatory variables which are taken into account are Oil and Gold. We first compute the *MGC* coefficients for both the response variables. Through the *MGC* coefficients we can detect which covariate provides the greater contribution in explaining the bitcoin price variability. The covariate with the greatest contribution is included into the linear regression model. The contribution of the remaining explanatory variable is assessed in terms of the *RGI* index. Results on the Lorenz zonoid measure referred to Coinbase Bitcoin and HitBtc Bitcoin, together with the *MGC* coefficients, are reposted in Table 2.

Table 2 Lorenz zonoid and *MGC* coefficient values

Response variable	$LZ_{d=1}(\cdot)$	$MGC_{(\cdot\|Gold)}$	$MGC_{(\cdot\|Oil)}$
Coinbase Bitcoin	0.554	0.339	0.332
HitBtc Bitcoin	0.554	0.407	0.341

From Table 2 it arises that both the Coinbase Bitcoin and HitBtc Bitcoin prices present the same variability, measured by the corresponding Lorenz zonoids. We can conclude that both the prices do not suffer from strong daily differences. Variable Gold provides a contribution equal to the 33.9% and 40.7% for the Coinbase Bitcoin and HitBtc Bitcoin variables, respectively. The covariate with the smallest contribution is the Oil variable. Thus, variable Gold is the first variable to be introduced into the model. The contribution provided by the Oil variable is measured by the *RGI* index. Results are displayed in Tables 3 and 4 for the response variables Coinbase Bitcoin and HitBtc Bitcoin, respectively.

Table 3 Results in terms of *RGI* indices for response variable Coinbase Bitcoin

Covariate	Ordering of inclusion	*RGI* values
Oil	2	$RGI_{Coinbase\|Gold} = 0.339$
$Gold$	3	$RGI_{Coinbase,Oil\|Gold} = 0.238$

Table 4 Results in terms of *RGI* indices for response variable HitBtc Bitcoin

Covariate	Ordering of inclusion	*RGI* values
$Gold$	2	$RGI_{Coinbase\|Gold} = 0.407$
Oil	3	$RGI_{Coinbase,Oil\|Gold} = 0.259$

5 Conclusions

In this paper we showed how Lorenz zonoids can be usefully employed to assess the relative contribution associated with single independent variables included in linear models.

The employment of the Lorenz zonoid allows to compare different phenomena by ranking them in terms of their underlying the variability.

Our approach presents similarities with the R^2-based approach. Both methods are built on a quantitative response variable and are addressed in detecting the variables which mainly impact on the phenomenon of interest.

References

1. Agresti, A.: Categorical Data Analysis. Edited by John Wiley and Sons (2002)
2. Dall'Aglio, M., Scarsini, M.: Zonoids, Linear Dependence, and Size-Biased Distributions on the Simplex. Advances in Applied Probability, 35 (2003)
3. Giudici, P.: Applied Data Mining: Statistical Methods for Business and Industry. Wiley, Hoboken (2003)
4. Giudici, P., Raffinetti, E.: On the Gini measure decomposition. Statistics and Probability Letters, Vol. 81, Issue 1, 133-139 (2011)
5. Giudici, P., Abu-Hashish, I.: What determines bitcoin exchange prices? A network VAR approach. Finance Research Letters, Vol. 28, 309-318 (2019)
6. Koshevoy, G.: Multivariate Lorenz majorization. Social Choice and Welfare, 12 (1995)
7. Koshevoy, G., Mosler, K.: The Lorenz Zonoids of a Multivariate Distribution. Journal of the American Statistical Association, 91, No. 434 (1996)
8. Koshevoy, G., Mosler, K.: Multivariate Lorenz dominance based on zonoids. AStA Advance in Statistical Analysis, 91, No. 1, 57–76 (2007)
9. Lerman, R., Yitzhaki, S.: A note on the calculation and interpretation of the Gini index. Economics Letters, 15(3–4), 363–368 (1984)
10. Lorenz, M.O.: Methods of Measuring the Concentration of Wealth. Journal of the American Statistical Association, **9(70)**, 209–219 (1905)
11. Mosler, K.: Majorization in economic disparity measures. Linear Algebra and its applications, 220 (1994)
12. Muliere, P., Petrone, S.: Generalized Lorenz curve and monotone dependence orderings. Metron Vol L, No. 3-4 (1992)