Multivariate Ranks-Based Concordance Indexes

Emanuela Raffinetti and Paolo Giudici

Abstract The theoretical contributions to a "good" taxation have put the attention on the relations between the efficiency and the vertical equity without considering the "horizontal equity" notion: only recently, measures connected to equity (iniquity) of a taxation have been introduced in literature. The taxation problem is limited to the study of two quantitative characters: however the concordance problem can be extended in a more general context as we present in the following sections. In particular, the aim of this contribution consists in defining concordance indexes, as dependence measures, in a multivariate context. For this reason a k-variate ($k > 2$) concordance index is provided recurring to statistical tools such as ranks-based approach and multiple linear regression function. All the theoretical topics involved are shown through a practical example.

1 An Introduction to Concordance Index Problem

The issue of defining a concordance index often recurs in the statistical and economical literature. Although the 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 amount (see e.g. Musgrave 1959).

The analysis is focused on considering n pairs of ordered real values, (x_i, y_i), $i = 1, 2, \ldots, n$, whose components describe measures of two quantitative variables referred to each element of a statistical population: let us denote by 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 by $r(x_i)$)

E. Raffinetti (✉) · P. Giudici
University of Pavia, Via Strada Nuova 65, Italy
e-mail: emanuela.raffinetti@unipv.it; giudici@unipv.it

A. Di Ciaccio et al. (eds.), *Advanced Statistical Methods for the Analysis of Large Data-Sets*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-642-21037-2_42, © Springer-Verlag Berlin Heidelberg 2012

and to variable Y (denoted by $r(y_i)$). 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

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

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

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

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 these ones 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 quantitative variables.

Statistical literature provides a wide set of association indicators such as the Kendall-τ, the Spearmann-ρ and the Gini index: as well known these indexes assume values between -1 and $+1$ and, in particular one can show that they are equal to

$$-1 \Leftrightarrow (\forall i) \text{ if } r(y_i) = n + 1 - r(x_i) \tag{3}$$

$$+1 \Leftrightarrow (\forall i) \text{ if } r(x_i) = r(y_i). \tag{4}$$

These indexes, however, even if they are 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 considered extent character: a possible solution to this problem can be identified in resorting to the Lorenz curve and the dual Lorenz curve. In the taxation context the analysis is limited to the bivariate case (see e.g. Muliere 1986): in the following sections we consider the extension of this problem in a more general case when one considers more than two variables.

Furthermore, R^2 is not suited in this context as it looks for linear relationships.

2 Concordance Problem Analysis in a Multivariate Context

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 $(Y, X_1, X_2, \ldots, X_k)$.

The followed procedure is very simple and consists in applying a model able to describe the relation among the target variable Y and the explanatory variables $X_1, X_2, \ldots, X_k$: in order to define a concordance index in the hypothesis that the dependent variable Y is conditioned by more than one explanatory variable, one can recur to the multiple linear regression model (see e.g. Leti 1983). Thus the estimated variable Y values can be obtained recurring to the following relation

$$E(Y_i|X_1, X_2, \ldots, X_k) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} = \hat{y}_i; \qquad (5)$$

obviously the gap between the observed value y_i and the estimated value $\hat{y}_i$, where $i = 1, 2, \ldots, n$, on the basis of the regression represents the residual deviance of the model that has to be minimized (see e.g. Leti 1983).

2.1 Proposal: Multivariate Ranks-Based Approach

The starting point is based on 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 $y_{(i)}$ denotes the y_i ordered in an increasing sense and M_Y is the Y mean) 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_i ordered in a decreasing sense) (see e.g. Petrone and Muliere 1992). 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 to 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, by 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" is provided only if

$$\sum_{j=1}^{i} y_{(j)} = \sum_{j=1}^{i} y_j^* \text{ for every } i = 1, 2, \ldots, n, \qquad (6)$$

that is if and only if $r(y_i) = r(\hat{y}_i)$: obviously, it implies that if the residual deviance of the model decreases the concordance setting is attained due to the fact that the y_i preserve their original ordering also with respect to $r(\hat{y}_i)$.

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

$$\sum_{j=1}^{i} y_{(n+1-j)} = \sum_{j=1}^{i} y_j^* \text{ for every } i = 1, 2, \ldots, n. \tag{7}$$

Recalling the following inequalities

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

and

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

provided that $\sum_{j=1}^{i} y_{(j)} \leq \sum_{j=1}^{i} y_j^* \leq \sum_{j=1}^{i} y_{(n+1-j)}$ we have that $L_Y \leq C(Y|r(\hat{y}_i)) \leq L_Y^{'}$, as also shown in Fig. 1.

A multivariate concordance index can be then provided: its expression is the following

$$C_{Y,X_1,X_2,\ldots,X_k} = \frac{\sum_{i=1}^{n-1} \left\{ i/n - (1/(nM_Y)) \sum_{j=1}^{i} y_j^* \right\}}{\sum_{i=1}^{n-1} \left\{ i/n - (1/(nM_Y)) \sum_{j=1}^{i} y_{(j)} \right\}} : \tag{8}$$

this index represents the ratio of Y and $(Y|r(\hat{y}_i))$ concentration areas (Gini indexes): the concordance index enable to express the contribution of the k explanatory variables to the variable concentration. In particular the numerator of (8) describes

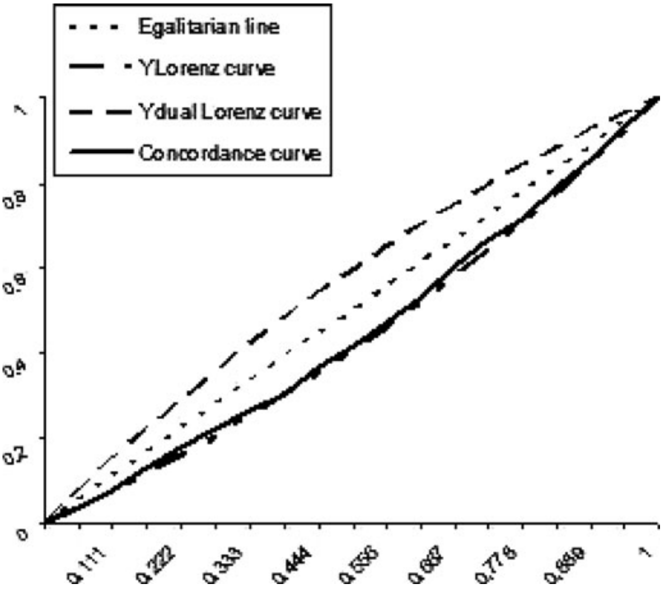

Fig. 1 Y Lorenz curve, Y dual Lorenz curve and concordance function

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 (8) 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.

Through some mathematical steps one can provide an alternative concordance index expression:

$$C_{Y,X_1,X_2,\ldots,X_k} = \frac{2\sum_{i=1}^n iy_i^* - n(n+1)M_Y}{2\sum_{i=1}^n iy_{(i)} - n(n+1)M_Y}. \tag{9}$$

Proof. Let's try to simplify (8) by operating both in the numerator and in the denominator in the same manner.

After multiplying both the numerator and the denominator for nM_Y, and by applying the products in (8) we get

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

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

$$C_{Y,X_1,X_2,\ldots,X_k} = \frac{n(n+1)M_Y - 2\sum_{i=1}^n\sum_{j=1}^i y_j^*}{n(n+1)M_Y - 2\sum_{i=1}^n\sum_{j=1}^i y_{(j)}}. \tag{10}$$

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

$$\sum_{i=1}^n\sum_{j=1}^i y_j^* = n(n+1)M_Y - \sum_{i=1}^n iy_i^* \tag{11}$$

which substituted in (10) gives directly

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

Now $\sum_{i=1}^n iy_i$ 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. Muliere 1986). We can conclude that:

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

Proof. 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 systems of equations of page 4: since

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

is intuitively true for all i, then 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 (11) we have

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

which gives $\sum_{i=1}^{n} i y_{(i)} \geq \sum_{i=1}^{n} i y_i^*$. □

Remark 2. $C_{Y,X_1,X_2,\ldots,X_{k-1}} = +1$ if and only if concordance function overlaps with the Lorenz curve.

Proof. Concordance function overlaps with the Lorenz curve if and only if $\sum_{j=1}^{i} y_{(j)} = \sum_{j=1}^{i} y_j^* \Rightarrow r(y_i) = r(y_i^*)$ for every $i = 1, 2, \ldots, n$. □

Remark 3. $C_{Y,X_1,X_2,\ldots,X_{k-1}} = -1$ if and only if concordance function overlaps with the dual Lorenz curve.

Proof. This remark can be proved, similarly to Remark 1, from the second system of page 4 by first 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 (8) we get an equivalent formulation of the concordance index based on $L_Y^{'}$:

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

Finally, since from the second system of equations of page 4 we have $\sum_{j=1}^{i} y_j^* \leq \sum_{j=1}^{i} y_{(n+1-i)}$, $\forall i$, then the result follows similarly to Remark 1 proof. □

An alternative concordance measure, which provides a measure of distance between concordance function and the Y Lorenz curve, is the Plotnick indicator (see e.g. Plotnick 1981) expressed by

$$I^*_{Y,X_1,X_2,\ldots,X_k} = \frac{\sum_{i=1}^n iy_{(i)} - \sum_{i=1}^n iy_i^*}{2\sum_{i=1}^n iy_{(i)} - (n+1)\sum_{i=1}^n y_{(i)}}. \tag{12}$$

Furthermore, one can verify that:

$$I^*_{Y,X_1,X_2,\ldots,X_k} = 0 \Leftrightarrow r(\hat{y}_i) = r(y_i) \Rightarrow \sum_{i=1}^n iy_{(i)} = \sum_{i=1}^n iy_i^*, \tag{13}$$

$$I^*_{Y,X_1,X_2,\ldots,X_k} = 1 \Leftrightarrow r(\hat{y}_i) = n+1-r(y_i) \Rightarrow \sum_{i=1}^n iy_i^* = \sum_{i=1}^n (n+1-i)y_{(i)}. \tag{14}$$

2.2 Some Practical Results

Suppose to have data concerning 18 business companies 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 1.

Table 1 Data describing Sales revenues, Selling price and Advertising investments expressed in Euros

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 Results

Ordered y_i	$r(y_i)$	$\widehat{y}_i$	Ordered $\widehat{y}_i$	$r(\widehat{y}_i)$	y_i ordered by $r(\widehat{y}_i)$
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	356.71	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

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 regression coefficients $\beta_0 \cong 98.48$, $\beta_1 \cong 0.63$, $\beta_2 \cong 4.57$ so the regression line is

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

Once getting the estimated Y values, we assign their ranks and finally order Y values according to $\hat{y}_i$ ranks. All the results are summarized in Table 2: through all these information we can compute concordance index in a multivariate context using (8) recalling that y_i^* represent the Y variable values ordered with respect to $\hat{y}_i$ ranks. Concordance index assumes 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 well clear in Fig. 1 where concordance curve (denoted with the continuous black line), is very close to Y variable Lorenz curve (denoted by the dash dot line). A further verification of this result is provided by the Plotnick indicator (12), whose numerical value is very close to 0, meaning that the distance between concordance function and Lorenz curve is minimum.

3 Conclusion

Through this analysis it has been proved that dependence study can be led in terms of concordance and discordance topics: the choice of a linear regression model is limited when one considers only quantitative variable. In the described context we referred to quantitative variables because we started from the source of the concordance problem involving the income amount before and after taxation intended as a quantitative character.

A future extension can regard the application of the concordance index analysis in cases when one of the considered variable is binary and the adopted model is a logistic regression.

Another important development is establishing if there exists a relation between the determination coefficient, intended as a dependence measure in a linear regression model, and the concordance index: our further research, focused on this topic, is in progress.

Acknowledgements The authors acknowledge financial support from the European grant EU-IP MUSING (contract number 027097). The paper is the result of the close collaboration among the authors, however, it has been written by Emanuela Raffinetti with the supervision of Prof. Paolo Giudici.

References

Leti, G.: Statistica descrittiva. Il Mulino (1983)

Muliere, P.: Alcune osservazioni sull'equità orizzontale di una tassazione. Scritti in onore di Francesco Brambilla. Ed. by Bocconi Comunicazione **2**, (Milano, 1986)

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

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

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