

Multivariate Ranks-based concordance indexes

Paolo Giudici

Emanuela Raffinetti

Univeristy of Pavia, Corso Strada Nuova 65, e-mail: emanuela.raffinetti@unipv.it

Abstract: 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 tool such as ranks-based approach and multiple linear regression function. All the theoretical topics involved are shown through a practical example.

Keywords: concordance and discordance indexes, Lorenz Curve, multiple linear regression model.

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.

Our analysis is focused on considering n pairs of ordered real values, (x_i, y_i) , $i = 1, 2, \dots, n$, 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 $r(x_i)$) 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

$$r(x_i) = r(y_i), \quad i = 1, 2, \dots, n. \quad (1)$$

Obviously, getting an equity index asks for the existence of an *ordering* among individuals before taxation and the knowledge of each individual income amount after taxation. The purpose is then identifying indexes able to stress potential *functional monotone relations* among the involved quantitative variables.

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

2. Ranks-based concordance index

The analysis is oriented to the achievement of a concordance index when considering a random vector $(Y, X_1, X_2, \dots, X_{k-1})$, by using a linear multiple regression model

$$E(Y_i | X_1, X_2, \dots, X_{k-1}) = b_0 + b_1 x_{1i} + b_2 x_{2i} + \dots + b_{(k-1)} x_{(k-1)i} = \hat{y}_i \quad (2)$$

where $b_0, b_1, b_2, \dots, b_{(k-1)}$ represent the ordinary least squares estimated regression coefficients.

The starting point is based on building the response variable Y Lorenz curve L_Y , (characterized by the set of ordered pairs $(i/n, S_i/nM_Y)$, where $S_i = \sum_{j=1}^i y_{(j)}$, $y_{(i)}$ denotes the y_i s ordered in an increasing sense and M_Y is Y mean) and the dual Y Lorenz curve L'_Y (characterized by the set of ordered pairs $(i/n, S'_i/nM_Y)$, where $S'_i = \sum_{j=1}^i y_{(n+1-j)}$ and $y_{(n+1-i)}$ denotes the y_i s ordered in a decreasing sense). The purpose is getting the estimated Y values, $\hat{y}_i$ s, in order to build the concordance function by ordering Y values, y_i s, with respect to $\hat{y}_i$ s ranks: let us denote this ordering by $(y_i | r(\hat{y}_i))$ and let us denote it by y_i^* . The set of pairs $(i/n, (1/nM_Y) \sum_{j=1}^i y_j^*)$ defines the concordance curve indicated by $C(Y | r(\hat{y}_i))$.

A multivariate concordance index can be provided by the following expression

$$C_{Y, X_1, X_2, \dots, X_{(k-1)}} = \frac{\sum_{i=1}^{n-1} \left\{ i/n - (1/nM_Y) \sum_{j=1}^i y_j^* \right\}}{\sum_{i=1}^{n-1} \left\{ i/n - (1/nM_Y) \sum_{j=1}^i y_{(j)} \right\}}. \quad (3)$$

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 $(k-1)$ explicative variables to the variable Y concentration.

Since $\sum_{i=1}^n i y_i$ is an arrangement increasing function we can conclude that:

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

Proof. It is sufficient to show that $C_{Y, X_1, X_2, \dots, X_{(k-1)}} \leq +1$ and $C_{Y, X_1, X_2, \dots, X_{(k-1)}} \geq -1$. The first inequality is justified by the following

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

And the second inequality is motivated by

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

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

Proof. Concordance function overlaps with Lorenz curve if and only if

$$\sum_{j=1}^i y_{(j)} = \sum_{j=1}^i y_j^*, \Rightarrow r(y_i) = r(y_i^*) \text{ for every } i = 1, \dots, n.$$

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

Proof. Concordance function overlaps with the dual Lorenz curve if and only if

$$\sum_{j=1}^n y_{(n+1-j)} = \sum_{j=1}^i y_j^*, \Rightarrow r(y_i^*) = n+1 - r(y_i) \text{ for every } i = 1, \dots, n.$$

An alternative concordance measure, which provides a measure of distance between concordance function and Y Lorenz curve, is the Plotnick indicator expressed by

$$I_{Y, X_1, X_2, \dots, X_{(k-1)}}^* = \frac{\sum_{i=1}^n i y_{(i)} - \sum_{i=1}^n i y_i^*}{2 \sum_{i=1}^n i y_{(i)} - (n+1) \sum_{i=1}^n y_{(i)}}. \quad (4)$$

3. A practical example

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): 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 $b_0 \cong 95.48, b_1 \cong 0.63, b_2 \cong 4.57$ so the regression line is

$$\hat{y}_i = 95.48 + 0.63x_{1i} + 4.57x_{2i}. \quad (5)$$

Once getting the estimated Y values, we assign their ranks and finally order Y values according to $\hat{y}_i$ s ranks. All the results are summarized in **Table 1**: through all these information we can compute concordance index in a multivariate context using (3) recalling that y_i^* s represent variable Y values ordered with respect to $\hat{y}_i$ s ranks. Concordance index assumes value 0.801 proving that there is a strong concordance relation among response variable Y and the explanatory variables X_1, X_2 : this conclusion is well clear in **Figure 1** where concordance function, is very close to Y Lorenz curve.

A further verification of this result is provided by the Plotnick indicator (4), whose numerical value is very close to 1, meaning that the distance between concordance curve $C(Y | r(\hat{y}_i))$ and Y Lorenz curve is minimum.

Ordered y_i s	$\hat{y}_i$ s	Ordered $\hat{y}_i$ s	$r(\hat{y}_i)$	y_i s ordered by $r(\hat{y}_i)$
202	231,30	231,30	1	202

216	291,66	234,30	2	253
244	270,70	245,80	3	304
253	234,30	251,79	4	275
263	282,99	270,70	5	244
274	310,72	282,99	6	263
275	251,79	291,66	7	216
302	310,73	308,36	8	385
304	245,80	310,72	9	274
308	373,60	310,73	10	302
346	363,36	357,05	11	350
350	357,05	361,31	12	430
374	381,32	363,36	13	346
385	308,36	373,60	14	308
395	413,79	381,01	15	404
404	381,01	381,32	16	374
430	361,31	411,42	17	451
451	411,42	413,79	18	395

Table 1: Results

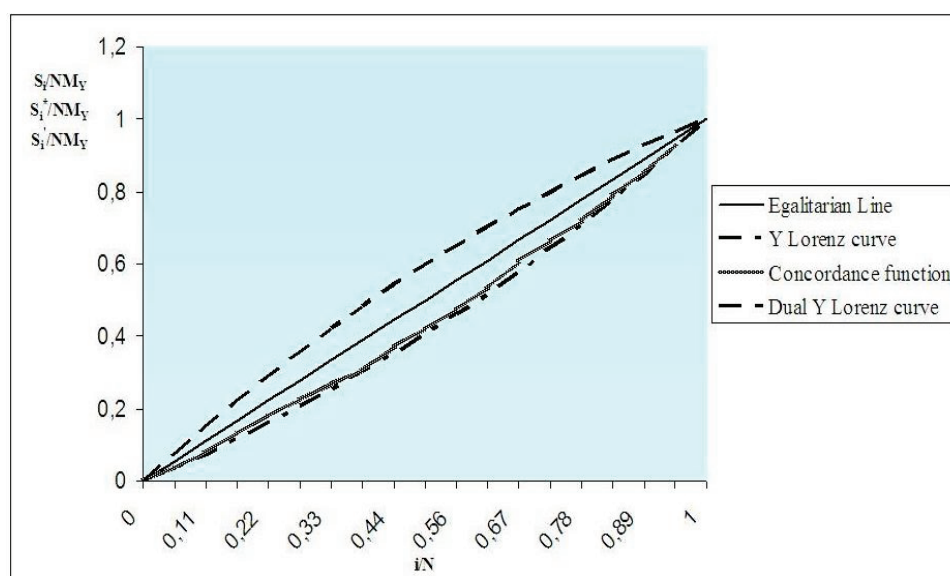


Figure 1: Y Lorenz curve, dual Y Lorenz curve and concordance function

References

- Leti G. (1983) Statistica descrittiva, Il Mulino, 901-909.
- Muliere P. (1986) Alcune osservazioni sull'equità orizzontale di una tassazione, *Edizioni di "Bocconi Comunicazione"*, Milano.
- Muliere P., Petrone S. (1992) Generalized Lorenz curve and monotone dependence orderings, *Metron*, L, 3-4.