Multivariate concordance measures: a proposal

Emanuela Raffinetti

Abstract We propose a novel multivariate measure of goodness of fit for a multiple linear regression model when the relevant involved explanatory variables assume mostly categorical nature. The proposed measure is based on the response variable Lorenz curve and its dual construction. Once the linear regression estimates are obtained, one can proceed by defining the concordance curve represented by the response variable original values ordered according to the ranks assigned to the corresponding estimated values. The concordance curve position explains the goodness of fit of the multiple linear regression model to the data. We also provide an extension of the Lorenz curve definition for ordinal variables in order to develop a further concordance index in this case.

Key words: concordance index, ranks, Lorenz curve for ordinal variables.

1 Background and current proposal

The issue of defining a concordance index, with the role of defining and measuring the existence of dependence relations among the analyzed involved variables, has been developed during the 80's with regard to the taxation problem in the bivariate context (see e.g. [5]). As well known, the statistical literature provides a wide set of association and concordance indicators, such as the Kendall-τ and the Spearman-ρ: however, these indices present a restriction. Even if invariant with respect to monotone transformations, they are unfortunately based on observations ranks and the same ranks can remain unchanged even if each individual income size has been substantially modified by the taxation process. In [5], in order to overcome the aforementioned restriction, the idea is based on resorting to the Lorenz curve, to its dual and to the concordance curve. In the last two years a contribution in terms

Emanuela Raffinetti
University of Pavia, Strada Nuova 65, Pavia (Italy), e-mail: emanuela.raffinetti@unipv.it

of concordance measures has been provided by [2] and [3]: the main interesting research topic described in [2] concerns the extension of the [5] concordance measure to the multivariate case. In particular, the purpose is based on defining a multivariate measure of goodness of fit when applying a multiple linear regression model and when the relevant involved explanatory variables assume mostly categorical nature and the response variable is quantitative. This aim can be reached through the Lorenz and concordance curves application. A relevant attention is devoted to the extension of the Lorenz curve in the case of ordinal variable: the definition of this topic represents the support for the further advances in the "ordinal" concordance analysis.

2 The multivariate concordance index based on Lorenz curves

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 resort to to the multiple linear regression model (see e.g. [4]). The "gap" between each observed value y_i and the corresponding estimated value $\hat{y}_i$ (where $i = 1,2,\ldots,n$) represents the contribution to the residual deviance of the model. The starting point of our proposal is to minimize rank differences between y and $\hat{y}$ rather than value differences (see [2]). To this aim we build the Lorenz curve of the variable Y, denoted with L_Y and characterized by the set of ordered pairs $(i/n,(1/(nM_Y))\sum_{j=1}^{i} y_{(j)})$, where $y_{(j)}$ represents the y_is ordered in an increasing sense and M_Y is the mean of Y. Furthermore one can build the so called dual Lorenz curve of the variable Y, denoted with $L_Y^{'}$ and characterized by the set of ordered pairs $(i/n,(1/(nM_Y))\sum_{j=1}^{i} y_{(n+1-j)})$, where $y_{(n+1-j)}$ represents the y_is ordered in a decreasing sense. Once the $\hat{y}_i$s are obtained, one can proceed with the construction of the concordance curve based on ordering the y_i values with respect to the ranks assigned to the corresponding $\hat{y}_i$s. Let us denote this ordering with $(y_i|r(\hat{y}_i))$, for each $i = 1,2,\ldots,n$, 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 that will be 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 if and only if $\sum_{j=1}^{i} y_{(j)} = \sum_{j=1}^{i} y_j^*$, for every $i = 1,2,\ldots,n$, that is, if and only if $r(y_i) = r(\hat{y}_i)$. On the other hand, 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 one to conclude that a perfect "overlap" occurs if and only if $\sum_{j=1}^{i} y_{(n+1-j)} = \sum_{j=1}^{i} y_j^*$ for every $i = 1,2,\ldots,n$. By proving that $L_Y \leq C(Y|r(\hat{y}_i)) \leq L_Y^{'}$, a multivariate concordance index, that will be named $C_{Y,X_1,X_2,\ldots,X_k}$, can thus be provided.

Its expression is the following:

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

where $-1 \leq C_{Y,X_1,X_2,\ldots,X_k} \leq +1$.

3 The Lorenz curve in the ordinal case

In Section 2 we have described the construction of a multivariate concordance index based on the employment of the Lorenz curves built on quantitative variables. Our aim now consists in building the Lorenz curve when the involved variables assume categorical nature: for this reason, a novel method is then needed.
Different definitions of inequality indices in the qualitative context have been developed in the economical literature: in [1] it has been deeply discussed about the problem related to the arbitrary choice of the scale explaining the different ordinal categories. The employment of different scales can lead to subjectivity and difficult interpretations of results. Our proposal can overcome the limits that arise with an arbitrary chosen scale by resorting to the ranks tool. Through the employment of ranks, one can obtain a more homogeneous interpretation of the measure given to different variable categories. If Y is an ordinal categorical variable, the value y_i does not correspond to a numerical value since it represents the category assumed by the ordinal categorical character: for this reason, our proposal is to substitute the ordinal categories with values able to represent their meaning. As well known, in the quantitative setting, the Lorenz curve construction satisfies the basic property of ordering in an increasing sense all the quantitative variable assumed values: our proposal, in the ordinal context, is to assign rank 1 to the smallest assumed category and value $(r_{k-1}+n_{k-1})$ to the largest one, where n_{k-1} corresponds to the absolute frequency associated to the $k-1$ ordinal considered category. Let us consider a categorical variable Y assuming k ordinal categories: the set of points characterizing the corresponding Lorenz curve is provided by the set of points

$$\left(\frac{\sum_{j=1}^{i} n_j}{\sum_{j=1}^{k} n_j}, \frac{\sum_{j=1}^{i} r_j n_j}{\sum_{j=1}^{k} r_j n_j}\right), \text{ where } i = 1,\ldots,k. \quad (2)$$

In particular r_j corresponds to the rank assigned to $j-th$ category: $r_1 = 1$ for the first ordinal category, $r_2 = (r_1 + n_1)$ for the second ordinal category and $r_k = (r_{k-1}+n_{k-1})$ for the last ordinal category. Let us denote the x-axis values with $F(r)$ and the y-axis values with $Q(r)$: in particular, we will call $F(r)$ as the *cumulative frequency percentage* and $Q(r)$ as the *cumulative rank percentage*. To illustrate our proposal we now introduce a simple example: suppose to consider 11 individuals and to collect information about their "Education degree" (Y) (see Table 1). The Lorenz curve graphical representation is provided in Fig. 1.

Table 1 Data

Y ="Education degree"	Absolute Frequency	Rank (r_j, $j = 1,\ldots,3$)
Secondary School Degree	$n_1 = 5$	1 ($r_1 = 1$)
High School Degree	$n_2 = 4$	6 ($r_2 = r_1 + n_1$)
University Degree	$n_3 = 2$	10 ($r_3 = r_2 + n_2$)

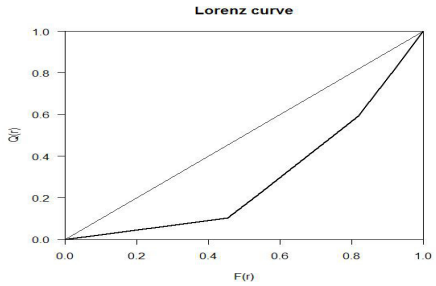

Fig. 1 The Lorenz curve in the ordinal context

4 Conclusions

A useful measure able to summarize the Lorenz curve is the Gini measure. When considering categorical ordinal values, the maximum homogeneity is obtained when all the statistical units are located in an unique category: in this context, the concentration area is null. This conclusion satisfies the normalization axiom: the Gini measure lower bound overlaps with the classical one since equivalent to 0. The maximum Gini measure value is not exactly equal to 1 but it can differ accordingly to each category absolute frequency.

Acknowledgements The author deeply wishes to acknowledge Prof. Paolo Giudici and Dr. Paola Cerchiello for their helpful suggestions.

References

1. Allison, R. A., Foster, J. E.: Measuring health inequality using qualitative data. Journal of Health Economics 23, pp 505-524 (2004)
2. Giudici, P., Raffinetti, E.: Multivariate Ranks-based concordance indexes. Volume of selected papers "Statistical Methods for the analysis of large data-sets". Springer-Verlag (2011)
3. Giudici, P., Raffinetti, E.: On the Gini measure decomposition. Statistics and Probability Letters, Vol. 81, Issue 1, 133–139 (2011)
4. Leti, G.: Descriptive Statistics (in Italian). Il Mulino (eds.) (1983)
5. Muliere, P.: Some remarks about the horizontal equity of a taxation (in Italian). In: Bocconi Comunicazione 2 (eds.), Milano (1986)