Measuring associations and evaluating forecasts of categorical and discrete variables

Jochem Huismans (University of Amsterdam)

Jan Willem Nijenhuis (University of Twente)

Andrei Sirchenko (Maastricht University)

June 10, 2022

To express anything important in mere figures is so plainly impossible that there must be endless scope for well-paid advice on how to do it.

— K. A. C. Manderville, The Undoing of Lamia Gurdleneck

Input:

- (i) the values of two categorical (or discrete) variables or
- (ii) the observed values of a categorical (discrete) variable and the predicted probabilities of each category

Stata command classify

Measuring associations and evaluating forecasts

Output:

- contingency table
- general measures of overall association and correlation (and also diagnostic scores of the accuracy of probabilistic forecast)
- class-specific measures for each class as well as their simple and weighted averages

Literature on measures of association

is poorly integrated across different fields

- a wide variety of scalar statistics have been developed and used in different fields
- a similarly wide variety of nomenclature has appeared in relation to these statistics
- some of these measures have been reinvented, duplicated and renamed on multiple occasions in other fields
- confusing terminology is confounded further by different notation

Literature on measures of association

is poorly integrated across different fields

- Cohen kappa coefficient (1960)
- Heidke skill score (1926)
- Doolittle association ratio (1887)
- Galton coefficient (1892)
- Hubert–Arabie adjusted Rand index (Hubert and Arabie 1985)

Diagnostic scores for probabilistic forecasts

- Brier score (half-Brier score, quadratic score, probability score)
- Power score
- Logarithmic score (ignorance score)
- Spherical score
- Pseudospherical score
- Zero-one score
- Ranked probability score (suitable only for ordinal outcomes)

Diagnostic scores for probabilistic forecasts

- Spherical score (Winkler 1967; Winkler and Murphy 1968; Friedman 1983; $[0 \leftarrow 1]$):

$$1 - \frac{1}{n} \sum_{i=1}^{n} \frac{\sum_{k=1}^{K} \delta_{ik} \Pr(y_i = k)}{\sqrt{\sum_{k=1}^{K} [\Pr(y_i = k)]^2}}$$

- Ranked probability score (suitable only for ordinal outcomes; identical to the Brier score for binary outcomes (Epstein 1969; Murphy 1971); $[0 \leftarrow 1]$):

$$\frac{1}{n(K-1)} \sum_{i=1}^{n} \sum_{k=1}^{K-1} \left(\sum_{j=1}^{k} \Pr(y_i = j) - \sum_{j=1}^{k} \delta_{ij} \right)^2$$

	$x=1$	$x=2$	...	$x=K$	Total
$y=1$	n_{11}	n_{12}	...	n_{1K}	$n_{1+}=\sum n_{1j}$
$y=2$	n_{21}	n_{22}	...	n_{2K}	$n_{2+}=\sum n_{2j}$
...	...	...	...	...	...
$y=K$	n_{K1}	n_{K2}	...	n_{KK}	$n_{K+}=\sum n_{Kj}$
Total	$n_{+1}=\sum n_{i1}$	$n_{+2}=\sum n_{i2}$	...	$n_{+K}=\sum n_{iK}$	$n=\sum\sum n_{ij}$

Measures of association

General and class-specific measures

- General measures of the overall performance — they include explicitly all concordant (matched) pairs n_{kk}

- Class-specific measures computed for each class k — they include explicitly only one matched pair n_{kk}.

Measures of association

Symmetric measures: two types of symmetry

- A measure is transpose symmetric if it treats both variables equivalently, and so it is invariant to relabelling of them — it remains unchanged if the row variable and column variable are interchanged (if any n_{ij} and $n_{ji}, i \neq j$ are swapped).
- A measure is complement symmetric if it treats all categories equivalently, and so it is invariant to relabelling of them — it remain unchanged if any two columns and the corresponding two rows are swapped).

Measures of association

Asymmetric measures

- Asymmetric measures have typically been developed in the binary context for rare and/or extreme events.
- So the occurrences get larger weights in their definitions than the nonoccurrences.
- The true positives and true negatives are not treated equally, and the false positives and false negatives are not treated equally either: the negative matches do not mean necessarily any similarity between two objects, and type 2 errors are often more serious than type 1 errors.
- To measure association between two variables x and y, it does matter which variable is x and which is y.

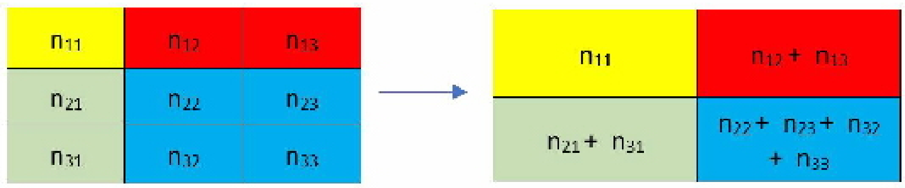
- Goodman-Kruskal λ_r coefficient (adjusted count R^2, Brennan and Prediger κ_b, Appleman (Goodman and Kruskal 1954; Brennan and Prediger 1981); $[0 \longrightarrow 1]$):

$$\frac{\sum_{k=1}^{K} n_{kk} - \max_{j=1}^{K} n_{+j}}{n - \max_{j=1}^{K} n_{+j}}$$

- Goodman-Kruskal symmetrical λ_r coefficient (Goodman and Kruskal 1954; $[-1 \longrightarrow 1]$):

$$\frac{2\sum_{k=1}^{K} n_{kk} - \max_{i=1}^{K} n_{i+} - \max_{j=1}^{K} n_{+j}}{2n - \max_{i=1}^{K} n_{i+} - \max_{j=1}^{K} n_{+j}}$$

- The class-specific measures include only one concordant pair n_{kk}, and designed for binary outcomes, mostly for a positive category.
- To compute them, any $K \times K$ contingency table can be converted (using a so-called one-vs-all binarisation strategy) to a series of K 2×2 contingency tables:



Measures of association

Class-specific contingency table

	$x = k$	$x \neq k$	Total
$y = k$	$n_{11}^{(k)} = n_{kk}$	$n_{12}^{(k)} = n_{k+} - n_{kk}$	$n_{1+}^{(k)} = n_{k+}$
$y \neq k$	$n_{21}^{(k)} = n_{+k} - n_{kk}$	$n_{22}^{(k)} = n - n_{k+} - n_{+k} + n_{kk}$	$n_{2+}^{(k)} = n - n_{k+}$
Total	$n_{+1}^{(k)} = n_{+k}$	$n_{+2}^{(k)} = n - n_{+k}$	n

- The `classify` command also computes the simple arithmetic and weighted arithmetic averages of all class-specific measures as:

$$
\begin{aligned}
Measure_{macro} &= \frac{1}{K}\sum_{k=1}^{K} Measure_k \\
Measure_{weighted} &= \sum_{k=1}^{K} Measure_k \frac{n_{+k}}{n}
\end{aligned}
$$

- The macro-averaged measures calculate unweighted (arithmetic) mean of class-specific coefficients.
- The weighted-averaged measures take a weighted mean. The weights for each class are the total number of observations of that class.

Measures of association

Class-specific measures

- Precision (positive predictive value, confidence, success ratio, post agreement, frequency of hits, correct alarm ratio (Grossmann 1898 cited in Muller 1944; Dice 1945; Wallace 1983); for negative category: negative predicted value, inverse precision, true negative accuracy; $R : [0 \longrightarrow 1]$):

$$\frac{n_{11_{(k)}}}{n_{1+_{(k)}}}$$

- F_1-score (harmonic mean of precision and recall, percent positive agreement, Gleason, Sørensen–Dice, Sørensen, Dice, Czekanowski, Nei-Li, Bray-Curtis, Upholt F, Burt, Lance–Williams, Pirlot, Tversky, Gower-Legendre T (Czekanowski 1913, 1932; Gleason 1920; Dice 1945; Sørensen 1948; Bray 1956; Bray and Curtis 1957; Lance and Williams 1966; Upholt 1977; Tversky 1977; Nei and Li 1979); $R : [0 \longrightarrow 1]$):

$$\frac{2n_{11_{(k)}}}{n_{1+_{(k)}} + n_{+1_{(k)}}}$$

Classify and be happy

" . . . there is no absolutely general measure of the degree of dependence. Every attempt to measure a conception like this by a single number must necessarily contain a certain amount of arbitrariness and suffer from certain inconveniences."

— *Cramér (1924)*