

300 measures of association, similarity, information and forecast evaluation for categorical data: the `catmetrics` command

Andrei Sirchenko (Nyenrode Business University)
Dragoş Bînzari (University of Amsterdam)
Konrad Wrebiak (University of Amsterdam)

September 17, 2026

A new Stata command catmetrics

Measuring associations and evaluating forecasts of categorical data

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*

Measuring associations and similarities

Measures based on binary contingency tables

	Male	Female	Total
Blonde	$n_{11} = 8$	$n_{12} = 16$	$n_{1+} = 24$
Brunette	$n_{21} = 14$	$n_{22} = 18$	$n_{2+} = 32$
Total	$n_{+1} = 22$	$n_{+2} = 34$	$n = 56$

- Pearson / Yule correlation φ : $\frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{1+}n_{+1}n_{+2}n_{2+}}} = -0.106$
- Gilbert index: $\frac{n_{11}}{n - n_{22}} = 0.211$
- Diagnostic odds ratio: $\frac{n_{11}n_{22}}{n_{12}n_{21}} = 0.643$
- Relative risk ratio: $\frac{n_{11}n_{2+}}{n_{21}n_{1+}} = 0.762$

Evaluating categorical forecasts

Measures based on confusion matrices

		Actual values	
		Positive	Negative
Predicted values	Positive	True positive (TP)	False positive (FP)
	Negative	False negative (FN)	True negative (TN)

- Accuracy: $\frac{TP+TN}{n} = 0.464$
- Peirce skill score: $\frac{TP \times TN - FP \times FN}{(TP+FN)(FP+TN)} = -0.107$
- F_1 -score: $\frac{2TP}{2TP+FP+FN} = 0.348$
- Heidke skill score: $\frac{2(n_{11}n_{22} - n_{12}n_{21})}{n_{+1}n_{2+} + n_{1+}n_{+2}} = -0.105$

Measures based on K by K tables

	$x = 1$	...	$x = K$	Total
$y = 1$	n_{11}	...	n_{1K}	$n_{1+} = \sum_{j=1}^K n_{1j}$
...	...	...	...	...
$y = K$	n_{K1}	...	n_{KK}	$n_{K+} = \sum_{j=1}^K n_{Kj}$
Total	$n_{+1} = \sum_{i=1}^K n_{i1}$	...	$n_{+K} = \sum_{i=1}^K n_{iK}$	$n = \sum_{i=1}^K \sum_{j=1}^K n_{ij}$

- Accuracy: $\frac{1}{n} \sum_{k=1}^K n_{kk}$
- Cohen κ coefficient: $\frac{n \sum_{k=1}^K n_{kk} - \sum_{k=1}^K n_{k+} n_{+k}}{n^2 - \sum_{k=1}^K n_{k+} n_{+k}}$
- Stuart τ_c coefficient:

$$\frac{2K \left[\sum_{i=1}^K \sum_{j=1}^K n_{ij} \left(\sum_{h>i} \sum_{k>j} n_{hk} \right) - \sum_{i=1}^K \sum_{j=1}^K n_{ij} \left(\sum_{h>i} \sum_{k<j} n_{hk} \right) \right]}{n^2 (K-1)}$$

Class-specific binary measures

- Many binary measures do not treat the positive and negative categories symmetrically.

$$G\text{-measure} = \frac{n_{11}}{\sqrt{n_{1+}n_{+1}}} = \sqrt{\frac{n_{11}}{n_{1+}} \frac{n_{11}}{n_{+1}}}$$

- The G-measure ignores true negatives (n_{22}) and evaluates performance on the positive class.
- In some applications, however, the negative class is substantively important rather than merely a background category.

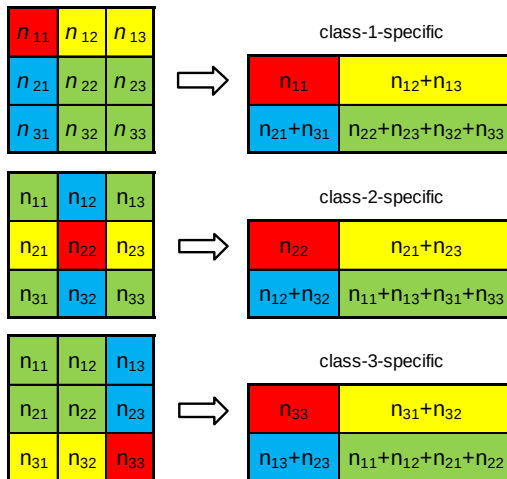
$$\text{Negative-class } G\text{-measure} = \frac{n_{22}}{\sqrt{n_{2+}n_{+2}}} = \sqrt{\frac{n_{22}}{n_{2+}} \frac{n_{22}}{n_{+2}}}$$

Class-specific evaluation for multiclass data

- For each class k , `catmetrics` constructs a one-versus-all binarized table and computes the class-specific metrics based on each of the 225 binary measures.

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

Conversion of a 3x3 classification table into three class-specific 2x2 tables



Class-specific evaluation for multiclass data

	x=1	x=2	x=3	Total
y=1	30	9	0	39
y=2	25	163	21	209
y=3	0	9	19	28
Total	55	181	40	276

- Overall accuracy is $\frac{1}{n} \sum_{k=1}^K n_{kk} = 0.77$
- What proportion of actual observations in each class is correctly predicted?

Class-specific evaluation for multiclass data

	Class-1-specific			Class-2-specific			Class-3-specific	
	x = 1	x not 1		x = 2	x not 2		x = 3	x not 3
y = 1	30	9	y = 2	163	46	y = 3	19	9
y not 1	25	212	y not 2	18	49	y not 3	21	227

- The class-specific hit rates for classes 1, 2 and 3 are:

$$\frac{30}{30+25} = 0.55, \frac{163}{163+18} = 0.90, \text{ and } \frac{19}{19+21} = 0.47$$

- Although class 2 has the highest hit rate, it also has the worst adjusted noise-to-signal ratio (Kaminsky and Reinhart 1999)
- The adjusted noise-to-signal ratio for a binary table is $\frac{n_{12}n_{+1}}{n_{+2}n_{11}}$
- The adjusted noise-to-signal ratios are 0.07, 0.54, and 0.08

Class-specific measures

Arithmetic and weighted averages

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

$$Measure_{macro} = \frac{1}{K} \sum_{k=1}^K Measure_k$$

$$Measure_{weighted} = \sum_{k=1}^K Measure_k \frac{n_k}{n}$$

- The macro-averaged measure calculates unweighted mean of class-specific coefficients
- The weighted-averaged measure makes a weighted mean

Evaluating probabilistic forecasts of categorical data

Strictly proper scoring rules

- Brier score (Brier 1950):

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

- Power score (Good 1971):

$$\frac{1}{n} \sum_{i=1}^n \left[\frac{1}{\beta} - \sum_{k=1}^K \delta_{ik} \Pr(y_i = k)^{\beta-1} + \frac{\beta-1}{\beta} \sum_{k=1}^K [\Pr(y_i = k)]^{\beta} \right]$$

- Ranked probability score (Epstein 1969):

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

Evaluating probabilistic forecasts of categorical data

Diagnostic measures

- The two-alternative forced choice score (Mason and Weigel 2009):

$$\frac{\sum_{l=1}^K \sum_{k \neq l}^K \sum_{i=1}^{n_{+k}} \sum_{j=1}^{n_{+l}} I[p_{k,i}(l), p_{l,j}(l)]}{\sum_{l=1}^K \sum_{k \neq l}^K n_{+k} n_{+l}}$$

where $p_{k,i}(l)$ is the forecast probability assigned to category l for observation i whose observed category is k , and

$$I[p_{k,i}(l), p_{l,j}(l)] = \begin{cases} 0 & \text{if } p_{l,j}(l) < p_{k,i}(l) \\ 0.5 & \text{if } p_{l,j}(l) = p_{k,i}(l) \\ 1 & \text{if } p_{l,j}(l) > p_{k,i}(l) \end{cases} .$$

Literature

is poorly integrated across different fields

- A wide variety of scalar statistics have been developed and used in different fields, typically under different names
- Algebraically identical measures have been independently rediscovered and renamed without recognition that they were already established elsewhere, with little or no cross-referencing
- Confusing terminology is confounded further by different labels and notations

Reinvention and duplication

is common when a measure addresses a universal methodological problem

- Cohen kappa coefficient (1960) is a chance-corrected agreement measure in biomedical, psychological and educational

studies:
$$\frac{n \sum_{k=1}^K n_{kk} - \sum_{k=1}^K n_{k+} n_{+k}}{n^2 - \sum_{k=1}^K n_{k+} n_{+k}}$$

Reinvention and duplication

is common when a measure addresses a universal methodological problem

- Cohen kappa coefficient (1960) is a chance-corrected agreement measure in biomedical, psychological and educational

studies:
$$\frac{n \sum_{k=1}^K n_{kk} - \sum_{k=1}^K n_{k+} n_{+k}}{n^2 - \sum_{k=1}^K n_{k+} n_{+k}}$$

- Heidke skill score (1926) in meteorological forecast

verification:
$$\frac{2(n_{11}n_{22} - n_{12}n_{21})}{n_{+1}n_{2+} + n_{1+}n_{+2}}$$

Reinvention and duplication

is common when a measure addresses a universal methodological problem

- Cohen kappa coefficient (1960) is a chance-corrected agreement measure in biomedical, psychological and educational studies:
$$\frac{n \sum_{k=1}^K n_{kk} - \sum_{k=1}^K n_{k+} n_{+k}}{n^2 - \sum_{k=1}^K n_{k+} n_{+k}}$$
- Heidke skill score (1926) in meteorological forecast verification:
$$\frac{2(n_{11}n_{22} - n_{12}n_{21})}{n_{+1}n_{2+} + n_{1+}n_{+2}}$$
- Doolittle association ratio (1887) in the context of a dispute over tornado forecasting

Reinvention and duplication

"No scientific discovery is named after its original discoverer" (Stigler 1980)

$$\frac{2n_{11}}{n_{1+} + n_{+1}}$$

Year	Name / attribution	Field
1909	Czekanowski	Anthropology, taxonomy, numerical classification
1920	Gleason	Plant ecology, quadrat comparison
1945	Dice	Ecology, species association
1948	Sørensen	Plant sociology, ecological similarity
1948	Burt	Psychometrics / classification catalogues
1956	Bray-Curtis	Ecology, ordination, community ecology
1966	Lance-Williams	Numerical taxonomy, cluster analysis
1974	F₁-score	Information retrieval; machine learning.
1977	Upholt F	Molecular genetics, restriction-fragment comparison
1979	Nei-Li coefficient	Molecular evolution, genetic similarity
1980s	Piriot/White	Binary-similarity catalogues, classification literature
1994	Zijdenbos	Medical image segmentation, neuroimaging
2000s	Fleiss-Levin-Paik	Inter-rater reliability, agreement analysis

Alternative terminology

$$\frac{1}{n} \sum_{k=1}^K n_{kk}$$

- Accuracy
- (Crude) agreement
- Causal support
- Classification rate
- Count R^2
- Hit score
- Holsti $C.R.$ coefficient of agreement
- Kendall coefficient
- Osgood coefficient
- Proportion correct
- Rand coefficient
- Ratio test discriminant
- Simple matching coefficient
- Sokal-Michener coefficient
- Success rate

A new Stata command catmetrics

- Computes 301 measures, together with their aliases, attainable ranges, neutral values, and directions of improvement
- Reports 225 binary-only measures, 63 multicategory nominal or ordinal measures, and 13 multicategory ordinal-only measures
- Reports 6 strictly proper scoring rules and 12 diagnostic measures for probabilistic forecasts
- Returns also class-specific quantities and their macro and weighted averages

Output of catmetrics command

Input: a pre-existing classification table

```
. catmetrics, mat(C)
```

Contingency Table

Actual	1	2	3	Total
Predicted				
1	11	2	3	16
2	3	16	6	25
3	7	5	22	34
Total	21	23	31	75

Selected measures of association & hard forecast evaluation

Measure	Range	Value
1. Accuracy	[0 → 1]	0.6533
9. Balanced accuracy	[0 → 0.5 → 1]	0.6430
42. Clayton skill score	[-1 → 0 → 1]	0.4766
44. Cohen kappa coefficient	[-1 → 0 → 1]	0.4672
52. Cramer concordance coefficient	[0 → 1]	0.2289
132. Goodman-Kruskal lambda coefficient	[0 → 1]	0.3659
134. Goodman-Kruskal tau coefficient	[0 → 1]	0.2262
138. Gorodkin Rk coefficient	[-1 → 0 → 1]	0.4697
163. Krippendorff alpha coefficient	[-.99 → 1]	0.4694
205. Peirce skill score	[-1 → 0 → 1]	0.4629
234. Scott pi coefficient	[-1 → 1]	0.4658
241. Sokal-Sneath similarity coefficient #3	[0 → .5 → 1]	0.6506
256. Theil uncertainty coefficient U	[0 → 1]	0.1987
260. Tschuprow T coefficient	[0 → 1]	0.4784

See the Excel file 'Catmetrics.xlsx' for the complete output.

Output of catmetrics command

Input: two observed categorical variables

```
. webuse lbw, clear  
(Hosmer & Lemeshow data)  
  
. catmetrics smoke low, metrics(54 196 217 256 266 280 281 282) talpha(0.2) tbeta(0.2)
```

Contingency Table

smoke=	Positive (0)	Negative (1)	Total
low=			
Positive (0)	86	44	130
Negative (1)	29	30	59
Total	115	74	189

Selected measures of association & hard forecast evaluation

Measure	Range	Value
54. Czekanowski index	[0 → 1]	0.7020
196. Odds ratio	[0 → 1 → 8742.25]	2.0219
217. Relative risk ratio	[0 → 1 → 188]	1.3459
256. Theil uncertainty coefficient U	[0 → 1]	0.0207
266. Tversky index (alpha=.2,beta=.2)	[0 → 1]	0.8549
280. Yule phi correlation coefficient	[-1 → 0 → 1]	0.1614
281. Yule Q coefficient	[-1 → 0 → 1]	0.3382
282. Yule Y coefficient	[-1 → 0 → 1]	0.1742

See the Excel file 'Catmetrics.xlsx' for the complete output.

Output of catmetrics command

Input: an observed variable and predicted probabilities

```
. usehsy sysdsn1, clear
(Health insurance data)

. quietly mlogit insure age male nonwhite i.site

. predict p*
(option pr assumed; predicted probabilities)
(1 missing value generated)

. catmetrics insure, probs(p1 p2 p3) classmetrics nonecel ///
> metrics(p1 p2 p6 p16 p15 p16 p17 1 9.44 54 120 138 148 205)
WARNING: 20 observation(s) with missing data were removed.
(Indices removed: 5 8 22 41 63 110 144 177 189 216 225 241 242 255 268 301 308 338 343 412 487 494 513 521 537 566 614 620 636)
```

Confusion Matrix

Actual \ Predicted	1	2	3	Total
1	203	132	28	363
2	90	145	17	252
3	0	0	0	0
Total	293	277	45	615

Selected proper scores for probabilistic forecasts

Score	Range	Value
P1 Brier score	[0 + 1]	0.2699
P2 Logarithmic score	[0 + ∞]	1.5835
P6 Spherical score	[0 + 1]	0.5225

Selected diagnostic measures for probabilistic forecasts

Score	Range	Value	Class (1)	Class (2)	Class (3)	Macro avg	Wgted avg
P14 Mean probability rate	[0 + 1]	0.460					
P15 Normalized area under CAP	[0 + 1]		0.236	0.260	0.300	0.266	0.252
P16 Normalized area under ROC	[0 + 1]		0.618	0.630	0.650	0.633	0.626
P17 ZARC score #1	[0 + 1]	0.627					

Selected measures of association & hard forecast evaluation

Measure	Range	Value	Class (1)	Class (2)	Class (3)	Macro avg	Wgted avg
1. Accuracy	[0 + 1]	0.566					
9. Balanced accuracy	[0 + 0.5 + 1]	0.405					
44. Cohen kappa coefficient	[-1 + 0 + 1]	0.187					
54. Cramér's V index	[0 + 1]		0.619	0.548	0.000	0.389	0.542
120. Gilbert skill score	[-.33 + 0 + 1]	0.107	0.107	0.116	0.000	0.075	0.104
138. Goodwin RK coefficient	[-1 + 0 + 1]	0.191					
148. Hit rate	[0 + 1]	0.693	0.693	0.523	0.000	0.405	0.566
205. Peirce skill score	[-1 + 0 + 1]	0.177					

Output of catmetrics command

Input: an observed variable & probabilities from the estimated model

```
. webuse tobacco, clear  
(Fictional tobacco consumption data)  
  
. quietly zioprobit tobacco education income i.female age, ///  
> inflate(education income i.parent age i.female i.religion)  
  
. catmetrics tobacco, pred metrics(ordinal)
```

Confusion Matrix

Actual \ Predicted	0	1	2	3	Total
0	8806	783	458	586	9833
1	1166	2941	162	0	4269
2	237	82	416	36	771
3	60	0	14	53	127
Total	9469	3806	1050	675	15000

Selected proper scores for probabilistic forecasts

Score	Range	Value
P5 Ranked probability score [ORD]	[0 + 1]	0.0814

ORD: suited for ordinal data only.

Selected diagnostic measures for probabilistic forecasts

Score	Range	Value
P18 2AFC score #2 [ORD]	[0 + 1]	0.8167

ORD: suited for ordinal data only.

Selected measures of association & hard forecast evaluation

Measure	Range	Value
118. Gervity skill score [ORD]	[-1 + 0 + 1]	0.2869
131. Goodman-Kruskal gamma coefficient [ORD]	[-1 + 0 + 1]	0.6397
156. Kendall tau-a coefficient [ORD]	[-.75 + 0 + .75]	0.2255
157. Kendall tau-b coefficient [ORD]	[-1 + 0 + 1]	0.4440
162. Koppen index [ORD]	[0 + 1]	0.8358
180. Mantel-Haenszel statistic [ORD]	[0 + 14999]	1.7e+03
209. Pollaczek-Geiringer coefficient [ORD]	[-1 + 0 + 1]	0.5960
245. Somers d statistics [ORD]	[-1 + 0 + 1]	0.4253
252. Stuart tau-c correlation [ORD]	[-1 + 0 + 1]	0.3007
258. Tonnes coefficient [ORD]	[-2 + 2]	1.2829
268. 2AFC statistic #2 [ORD]	[0 + .5 + 1]	0.7127

ORD: suited for ordinal data only

See the Excel file 'Catmetrics.xlsx' for the complete output.

Concluding remarks

No single measure is uniformly optimal

" . . . 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)

- Different measures emphasize different aspects of the joint distribution
- The choice of one measure over another implicitly determines which information is prioritized and which is ignored