A test for variable importance

Sulla scelta di variabili esplicative rilevanti

Rosaria Simone

Abstract Statistical literature is being more and more concerned with debates about hypothesis testing and p-values supporting the significance of a given variable specification. Specifically, if on one hand statistical foundations about significance are not arguable, scholars should be able to distinguish between significance and variable importance. This is a matter of serious concern in questionnaire analysis to derive respondents' profiles and develop targeted marketing strategies, for instance. To this aim, this contribution proposes a hypothesis system that considers the normalized dissimilarity measure to assess the importance of explanatory variables in the setting of mixture models for ordinal data to account for uncertainty of choice.
Abstract *Un dibattito sempre più presente nella letteratura statistica moderna riguarda lo studio della effettiva importanza di covariate significative rispetto ai risultati dei classici test di ipotesi e tecniche di selezione del modello. Se l'applicazione delle procedure standard non è in discussione, d'altra parte la distinzione tra variabili significative e variabili rilevanti assume un ruolo fondamentale ai fini decisionali. Tale problematica è di particolare rilievo nell'dei dati provenienti da questionari, ad esempio per la profilazione dei consumatori al fine di individuare specifiche strategie di marketing. In questo contesto, un sistema di ipotesi basato su una misura di dissimilarità viene proposto per testare la rilevanza di variabili esplicative nel caso di una classe di misture per dati ordinali.*

Key words: Dissimilarity, Variable Selection and Importance, CUB mixture models, Rating Data

Rosaria Simone
Department of Political Sciences, University of Naples Federico II, Via L. Rodinò, 22, 80138 Naples, Italy e-mail: rosaria.simone@unina.it

1 Motivations

The opening lines of [3] invite readers to critically use p-values in the era of big data:

> There is growing frustration with the concept of the p-value. Besides having an ambiguous interpretation, the p-value can be made as small as desired by increasing the sample size, n. The p-value is outdated and does not make sense with big data: Everything becomes statistically significant.

The purpose of this contribution is to investigate such concern and propose a methodology for variable selection in the setting of statistical models for rating data. Our discussion stems from a well consolidated idea to measure separation of probability distributions relying on the concept of Gini's *Transvariation* [5] which has been applied in several circumstances (see [1], for instance). Departing from the identity: $\min(a,b) = \frac{1}{2}\big(a+b-|a-b|\big)$, an inverse indicator of how far apart two (discrete) probability distributions $\mathbf{p} = (p_1,\dots,p_m)', \mathbf{q} = (q_1,\dots,q_m)'$ are, is given by:

$$\sum_{r=1}^{m} \min(p_r,q_r) = 1 - \frac{1}{2}\sum_{r=1}^{m} |p_r - q_r|. \qquad (1)$$

This is a measure of their overlapping. For instance, consider the conditional distributions of a discrete response given a dichotomous factor (dashed lines joining mass probabilities are chosen to enhance visualization in Figure 1). Albeit statistically significant differences are found, the discrimination of response patterns becomes more meaningful from left to right as the overlapping gets smaller.

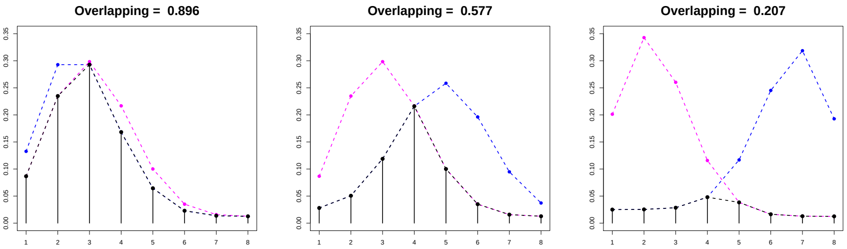

Fig. 1 Visualization of overlapping between discrete probability distributions

As an application of how this measure could work in discriminating significant covariates for a given ordinal response, here the focus will be on CUB models [2]. The original paradigm is the weighted combination of a (shifted) Binomial distribution $g_r(\xi_i)$ for the *feeling* component and a (discrete) Uniform for the *uncertainty* component, meant as the fuzziness derived from the discretization of the continuous latent perception. For a sample $(R_1,\dots,R_n)$ of ordinal responses, say on the support $\{1,\dots,m\}$ for a given $m > 3$, a CUB regression model is specified via:

$$Pr(R_i = r|\boldsymbol{y}_i, \boldsymbol{w}_i) = \pi_i\, g_r(\xi_i) + (1-\pi_i)\frac{1}{m}, \qquad (2)$$

where feeling ξ_i and uncertainty π_i parameters are linked to values of subjects' covariates $\boldsymbol{w}_i, \boldsymbol{y}_i$ by a logit link:

$$logit(\xi_i) = \boldsymbol{w}_i\,\boldsymbol{\gamma}, \qquad logit(\pi_i) = \boldsymbol{y}_i\,\boldsymbol{\beta}.$$

This full model specification is customarily abbreviated as CUB (p,q), where $\boldsymbol{\gamma}^{'} = (\gamma_0, \dots, \gamma_q)^{'}, \boldsymbol{\beta}^{'} = (\beta_0, \dots, \beta_p)^{'}$ are the estimable parameters: when $p = q = 0$, then model fitting assumes constant feeling ξ and uncertainty π parameters. Estimation of CUB models relies on likelihood methods and, specifically, on the implementation of the Expectation-Maximization algorithm. Fit improvements yielded by the specification of covariates can be tested via a Likelihood Ratio Test if models are nested: in general, the significance of an explanatory variable can be checked via standard Wald test. In the following, let D be a dichotomous variable included in the model specification to explain patterns of responses in terms of feeling and/or uncertainty, so that $\boldsymbol{\theta}_0 = (\pi_0, \xi_0)$, $\boldsymbol{\theta}_1 = (\pi_1, \xi_1)$ are the parameters of the conditional CUB distributions $\big(R_i|D_i = 0\big) \sim$ CUB (π_0, ξ_0) and $\big(R_i|D_i = 1\big) \sim$ CUB (π_1, ξ_1). If D implies statistically significant differences in model parameters, then two subgroups of respondents are identified and one should establish if the resulting clustering is actually relevant. This issue is particularly common when the sample size n is large. Here we wish to discuss a system of hypothesis:

H_0 : D should not be retained in the model (D is not important)

versus

H_1 : D should be retained in the model (D is important).

Thus, significant differences in model parameters will be investigated in order to disclose to which extent the clustering variable is actually relevant and should be retained in the model.

2 A test for variable importance

In order to test if the inclusion of a significant factor leads to relevant improvement of the fit, the (normalized) dissimilarity measure [6, 8]:

$$Diss(\boldsymbol{p}, \boldsymbol{q}) = \frac{1}{2}\sum_{r=1}^{m} |p_r - q_r| \ \in (0,1) \qquad (3)$$

stems quite naturally from the motivating discussion. If $\boldsymbol{p}, \boldsymbol{q}$ are two probability distributions, it assesses the proportion of cases in which the two distributions differ. Thus, if a CUB $(1,1)$ is fitted to the data with a dichotomous factor D for both components:

$$logit(\pi_i) = \beta_0 + \beta_1 D_i, \qquad logit(\xi_i) = \gamma_0 + \gamma_1 D_i,$$

and, accordingly, two clusters are identified, the dissimilarity between the estimated conditional response probabilities of $R_i|D=0$, $R_i|D=1$ indicates how far apart the groups $D=0$ and $D=1$ are in terms of the corresponding estimated CUB $(0,0)$ probability distributions. Similar considerations hold if the dichotomous variable D is specified only for one of the components.

3 A simulation experiment

The validation of the proposed approach to test variable importance will be run with a Monte Carlo experiment. For illustrative purposes, we shall consider the simplest case of a dichotomous variable D, with levels 0,1 (for instance, males and females, smokers and non smokers, etc.), able to discriminate feeling, by assuming heterogeneity constant among subjects. Thus, in the end we shall have two separate groups of respondents, corresponding to feeling parameters ξ_0 and ξ_1 if $D=0$ or $D=1$, respectively. We derive the empirical critical values c_α under the null $H_0 : \xi_0 = 0.30,\ \xi_1 = 0.35$ by generating a sample of data in which a dummy covariate is significant but the difference in parameter values is very small, thus it may raise doubts about importance of the implied classification. To this aim, we sample 1000 times from the null distribution for varying $\pi \in (0.2, 0.4, 0.6, 0.8)$, different numbers of categories and sample sizes for the two groups.

Empirical critical values for the dissimilarity statistics corresponding to nominal level $\alpha = 0.05$ are summarized in Table 1: lower and upper bounds (l_b and u_b, resp.) of 80% bootstrap confidence intervals (1000 replicates) are also reported as an instance of a measure of uncertainty of the test statistics. Thus, at level α, a value of dissimilarity between the implied conditional distributions lower than the corresponding critical value indicates that D has a weak importance for the purpose of discrimination of response patterns and its specification in the model could be matter of discussion.

As a by product of the simulation experiment, new evidence is found to support the specification of uncertainty for ordinal data models: indeed, critical values decrease for higher values of heterogeneity (that is, larger weights for the Uniform distribution), indicating that this component has a not-negligible effect in the analysis of variable importance. In order to enhance the purpose of the test, an additional simulation experiment has been planned: for each run and for the chosen parameter values, a sample has been generated with a significant dummy variable splitting the observations into two groups of sizes n_0 and n_1 respectively. Then the dissimilarity test has been applied to check for variable importance according to the proposal. Results are summarized in Table 2 and highlight that, especially for large samples,

Table 1 Empirical critical values with increasing level of uncertainty parameter

	$m=5$			$m=7$			$m=10$		
	c_α	l_b	u_b	c_α	l_b	u_b	c_α	l_b	u_b
	$n_0=300,\quad n_1=400,\quad \xi_0=0.30,\quad \xi_1=0.35$								
$\pi=0.2$	0.104	0.102	0.106	0.103	0.101	0.106	0.102	0.099	0.104
$\pi=0.4$	0.117	0.115	0.119	0.112	0.109	0.114	0.116	0.112	0.119
$\pi=0.6$	0.122	0.119	0.130	0.125	0.121	0.129	0.135	0.132	0.138
$\pi=0.8$	0.137	0.132	0.142	0.135	0.131	0.139	0.147	0.142	0.150
	$n_0=1300,\quad n_1=1400,\quad \xi_0=0.30,\quad \xi_1=0.35$								
$\pi=0.2$	0.058	0.057	0.061	0.053	0.052	0.055	0.054	0.053	0.055
$\pi=0.4$	0.064	0.063	0.065	0.065	0.063	0.067	0.072	0.070	0.073
$\pi=0.6$	0.075	0.073	0.076	0.079	0.078	0.080	0.094	0.093	0.096
$\pi=0.8$	0.090	0.088	0.093	0.099	0.098	0.101	0.119	0.118	0.121
	$n_0=13000,\quad n_1=14000,\quad \xi_0=0.30,\quad \xi_1=0.35$								
$\pi=0.2$	0.024	0.023	0.024	0.026	0.025	0.026	0.030	0.030	0.031
$\pi=0.4$	0.042	0.041	0.042	0.046	0.046	0.047	0.056	0.055	0.056
$\pi=0.6$	0.060	0.059	0.060	0.066	0.066	0.067	0.081	0.081	0.081
$\pi=0.8$	0.077	0.076	0.077	0.086	0.086	0.086	0.106	0.105	0.106

the classical concept of statistical significance has to be accompanied by a more specific analysis of variable importance.

Table 2 Importance rates for a significant dummy variable

	$\pi=0.2$		$\pi=0.4$		$\pi=0.6$		$\pi=0.8$	
Importance	Yes	No	Yes	No	Yes	No	Yes	No
$n_0=300,\quad n_1=400,\quad \xi_0=0.3,\quad \xi_1=0.35$								
$m=5$	0.09	0.91	0.06	0.94	0.04	0.96	0.04	0.96
$m=7$	0.08	0.92	0.07	0.93	0.05	0.95	0.05	0.95
$m=10$	0.06	0.94	0.09	0.91	0.04	0.96	0.11	0.89
$n_0=1300,\quad n_1=1400,\quad \xi_0=0.3,\quad \xi_1=0.35$								
$m=5$	0.03	0.97	0.05	0.95	0.06	0.94	0.11	0.89
$m=7$	0.07	0.93	0.05	0.95	0.14	0.86	0.08	0.92
$m=10$	0.08	0.92	0.13	0.87	0.12	0.88	0.12	0.88
$n_0=13000,\quad n_1=14000,\quad \xi_0=0.3,\quad \xi_1=0.35$								
$m=5$	0.16	0.84	0.12	0.88	0.09	0.91	0.17	0.83
$m=7$	0.14	0.86	0.12	0.88	0.10	0.90	0.13	0.87
$m=10$	0.14	0.86	0.11	0.89	0.11	0.89	0.14	0.86

The proposed variable-importance test has shown perfect agreement with the corresponding one using the (symmetrized) Kullback-Leibler divergence (here not reported for the sake of brevity), but it is more advantageous since the dissimilarity measure is a proper (normalized) distance and it is able to foster interpretation and visualization of results. Similar conclusions are found when considering $\xi_0 = 0.1$ or $\xi_1 = 0.5$ for the feeling under the null (it is not necessary to test for values $\xi > 0.5$

since CUB distributions are reversible) and for increasing differences between ξ_0 and ξ_1. Dually, the proposed test can be run in case the clustering covariate is tested to explain heterogeneity for samples with homogeneous feeling or both components.

4 On-going developments

The proposed testing procedure for variable importance prescribes that, once a dichotomous factor is specified in the model to explain the response (in terms of feeling and uncertainty in case one assumes the CUB paradigm), then the dissimilarity between the estimated conditional distributions can reveal its discrimination ability in an effective and insightful way. The topics here investigated are being subject to more in-depth analysis stemming from real case-studies; further studies are tailored to the application of the approach to other classes of models, as well as to the study of properties of the dissimilarity estimator -also known as Duncan segregation index in other contexts- in the vein of [4, 10]. Notice that the same approach can be exploited to design a proper test of significance for difference in parameter values:

$$H_0 : \boldsymbol{\theta}_0 = \boldsymbol{\theta}_1 \quad \textit{versus} \quad H_1 : \boldsymbol{\theta}_0 \neq \boldsymbol{\theta}_1$$

with test statistics based on the dissimilarity between the conditional distributions. From some preliminary investigations in this perspective one obtains a test that is as powerful as the corresponding one using the Kullback-Leibler divergence to assess distances between distributions. This approach has been investigated in [7, 9] to design a homogeneity test in case of continuous populations and small sample sizes.

References

1. Bragoli, D., Ganugi, P., Ianulardo, G.: Gini's transvariation analysis: an application on financial crises in developing countries. Empirica **40**, 153–174 (2013)
2. D'Elia A., Piccolo D.: A mixture model for preference data analysis. Comput. Stat. Data An. **49**, 917–934 (2005)
3. Demidenko, E.: The p-values You Can't Buy. The American Statistician. **70**, 33–38 (2016)
4. Forcina, A., Galmacci, G.: On the distribution of the Index of Dissimilarity. Metron **32**, 361–374 (1974)
5. Gini, C.: Il concetto di transvariazione e le sue prime applicazioni. in: Transvariazione, Gini, C. ed., Libreria Goliardica, Roma (1916)
6. Gini, C.: La dissomiglianza. Metron, 85–215 (1965)
7. Girone, G., Nannavecchia, A.: The distribution of an Index of Dissimilarity for two samples from a Uniform Population. Applied Mathematics **4**, 1028–1037 (2013)
8. Leti, G. (1983). *Statistica descrittiva*. Il Mulino, Bologna.
9. Manca, F., Marin, C.: Simulated Sample Behaviour of a Dissimilarity Index when Sampling from Populations differing by a location parameter only. Applied Mathematics **5**, 2199–2208 (2014)
10. Mazza, A., Punzo, A.: On the Upward Bias of the Dissimilarity Index and Its Corrections. Sociological Methods and Research **44**(1), 80– 107 (2015)