Statistica & Applicazioni - Vol. XIV, n. 2, 2016, pp. 107-122

RELATIVE-IMPORTANCE ASSESSMENT OF EXPLANATORY VARIABLES IN GENERALIZED LINEAR MODELS: AN ENTROPY-BASED APPROACH

Nobuoki Eshima*
Claudio Giovanni Borroni**
Minoru Tabata***

SUMMARY

The object of the present paper is to propose a method for relative-importance assessment of explanatory variables in generalized linear models, through an analysis of the variation of entropy of the response variable. First, the problem is reviewed in the ordinary regression model and some criteria to be met by a suitable measure are emphasized. Second, the logic of variation in entropy is introduced, for the assessment both of the predictive power of the whole model and of the relative importance of each variable. Third, the occurrence of a causal order of variables is discussed and a new approach is proposed to deal with cases where this order lacks. Finally, the ability to meet the listed criteria is checked for the proposed measure and two relevant examples (logit model and two-way ANOVA model) are provided, both with numerical applications.

Keywords: *Entropy Coefficient of Determination; Generalized Linear Models; Variable Importance.*

1. INTRODUCTION

Generalized linear models (GLMs) can be flexibly designed by choosing random, systematic and link components to make useful regression analyses of both continuous and categorical response variables (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989). Above all, GLMs play an important role for the non-normal response variables, which arise in many fields, e.g. in biomedical researches, behavioral sciences, economics, etc. When the GLM reduces to the ordinary linear regression, practical data analyses usually face other statistical problems beyond those of estimating and testing parameters. Specifically, the fit of the model is considered and often used as a basis to search for outliers or leverage points and to assess the importance of each explanatory variable. While fit analysis is obviously conducted by the coefficient of determination R^2, there are several attempts to discuss the relative importance of explanatory variables and no unified and satisfactory approach for this topic exists. A recent review by Grömping (2015) sheds light to the vast set

* Center for Educational Outreach and Admissions - Kyoto University - JAPAN (email: eshima.nobuoki.2z@kyoto-u.ac.jp).
** Dipartimento di Statistica e Metodi Quantitativi - Università degli Studi di Milano-Bicocca - via Bicocca degli Arcimboldi, 8 - 20126 MILANO (email: ✉ claudio.borroni@unimib.it).
*** Department of Mathematical Sciences - Graduate School of Engineering - Osaka Pref. University - JAPAN (email: mnrtabata@luck.ocn.ne.jp).

of proposed measures (see also Grömping, 2007, 2009). The main problem is that the explanatory variables are likely to be correlated. The regression coefficients provide a first insight of the explanatory power of each variable, indeed, but they might not be suitable in case of correlation, being that their interpretation needs to assume that the other variables are held constant. Moreover, the regression coefficients are not scale invariant, so that the judged importance may largely depend on the chosen unit of measurement of the explanatory variables. Some Authors revert then to suitable functions of the regression coefficients, like standardized coefficients or their products with raw correlation coefficients (Hoffman, 1960; Pratt, 1987). However, some further problems of interpretation arise, mainly due to the possibility of assigning a negative contribution to some variables (see, above all, Darlington, 1968). A common idea of most proposed approaches is that the assessment of importance must be regarded as a decomposition of the R^2. In effect, such a requirement can be easily met when the regressors are uncorrelated, while most measures do not give this result in the case of correlation. Kruskal (1987) pointed out that such a decomposition can be provided, even in the correlated case and with non-negative contributions, by the square of the partial correlation coefficient or, equivalently, by the share of the total variance which is additionally explained when a new regressor is introduced in the model. This logic needs to assume that regressors have a natural ordering, however, which is not usually the case, apart from some few examples. By using a logic already provided in Lindeman, Merenda and Gold (1980), the Author proposes then to take the average of the contribution of a given variable, over all possible orderings of the set of regressors. This simple idea, which was further generalized (Theil and Chung, 1988; Feldman, 1999), will be used is the present paper.

Before going back to GLMs, it has to be emphasized that many Authors proposed criteria to be met by a relative-importance measure in linear regression (for a full list of them, see Grömping, 2015). We think that the most important criteria should be as follows:

(a) *Proper decomposition:* the model variance is to be decomposed into shares, that is the sum of all shares has to be the model variance.
(b) *Non-negativity:* all shares have to be non-negative.
(c) *Exclusion:* the share allocated to a regressor whose coefficient is zero should be zero as well.
(d) *Inclusion:* a regressor with nonzero coefficient should receive a nonzero share.

While criteria (a), (b) and (d) are natural, the need for criteria (c) (Feldman, 1999) is quite controversial. As pointed out by Grömping (2007), in effect, in some causal structure a zero coefficient does not necessarily indicate an unimportant explanatory variable.

Apart from a paper by Chevan and Sutherland (1991), it seems that the problem of the relative-importance assessment of explanatory variables in the general framework of GLMs has not been sufficiently discussed in the literature. This fact can be perhaps ascribed to the lack of an unified view even for the measurement of the total fit of the model. To this aim, Eshima and Tabata (2007, 2010, 2011) gave further in-

sight by developing an entropy-based approach to the predictive power of GLMs. This approach leads to a new measure of fit, the *entropy coefficient of determination* (ECD), whose details will be provided in the next section. As a matter of facts, the ECD is a valid alternative to the R^2, which is known not to be completely well-suited in GLMs analyses. Recently, Eshima, Tabata, Borroni and Kano (2015) applied the ECD to the problem of relative-importance assessment of explanatory variables. Their contribution can be framed in the field of path analysis, essentially because a causal ordering of the variables is assumed. As above discussed, however, in GLMs such an assumption is met very rarely. This paper will then try to fill this gap by considering an unordered set of explanatory variables.

The paper is organized as follows. The next section reviews the concept of ECD in a GLM. Moreover, the path analysis approach in Eshima *et al.* (2015) is reviewed and the measurement of the total, direct and indirect effects of explanatory variables is discussed when no causal order exists. Further, a new measure of relative importance is proposed, by using an average over all possible orderings of the set of variables. Such a measure can be applied for all GLMs, but Section 3 gives two examples of relevant applications, the logit model and the two-way ANOVA model. Both examples are supported by computations on real data. A final summary is given in Section 4.

2. ENTROPY-BASED ANALYSIS IN GLMS

Let $\boldsymbol{X} = \left(X_1, X_2, ..., X_p\right)^T$ and Y be a $p \times 1$ vector of explanatory variables and a response variable, respectively. Let $f(y|\boldsymbol{x})$ be the conditional probability or density function of Y given $\boldsymbol{X} = \boldsymbol{x}$. The function $f(y|\boldsymbol{x})$ is assumed to be a member of the following family:

$$f(y|\boldsymbol{x}) = \exp\left(\frac{y\theta - b(\theta)}{a(\varphi)} + c(y, \varphi)\right), \tag{1}$$

where θ and φ are parameters, and $a(\varphi)$ (> 0), $b(\theta)$ and $c(y, \varphi)$ are specific functions. Finally, let $\boldsymbol{\beta}^T = (\beta_1, \beta_2, ..., \beta_p)^T$ be a vector of parameters. Given a link function $h(u)$ and the linear predictor $\eta = \boldsymbol{\beta}^T\boldsymbol{x}$, in a GLM the parameter θ is a function of $\boldsymbol{\beta}^T\boldsymbol{x}$, say $\theta(\boldsymbol{\beta}^T\boldsymbol{x})$. For continuous random variables $\boldsymbol{X}$ and Y, denote by $g(\boldsymbol{x})$ and $f(y)$ their marginal density functions and define the following average distances

$$\begin{aligned} D(f(y)\,||f(y|\boldsymbol{x})) &= \int\int f(y)\, g(\boldsymbol{x})\, \log\left(\frac{f(y)}{f(y|\boldsymbol{x})}\right) d\boldsymbol{x}dy \\ D(f(y|\boldsymbol{x})\,||f(y)) &= \int\int f(y|\boldsymbol{x})\, g(\boldsymbol{x})\, \log\left(\frac{f(y|\boldsymbol{x})}{f(y)}\right) d\boldsymbol{x}dy. \end{aligned} \tag{2}$$

It can be easily proved that

$$\frac{\mathrm{Cov}(\theta, Y)}{a(\varphi)} = D(f(y)||f(y|\boldsymbol{x})) + D(f(y|\boldsymbol{x})||f(y)), \tag{3}$$

so that the ratio $\frac{\mathrm{Cov}(\theta,Y)}{a(\varphi)}$ can be expressed by the Kullback-Leibler (KL) information. Moreover, the same ratio can be regarded as the average change of uncertainty of the response variable Y through the explanatory variables vector $\boldsymbol{X}$ (Eshima and Tabata, 2007, 2010). In this paper, such a ratio will be denoted by $KL(\boldsymbol{X}, Y)$ and referred to as the *model variation in entropy*. In (3), the numerator $\mathrm{Cov}(\theta, Y)$ can be interpreted as the explained variation of Y in entropy and the denominator $a(\varphi)$ as the error variation in entropy. The entropy coefficient of determination (ECD) is then defined by

$$\mathrm{ECD}(\boldsymbol{X}, Y) = \frac{KL(\boldsymbol{X}, Y)}{KL(\boldsymbol{X}, Y) + 1} \tag{4}$$

and it is interpreted as the proportion of explained *variation* of Y in entropy.

Notice that the integrals in (2) can be replaced by summations if $\boldsymbol{X}$ and/or Y is discrete. In addition, the explanatory variables in the model are not necessarily random; in the case of fixed factors (as detailed in Eshima and Tabata, 2007 and in the example of Section 3.2 below), indeed, one just needs to replace expectations with arithmetic means over the levels of each explanatory variable.

Eshima *et al.* (2015) developed an approach to assess the importance of the explanatory variables of a GLM when they follow a given *ordering*, a situation which is often encountered in applications and that is fully described in the literature (see for instance Williams, 1978). This assumption means that a given explanatory variable can have both a direct effect on the response and an indirect effect through the other *following* variables in the ordering; however, an indirect effect of that variable through the *preceding* variables in the ordering does not exist. Equivalently, one can claim that a path leading a variable to the response exists or that a *causal* ordering of the explanatory variables holds.

Of course, the causal ordering of the p variables can be described by a given permutation of the set $\{1, 2, \ldots, p\}$ of indices. Suppose now, without loss of generality, that the natural ordering $(1, 2, \ldots, p)$ applies, so that X_i precedes X_{i+1} $(i = 1, 2, \ldots, p-1)$. One can first define the total effect of X_1 on Y as

$$e_T(X_1 \longrightarrow Y) = \frac{KL\big((X_1, X_2, ..., X_p), Y\big) - KL\big((X_2, X_3, ..., X_p), Y|X_1\big)}{KL(\boldsymbol{X}, Y) + 1}$$

where $KL\big((X_2, X_3, ..., X_p), Y|X_1\big)$ is the conditional KL information given X_1. Second, the total effect of X_2 is defined as

$$e_T(X_2 \longrightarrow Y) = \frac{KL\big((X_2, X_3, ..., X_p), Y|X_1\big) - KL\big((X_3, X_4, ..., X_p), Y|X_1, X_2\big)}{KL(\boldsymbol{X}, Y) + 1}.$$

So, the total effects of X_i $(i = 1, 2, ..., p)$ can be derived by induction as

$$e_T(X_i \longrightarrow Y) = \tag{4}$$
$$= \frac{KL\big((X_i, X_{i+1}, ..., X_p), Y | X_1, X_2, ..., X_{i-1}\big) - KL\big((X_{i+1}, X_{i+2}, ..., X_p), Y | X_1, X_2, ..., X_i\big)}{KL(\boldsymbol{X}, Y) + 1}$$

Notice that

$$e_T(X_i \longrightarrow Y) \geq 0 \qquad (i = 1, 2, ..., p) \tag{5}$$

and that

$$\mathrm{ECD}(\boldsymbol{X}, Y) \equiv e_T(\boldsymbol{X} \longrightarrow Y) = \sum_{i=1}^{p} e_T(X_i \longrightarrow Y). \tag{6}$$

One can then look at the ratio

$$CR_{(1,2,...,p)}(X_i) = \frac{e_T(X_i \longrightarrow Y)}{e_T(\boldsymbol{X} \longrightarrow Y)} \tag{7}$$

as the contribution of each explanatory variable X_i $(i = 1, 2, ..., p)$. Every such ratio will range in $[0, 1]$ and their sum will be equal to 1. This means that, when the explanatory variables follow the order in $(1, 2, \ldots, p)$, the contribution ratios in (7) can be also re-interpreted as measures of importance of a single explanatory variable, relative to the whole set. If we denote as $RI(X_i)$ $(i = 1, 2, ..., p)$ such measures, we can then simply set

$$RI(X_i) = CR_{(1,2,...,p)}(X_i) \qquad i = 1, 2, ..., p.$$

Notice that the ratios in (7) can be computed for every fixed ordering, that is for every permutation of the set $\{1, 2, \ldots, p\}$; however, only when the explanatory variables follow *that* causal ordering, (5) and (6) will hold, in such a way that (7) are also measures of relative importance.

REMARK 1

The approach above can be applied for every GLM with an arbitrary link function. However, in the case of canonical link, i.e when $\theta = \sum_{i=1}^{p} \beta_i x_i$, *the computation of the effect of the explanatory variables on Y simplifies as detailed in the following. Denote as* $S_i = \{X_1, X_2, \ldots, X_{i-1}\}$ *the set of variables preceding* X_i *and as* $T_i = \{X_i, X_{i+1}, \ldots, X_p\}$ *the set of those following and including* X_i $(i = 1, \ldots, p)$.

$$e_T(X_i \longrightarrow Y) = \frac{\beta_i \,\mathrm{Cov}(X_i, Y | S_i)}{a(\varphi)(KL(\boldsymbol{X}, Y) + 1)} + \frac{\sum_{j=i+1}^{p} \beta_j \left(\mathrm{Cov}(X_j, Y | S_i) - \mathrm{Cov}(X_j, Y | S_{i+1})\right)}{a(\varphi)(KL(\boldsymbol{X}, Y) + 1)}$$

where $\mathrm{Cov}(X_k, Y | S_i)$ *is the conditional covariance between* X_k *and Y given the variables* $X_1, X_2, \ldots, X_{i-1} (k = i, i + 1, .., p; i = 1, 2, ..., p - 1)$. *Under multivariate normality, that is for the usual linear regression, let* $R^2(S_i | T_i)$ *be the partial coefficient*

of determination for regressors $\{X_i, X_{i+1}, ..., X_p\}$, *given* $\{X_1, X_2, ..., X_{i-1}\}$ $(i = 1, 2, \ldots, p)$. *Then, from (4) and (6),*

$$e_T(X_i \longrightarrow Y) = \frac{\dfrac{R^2(T_i|S_i)}{1 - R^2(T_i|S_i)} - \dfrac{R^2(T_{i+1}|S_{i+1})}{1 - R^2(T_{i+1}|S_{i+1})}}{\dfrac{R^2}{1 - R^2} + 1}$$

$$= \left(1 - R^2\right)\left(\frac{1}{1 - R^2(T_i|S_i)} - \frac{1}{1 - R^2(T_{i+1}|S_{i+1})}\right)$$

and

$$e_T(\boldsymbol{X} \longrightarrow Y) = R^2$$

where $R^2 = R^2(T_1|S_1)$. *The factor contribution can then be calculated from (7).*

The ideas above are now used to measure the contribution of a set of explanatory variables when there is no causal ordering among them. As usual, the total effect of all explanatory variables is given by

$$e_T(\boldsymbol{X} \longrightarrow Y) = \frac{KL(\boldsymbol{X}, Y)}{KL(\boldsymbol{X}, Y) + 1} \tag{8}$$

After denoting as $\boldsymbol{X}^{/i} = \{X_1, X_2, ..., X_{i-1}, X_{i+1}, ..., X_p\}$ the set of all explanatory variables excluding X_i $(i = 1, 2, \ldots, p)$, the direct effect of this set can be measured as

$$e_D\left(\boldsymbol{X}^{/i} \longrightarrow Y\right) = \frac{KL\left(\boldsymbol{X}^{/i}, Y|X_i\right)}{KL(\boldsymbol{X}, Y) + 1}. \tag{9}$$

By subtracting (9) from (8), we can thus define the total effect of X_i on Y as

$$e_T(X_i \longrightarrow Y) = e_T(\boldsymbol{X} \longrightarrow Y) - \ e_D\left(\boldsymbol{X}^{/i} \longrightarrow Y\right) =$$

$$= \frac{KL(\boldsymbol{X}, Y) - KL\left(\boldsymbol{X}^{/i}, Y|X_i\right)}{KL(\boldsymbol{X}, Y) + 1}.$$

Similarly, the direct effect of X_i on Y is defined as

$$e_D(X_i \longrightarrow Y) = \frac{KL\left(X_i, Y|\boldsymbol{X}^{/i}\right)}{KL(\boldsymbol{X}, Y) + 1},$$

and $e_D(X_i \longrightarrow Y) = 0$ when the regression coefficient $\beta_i = 0$ $(i = 1, 2, \ldots, p)$. The indirect effect of X_i can be thus calculated as

$$e_I(X_i \longrightarrow Y) = e_T(X_i \longrightarrow Y) - \ e_D(X_i \longrightarrow Y) =$$

$$= \frac{KL(\boldsymbol{X}, Y) - KL\left(\boldsymbol{X}^{/i}, Y|X_i\right) - KL\left(X_i, Y|\boldsymbol{X}^{/i}\right)}{KL(\boldsymbol{X}, Y) + 1}.$$

Moreover, by using the same logic as in (7), the contribution ratio of X_i is now

$$\widetilde{CR}(X_i) = \frac{e_T(X_i \longrightarrow Y)}{e_T(\boldsymbol{X} \longrightarrow Y)} = \frac{KL(\boldsymbol{X}, Y) - KL\left(\boldsymbol{X}^{/i}, Y | X_i\right)}{KL(\boldsymbol{X}, Y)}. \tag{10}$$

Notice that the measures in (10) do not generally coincide with those in (7). In effect, such a coincidence will only hold when the explanatory variables follow a causal ordering and if X_i is the first variable of that ordering. As a consequence, in lack of a causal ordering, $\sum_{i=1}^{p} \widetilde{CR}(X_i) \neq 1$ and the quantity in (10) cannot be regarded as a measure of the importance of X_i relative to the whole set. That is, the concepts of factor contribution and of relative importance must be separated when no causal ordering of the explanatory variables can be invoked.

REMARK 2

Again, a kind of simplification is obtained for GLMs with canonical link; here, the contribution ratio of the explanatory variables are calculated by using covariances between the explanatory and response variables. Let $\mathrm{Cov}\left(Y, \boldsymbol{\beta}^T \boldsymbol{X} | X_k = x_k\right)$ *be the conditional covariance between* Y *and* $\boldsymbol{\beta}^T \boldsymbol{X}$ *given* X_i $(i = 1, 2, \ldots, p)$. *The total effect of* X_i *can be measured as*

$$\mathrm{Cov}\left(Y, \boldsymbol{\beta}^T \boldsymbol{X}\right) - \mathrm{Cov}\left(Y, \boldsymbol{\beta}^T \boldsymbol{X} | X_i\right) = \beta_i \mathrm{Cov}(Y, X_i) + \sum_{\substack{j=1 \\ j \neq i}}^{p} \beta_j \left(\mathrm{Cov}\left(Y, X_j\right) - \mathrm{Cov}\left(Y, X_j | X_i\right)\right)$$

Thus, the contribution ratio of X_i can be calculated by

$$\widetilde{CR}(X_i) = \frac{\beta_i \mathrm{Cov}(Y, X_i) + \sum_{\substack{j=1 \\ j \neq i}}^{p} \beta_j \left(\mathrm{Cov}\left(Y, X_j\right) - \mathrm{Cov}\left(Y, X_j | X_i\right)\right)}{\mathrm{Cov}\left(Y, \boldsymbol{\beta}^T \boldsymbol{X}\right)}.$$

To define a measure of relative importance when no ordering of the explanatory variables exists, one can consider a mean over all possible orderings of such variables. This is a well-accepted logic in such kinds of measurement, as testified (with variants) by many papers in the literature (Kruskal, 1987; Lindeman, Merenda and Gold, 1980; Feldman, 1999, above all). It has to be emphasized that another well-established logic is the one based on hierarchical partitioning (Chevan ans Sutherland, 1991) which, however, will not be applied here. Consider a permutation $\boldsymbol{r} = \left(r_1, r_2, ..., r_p\right)$ of the set of indices $\{1, 2, ..., p\}$. Let $S_i(\boldsymbol{r})$ denote the set of such variables appearing before X_i in the permutation $\boldsymbol{r}$ and let $T_i(\boldsymbol{r})$ denote the set of all other explanatory variables, including X_i $(i = 1, 2, \ldots, p)$. Then, as in (7), the contribution ratio of X_i for the ordering $\boldsymbol{r}$ can be defined as

$$CR_{\boldsymbol{r}}(X_i) = \frac{KL(T_i(\boldsymbol{r}), Y \mid S_i(\boldsymbol{r})) - KL((T_i(\boldsymbol{r}) \backslash \{X_i\}), Y \mid (S_i(\boldsymbol{r}) \cup \{X_i\}))}{KL(\boldsymbol{X}, Y)} \tag{11}$$

and the relative importance of X_i $(i = 1, 2, \ldots, p)$ is now defined as

$$RI(X_i) = \frac{1}{p!}\sum_{\boldsymbol{r}} CR_{\boldsymbol{r}}(X_i) \tag{12}$$

where the summation is taken over the $p!$ permutations of the set $\{1, 2, ..., p\}$. As an example, when $p = 3$, one gets:

$$\begin{aligned} RI(X_i) = {} & \frac{1}{6KL((X_1, X_2, X_3), Y)} \cdot \\ & \big\{2KL((X_1, X_2, X_3), Y) - 2KL((X_2, X_3), Y|X_1) + KL((X_1, X_3), Y|X_2) \\ & - KL(X_3, Y|X_1, X_2) + KL((X_1, X_2), Y|X_3) - KL(X_2, Y|X_1, X_3) \\ & + 2KL(X_1, Y|X_2, X_3)\big\} \qquad (i = 1, 2, 3). \end{aligned}$$

As an explanatory-power measure, the ECD is a variation function just of the explanatory variables X_k for which $\beta_k \neq 0$ $(k = 1, 2, \ldots, p)$. So it is meaningful to consider the relative-importance measure in (12) only for the explanatory variables with non-zero coefficients. Apart from this limitation, the proposed measure is quite general, especially because is easily applied to all GLMs without the need for further assumptions. In the next section, two classical cases of application will be discussed, both complemented with the analysis and interpretation of real datasets.

TABLE 1. - *Data from 2276 high school seniors about whether they have ever used alcohol, cigarettes, or marijuana*

Marijuana (Y)	Cigarettes (X_2)	Alcohol (X_1)	
		Yes	No
Yes	Yes	911	3
	No	44	2
No	Yes	538	43
	No	456	279

REMARK 3

After substituting the ECD and the entropy model variation $KL(\boldsymbol{X}, Y)$, respectively with R^2 and the total model variance, the proposed method satisfies all criteria (a) to (d) listed in Section 1 for linear regression. However, if the method is applied to a set of explanatory variables which includes variables with zero coefficients, then it does not satisfy criterion (c).

3. EXAMPLES

3.1 *Example 1 - Logit model*

A common application for a dichotomic response Y is the logit model. As known, such a model can be easily framed in the GLM class by assuming that, conditionally on a set of p explanatory variables, Y has a Bernoulli distribution with success

probability $\delta \in [0,1]$ and by considering the logit as a link function. In the following, we will look at an example where $p = 2$, so that the model can be written as

$$\text{logit}(\delta) = \log \frac{\delta}{1-\delta} = \mu + \beta_1 x_1 + \beta_2 x_2.$$

By looking at the general structure in (1), the function $a(\varphi) \equiv 1$ so that $\text{Cov}(\theta, Y) = KL((X_1, X_2), Y)$. Table 1 shows data from 2276 high school seniors about whether they have ever used alcohol, cigarettes, or marijuana (Agresti, 2002, pp. 322-323). A logit model can be applied for the response variable Marijuana Use Y with Alcohol Use X_1 and Cigarette Use X_2 as explanatory variables. Notice that, in this case, it is not appropriate to assume any causal order in the variables, because some people can use alcohol before cigarettes, and others cigarettes before alcohol. In this logit model, $\theta = \mu + \beta_1 x_1 + \beta_2 x_2$ and the estimates are $\widehat{\mu} = -5.309$, $\widehat{\beta_1} = 2.986$ and $\widehat{\beta_2} = 2.848$. The usual interpretation of the results is given by using odds with respect to Marijuana Use. The partial odds in Marijuana Use Y given Cigarette Use X_2 is $\exp(2.986) = 19.806$ times higher at Alcohol Use $X_1 =$ ''yes'' than that at Alcohol Use $X_1 =$ ''no'', and the partial odds in Marijuana Use Y given Alcohol Use X_1 is $\exp(2.848) = 17.253$ times higher at Cigarette Use $X_2 =$ ''yes'' than that at Cigarette Use $X_2 =$ ''no''. According to the usual interpretation, the effects of Alcohol Use and Cigarette Use on Marijuana Use might then be thought as equivalent, provided that Alcohol Use and Cigarette Use are completely controlled. However Alcohol Use and Cigarette Use are associated and cannot be completely controlled. Hence, in addition to the usual method based on odds, it may be sensible to use a new method for assessing the contributions and the relative-importance degrees of the explanatory variables. The direct effect of a variable should be defined given the other explanatory variable, and the indirect effect is to be defined through the association between them.

By using the ML estimates above, one can compute

$$\text{Cov}(\theta, Y) = 0.529 \,(= KL((X_1, X_2), Y)),$$

so that $ECD = 0.346$ and 34.6% of variation of Y in entropy is explained by the two explanatory variables. Moreover

$$\begin{aligned}
\text{Cov}(\theta, Y) - \text{Cov}(\theta, Y|X_1) &= KL((X_1, X_2), Y) - KL((X_1, X_2), Y|X_1) = \\
&= 0.529 - 0.288 = 0.241 \\
\text{Cov}(\theta, Y) - \text{Cov}(\theta, Y|X_2) &= KL((X_1, X_2), Y) - KL((X_1, X_2), Y|X_2) = \\
&= 0.529 - 0.072 = 0.457,
\end{aligned}$$

so that

$$\begin{aligned}
e_T((X_1, X_2) \longrightarrow Y) &= \frac{0.529}{0.529 + 1} = 0.346 \\
e_T(X_1 \longrightarrow Y) &= \frac{0.241}{0.529 + 1} = 0.158 \\
e_D(X_1 \longrightarrow Y) &= \frac{0.072}{0.529 + 1} = 0.047
\end{aligned}$$

$$e_T(X_2 \longrightarrow Y) = \frac{0.457}{0.529 + 1} = 0.299$$
$$e_D(X_2 \longrightarrow Y) = \frac{0.288}{0.529 + 1} = 0.188.$$

The contribution ratio of both explanatory variables are

$$CR(X_1) = \frac{0.241}{0.529} = 0.456$$
$$CR(X_2) = \frac{0.457}{0.529} = 0.864,$$

so that, the contribution of Alcohol Use X_1 on Marijuana Use Y is 45.6%, whereas that of Cigarette Use X_2 on Marijuana Use Y is 86.4%. The effect of Cigarette Use X_2 on Marijuana Use Y is about two times greater than that of Alcohol Use X_1. The relative-importance degrees of the explanatory variables are as follows:

$$RI(X_1) = \frac{1}{2}\left(\frac{\mathrm{Cov}(\theta, Y) - \mathrm{Cov}(\theta, Y|X_1)}{\mathrm{Cov}(\theta, Y)} + \frac{\mathrm{Cov}(\theta, Y|X_2)}{\mathrm{Cov}(\theta, Y)}\right) =$$
$$= \frac{1}{2}\left(\frac{0.241}{0.529} + \frac{0.072}{0.529}\right) = 0.296$$
$$RI(X_2) = \frac{1}{2}\left(\frac{0.457}{0.529} + \frac{0.288}{0.529}\right) = 0.704,$$

so that the degree of relative importance of X_2 is about three times that of X_1.

3.2 *Example 2 - Two-way ANOVA model*

Suppose to consider a two-factors completely randomized experiment. The proposed method can be applied to measure the contribution of each factor in a fixed-effect model for the response variable Y. Denote the two factors by X_1 and X_2, with levels $\{1, 2, \ldots, I\}$ and $\{1, 2, \ldots, J\}$ respectively. For each combination (i, j) of the two factors, Y is assumed to be normally distributed with constant variance σ^2 and with expectation

$$E(Y|X_1 = i, X_2 = j) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} \qquad i = 1, \ldots, I; j = 1, \ldots, J,$$

where the parameters of the model are subjected to the following constraints

$$\sum_{i=1}^{I} \alpha_i = \sum_{j=1}^{J} \beta_j = \sum_{i=1}^{I} (\alpha\beta)_{ij} = \sum_{j=1}^{J} (\alpha\beta)_{ij} = 0.$$

Given that the explanatory variables are fixed, expectations and covariances are defined here as arithmetic means upon the levels of the considered categorical variables (see Eshima and Tabata, 2007). Moreover, the model can be easily framed in

the class of GLMs, after some simple substitutions. Specifically, define the following dummy variables

$$X_{1i} = \begin{cases} 1 & \text{if } X_1 = i \\ 0 & \text{if } X_1 \neq i \end{cases} \qquad X_{2j} = \begin{cases} 1 & \text{if } X_2 = j \\ 0 & \text{if } X_2 \neq j \end{cases}$$

$(i = 1, \ldots, I; j = 1, \ldots, J)$ and use them to define the vectors

$$\boldsymbol{X}_1 = (X_{11}, X_{12}, \ldots, X_{1I})^T \qquad \text{and} \qquad \boldsymbol{X}_2 = (X_{21}, X_{22}, \ldots, X_{2J})^T.$$

The linear predictor of the model is then

$$\theta = \mu + \boldsymbol{\alpha}^T \boldsymbol{X}_1 + \boldsymbol{\beta}^T \boldsymbol{X}_2 + \operatorname{tr}\left[\boldsymbol{\lambda}^T \boldsymbol{X}_1 \boldsymbol{X}_2^T\right],$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_I)^T$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_J)^T$ and $\boldsymbol{\lambda}$ is the matrix of the interaction parameters, whose $(i,j)-$th element is $(\alpha\beta)_{ij}$ $(i = 1, \ldots, I; j = 1, \ldots, J)$. Simple calculations show that

$$\operatorname{Cov}(Y, \boldsymbol{X}_1) = (1/I)\,\boldsymbol{\alpha} \qquad \operatorname{Cov}(Y, \boldsymbol{X}_2) = (1/J)\,\boldsymbol{\beta} \qquad \operatorname{Cov}(Y, \boldsymbol{X}_1 \boldsymbol{X}_2^T) = (1/IJ)\,\boldsymbol{\lambda},$$

so that, being that the link is canonical,

$$\operatorname{Cov}(Y, \theta) = \boldsymbol{\alpha}^T \operatorname{Cov}(Y, \boldsymbol{X}_1) + \boldsymbol{\beta}^T \operatorname{Cov}(Y, \boldsymbol{X}_2) + \operatorname{tr}\left[\boldsymbol{\lambda}^T \operatorname{Cov}(Y, \boldsymbol{X}_1 \boldsymbol{X}_2^T)\right]$$

and the total effect of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ is

$$\frac{1}{\sigma^2} \operatorname{Cov}(Y, \theta) = \frac{1}{\sigma^2} \left[\frac{1}{I} \sum_{i=1}^{I} \alpha_i^2 + \frac{1}{J} \sum_{j=1}^{J} \beta_j^2 + \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} (\alpha\beta)_{ij}^2 \right].$$

The three terms into square brackets may be referred to as the main effect of X_1, that of X_2 and the interactive effect, respectively. The ECD, which coincides with R^2 in this case, is computed as:

$$\text{ECD} = \frac{(1/I) \sum_i \alpha_i^2 + (1/J) \sum_j \beta_j^2 + (1/IJ) \sum_i \sum_j (\alpha\beta)_{ij}^2}{(1/I) \sum_i \alpha_i^2 + (1/J) \sum_j \beta_j^2 + (1/IJ) \sum_i \sum_j (\alpha\beta)_{ij}^2 + \sigma^2}$$

The effects of the factor X_1 on the response variable Y can then be described as follows:

$$e_T(X_1 \longrightarrow Y) = \frac{(1/I) \sum_i \alpha_i^2 + (1/IJ) \sum_i \sum_j (\alpha\beta)_{ij}^2}{(1/I) \sum_i \alpha_i^2 + (1/J) \sum_j \beta_j^2 + (1/IJ) \sum_i \sum_j (\alpha\beta)_{ij}^2 + \sigma^2}$$

$$e_D(X_1 \longrightarrow Y) = \frac{(1/I) \sum_i \alpha_i^2}{(1/I) \sum_i \alpha_i^2 + (1/J) \sum_j \beta_j^2 + (1/IJ) \sum_i \sum_j (\alpha\beta)_{ij}^2 + \sigma^2}$$

$$e_I(X_1 \longrightarrow Y) = \frac{(1/IJ)\sum_i \sum_j (\alpha\beta)_{ij}^2}{(1/I)\sum_i \alpha_i^2 + (1/J)\sum_j \beta_j^2 + (1/IJ)\sum_i \sum_j (\alpha\beta)_{ij}^2 + \sigma^2}$$

so that the contribution ratio of X_1 is

$$CR(X_1) = \frac{(1/I)\sum_i \alpha_i^2 + (1/IJ)\sum_i \sum_j (\alpha\beta)_{ij}^2}{(1/I)\sum_i \alpha_i^2 + (1/J)\sum_j \beta_j^2 + (1/IJ)\sum_i \sum_j (\alpha\beta)_{ij}^2}$$

and, similarly,

$$CR(X_2) = \frac{(1/J)\sum_j \beta_j^2 + (1/IJ)\sum_i \sum_j (\alpha\beta)_{ij}^2}{(1/I)\sum_i \alpha_i^2 + (1/J)\sum_j \beta_j^2 + (1/IJ)\sum_i \sum_j (\alpha\beta)_{ij}^2}$$

The two contribution ratio do not sum to one, as no causal ordering between the factors exist. To measure the relative importance of the two explanatory variables, relative to the whole set, the approach in (12) can then be applied:

$$RI(X_1) = \frac{(1/I)\sum_i \alpha_i^2 + \frac{1}{2IJ}\sum_i \sum_j (\alpha\beta)_{ij}^2}{(1/I)\sum_i \alpha_i^2 + (1/J)\sum_j \beta_j^2 + (1/IJ)\sum_i \sum_j (\alpha\beta)_{ij}^2}$$

$$RI(X_2) = \frac{(1/J)\sum_j \beta_j^2 + \frac{1}{2IJ}\sum_i \sum_j (\alpha\beta)_{ij}^2}{(1/I)\sum_i \alpha_i^2 + (1/J)\sum_j \beta_j^2 + (1/IJ)\sum_i \sum_j (\alpha\beta)_{ij}^2}.$$

To illustrate the indexes above, consider the dataset in Table 2, used by Daniel (1999). The time Y (minutes) spent on individual home visits by a sample of 80 public health nurses was recorded, along with the kind of disease X_1 of the patient and the age group X_2 of the nurse. Both factors X_1 and X_2 have four levels. Moreover, there is no causal ordering between the factors and each of them is likely to impact on the response variable and to interact with the other factor. Then, the direct effect of each factor can be defined as that given the other factor and the indirect effect is that through their interaction. After estimating the model, simple calculations show that

$$\frac{1}{4}\sum_{i=1}^{4} \alpha_i^2 = 37.4056 \qquad \frac{1}{4}\sum_{j=1}^{4} \beta_j^2 = 15.0131 \qquad \frac{1}{16}\sum_{i=1}^{4}\sum_{j=1}^{4} (\alpha\beta)_{ij}^2 = 7.6056.$$

Moreover, the variance of the model σ^2 is estimated as 11.7450. The total effect of both factors on the response variable, i.e. the ECD, is then

$$e_T((X_1, X_2) \longrightarrow Y) = ECD = \frac{37.4056 + 15.0131 + 7.6056}{71.7693} = 0.8363,$$

TABLE 2. - *Data from 80 public health nurses about the time (minutes) spent on individual home visits, by age group of the nurse and kind of disease of the patient*

	X_2 = age group			
	(20 to 29)	(30 to 39)	(40 to 49)	(50 and over)
X_1 = type of patient				
Cardiac	20	25	24	28
	25	30	28	31
	22	29	24	26
	27	28	25	29
	21	30	30	32
Cancer	30	30	39	40
	45	29	42	45
	30	31	36	50
	35	30	42	45
	36	30	40	60
C.V.A.	31	32	41	42
	30	35	45	50
	40	30	40	40
	35	40	40	55
	30	30	35	45
Tuberculosis	20	23	24	29
	21	25	25	30
	20	28	30	28
	20	30	26	27
	19	31	23	30

that is 83.63% of entropy is explained by the two factors. The contributions of the factors are calculated as follows:

$$e_T(X_1 \longrightarrow Y) = \frac{37.4056 + 7.6056}{71.7693} = 0.6272$$
$$e_D(X_1 \longrightarrow Y) = \frac{37.4056}{71.7693} = 0.5212$$
$$e_T(X_2 \longrightarrow Y) = \frac{15.0131 + 7.6056}{71.7693} = 0.3152$$
$$e_D(X_2 \longrightarrow Y) = \frac{15.0131}{71.7693} = 0.2092.$$

Hence, the factor contribution ratios are:

$$CR(X_1) = \frac{37.4056 + 7.6056}{37.4056 + 15.0131 + 7.6056} = 0.7499$$
$$CR(X_2) = \frac{15.0131 + 7.6056}{37.4056 + 15.0131 + 7.6056} = 0.3768.$$

The contribution of X_1 on Y is then about twice greater than that of X_2. The relative importance of the explanatory variables are instead measured as:

$$RI(X_1) = \frac{37.4056 + (1/2)(7.6056)}{37.4056 + 15.0131 + 7.6056} = 0.6865$$
$$RI(X_2) = \frac{15.0131 + (1/2)(7.6056)}{37.4056 + 15.0131 + 7.6056} = 0.3135.$$

Notice that, since the interactive effect is relatively small, the contribution ratios and the corresponding relative-importance degrees are quite similar here. However, only the latter measures sum to one.

4. SUMMARY AND CONCLUSIONS

In most data analyses, GLMs are widely used to explain the effect of a set of explanatory variables on a response variable, under several distributional assumptions. Except for the ordinary linear regression model, however, only the estimation of the coefficients of variables and the related statistical tests are usually performed in practical analyses using GLMs. The assessment of the predictive power of the model and, above all, of the relative importance of explanatory variables is rarely carried out, probably due to the lack of a widely accepted approach in the literature. The present paper proposes to use an entropy-based approach for both kinds of assessment. The predictive power of the model is measured by the ECD in Eshima and Tabata (2007, 2010, 2011), which can be interpreted as the proportion of the variation in entropy of the response variable, explained by the chosen set of regressors.The variation in entropy is also used to assess the total effect of a single variable, so that a relative contribution measure can be defined as the ratio of this effect to the joint effect of all explanatory variables. To isolate the direct effect of a single variable, however, one needs to consider a causal order in the set of regressors, so as to develop a kind of a path analysis (Eshima *et al.*, 2015). Being that the occurrence of a causal order is quite rare in applications, then, this paper proposes to take an average over all possible orderings, a commonly used method in the literature for such kinds of problems (see Kruskal, 1987; Lindeman *et al.*, 1980). The resulting relative-importance assessment provides a decomposition of the ECD with regard to the explanatory variables with non-zero regression coefficients and it is able to satisfy criteria (a)-(d), commonly required for a relative-importance measure in linear regression, as listed in Section 1. Moreover, as shown by the examples presented in Section 3, the proposed method can be applied to all GLMs, not only when the response variable is categorical rather than continuous, but also when the explanatory variables are fixed factors.

ACKNOWLEDGEMENT

This research was supported by Grant-in-aid for Scientific Research 26330045, Ministry of Education, Culture, Sports, Science and Technology of Japan.

REFERENCES

Agresti A. (2002). *Categorical Data Analysis, Second Edition*. John Wiley & Sons, New York.

Chevan A., Sutherland M. (1991). Hierarchical Partitioning. *The American Statistician*, **45**, 90-96.

Daniel W.W. (1999). *Biostatistics: A Foundation for Analysis in the Health Sciences, Seventh Edition*. John Wiley & Sons, New York.

Darlington R.B. (1968). Multiple regression in psychological research and practice. *Psychological Bulletin*, **69**, 161-182.

Eshima N., Tabata M. (2007). Entropy correlation coefficient for measuring predictive power of generalized linear models. *Statistics and Probability Letters*, **77**, 588-593.

Eshima N., Tabata M. (2010). Entropy coefficient of determination for generalized linear models. *Computational Statistics and Data Analysis*, **54**, 1381-1389.

Eshima N., Tabata M. (2011). Three predictive power measures for generalized linear models: the entropy coefficient of determination, the entropy correlation coefficient and the regression correlation coefficient. *Computational Statistics and Data Analysis*, **55**, 3049-3058.

Eshima N., Tabata M., Borroni C.G., Kano Y. (2015). An entropy-based approach to path analysis of structural generalized linear models: a basic idea. *Entropy*, **17**, 5117-5132.

Feldman B. (1999). *Manuscript for a Contributed Paper at the Econometric Society World Congress 2000*. Available at http://fmwww.bc.edu/RePEc/es2000/1140.pdf

Hoffman P.J. (1960). The paramorphic representation of clinical judgment. *Psychological Bulletin*, **57**, 116-131.

Kruskal W. (1987). Relative importance by averaging over orderings. *The American Statistician*, **41**, 6-10.

Lindeman R.H., Merenda P.F., Gold R.Z. (1980). *Introduction to Bivariate and Multivariate Analysis*. Scott Foresman, Glenview (IL).

McCullagh P., Nelder J.A. (1989). *Generalized Linear models, Second Edition*. Chapman and Hall, London.

Nelder J.A., Wedderburn R.W.M. (1972). Generalized linear model. *Journal of the Royal Statistical Society - Series A*, **135**, 370-384.

Pratt J.W. (1987). Dividing the indivisible: using simple symmetry to partition variance explained. In T. Pukkila and S. Puntanen (Eds.), *Proceedings of the Second Tampere Conference in Statistics* (pp. 245-260). University of Tampere, Finland.

Theil H., Chung C.F. (1988). Information-theoretic measures of fit for univariate and multivariate linear regressions. *The American Statistician*, **42**, 249-253.

Williams E.J. (1978). Postscript to 'Linear hypotheses: regression'. In W.H. Kruskal and J.M. Tanur (Eds.), *International encyclopedia of statistics* (pp. 537-541). Free Press, New York.