

DECOMPOSITION OF VARIANCE IN TERMS OF CONDITIONAL MEANS

A. Figà Talamanca, A. Guerriero, A. Leone, G. P. Mignoli, E. Rogora

1. INTRODUCTION AND METHODOLOGY

Let $X = (x_1, \dots, x_N)$ be a numerical variable defined on a population P of N individuals. We may think of X as an element of a real vector space \mathbf{L} of dimension N . We equip \mathbf{L} with a real, normalized scalar product: for $X, Y \in \mathbf{L}$, and $Y = (y_1, \dots, y_N)$, we define:

$$\langle X, Y \rangle = \frac{1}{N} \sum_{i=1}^N x_i y_i.$$

The length or norm of a vector is defined in terms of the scalar product:

$$\|X\|^2 = \langle X, X \rangle.$$

The mean value of a vector X is of course the scalar

$$\bar{X} = \frac{1}{N} \sum_{i=1}^N x_i.$$

We may also think of the mean value as a vector $E_0(X)$ of \mathbf{L} having all its components equal to the scalar \bar{X} . In this context E_0 may be thought of as a linear operator defined on \mathbf{L} and mapping \mathbf{L} into the subspace of constant vectors. The variance of X can be written then as:

$$V(X) = \|X - E_0(X)\|^2 = \langle X - E_0(X), X - E_0(X) \rangle.$$

We now suppose that the indices $i = 1, \dots, N$, correspond to individuals of a population P , and that X is a numerical variable defined on the population P . We further suppose that π is a partition of the population P into q disjoint classes

P_1, P_2, \dots, P_q . Denote by $|P_j|$ the number of elements of P_j , so that $N = |P_1| + \dots + |P_q|$. We can then define a vector $E_\pi(X)$ with components:

$$E_\pi(X)_i = \frac{1}{|P_k|} \sum_{j \in P_k} x_j \quad (i \in P_k). \quad (1)$$

Observe that two components of this vector are identical if their indices belong to the same class P_k of the partition π . The trivial identity:

$$X - E_0(X) = E_\pi(X) - E_0(X) + X - E_\pi(X),$$

implies

$$V(X) = \|X - E_0(X)\|^2 = \|E_\pi(X) - E_0(X)\|^2 + \|X - E_\pi(X)\|^2,$$

because, as it is easily seen, $E_\pi(X) - E_0(X)$ and $X - E_\pi(X)$ are orthogonal vectors.

Suppose now that $\pi_1, \pi_2, \dots, \pi_n$ is a finite sequence of partitions of the population P , into respectively q_1, q_2, \dots, q_n , classes. Suppose further that each partition π_j is a refinement of the partition π_{j-1} . (This means that each class of the partition π_j is contained in a class of the partition π_{j-1}). Define for completeness the trivial partition π_0 consisting of the full population P . Let P_k^j , for $k=1, \dots, q_j$ be the disjoint classes of the population P relative to the partition π_j . With reference to the partition π_j define the operator

$$E_j(X) = E_{\pi_j}(X).$$

In this fashion (1) reads:

$$E_{\pi_j}(X)_i = \frac{1}{|P_k^j|} \sum_{b \in P_k^j} x_b, \quad (i \in P_k^j).$$

Observe that this definition makes sense also in the case $j=0$. The trivial identity

$$X - E_0(X) = \sum_{j=1}^n [E_j(X) - E_{j-1}(X)] + X - E_n(X). \quad (2)$$

implies, because of the orthogonality of the terms on the right hand side of (2),

$$V(X) = \sum_{j=1}^n \|E_j(X) - E_{j-1}(X)\|^2 + \|X - E_n(X)\|^2. \quad (3)$$

We are interested in the case in which the sequence of partitions π_j is defined by a sequence of qualitative characters C_1, C_2, \dots, C_n of the population P . We can define the partition π_j by considering the classes of the population formed by individuals with identical values of the first j characters.

In this case the first n summands on the right hand side of (3) represent the contributions to the variance of the n qualitative characters C_1, \dots, C_n within the population considered.

Observe however that, while the sum of the first n terms of the right hand side of (3) is independent of the order in which the characters C_1, \dots, C_n are considered, the operators E_j , for $0 < j < n$ are defined with respect to partitions which strongly depend on the order in which the characters are taken. As an obvious consequence, the value of each term $\|E_j(X) - E_{j-1}(X)\|^2$ also depends on the order of the characters. In a different order the characters would define a different set of partitions; only π_0 and π_n , and consequently E_0 and E_n are independent of the chosen order.

We are led therefore to look for a natural order of the qualitative characters considered. We propose an ordering based on systematic, step by step, comparisons of the conditional means with respect to the variables considered. This ordering, which we call *Stepwise Optimal Ordering* (SOO) is defined as follows:

We choose the character C_1 and the corresponding partition π_1 which maximizes $\|E_1(X) - E_0(X)\|^2$. If C_1, \dots, C_k are chosen, the character C_{k+1} is chosen so that it refines the partition π_k into the partition π_{k+1} in such a way that the value $\|E_{k+1}(X) - E_k(X)\|^2$ is largest.

The order C_1, \dots, C_n determined in this fashion may be considered as a ranking of the variables. One should be aware, however, that this ranking cannot be interpreted in terms of relative importance in determining the phenomenon measured by the variable X . As will be seen in the applications below, the qualitative characters considered may be far from independent. This may imply that a character which is recognized as a primary cause of the intensity of the phenomenon measured by X , may be mediated by other characters to whom it is associated, and therefore appear in the last positions of the ranking.

We do not propose a clear cut interpretation of the significance of the ranking obtained by our method, nor of the relative size of the first n addends which appear in (3), when the qualitative characters are ordered according to our prescription. On the contrary, rather than expecting straight answers, we expect that both

the ranking and the relative size of the addends in the expression (3) would solicit questions concerning the dependence of the variable X on the qualitative variables and the interdependence of the qualitative variables themselves (with all the cautions regarding the possibility to consider causal relations between the variables, (Blalock, 1961; Sobel, 1996; Sobel, 1998; Sobel, 2000)).

Nevertheless, in the very special case considered in the simulated experiment of section 4, our method yields a ranking that reflects the relative weight of the characters.

In the following two sections we apply our method and discuss the “ranking” of the qualitative characters, thus obtained to two sets of data. The fourth section is dedicated to a simulated experiment.

We should mention that the ideas contained in chapter 8D of Diaconis (1988) were influential in the inception of this work, which started as an attempt to apply Diaconis' ideas to the case of tree-structured data, under the action of the group of tree-automorphisms. Under this action the ranges of the operators $E_j - E_{j-1}$ turn out to be irreducible subspaces of \mathbf{L} .

2. THE SCORE ON AN ENTRANCE EXAMINATION

Entering students of the University of Rome “La Sapienza” in scientific and technical fields take a multiple choice test in mathematics, which consists of 30 questions¹. At the moment the purpose of the test is to discourage students who do not have an adequate background, and to make students aware of their potential weaknesses.

We consider a population of 2,451 students who took the test in 2005, and we let X be the score achieved by each student, that is the number of correct answers. The variable X depends on the 30 dyadic characters, corresponding to the correct or incorrect answer to each question. Of course, in this case, $E_{30}(X) = X$, and

$$V(X) = \|X - E_0(X)\|^2 = \sum_{j=1}^{30} \|E_j(X) - E_{j-1}(X)\|^2.$$

The variable X takes values between 0 and 30. Its mean value is 12.9 and the variance is $V(X) = 29.8$. The histogram of X is in Fig. 1.

¹ The test, in Italian, is available for downloading at the internet address <http://www.mat.uniroma1.it/people/rogora/pdf/test.pdf>.

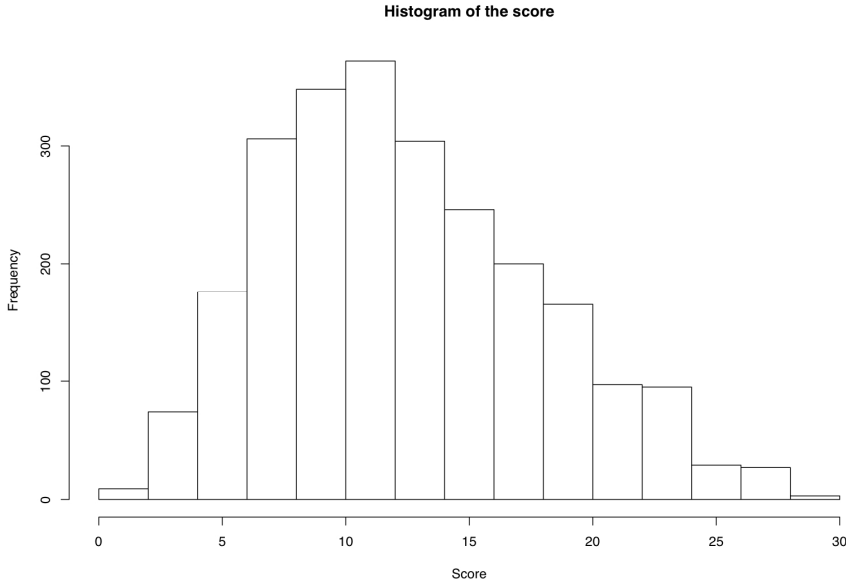


Figure 1 – The histogram of the score.

An application of our method shows that just ten questions, chosen according to the ranking we propose, “explain” 88% of the variance. In other words, if we write

$$V(X) = \sum_{j=1}^{10} \|E_j(X) - E_{j-1}(X)\|^2 + \|X - E_{10}\|^2.$$

the remainder term $\|X - E_{10}\|^2 = 3.58$ amounts to just 12% of $V(X) = 29.8$.

We presently list the remainders $\|X - E_k\|^2$, for $k=1, \dots, 10$, obtained by applying our method, as percentage of $V(X)$. To wit the values

$$c_k = \|X - E_k\|^2 / V(X),$$

$$c_1 = \frac{75}{100}, \quad c_2 = \frac{59}{100}, \quad c_3 = \frac{48}{100}, \quad c_4 = \frac{40}{100}, \quad c_5 = \frac{34}{100},$$

$$c_6 = \frac{29}{100}, \quad c_7 = \frac{25}{100}, \quad c_8 = \frac{20}{100}, \quad c_9 = \frac{16}{100}, \quad c_{10} = \frac{12}{100}.$$

We do not claim that our method necessarily chooses the 10 characters for which $\|X - E_{10}(X)\|^2$ is lowest. In general, with arbitrary data, this may not be the case.

However, in this particular case, our choice compares well with other possible choices, as shown by the experiment which we presently describe. We selected, at random, 300 subsets of ten elements of the original thirty questions and we computed the conditional mean $E_\pi(X)$ with respect to the partition π obtained by grouping together the students with identical performance on each of the ten question chosen. We computed then

$$\|X - E_\pi(X)\|^2, \quad (4)$$

relative to each “ten element” choice. The results are summarized in Fig. 2.

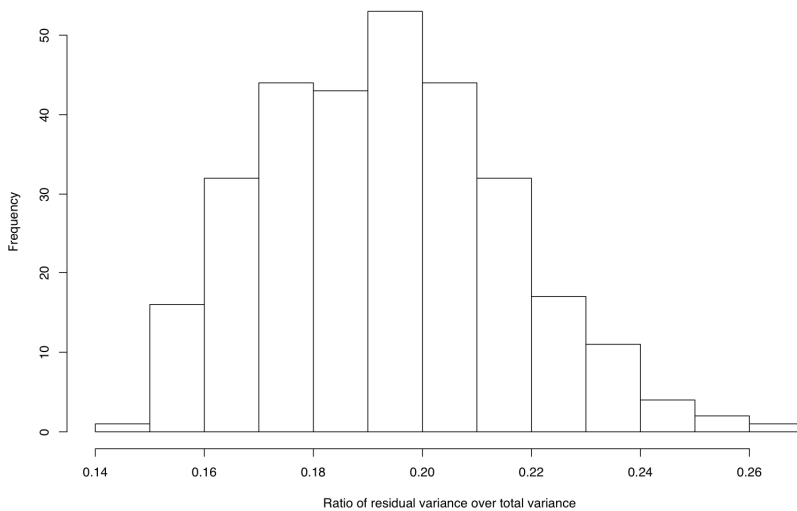


Figure 2 – Histogram of the values of residual variance (4), as percentage of total variance for 300 randomly selected subsets of 10 questions.

Observe that the lowest value of quantity (4) achieved by one of the 300 subsets we selected, is higher than 0.14, while with the SOO choice of ten characters we achieved the value 0.12.

The experiment shows that the algorithm we propose performs decidedly better than a random choice if we want to choose ten out of thirty questions, in such a way that the total variance of the variable X is best explained. In conclusion there is at least some experimental evidence that our method may be used to select a small number of characters which account for most of the variance.

The variables selected according to SOO discriminate the students better than the other variables, Indeed, if we rank the items according to the item discrimination index of classical test theory, the ten items selected by SOO find place among the first 11 items. We also estimated, according to Rasch model, the difficulty of the 30 items. It turns out that the ten items chosen according to SOO occupy a middle position. Indeed the 10 items thus chosen place themselves be-

tween the eight and the twentieth position in the ranking. Finally, we note that eight of the ten selected variables are among the ten most important variables in terms of linear regression and the order of the first five variables coincides under both methods.

3. THE VARIABLE “DELAY IN COMPLETING A DEGREE”

The Italian system of higher education is characterized by the marked difference between the time employed by most students to complete a degree and the number of years formally required to graduate. The average delay in completing a degree is well above two years for most fields of study². In this section we consider a population of Italian university graduates obtained using the data bank “AlmaLaurea” which collects data of university graduates from a set of Italian universities³. The population amounts to 58,091 graduates of 27 universities in 2003. On this population the variable X represents the delay in completing the degree, computed in years, starting from a conventional date (November 1st) in which according to formal regulations the degree should have been completed. We excluded delays above ten years, which should be better interpreted as leaving and resuming the studies after several years. We study the dependence of X on seven possible characters, which are the following:

- (UN) University where the degree was obtained
- (PE) Parent's level of education
- (HS) Type of high school attended
- (GD) Grade in the final year of high school
- (MA) Degree major
- (WO) Working or not working during the studies
- (GN) Gender

Proceeding as outlined in the introduction, we obtain the following ranking of the seven variables:

GD, UN, MA, HS, PE, WO, GN.

Accordingly we consider the operators

$$E_0, E_1, E_2, E_3, E_4, E_5, E_6, E_7,$$

and write

² The recent reform of the university system may hopefully change this in the near future.

³ AlmaLaurea Consortium is an association of 49 Italian universities which, since 1994 collects statistical data about the scholastic and employment records of university graduates (Cammelli, 2005; Cammelli, 2006). The data bank of AlmaLaurea is also made available, under certain conditions, to prospective employers.

$$V(X) = \sum_{j=1}^7 \|E_j(X) - E_{j-1}(X)\|^2 + \|X - E_7(X)\|^2 \quad (5)$$

The variance of the variable X is $V(X) = 4.61$, while the residual variance, not “explained” by the qualitative variables under consideration is $\|X - E_7(X)\|^2 = 1.94$. The decomposition of the variance (3) is:

$$4.61 = (0.30 + 0.28 + 0.49 + 0.45 + 0.51 + 0.33 + 0.31) + 1.94 = 2.67 + 1.94$$

Thus 2.67 represents the portion of the variance which is “explained” by the characters considered. We may say, therefore, that these characters explain 62% of the variance.

In this case the ranking obtained by our method is relatively “robust”. Indeed if we omit consideration of one of the characters, the relative ranking of the other characters remains unchanged. We do not claim of course that this type of “robustness” is inherent in our method. It may very well occur, with different data, that omitting one character would determine a change in the order of the remaining ones.

We compared our results with those obtained by using the binomial logistic regression and the multivariate analysis of variance. Notice, however, that these methods, unlike SOO, require hypothesis on the distribution of data which are not met by our example. On the other hand, SOO is not suitable for inferential purposes.

As for the binomial logistic regression, the response variable was dichotomized assigning the value zero to the population of graduates with a delay of less than one year (34.1%) and value one to the others (65.9%). This is of course an arbitrary choice, which has to be made to apply the method. As a measure of the effect produced by each independent variable we use the standard deviation of the theoretic probabilities associated to the values of the independent variables⁴. The results of our computations are summarized in table 1.

TABLE 1
*Effect of the 7 independent variables on the probability to graduate within one year
logistic binomial regression analysis*

Variable	SOO	Effect
GD	1	0.11
UN	2	0.09
MA	3	0.10
HS	4	0.06
PE	5	0.04
WO	6	0.03
GN	7	0.01

⁴ More precisely, let $p(j,i)$ be the mean of the probabilities of graduating within one year after forcing, in our population, the j variable to take the value i . The effect of the variable j is then defined to be the standard deviation of the numbers $p(j,i)$, weighted by the numbers $n(j,i)$ of individuals for which the j -th variable takes the value i in the original population.

We observe that the ranking of the variables determined by the effects on the dependent variable coincides, except for one inversion, with the ranking obtained by SOO. It should be noted that in the regression model we inserted only the principal effects of the independent variables and not the possible interaction between them.

The comparison of SOO with multivariate analysis of variance yields almost exactly the same results.

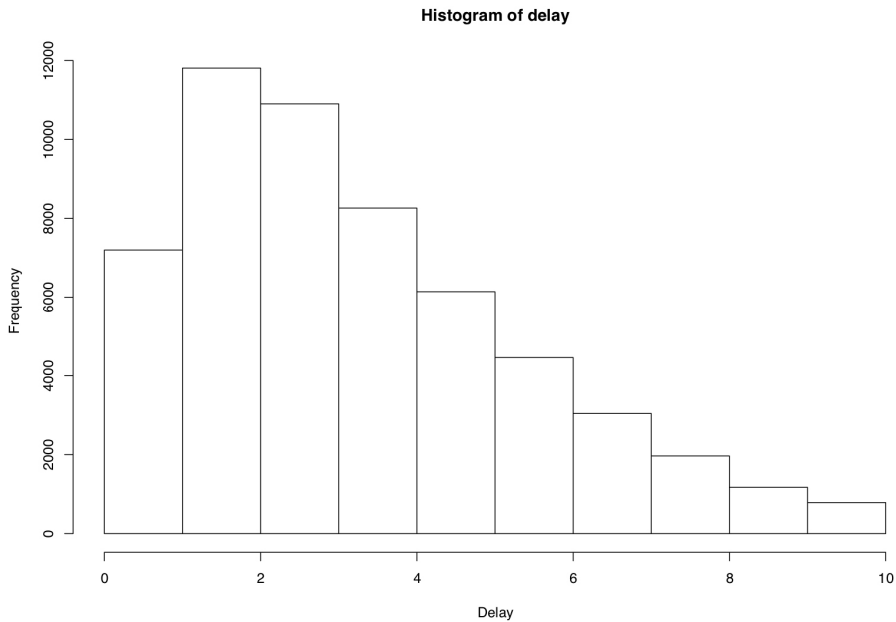


Figure 3 – Histogram of the delay.

4. A SIMULATED EXPERIMENT

In order to better understand the properties of our Stepwise Optimal Order, we performed a simulation, repeating 20 times the following experiment.

First we constructed 10 vectors x_1, \dots, x_{10} each of 100 components and each component extracted from a simulated Bernoulli variable. Then we considered the variable

$$x = c_1 x_1 + c_2 x_2 + \dots + c_{10} x_{10} + \varepsilon \quad (6)$$

with $c_1 = 1, c_2 = 0.9, \dots, c_{10} = 0.1$ and ε consisting of 100 independent realizations of a simulated Gaussian variable with mean 0 and standard deviation 0.03.

In 18 cases out of the 20 observed experiments, SOO was exactly 1, 2, 3, ..., 10, i.e. for the variable x this order reflected, most of the time, the size of the co-

efficients c_1, \dots, c_{10} which enter formula (6). In the remaining two cases the difference between SOO and the increasing order was just one inversion.

5. CONCLUSIONS

We conclude that the application of our method to numerical variables which depend on qualitative variables, may yield valuable information and insights on the dependence of the numerical variable on the qualitative variables, and the interdependence of the qualitative variables themselves. We observe that in the simulated experiment described in the last section the ranking obtained by our method reflects the relative weight of the variables. Nevertheless we do not propose, in general, a clear cut interpretation of the significance of this ranking. Finally we observe that the role of the qualitative variables in the two sets of data on which we test our method is different. In the first example the 30 variables are measures of the response variable, rather than causes of it, while in the second example the qualitative variables may be interpreted as “causes” of the numerical variable and the interpretation of the results may very well be different.

Department of Mathematics
University of Rome “La Sapienza”

ALESSANDRO FIGÀ TALAMANCA

Alma Laurea

ANGELO GUERRIERO
ALBERTO LEONE
GIAN PIERO MIGNOLI

Department of Mathematics
University of Rome “La Sapienza”, rogora@mat.uniroma1.it

ENRICO ROGORA

REFERENCES

- H.M.JR. BLALOCK, (1961), *Causal inferences in Nonexperimental Research*, Chapel Hill, University of North Carolina Press, 1961.
- A. CAMMELLI, (2005), La qualità del capitale umano dell'università. Caratteristiche e performance dei laureati 2003, in *La qualità del capitale umano dell'università in Europa e in Italia*, ed. Cammelli A., Bologna, Il Mulino, 2005.
- A. CAMMELLI, (2006), La riforma alla prova dei fatti, in *Settimo profilo dei laureati italiani*, ed. Consorzio interuniversitario AlmaLaurea, Bologna, Il Mulino, 2006.
- P. DIACONIS, (1988), *Group Representations in Probability and Statistics*, Lecture notes - Monograph series volume 11, Institute of Mathematical Statistics, Hayward, California, 1988.
- M.E. SOBEL, (1996), An Introduction to Causal Inference, in *Sociological Methods and Research*, 24, 1996, pp. 353-379.
- M.E. SOBEL, (1998), Causal Inference in Statistical Models of the Process of Socioeconomic Achievement, in *Sociological Methods and Research*, 27, 1998, pp. 318-348.
- M.E. SOBEL, (2000), Causal Inference in the Social Sciences, in *Journal of the American Statistical Association*, 95, 2000, pp. 647-651.

SUMMARY

Decomposition of variance in terms of conditional means

Two different sets of data are used to test an apparently new approach to the analysis of the variance of a numerical variable which depends on qualitative variables. We suggest that this approach be used to complement other existing techniques to study the interdependence of the variables involved. According to our method, the variance is expressed as a sum of orthogonal components, obtained as differences of conditional means, with respect to the qualitative characters. The resulting expression for the variance depends on the ordering in which the characters are considered. We suggest an algorithm which leads to an ordering which is deemed natural. The first set of data concerns the score achieved by a population of students on an entrance examination based on a multiple choice test with 30 questions. In this case the qualitative characters are dyadic and correspond to correct or incorrect answer to each question. The second set of data concerns the delay to obtain the degree for a population of graduates of Italian universities. The variance in this case is analyzed with respect to a set of seven specific qualitative characters of the population studied (gender, previous education, working condition, parent's educational level, field of study, etc.)