Original Articles:
Quantitative Approach in Occupational Health

International Physiology
Vol. 3 No. 1, January - June 2015
DOI: https://dx.doi.org/10.21088/ip.2347.1506.3115.7

Predictability of a Set of Physiological Variables form a Set of Anthropometric Variables for a Data Set from Fishermen

Babulal Seal*, Sanchayita Sadhu, Baidyanath Pal*****

Abstract

In univariate case, general measure of the degree of dependency of one variable on another is a correlation ratio. In multivariate setup earlier it was defined such through trace of conditional covariance matrix of a sub-vector given the other sub-vector divided by the trace of the variance-covariance matrix of the former and also obtained some results on such measure for predictability. But more appropriate generalization of such measure is defined here through eigenvalues of the dispersion matrix of the predictor relative to whole dispersion matrix. Then it is applied to see how far physiological variables depend on anthropometric variables. It is found that chosen physiological variables have higher degree of dependency on the selected anthropometric variables.

Key words: Correlation Ratio; Measure of Predictability; AMS Classification: 62H, 62P.

Introduction

To guess the physiological pattern or many psychological pattern of a human being from the constitution and structure of his or her body is an old custom. Also this is one of the important principles in Organon of Medicine. That is why it is quite reasonable to understand the physiological variables from anthropometric variables. This work tries to find out how much we may understand the physiological characteristics depending on some anthropometric variables. For diseases or smooth function of a body it is necessary to understand many physiological variables and here it is to see how these variables are susceptible to anthropometric variables. Here it is an attempt to see this. For this purpose, data from Digha Fishermen are used. The reason is that they are localised in that area for many years and they have their own identity, structure, similar professions etc. Similar kind of things can be used for other populations also, where variations in physiological characteristic are not too different for a given anthropometric variables. As for each such set we estimate the physiological variables by their means and that is why for too much mixing population this means vector will be badly used.

Here we have three physiological variables which are usually evaluated from any routine check up of the body conditions and we have many anthropometric variables but some of these are collinear and after filtering collinearity from these collections we have height (ht), weight (wt), chest circumference normal (ccn) and fat mass (fm) as anthropometric variables. So, the first thing is to know about the relationship between these two multivariate sub-vectors. The simpler way to get the relationship is to get by multivariate linear regression. But this relationship may not be suitable. There may be other type of relation.

Now, the question is how far the relationship or the "degree of relationship" can be assured from the data set. Best thing is to calculate or develop correlation ratio in this case. In bivariate case we all know this and this was developed by Pearson long back. But in multivariate set up, though there is a kind of work by Sampson (1984), we hope that it can be modified. The present paper gives a modification of that and by using the modified formula the "degree of such prediction" to the data set is obtained.

In section 2, a discussion on preliminaries on earlier works is given. In section 3, the concept of the revised present works is given. In section 4, the

Author's Affiliations: * Department of Statistics, ** INSPIRE Fellow, Department of Statistics, The University of Burdwan. *** Biological Anthropology Unit, Indian Statistical Institute, Kolkata.

Corresponding Author: Babulal Seal, Department of Statistics, The University of Burdwan, Rajbati, Burdwan, West Bengal - 713104

Email: babulal_seal@yahoo.com

revised measure obtained in section 3 is applied to a real life data set on physiologic and anthropometric variables.

Discussion on Some Preliminaries

In bivariate case correlation ratio is a measure of relationship between the statistical dispersion within categories and the dispersion across the whole population or sample. The measure is taken as the ratio of two standard deviations representing these types of variation as in the following.

$y_{11,}y_{12},...,y_{1n_1}$ *for* x_1,

If the observations on responses are:

$y_{21}, y_{22},...,y_{2n_2}$ *for* x_2,

. .
. .
. .

$y_{k1}, y_{k2},...,y_{kn_k}$ *for* x_k

Then,

$$\eta^2 = \frac{\sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_k}(y_{ij} - \bar{y})^2} \qquad (2.1)$$

is a measure.

It is to be noted that if the relationship between *x* and *y* is linear, this will give the same result as the square of Pearson's correlation coefficient. Otherwise correlation ratio will be larger in magnitude. It can therefore be used for deciding non-linear relationship.

The correlation ratio was introduced by Karl Pearson as a part of analysis of variance. A multivariate version of this was done by Allan R. Sampson (1984). A multivariate correlation ratio of a random vector **Y** upon a random vector **X** was defined by him as:

$$\eta_{\Lambda}(\mathbf{Y};\mathbf{X}) = \{tr(\Lambda^{-1}CovE(\mathbf{Y} \mid \mathbf{X})\}^{1/2}\{tr(\Lambda^{-1}\Sigma_y)\}^{-1/2}, \qquad (2.2)$$

where, Λ is a fixed positive definite matrix related to the relative importance of predictability for entire variation of **Y,** and Σ_y **is** dispersion matrix of Y.

The previous formula enjoys the important properties

$$\min_g E(\mathbf{Y} - g(\mathbf{X})^2 = (1 - \eta^2(\mathbf{Y};\mathbf{X})Var\mathbf{Y}$$

where, $g(\mathbf{X}) = E(\mathbf{Y} \mid \mathbf{X} = x)$ $g(\mathbf{X}) = E(\mathbf{Y} \mid \mathbf{X} = x)$

Concept of the Present Work

Here anthropometric variables are used to predict some important physiological variables. This is a routine analysis through multivariate regression. A close work to have a measure of importance of a variable to predict the dependent variable was considered theoretically by Seal et al, 2015 in a different and vivid manner, although traditionally one uses stepwise regression, AIC etc. That method was used to find most important variables for weight.

But we want to find out a degree of maximum prediction. That may be linear or non-linear. In univariate case i.e. one univariate variable depending on another univariate variable, the measure of such degree of prediction is correlation ratio. When both **X** (anthropometric) and **Y** (physiological) are multivariate, we want to find out a generalisation of the correlation ratio.

Though, in literature a kind of work (i.e. multivariate correlation ratio, as states earlier) is available through trace of a variance-covariance matrix of conditional expectation divided by the trace of variance covariance matrix. It is evident that the expression (3.1) is more close to (2.1) than (2.2). So we change the definition a little which is given in (3.1). Now looking at this matrix may be simplified by its eigenvalues separately, trace, determinant or the simplest form in terms of its maximum eigenvalue. These can be interpreted in terms of spread of data. We have data set on physiologic variables and anthropometric variables of persons from fishermen of Digha. For this data set the degree of prediction is obtained.

Thus, the modified degree of prediction along this line is defined by the maximum eigen- value of

$$S_2^{-1/2} S_1 S_2^{-1/2} \qquad (3.1)$$

where, S_1 is the within sum of square and product matrix and S_2 is the sum of square and product matrix Now let us consider the data set from Digha fishermen containing 252 observations including anthropometric, demographic and physiological

variables (source of data: Plan project survey data of BAU, ISI, Kolkata-2008). The independent variables are (i) age (demographic variable) and 4 anthropometric variables (ii) height (ht), (iii) weight (wt), (iv) chest circumference normal (ccn), and (v) fat mass (fm). These five variables are considered as covariates. Systolic blood pressure (sbp), diastolic blood pressure (dbp) and pulse rate (pr) are the three response variables. Some other covariates were there but due to multi-colinearity others were discarded.

Covariates are divided into some intervals such that within each interval persons have same type of physical characteristics as in the following. However, other appropriate choices may be applied. Let us demonstrate our formula for the present data.

Age intervals are:

[18, 22], [23, 30], [31, 40], [41, 50], [51, 60], [61, 70], [71, 77]

Height (cm) intervals are:

[142, 155], [155.001, 170], [170.001, 180]

Weight (kg) intervals are:

[35, 45], [45.001, 55], [55.001, 65], [65.001, 75], [75.001, 85]

Chest Circumference Normal intervals are:

[66, 80], [80.001, 100], [100.001, 110]

Fat Mass intervals are:

[2, 10], [10.001, 20], [20.001, 28]

Then all possible combinations (7×3×5×3×3) are considered. We have taken readings on sbp, dbp, and pr for each of the combinations. Some of the combinations do not have readings from our practical data. These are ignored and after that we get 50 groups where observations are available.

Now to obtain within variance covariance matrix we consider from all these 50 groups 50 matrices such that each is obtained from the columns of sbp, dbp, and pr respectively. So we get $\sum_{i=1}^{50} n_i(\bar{y}_i - \bar{y})(\bar{y}_i - \bar{y})'$ where n_i is number of observations in i^{th} group.

From the data set $\bar{y}_i$ are summarized in the following table.

We find the grand mean vector $\bar{y}$ by finding the mean of each column of the mentioned mean vectors. Let us consider the difference between each group vectors and the grand mean i.e. $\bar{y}_i - \bar{y}$. After that $(y_i - \bar{y})(y_i - \bar{y})'$ is obtained and then multiplying the group matrix by the corresponding number of observations and then taking their sum we get S_1 (say).

Now consider the difference between sbp, dbp, pr and the grand mean for each individual and then consider their sum of square and product matrix S_2 (say).

So,

$$S_1 = \begin{pmatrix} 14282.231 & 6719.695 & 1374.334 \\ 6719.695 & 6891.891 & 1433.982 \\ 1374.334 & 1433.982 & 6465.473 \end{pmatrix}$$

and

$$S_2 = \begin{pmatrix} 53753.2170 & -3092.379 & 636.7145 \\ -3092.3787 & 31533.343 & 21135.2275 \\ 636.7145 & 21135.228 & 28560.6688 \end{pmatrix}$$

So, computing the matrix $(S_2)^{-1/2} S_1 (S_2)^{-1/2}$, the final matrix to draw conclusion is obtained as

$$\begin{pmatrix} 0.28373878 & 0.2307312 & -0.07153218 \\ 0.23073124 & 0.3982566 & -0.23350910 \\ -0.07153218 & -0.2335091 & 0.39277804 \end{pmatrix}$$

Now the degree of prediction can be expressed by

$$\frac{\max_{eigenvalue} y' S_1 y}{\max_{eigenvalue} y' S_2 y} \; or \; \max_{eigenvalue} S_2^{-1/2} S_1 S_2^{-1/2} \qquad (4.1)$$

The eigenvalues of the above matrix are 0.73581317, 0.27018538, and 0.06877491. So, the maximum eigen is 0.73581317 and this is nearer to 1. So, it can be said that the covariates can predict more than 70% of the "degree of relationship" as the maximum eigenvalue should be one for full predictability and this holds if S_1=S_2 i.e. 4.1 becomes identity matrix.

Table 1: The Mean Matrix

Groups	Mean of sbps	Mean of dbps	Mean of prs
Group 1	106.6667	65.55667	64.00000
Group 2	110.5189	73.70333	72.00000
Group 3	106.8325	73.66750	75.00000
group 4	113.3325	77.33500	73.50000
group 5	118.0000	84.33000	73.00000
group 6	92.6700	69.33000	72.00000
group 7	128.0000	61.33000	66.00000
group 8	115.2589	69.40889	82.44444
group 9	115.5000	74.44417	75.00000
group 10	114.5850	73.83375	81.75000
group 11	109.5567	74.22333	84.00000
group 12	125.8086	83.61857	84.42857
group 13	114.0000	70.55667	81.33333
group 14	112.0000	79.33000	78.00000
group 15	98.6700	70.00000	90.00000
group 16	123.3300	84.00000	72.00000
group 17	116.0000	78.00000	72.00000
group 18	108.6700	69.67000	78.00000
group 19	112.0000	82.67000	66.00000
group 20	110.8600	69.41176	72.70588
group 21	115.8892	75.89000	78.33333
group 22	110.0394	73.58941	72.00000
group 23	114.4000	78.26600	73.60000
group 24	106.3350	70.66500	72.00000
group 25	106.0000	64.00000	72.00000
group 26	180.6700	100.67000	72.00000
group 27	105.6575	70.16500	75.00000
group 28	116.0000	70.67000	66.00000
group 29	105.3300	70.67000	66.00000
group 30	160.0000	80.00000	72.00000
group 31	110.0000	72.00000	66.00000
group 32	120.0000	85.33000	60.00000
group 33	130.6700	94.00000	90.00000
group 34	100.0000	60.00000	78.00000
group 35	109.6650	64.66500	66.00000
group 36	125.0000	94.00000	84.00000
group 37	113.3300	79.33000	72.00000
group 38	112.0000	74.67000	69.00000
group 39	127.3300	96.00000	78.00000
group 40	121.1785	78.61538	83.07692
group 41	115.2882	72.97000	73.09091
group 42	123.7542	80.31632	72.94737
group 43	111.4450	74.66667	78.00000
group 44	119.5000	76.91750	70.50000
group 45	110.6700	80.67000	78.00000
group 46	138.0000	76.00000	66.00000
group 47	114.6650	75.00000	60.00000
group 48	115.3325	73.00000	72.00000
group 49	108.0000	70.67000	72.00000
group 50	112.0000	84.00000	66.00000

Conclusion

Though, theoretically we have modified the formula of correlation ratio and obtained the degree of relationship but we are to see how it is related to multiple correlation coefficient. This is a new technique which can be used for data from various populations especially for the tribes who have special anthropological characteristics.

References

[1] Blackman, J. (1905), On Tests of Linearity of Regression in Frequency Distribution, Biometrika 4, 332–350.

[2] Draper, N.R. and H. Smith (1981), Applied Regression Analysis, Wiley New Yeork, 2nd ed.

[3] Fisher, R.A. (1925), Statistical Methods for Research Works, Oliver and Boyed, Ltd., 14th ed. (1970P, Hafner Publishing Company.

[4] Hall, W.J. (1970), On Characterizing Dependence in joint Distributions, in: Essayes in Probability and statistics, Univ. of North Carlina Press, pp. 339–376.

[5] Pearson, K. (1903), Mathematical Contributions to Theory of Evaluation. On Homotyposis in Homologous but Differentiated Organs, Roy, Soc. London Proc. 71, 288–313.

[6] Pearson, K. (1905), Mathematical Contributions to the Theory of Evolution XIV. On the General Theory of Skew Correlation and Non-linear Regression, in: Drapers' Company Research Memories, Biometric Ser.II.

[7] Pearson E.S. (1926), Review of Statistical Methods for research workers, Science Progres, 20, 733–734.

[8] Sampson Allan R. (1984), A Multivariate Correlation Ratio, Statistics and Probability Letters, 2 , 77–81.

[9] Seal B., P. Hor, B. Pal, A Method of Selecting Important Variables in Multivariate Anthropometric Data, Accepted for publication in March Issue, 2015.

[10] Software used: R and SPSS.