Rev. Port. Cardiol. 16(2):141-146, 1997

Artigos Originais

Factor analysis, a more accurate method to be used in epidemiological studies of blood pressure in children [16]

Mário Alberto Espiga Macedo*,§,§§, Duarte Trigueiros**§§§, A. Falcão de Freitas***§

PALAVRAS-CHAVE: Epidemiologia; Pressão arterial; Crianças; Análise de factores.
KEY-WORDS: Epidemiology; Blood pressure; Children; Factor analysis.

Introduction

When comparing blood pressure (BP) in children from the same or different populations, one faces the problem of having to consider other characteristics along with BP measurements. Age and weight, for instance, dictate whether a given BP measurement should indicate high blood pressure or not. Sexual maturity and skinfold also have a role in setting standards or expected values for the BP, as different ponderal development levels determine different BP expected in children of the same age[1]. Thence, prior to any comparative study, tables or indices must be obtained to convert BP measurements into standard values, where characteristics such as age, weight and sexual maturity have been accounted for.

Hitherto, the only tables available which provided standard BP values in children were those published in the Report of the Second Task Force on Blood Pressure Control in Children[2]. Such tables apply to a particular population, thereby being of little interest when assessing other populations.

This study describes a general methodology which allows the construction of indices (scores) to compare the cardiovascular characteristics of children. In producing such an index, a new viewpoint is explored, leading to the use of factor analysis (FA) instead of regressions as statistical tools[3,4]. We show that the index obtained effectively removes the influence of age, weight, height and other characteristics specific to each case, while displaying high sensitivity to cardiovascular characteristics. This index yields, for each subject, a unique standardised value between -3 and +3 showing, in standard deviation units, the deviation of BP from that expected for age, weight and other characteristics.

Material and methods

Schoolchildren aged 5 to 18 years old from regions of the North of Portugal with different and very specific socio-economic and geographical characteristics were screened, along with their siblings and parents during a two year period. The sample of 889 children consisted of 389 boys and 500 girls. Blood pressure (BP), age (A), weight (W), height (H), skinfold (Sk), sexual maturation (SM), body mass index (BMI), rurality (Z) and genital development (GD), were the variables studied. Blood pressure was measured with a mercury sphygmomanometer on the right arm, the subject seated and the arm extended over a table at heart level. A set of different-sized cuffs were used. The cuff bladder used was wide enough to cover at least two-thirds of the arm completely, without overlapping[5]. The mean of six readings taken by two observers was considered the BP value of each patient[6]. First and fifth phases of Korotkoff sounds were recorded as systolic blood pressure (SBP) and diastolic blood

§ Serviço de Medicina II, Faculdade de Medicina do Porto;

§§ Centro de Citologia Experimental, Universidade do Porto;

§§§ ISCTE, Universidade de Lisboa, Portugal

* Professor Auxiliar da Faculdade de Medicina da Universidade do Porto e Assistente Graduado do Serviço de Medicina II do Hospital S. João, Porto.

** Professor Auxiliar do ISCTE, Universidade de Lisboa.

*** Professor Catedrático da Faculdade de Medicina da Universidade do Porto e Director do Serviço de Medicina II do Hospital S. João, Porto.

Recebido para publicação em Setembro de 1996.
Aceite para publicação em Dezembro de 1996.

pressure (DPB), respectively[7]. Height was measured by a standard anthropometric method with the subject standing barefoot. Weight was obtained with the subject wearing only shorts and without shoes, on a balanced standard scale. Skinfold was measured with a Lange skinfold caliper[8]. To assess relative body heaviness, BMI (weight/ height^2, in Kg/m^2) was calculated for each individual[9]. Sexual maturation was classified according to the Tanner criteria[10,11]. Z is an index reflecting each region's degree of rurality. Genital development reflects progress in sexual maturity: both sexually mature subjects and young children exhibit GD=0 whereas adolescents will have positive GD in different degrees.

All calculations were carried out with the use of an SPSS statistical package[12]. Distributional characteristics of the above variables were assessed prior to any statistical manipulation. Variables showing lognormality rather than normality were applied to the appropriate logarithmic transformation (all of them except DBP and Z). Lognormality is expected in children in variables such as weight and height because they are in a growing phase. After logarithmic transformation, variables reflect growth and exhibit skeweness and kurtosis consistent with the normality hypothesis[13].

Regression Analysis

Regression analysis is a statistical technique which assesses the relationship between one dependent variable (DV) and several independent variables (IVs)[3,14-17]. Regressions are often used when the intent of the analysis is the prediction of DV, or the assessement of deviations from prediction.

The result of a regression analysis is an equation representing the best prediction of a DV from several continuous IVs. Regressions will be appropriate when each IV is strongly correlated with the DV, but weakly correlated with other IVs. Each IV is expected to predict a substantial and independent segment of the variability in the DV. Only statistically significant IVs are accepted by the algorithm and, as a result of such a selection of variables, the final regression model will contain only those variables able to provide good prediction of DV on purely statistical grounds.

When trying to build BP indices using regressions, one faces two major difficulties. Firstly, since there are two cardiovascular measurements per subject (SBP and DBP), two indices are obtained, not just one. This is awkward since, underlying those two variables, there is only a single physiological characteristic. Indeed, it would be desirable to obtain, only one value reflecting such an underlying physiological characteristic instead of two indices. Secondly, as shown in the results, regressions discard many meaningful variables with the potential to explain BP and consequently show little ability to explain variability in SBP and DBP. Clearly, when trying to explain the variability of SBP and DBP separately, some information is lost, mostly that reflecting the common physiological principle underlying the above two measurements.

Factor Analysis

In order to overcome the above two limitations, our study proposes adopting a statistical methodology called factor analysis (FA) [3,4,17,18] instead of regressions.

Factor analysis uncovers coherent subsets of variables. Groups of variables that are correlated with one another, but largely independent of other variables, are combined into the same factors. The specific goal of FA is to summarize patterns of correlations among observed variables so as to reduce a large number of observed variables to a smaller number of factors, thus providing an operational definition, based on observed variables, for an underlying common principle[19].

When reducing numerous variables down to a few factors, FA produces linear combinations of observed variables, each combination being a factor. The first factor to be computed is the linear combination of observed variables which is able to explain the maximum amount of variance and co-variance of those observed variables. The second factor then performs the same operation on the variance and co-variance left unexplained by the first factor. Thus, the second factor is the linear combination of observed variables which explains maximum variability uncorrelated with the first factor, and so on. In a good FA, a high percentage of the variance and co-variance present in the observed variables is accounted for by the first few factors.

Since factors summarize patterns of correlations in the observed correlation matrix, they can be used to reproduce such an observed correlation matrix. Inclusion of many factors in a given FA would improve the similarity between observed and reproduced correlation matrices since the more factors extracted, the better the fit and the greater the percentage of variance in the data «explained» by the factor solution. However, the more factors extracted, the less parsimonious the solution. An estimation of the number of factors to extract is obtained from observing the amount of variance and co-variance explained by each factor. This amount is called the «Eigenvalue» of the factor. Since Eigenvalues represent standardised variance, the variance that each standardised observed variable contributes to the overall variability is 1, thus a factor exhibiting an Eigenvalue less than 1 is not as important, from a variance perspective, as an observed variable. (Table I, II and III).

After extraction, orthogonal rotation of factors [20] is often used to improve the interpretability and scientific utility of the solution. In our case, epidemiological variables listed above were the object of FA and, as a result, three factors emerged, each of them reproducing a distinctive physiological feature. One such feature is blood pressure. Therefore, the factor reproducing BP can be used as an index where the other physiological features are accounted for. Interestingly, improvements in explained variability over regressions are very significant. Both SBP and DBP are explained to a large extent, without losing the information contained in our set of epidemiological variables.

Results

In this section, the use of factor analysis to build our cardiovascular index is confronted with the same attempt using regressions.

The following tables present the results of applying factor analysis to our data, for both sexes separately. Table I represents the Eigenvalues of the ten factors which might be extracted. These tables also show Eigenvalues in percentage and cumulative percentage of the overall variability present in our data.

TABLE I

FACTOR ANALYSIS. FACTORS AND THEIR EIGENVALUES. MALES AND FEMALES

Factors	Eigenvalue		Eigenvalue in Pct		Cum. Pct	
	M	F	M	F	M	F
1	4.66	5.18	51.8	57.6	51.8	57.6
2	1.50	1.29	16.8	14.4	68.6	72.0
3	1.08	1.03	12.1	11.5	80.7	83.5
4	0.71	0.59	7.9	6.6	88.6	90.3
5	0.40	0.56	4.5	5.6	93.1	91.2
6	0.28	0.34	3.1	3.9	98.9	94.0
7	0.23	0.26	2.6	2.9	97.1	96.9
8	0.09	0.18	1.9	2.0	99.0	99.0
9	0.09	0.09	1.1	1.0	99.9	99.9
10	0.006	0.006	0.1	0.1	100.0	100.0

TABLE II

FACTOR ANALYSIS. CORRELATION BETWEEN THE VARIABLES AND THE 3 FACTORS EXTRACTED. MALES.

Variables	Factor 1	Factor 2	Factor 3
ln weight	0.96		
ln sex. mat.	0.90		
ln height	0.89		
ln age	0.87		
ln BMI	0.81		
ln skinfold		0.86	
Z		0.79	
ln GD		0.61	
DBP			0.98
ln SBP			0.66

ln - logaritmo natural.

TABLE III

FACTOR ANALYSIS. CORRELATION BETWEEN THE VARIABLES AND THE 3 FACTORS - FEMALE

Variables	Factor 1	Factor 2	Factor 3
ln sex. mat.	0.96		
ln weight	0.94		
ln age	0.88		
ln height	0.86		
ln BMI	0.83		
Z		0.89	
ln skinfold		0.84	
ln GD		0.62	
DBP			0.97
ln SBP			0.71

ln - logaritmo natural.

In males, the accumulated percentage for the first factor is 51,8% and for the first three factors 80,7%. In females, the accumulated percentage for the first factor is 57,6% and for the first three factors 83,5%. Based upon results displayed in Tables I and IV we decided that only three factors should be extracted.

Table II and III, after the extraction of the three factors, represent the correlation obtained between our set of variables and those three factors. Only values higher than 0,35 are presented. Clearly, each factor reproduces a different set of variables.

Table IV represents, for each variable, the percentage of its variability explained by the three factors. For systolic and diastolic blood pressure, in both sexes, more than 80% of

TABLE IV

VARIABILITY EXPLAINED BY THE 3 FACTORS EXTRACTED FOR EACH VARIABLE - MALES AND FEMALES

Variables	Variability Explained (Communality) Male	Female
Weight	97.2	96.6
Height	86.7	85.1
Age	86.3	86.2
Sexual Mat	83.9	66.2
BMI	69.4	66.2
SBP	80.3	79.3
DBP	88.1	90.7
Z	63.3	55.2
Skinfold	54.9	71.1
GD	42.8	46.2

the total variability is explained by this method.

Distributional characteristics of factor n.° 3 were also assessed after extraction. In both groups (boys and girls) this factor exhibits skewness and kurtosis consistent with the normality hypothesis. We recall that extracted factors have a mean of zero thus, a standard deviation of one is a standardised score.

In order to understand the extent to which FA improves the amount of explained variability in indices when compared with regressions, a regression analysis was also undertaken, using the same set of data. The results obtained are displayed in Tables V and VI.

The overall accurancy of prediction is reflected by R^2 (proportion of explained variability). The R^2 values vary between 40.9% and 47.2%. Only a few variables from the original set are accepted by the algorithm (weight, skinfold), whereas other meaningful variables do not attain statistical significance.

When comparing the results of applying FA and regressions, we may conclude that the first method is more efficient in explaining overall variability.

TABLE V

SYSTOLIC BLOOD PRESSURE. MULTIPLE LINEAR REGRESSION - BOTH SEXES: EXPLAINED VARIABILITY

Sex	N	Const.	Coefficient β weight	R^2	F
Male	394	56.4	0.64	41.2	240.0****
Female	500	-28.0	0.65	41.9	347.2****

****p<0.0005 weight - In males square root and in females, natural logarithm; R^2 - percentage of variance explained, coefficient of determination; F - value F of multiple regression.

TABLE VI

DIASTOLIC BLOOD PRESSURE. MULTIPLE LINEAR REGRESSION - BOTH SEXES: EXPLAINED VARIABILITY

Sex	N	Const.	Coefficient β SBP	SK	BMI	SM	H	R^2	F
Male	394	56.4	-0.70****	0.15***	-0.17***	-	-	40.9	78.05(*)
Female	500	-28.0	0.67****	0.07*	-	0.21****	0.14	47.2	107.4(*)

*p<0.05; **p<0.001; ***p<0.0001; ****p<0.0005; (*) p=0.0

Discussion and Conclusions

Tables II and III show that the obtained factors are strongly correlated with meaningful groups of variables. Namely, factor n.° 3 relates to SBP and DBP, exhibiting negligible correlations with the other variables. Therefore, factor n.° 3 reflects, almost exclusively, the cardiovascular characteristics of children. Once extracted, factor n.° 3 is a valuable instrument for the assessment of deviations of BP from values expected for a given age, weight, and so on. Children exhibiting values of factor n.° 3 near zero have BP according to what is expected. Values of factor n.° 3 above two standard deviations denote a trend towards high blood pressure. Normality of factor n.° 3 allows us to link values of this index with the corresponding probabilities. For instance, the likelihood of finding, by random sampling, children exhibiting values of factor n.° 3 of +2 or smaller is 0.95. Besides factor n.° 3, two other factors were extracted, corresponding to other physiological features. Factor n.° 1 highlights ponderal development. Factor n.° 2 captures specific environmental and genetic characteristics impinging upon particular regions. Such characteristics clearly affect sexual development and skinfold, amongst other less important variables, being markedly different for boys and girls. Since factors n.° 1 and 2 are uncorrelated with factor n.° 3, measurements based upon factor n.° 3 will be independent of any characteristic already accounted for by the first two factors.

Factor n.° 3 can be easily assessed via a formula which is a linear combination of its component variables. This formula may be obtained from the output of the package used to perform the factor analysis[12].

Compared to regressions, factor analysis has shown the potential to explain a large amount of variability in this specific task. This improvement stems from SBP and DBP being assessed as two correlated variables, not as two independent characteristics as in the case of regressions. Indeed, factor n.° 3 should be seen as assessing a physiological characteristic underlying both SBP and DBP when other characteristics are accounted for. Our methodology is robust regarding the number of variables used. For instance, in case less ponderal variables were present in the analysis, the factor structure obtained would be the same. It is probably desirable to use a smaller set of ponderal variables so as to avoid redundant variability in factor n.° 1. Also, when the index to be obtained is aimed at assessing homogeneous populations (e.g. single a region), factor n.° 2 and the corresponding component variables may be omitted from the analysis.

Besides yielding unique, easy to calculate and flexible cardiovascular score for children, thus circumventing the problem of making decisions based on two variables (SBP and DBP), a major strength of the proposed methodology is that such an index is the result of a process where variability specific to BP is *isolated* rather than *explained.* Other more sophisticated statistical methodologies, such as canonical correlation, may also be used to obtain one unique index from two explained variables. However, such an index would reflect BP as explained by a specific set of variables, thus being sensitive to the *a-priori* decision on which variables might be important to explain BP on children. As mentioned above, results of factor analysis are robust regarding the set of variables used, mainly because this methodology is aimed at isolating specific variability rather than explaining it in terms of independent variables. FA probably makes more sense in epidemiological studies of this kind.

In a forthcoming paper the results will be presented the potential and applicability of the devised index shown.

Resumo

O objectivo do nosso estudo, foi verificar se a análise de factores era capaz de explicar de melhor forma a variabilidade da pressão arterial sistólica (PAS) e da pressão arterial diastólica (PAD), quando comparada com a análise de regressão, que é o método habitualmente usado para estudar um conjunto de variáveis relacionadas com a pressão arterial (PA). PAS, PAD, peso, altura, índice de massa corporal, prega cutânea tricipital, maturação sexual e ruralidade, foram estudados em 889 crianças com a idade compreendidas entre os 5 e os 18 anos (389 rapazes e 500 raparigas). O método proposto transforma um conjunto de variáveis num novo conjunto de variáveis (os factores) que não estão correlacionadas entre si. Um dos factores obtidos explica claramente a variância da PA. Com este método, o algoritmo aceita todas as variáveis, enquanto a regressão rejeita a maioria delas. Este método explica a maior parte da variabilidade da PA, perdendo a menor informação possivel. Na nossa amostra a percentagem da variância total explicada pelos três factores, foi de 80,3% para a PAS e 88,1% para a PAD no sexo masculino; e de 79,3% para a PAS e 90,7% para a PAD no sexo feminino. Para a mesma amostra, a regressão só explica 41,2% nos rapazes e 41,9% nas raparigas para a PAS, e 40,9% nos rapazes e 47,2% nas raparigas da PAD.

Em conclusão, este método é mais exacto em estudos epidemiológicos, conseguindo um melhor resultado do que a regressão e perdendo muito pouca informação. Existem duas importantes razões para a utilização da metodologia proposta: primeiro porque calcula facilmente um índice cardiovascular único e flexível, em segundo lugar este índice é o resultado de uma metodologia onde a variabilidade específica da PA é isolada.

Summary in English: on page 123.

Bibliography

1. Espiga Macedo MA. Estudo epidemiológico da pressão arterial em crianças portuguesas. Tese de Doutoramento. Porto 1989.

2. Task Force on Blood Pressure Control in Children. Report of the second task force on blood pressure control in children - 1987. Pediatrics 1987;79:1-25.

3. Tabachnick BG, Fidell LS. Using Multivariate Statistics, ed 2. California State University, Northridge. Harper and Row, New York,1989.

4. Comrey AL. A first Course in Factor Analysis. Academic Press, New York, 1973.

5. Frohlich ED, Grim C, Labarthe DR, Max well MH, Perloff D, Weidman WH. Recommendations for human blood pressure determination by sphygomanometers. Report of a special task force appointed by the Steering Committee, American Heart Association. Circulation 78:1988;502A-14A.

6. Souchek J, Stamler J, Dyer AR, Paul O, Lepper MH. The value of two or three versus a single reading of blood pressure at a first visit. J Chron Dis 1979;32:197-210.

7. Rosner B, Prineas RJ, Loggie JM, Daniels SR. Blood pressure nomograms for children and adolescents, by height, sex, and age, in the United States. J. Pediatr 1993;123:871-86.

8. Waaler PE. Anthropometric studies in Norwegia children. Acta Paediatr Scand 1983;303:Suppl.3-39.

9. Benn RT. Some mathematical properties of weight-for-height indices used as measures of adiposity. Br J Prev Soc Med 1971;25:42-50.

10. Michaud PA, Wilkins J. Les Stades de Tanner: Leur utilisation en clinique et en recherche. Med et Hyg. 1982;40:877-88.

11. Daniel WA Jr. Evaluation of adolescents. Textbook of Pediatrics, ed. by Nelson W. WB Saunders Company, London 1983;38-46.

12. Nie NA, Hull CH, Jenkins JG. SPSS Statistical Package for the Social Sciences, ed. McGraw-Hill, New York;1984.

13. Box GEP, Cox DR. An Analysis of Transformations. J Royal Statist Society 1964;26(series B):211 -43.

14. Wonnacott RJ & Wonnacott TH. Introductory statistics, Ed. John Wiley and Sons, London;1985.

15. Snedecor GW & Cochram NG. Statistical methods, ed. The Iowa State Univ. Press,6th;1967.

16. Cook RD. Detection of influential observations in linear regression. Technometrics 1977;19:15-8.

17. Mardia KV, Kent JT & Bibby JM. Multivariate analysis, ed. Academic Press, London; 1979.

18. Harman HH. Morden factor analysis. Univ Chicago Press, 2d ed;1967.

19. Gorsuch SA. Factor analysis. Hillsdale NJ. Erlbaum, 1983.

20. Mulaik SA. The foundation of factor analysis. MacGraw-Hill, New York, 1972.

Pedido de separatas para:

M. ESPIGA DE MACEDO,
Serviço de Medicina II,
Faculdade de Medicina do Porto,
Alameda Prof. Hernâni Monteiro,
4200 Porto - Portugal. Fax - 02-6099157.