Journal of Modern Applied Statistical Methods

Volume 18 | Issue 1 Article 10

3-9-2020

A Study Verifying the Dimensioning of a Multivariate Dichotomized Sample in Exploratory Factor Analysis

Rosilei S. Novak
Federal University of Paraná (UFPR), rosileisouzanovak@gmail.com

Jair M. Marques
Federal University of Paraná (UFPR), jair.m.marques@gmail.com

Recommended Citation

Novak, R. S., & Marques, J. M. (2019). A study verifying the dimensioning of a multivariate dichotomized sample in exploratory factor analysis. Journal of Modern Applied Statistical Methods, 18(1), eP2993. doi: 10.22237/jmasm/1556669760

Journal of Modern Applied Statistical Methods
May 2019, Vol. 18, No. 1, eP2993.
doi: 10.22237/jmasm/1556669760

בס"ד

ISSN 1538 − 9472

EMERGING SCHOLARS

A Study Verifying the Dimensioning of a Multivariate Dichotomized Sample in Exploratory Factor Analysis

Rosilei S. Novak
Federal University of Paraná
Paraná, Brazil

Jair M. Marques
Federal University of Paraná
Paraná, Brazil

The sample size dichotomized was related to the measure of sampling adequacy, considering the explanations provided by factors and commonalities. Monte Carlo simulation generated multivariate normal samples and varying the number of observations, the factor analysis was applied in each sample dichotomized. Results were modeled by polynomial regression based on the sample sizing.

Keywords: Exploratory factor analysis, dichotomized data, sample size, polynomial regression.

Introduction

Exploratory factor analysis (EFA) is an effective method that can provide valuable data on the multivariate structure of a measurement instrument, identifying the theoretical constructs (Laros, 2005). It is applied to evaluate the correlation patterns existing on a large set of original variables and utilizes those correlation patterns to group a relatively smaller number of factors that can be used to recognize relations of variables interrelated among themselves. However, it is important to understand the nature of the dataset in order to make important decisions in the analysis process.

One consideration is the dimensioning of normal multivariate samples involving dichotomized variables. Other factors include the relation, for the same sample size, among the results of the exploratory factor analysis (EFA) for dichotomized variables is unknown.

doi: 10.22237/jmasm/1556669760 | Accepted: August 11, 2019; Published: March 9, 2020.
Correspondence: Rosilei S Novak, rosileisouzanovak@gmail.com

Many studies have been conducted using exploratory factor analysis (EFA) as an investigative tool with normal multivariate data, where this data is dichotomized, with the objective of assisting the researcher to clarify this question. However, there are still no conclusive studies on the relation between sample size of dichotomized data and the results of the exploratory factor analysis (EFA).

Everitt (1975) and Nunnally (1978) recommended sampling 1/10 (ten subjects per variable). Cattell (1978) suggested 3/6 (6 subjects to 3 variables). Gorsuch (1983) indicated the relation was at least 3/5 (5 subjects to 3 variables). MacCallum et al. (1999) have demonstrated, mathematically and empirically, that the sample size requirements are dependent on two aspects, factor and structure. They also showed that, as the common factors are sufficiently represented by an adequate number of variables, the proportion of the communalities have a considerable effect over the adjustment between sample and factorial loads. Mundfrom and Shaw (2005) recommended the sample size of 180 observations using the Monte Carlo method, varying the number of factors, the ratio of factors and the communalities. This question becomes more complex when the data studied by factor analysis are dichotomized.

Methods

For the execution of the study that verified the influence of the sample size of dichotomized data on an EFA the Matlab software was used, with the implement of three programs: Matrizc5, Simula5 and Regrespoli1.

Matrizc5 was used to generate multivariate normal random samples using the Monte Carlo simulation, from a phi correlation matrix, considering a distribution $Z \sim N(0, 1)$ the dichotomization followed the condition $P(z \leq z_c) = 0.50$, obeying the proportion of fifty percent of zero and fifty percent of one. From those samples, its corresponding dichotomized samples have been generated, all obeying the pre-requirements where the generated samples would have the $MSA > 0.5$ and the communalities ≥ 0.7. The samples not fitting the pre-requirements stablished were discarded and substituted.

For the analysis of correlation, the phi correlation coefficient is a technique of great importance in a statistical study that uses dichotomous data, but when dichotomized data is used, the use of the tetrachoric correlation coefficient is ideal.

Dichotomized multivariate normal data was used, and therefore it would be adequate the utilization of the tetrachoric correlation matrix, although many times this matrix is singular, not being appropriate for the use of factor analysis (Embreson & Reise, 2013, p. 37). The tetrachoric correlations matrix was

substituted by the phi correlation matrix, so the effect of this substitution over the factor analysis can be evaluated.

The sampling simulations have been generated with 30 variables and 4 factors. The sample sizes were considered equal to 2, 3, 4, 5, 6,…, 50 times the number of variables.

Described in Table 1 are the details of the 8 simulations carried out. The first column represents the simulation number and the second column the vectors representing the number of variables per factor, where the sum of elements from the vector indicates the number of variables and each column represents a factor.

The second program, Simula5, performed the factor analysis at each normal sample and to its dichotomized correspondent, individually oscillating the observations number, obtaining the MSA mean values, the proportion of variance explained by the first factor, the total proportion of variance explained and the communalities. In the factor analysis, the principal component analysis was used to estimate the model parameters. The Kaiser criterion was used to select the number of factors. Varimax rotation was used as rotation method, in order to simplify the data structure.

The third program, Regrpoli1, performed the modelling of the results only at the dichotomized samples. The results obtained from the MSA mean values, the proportion of variance explained by the first factor, the total proportion of variance explained, and the vector of the mean communalities values were modelled in function of the Naperian logarithms of the sample sizes, in order to decrease the variation. Polynomial models were used as the regression models.

Table 1. Classification of the variables per factor

Simulation	Variables Per Factor
1	[8 8 8 6]
2	[9 7 7 7]
3	[10 10 5 5]
4	[11 7 6 6]
5	[12 6 6 6]
6	[13 6 6 5]
7	[14 6 5 5]
8	[15 5 5 5]

The regression model evaluation was carried out making use of the following indicators: coefficient of determination (R^2), chi-square statistics for the adherence, and standard deviation of the adjustment. To each regression model used, residual analyses have been performed (null mean, homoscedasticity, Kolmogorov-Smirnov test for normality and independence tested through the Durbin-Watson test) being those conditions satisfied.

Results

The influence of the sample size of dichotomized data was verified on an EFA obtained tables containing the results of the polynomial regression models for the MSA, proportion of variance explained by the factor 1, total proportion of variance explained by the factors and the communalities, as its adjustment indicators.

Results Obtained for the MSA

In Table 2 are represented the polynomial regression models for 8 cases of factor analysis, considering the MSA as the dependent variable (y) and the sample size Naperian logarithm as independent variable (x).

In all the cases simulated, the best adjusted model corresponds to the fifth-degree polynomial model.

Table 3 shows the indicators for each of the performed regressions, in all cases the coefficient of determination is higher than 99%, and the value of the chi square statistics presents a significant result for the adherence of the adjustments. The standard deviations of the adjustments (S_Y) are all too small.

Table 2. Regression models for the MSA

Simulation	Vector	Model $y = a_0 + a_1x + a_2x^2 + \ldots + a_nx^n$
1	[8 8 8 6]	$y = -21.5511+41.2623x-30.6581x^2+11.4442x^3-2.1394x^4+0.1599x^5$
2	[9 7 7 7]	$y = -15.4156+28.7041x-20.4691x^2+7.3558x^3-1.3273x^4+0.0960x^5$
3	[10 10 5 5]	$y = -18.9449+36.2168x-26.7230x^2+9.9133x^3-1.8427x^4+0.1370x^5$
4	[11 7 6 6]	$y = -15.0365+28.6548x-20.9585x^2+7.7331x^3-1.4331x^4+0.1064x^5$
5	[12 6 6 6]	$y = -15.7350+30.1716x-22.2007x^2+8.2335x^3-1.5331x^4+0.1144x^5$
6	[13 6 6 5]	$y = -10.3839+19.0753x-13.0924x^2+4.5424x^3-0.7938x^4+0.0558x^5$
7	[14 6 5 5]	$y = -21.4265+41.3447x-30.8745x^2+11.5713x^3-2.1702x^4+0.1627x^5$
8	[15 5 5 5]	$y = -21.7826+41.7090x-30.9973x^2+11.5696x^3-2.1619x^4+0.1615x^5$

Table 3. Indicators for the MSA regression

Simulation	Vector	R^2	$\chi 2$	S_Y
1	[8 8 8 6]	0.9994	0.00003	0.00086
2	[9 7 7 7]	0.9997	0.00002	0.00065
3	[10 10 5 5]	0.9997	0.00001	0.00057
4	[11 7 6 6]	0.9998	0.00001	0.00050
5	[12 6 6 6]	0.9997	0.00001	0.00005
6	[13 6 6 5]	0.9999	0.00000	0.00038
7	[14 6 5 5]	0.9995	0.00002	0.00075
8	[15 5 5 5]	0.9996	0.00002	0.00074

Results Obtained for the Proportion of Variance Explained by the First Factor

In the Table 4 are represented the polynomial regression models for 8 cases of factor analysis, considering the proportion of variance explained by the first factor as the dependent variable (y) and the sample size Naperian logarithm as independent variable (x).

In all the cases simulated, the most adequate adjusted model corresponds to the fifth-degree polynomial model.

Table 5 shows indicators for each of the performed regressions. It can be verified that the determination coefficient is unstable, varying from approximately 53% to 97%, the chi square statistics presents significant results for the adherence of adjustments. The standard deviations of the adjustments (S_Y) are higher than the values obtained for the MSA, as shown in Table 3.

Table 4. Regression models adjusted to the proportion of variance explained by the first factor

Simulation	Vector	Model $y = a_0 + a_1x + a_2x^2 + \ldots + a_nx^n$
1	[8 8 8 6]	$y = -214.9676+487.2247x-397.6769x^2+159.3493x^3-31.4406x^4+2.4493x^5$
2	[9 7 7 7]	$y = 69.0479-99.3620x+84.0601x^2-35.5341x^3+7.4710x^4-0.6237x^5$
3	[10 10 5 5]	$y = -308.3474+709.5802x-594.8604x^2+245.1407x^3-49.8066x^4+3.9983x^5$
4	[11 7 6 6]	$y = -60.2234+192.9703x-172.5162x^2+74.5397x^3-15.6904x^4+1.2938x^5$
5	[12 6 6 6]	$y = -195.5082+462.8827x-380.1304x^2+380.1304x^3-30.9207x^4+2.4571x^5$
6	[13 6 6 5]	$y = 110.3694-170.6008x+144.4636x^2-60.7441x^3+12.6325x^4-1.0377x^5$
7	[14 6 5 5]	$y = 14.3890+35.4019x-30.0291x^2+12.0569x^3-2.3342x^4+0.1760x^5$
8	[15 5 5 5]	$y = -409.1023+872.4216x-697.7957x^2+276.1447x^3-54.1198x^4+4.2053x^5$

Table 5. Indicators for the regression of the proportion of variance explained by the first factor

Simulation	Vector	R^2	$\chi 2$	S_Y
1	[8 8 8 6]	0.8188	0.0073	0.0560
2	[9 7 7 7]	0.5938	0.0070	0.0596
3	[10 10 5 5]	0.9677	0.0043	0.0479
4	[11 7 6 6]	0.7680	0.0055	0.0530
5	[12 6 6 6]	0.6622	0.0054	0.0578
6	[13 6 6 5]	0.7065	0.0044	0.0552
7	[14 6 5 5]	0.5605	0.0075	0.0721
8	[15 5 5 5]	0.5265	0.0064	0.0576

Results Obtained for the Proportion of the Total Variance Explained

In Table 6 are represented the polynomial regression models for 8 cases of factor analysis, considering the proportion of the total variance explained as the dependent variable (y) and the sample size Naperian logarithm as independent variable (x).

Table 6. Regression models adjusted to the total proportion of variance explained by the factors

Simulation	Vector	Model $y = a_0 + a_1x + a_2x^2 + \ldots + a_nx^n$
1	[8 8 8 6]	$y = -318.2401+821.9924x-684.5084x^2+279.4772x^3-56.1966x^4+4.4641x^5$
2	[9 7 7 7]	$y = 651.9000-1151.6000x+909.400x^2-358x^3+70.2000x^4-5.5000x^5$
3	[10 10 5 5]	$y = -46.4662+293.9963x-276.3715x^2+123.2896x^3-26.5760x^4+2.2351x^5$
4	[11 7 6 6]	$y = 67.3391+55.3461x-80.5261x^2+43.0299x^3-10.1817x^4+0.9018x^5$
5	[12 6 6 6]	$y = 43.8470+114.8543x-136.0102x^2+69.0988x^3-16.2540x^4+1.4583x^5$
6	[13 6 6 5]	$y = 398.8664-641.9936x+502.0180x^2-195.6648x^3+37.8994x^4-2.9143x^5$
7	[14 6 5 5]	$y = -273.7632+724.5202x-595.2926x^2+239.3083x^3-47.3217x^4+3.6945x^5$
8	[15 5 5 5]	$y = -14.3238+194.4189x-172.7246x^2+72.6408x^3-14.7524x^4+1.1691x^5$

Table 7. Indicators for the regression of the total proportion of variance explained by the factors

Simulation	Vector	R^2	$\chi 2$	S_Y
1	[8 8 8 6]	0.9352	0.0077	0.1088
2	[9 7 7 7]	0.9659	0.0037	0.0766
3	[10 10 5 5]	0.9637	0.0041	0.0803
4	[11 7 6 6]	0.9805	0.0027	0.0639
5	[12 6 6 6]	0.9642	0.0039	0.0784
6	[13 6 6 5]	0.9602	0.0034	0.0741
7	[14 6 5 5]	0.9794	0.0027	0.0658
8	[15 5 5 5]	0.9563	0.0053	0.0893

The best adjusted model corresponds to the fifth-degree polynomial model for all the simulated cases.

Table 7 shows the indicators for each of the performed regressions. It can be verified that the determination coefficient is always higher than 93%, the chi-square statistics present significant results for the adherence of the adjustments. The standard deviations of the adjustments (S_Y) are also higher than the values obtained for the MSA, as shown in Table 3.

Results Obtained for the Communalities

In Table 8 are represented the polynomial regression models for 8 cases of factor analysis, considering the communality mean as dependent variable (y) and the sample size Naperian logarithm as independent variable (x).

It can be verified that in all simulated cases the better adjusted model corresponds to the fifth-degree polynomial model.

Table 8. Regression models adjusted to the communalities

Simulation	Vector	Model $y = a_0 + a_1x + a_2x^2 + \ldots + a_nx^n$
1	[8 8 8 6]	$y = -4.8152+11.6841x-9.7178x^2+3.9597x^3-0.7933x^4+0.0627x^5$
2	[9 7 7 7]	$y = 8.9975-16.3317x+12.8092x^2-5.0046x^3+0.9732x^4-0.0753x^5$
3	[10 10 5 5]	$y = 3.4162+4.9483x+3.5994x^2-1.3114x^3+0.2385x^4-0.0173x^5$
4	[11 7 6 6]	$y = -2.2536+6.5603x-5.6869x^2+2.3933x^3-0.4923x^4+0.0398x^5$
5	[12 6 6 6]	$y = 1.6673-1.9025x+1.5143x^2-0.6110x^3+0.1234x^4-0.0099x^5$
6	[13 6 6 5]	$y = -4.3464+9.8290x-7.5191x^2+2.8354x^3-0.5287x^4+0.0391x^5$
7	[14 6 5 5]	$y = 5.1505-8.7622x+6.9154x^2-2.7317x^3+0.5382x^4-0.0422x^5$
8	[15 5 5 5]	$y = 1.6175-1.3603x+0.6928x^2-0.1383x^3+0.0031x^4-0.0015x^5$

Table 9. Indicators for the regression of the communalities

Simulation	Vector	R^2	$\chi 2$	S_Y
1	[8 8 8 6]	0.8445	0.00019	0.0018
2	[9 7 7 7]	0.8889	0.00013	0.0015
3	[10 10 5 5]	0.9249	0.00013	0.0014
4	[11 7 6 6]	0.9009	0.00012	0.0014
5	[12 6 6 6]	0.8060	0.00020	0.0018
6	[13 6 6 5]	0.8211	0.00016	0.0016
7	[14 6 5 5]	0.9025	0.00015	0.0016
8	[15 5 5 5]	0.8367	0.00018	0.0017

Table 9 shows the indicators for each of the performed regressions, showing that the determination coefficient is higher than 80%, the chi square statistic presents significant values for the adherence of the adjustments. The standard deviations of the adjustments (S_Y) are lower than the values obtained for the regressions of the proportion of variance explained by factor 1 and by the proportion of variance explained by the factors (Tables 5 and 7).

Graphics Obtained Through Polynomial Regression

The graphics shown represent the tables of the MSA regression models, variance explained by the first factor, total variance explained, and communalities means in comparison to the sample size Naperian logarithm of the sample sizes for the simulations 1, 4 and 8, which represent the group behavior. Those graphics are shown in Figures 1, 2, and 3.

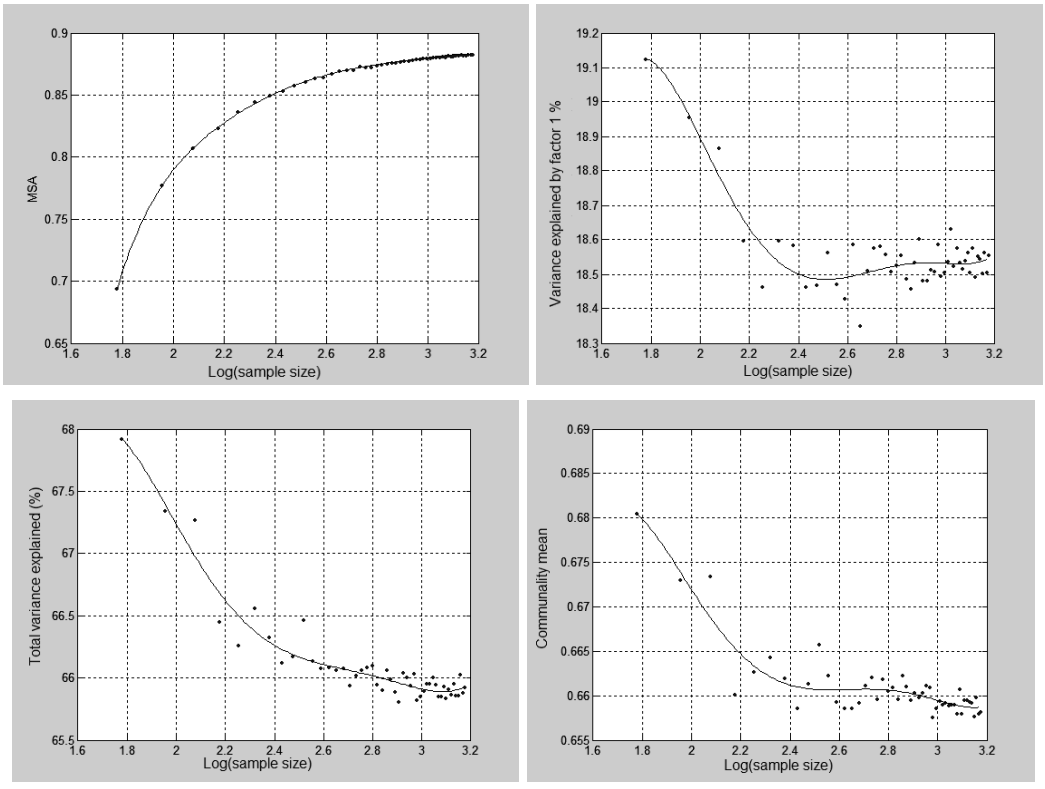

Figure 1. MSA regression models, variance explained by the factor 1, total variance explained, and communalities means in relation to the sample size logarithm of the vector for sample [8 8 8 6]

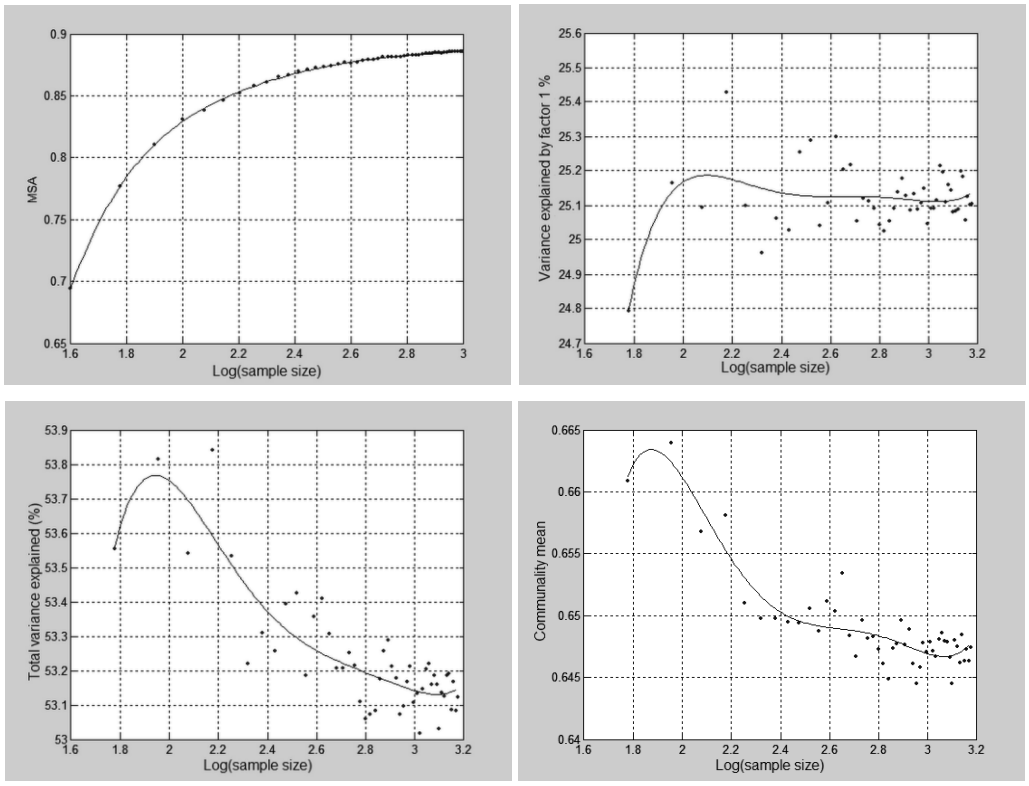

Figure 2. MSA regression models, variance explained by the factor 1, total variance explained, and communalities means in relation to the sample size logarithm of the vector for sample [11 7 6 6]

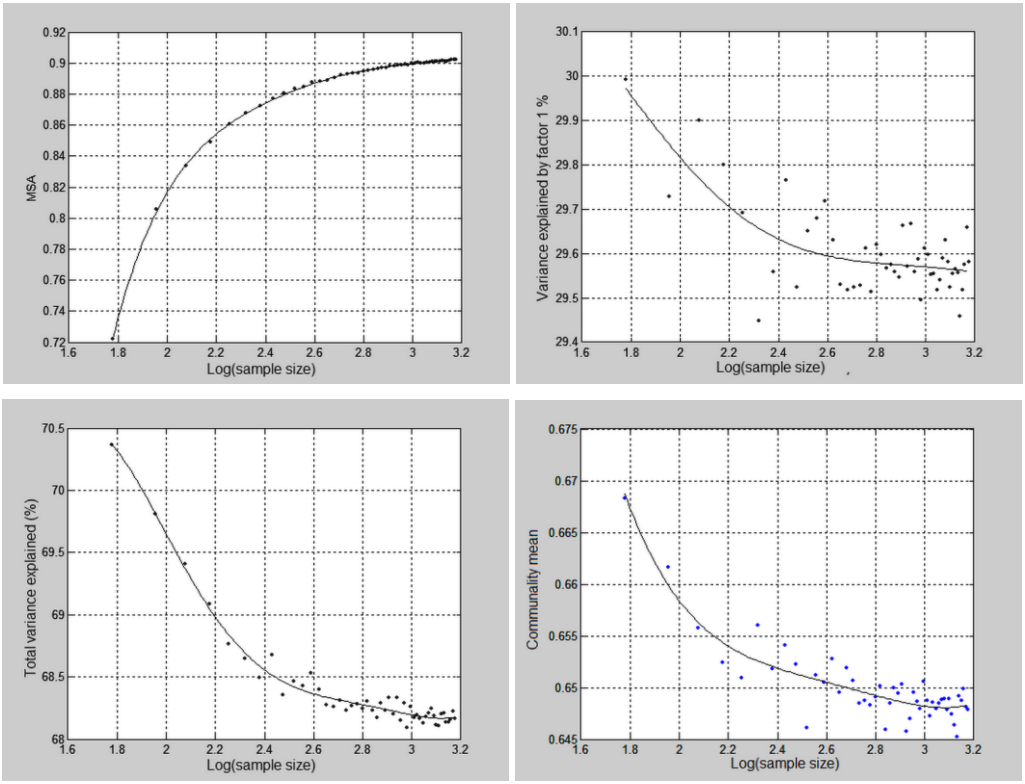

Figure 3. MSA regression models, variance explained by the factor 1, total variance explained, and communalities means in relation to the sample size logarithm of the vector for sample [15 5 5 5]

Conclusion

The influence of the sample size from dichotomized data on an EFA, for the studied cases, leads to the following conclusions:

I. For all the studied variables (MSA, proportion of variance explained by the first factor, total proportion of variance explained, and communalities means) the adequate polynomial regression model, in relation to the logarithm of the sample sizes, is the fifth-degree model.

II. The better adjustment was verified for the MSA, with coefficient of determination always higher than 0.99. It can also be verified that the MSA grows as the sample size gets larger but tends towards stabilization.

III. The worst adjustment was verified for the proportion of variance explained by the first factor, with great variability on the coefficient of determination, in some cases close to 0.50. On the corresponding graphics this result is very clear.

IV. The adjustment for the total determination also presented a good result, according to what is suggested by the indicators found, with coefficient of determination higher than 0.93.

V. The adjustment for the communalities means presented a coefficient of determination higher than 0.80, a result that is lower than the total determination.

References

Cattell, R. (1978). *The scientific use of factor analysis in behavioral and life sciences*. New York: Plenum Press. doi: 10.1007/978-1-4684-2262-7

Embreson, S. E., & Reise, S. P. (2013) *Item response theory*. New York: Psychologists Press. doi: 10.4324/9781410605269

Everitt, B. (1975). Multivariate analysis: The need for data, and other problems. *British Journal of Psychiatry*, *126*(3), 237-240. doi: 10.1192/bjp.126.3.237

Gorsuch, R. L. (1983). *Factor analysis* (2nd edition). Hillsdale, NJ: L. Erlbaum Associates.

Laros, J. A. **(**2005). O uso da análise fatorial: Algumas diretrizes para pesquisadores. In L. Pasquali (Ed.), *Análise fatorial para pesquisadores* (pp. 141-160). Brasília, Brazil: Universidade de Brasília LabPAM.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*(1), 84-99. doi: 10.1037/1082-989X.4.1.84

Mundfrom, D. & Shaw, D. T. (2005). Minimum sample size recommendations for conducting factor analyses. *International Journal of Testing, 5*(2), 159-168. doi: 10.1207/s15327574ijt0502_4

Nunnally, J. (1978). *Psychometric theory* (2nd edition). New York: McGraw-Hill.