FACTOR ANALYSIS MODEL FOR QUANTITATIVE AND VERBAL ABILITY ON STUDENTS PERFORMANCE

Augustine Michael Okutachi & Lami Edibo
Department of Mathematics and Statistics,
Federal Polytechnic, Idah, Kogi state, Nigeria
austineokutachi@gmail.com; edibolami@gmail.com

Abstract

In this study, we developed a factor analysis model that describes the impact of quantitative and verbal ability on student performance. Two courses namely Quantitative Techniques/Operations Research, (QT/OR) and General Studies/Communication in English, (GNS) whose scope requires quantitative and verbal ability were sampled for a number of students. Results obtained showed that quantitative ability explains substantial part of QT/OR scores while verbal ability explains substantial part of GNS scores.

Keywords: Factor analysis, Quantitative and verbal Ability, Factor loadings, Orthogonal factor model, Principal factor model.

Introduction

Factor Analysis, (FA) is a method for investigating whether a number of variables of interest, $Y_1,\dots,Y_p$ are linearly related to a smaller number of unobservable factors, $F_1,\dots,F_m$. $Y_1,\dots,Y_p$ are observable variables of interest and hence we have data on whereas $F_1,\dots,F_m$ are unobservable factors linearly related to $Y_1,\dots,Y_p$ and so we don't have data on them and hence we are required to build a mathematical model that can explain the effect of these unobservable factors on the corresponding responses, (see Bendangnuksung, and Prabu 2018).

The basic objective of factor analysis is to describe, if possible, the covariance relationship among many variables in terms of a few underlying, but unobservable random quantities called factors. Basically, the factor model is motivated by the following argument. Suppose variables can be grouped by their correlations, that is, all variables within a particular group are highly correlated among themselves but have relatively small correlations with variables in different group. Then it is conceivable that each group of variables represent a single factor (underlying construct) that is responsible for the observed correlations.

There are basically two types of factor analysis namely: The Exploratory factor analysis, (EFA) and the Confirmatory factor analysis (Nkiruka et al, 2017, Ahamed et al, 2017). While the EFA is used to find the underlying structure of a large set of variables, the CFA which is the main focus of this study allows a researcher to figure out if a relationship between a set of observed variables and their underlying constructs exist.

The FA as a popular statistical method has been extensively used in such areas as psychology, psychometrics, and educational testing. It has proven to be a useful tool in big data analysis. Examples of its most recent applications include Bosse et al. (2007), Banerjee and Gupta (2012), and Dickinson et al. (2011). It should be noted that the method of principal component analysis, which plays the role of exploratory data analysis for formal factor analysis, is used routinely in big data analysis (see, e.g., Witten et al., 2010).

Technically, the starting point of FA is a linear model in which the observed variables are expressed as linear functions of a vector of unobservable factors and random errors. The number of underlying factors is strictly less than the number of observed variables. Following Johnson and Wichern (2001), Öğütcü, (2020) and Anderson (2003), the most commonly used FA model for continuous response variables, namely the Gaussian FA (GFA) model, can be written as

$$y_i = l_{i0} + l_{ij}F_j + \varepsilon_i \qquad (1)$$

Where y_i is the p-dimensional i^{th} observation, l_{i0} is a p-dimensional column vector, l_{ij} is the $p \times m\ (m < p)$ factor loading matrix, F_j is a m-dimensional vector of unobserved factor scores, and $F_j \sim N_m(0, I_m)$, where I_m denotes the $(m \times m)$ identity matrix. The error term $\varepsilon_i \sim N_p(0, \Psi)$, where $\Psi = Diag(\psi_1^2, \ldots, \psi_p^2)$ is a diagonal matrix whose components are called specific variance. The parameters to be estimated are $\theta = (l_{i0}, l_{ij}, \Psi)$.

Since the unobserved factor scores and errors in GFA are assumed to be Gaussian, the usual maximum likelihood, (ML) or Bayesian approach, (BA) can be thought of as the appropriate estimation method for the data. The classical technique can be thought of as computing the sample covariance matrix or the sample correlation matrix and making inference based on the matrix obtained.

2. RESEARCH METHODOLOGY

The aim of this study is to measure the contribution of quantitative and verbal ability on students' performance in some selected courses with the following objectives: (i) To build a factor analysis model describing the relationship between the observed scores of students and the unobserved factors that influences the scores. (ii) To estimate the factor loadings of the factor analysis model using the principal factor approach and (iii) To estimate the percentage of students' performance accounted for by quantitative and verbal ability using the estimated factor loadings and hence the orthogonal factor model as well as principal component method comes to focus.

The Orthogonal Factor Model

Let the observable random vector Y, with $P-$components (dimensions) has mean μ and covariance matrix Σ .The orthogonal factor model postulates that X is linearly dependent upon a few unobservable random variables $F_1, \ldots, F_m$ called common factors and $p-$additional sources of variation $e_1, \ldots, e_p$ called error term or specific factors. In particular, the orthogonal factor analysis model is

$$X_i - \mu_i = \ell_{ij}F_j + \cdots\cdots + \ell_{im}F_m + \varepsilon_i \qquad (2)$$

with specific form given as:

$$\left.\begin{array}{rcccccc} X_1 - \mu_1 &=& \ell_{11}F_1 &+& \ell_{12}F_2 &+ \cdots\cdots +& \ell_{1m}F_m + e_1 \\ X_2 - \mu_2 &=& \ell_{21}F_1 &+& \ell_{22}F_2 &+ \cdots\cdots +& \ell_{2m}F_m + e_2 \\ \vdots & & \vdots & & \vdots & & \vdots \\ X_p - \mu_p &=& \ell_{p1}F_1 &+& \ell_{p2}F_2 &+ \cdots\cdots +& \ell_{pm}F_m + e_p \end{array}\right\} \qquad (3)$$

Equation (3) can be written in matrix form as:

$$X - \mu = LF + e \qquad (4)$$

where ℓ_{ij} = loadings of the i^{th} variable on the j^{th} factor, L =the matrix of the factor loadings, the i^{th} = specific factor, e_i is the error term associated only with the i^{th} response on X_i. The p deviations, $(X_1 - \mu_1), (X_2 - \mu_2), \cdots, (X_p - \mu_p)$ are expressed in terms of $p + m$ random variables $F_1, \ldots, F_m$; $e_1, e_2, \cdots\cdots, e_p$. The estimation of factor loadings and specific variances in the orthogonal factor model approach is based on a number of assumptions relating $X, F,$ and Ψ which is specified in Equation (5)

$$\left.\begin{array}{ll} E(F) = 0_{(m\times 1)}; & E(e) = 0_{(p\times 1)}; \quad COV(e,F) = E(e,F') = 0_{(p\times m)}; \\ COV(F) = E(FF') = I_{(m\times m)}; & COV(e) = E(ee') = \Psi_{(p\times p)}. \end{array}\right\} \qquad (5)$$

Now, squaring both sides of Equation (4) and take the expected value of it yields

$$\Sigma = E(X-\mu)(X-\mu)' = Cov(X) = E[LF(LF)' + e(LF)' + LFe' + ee']. \qquad (6)$$

Furthermore, multiplying Equation (4) by F and taking the expected value yields

$$Cov(X,F) = E(X-\mu)F' = LE(FF') + E(eF') \qquad (7)$$

Substituting Equation (5) in (6) and (7) yields

$$\left.\begin{aligned} var(X_i) &= \ell_{i1}^2 + \cdots + \ell_{im}^2 + \psi_i \\ cov(X_i X_k) &= \ell_{i1}\ell_{k1} + \cdots + \ell_{im}\ell_{km} \\ Cov(X_{ij} F_j) &= \ell_{ij} \end{aligned}\right\} \qquad (8)$$

Equation (8) is gives a consistent estimates of the factor loadings and specific variances based on the orthogonal factor model technique.

The Principal Factor Model/Principal Component Method

This method adapts the spectral decomposition theorem by factoring out the covariate matrix inform of its eigenvalues and eigenvectors. The theorem states that

$$\Sigma = \lambda_1 e_1 e_1' + \cdots + \lambda_p e_p e_p' \qquad (9)$$

Equation (9) can be written as

$$\Sigma = \begin{bmatrix} \sqrt{\lambda_1}e_1 & \sqrt{\lambda_2}e_2 & \sqrt{\lambda_3}e_3 & \cdots & \sqrt{\lambda_p}e_p \end{bmatrix} \begin{bmatrix} \sqrt{\lambda_1}e_1' \\ \sqrt{\lambda_2}e_2' \\ \sqrt{\lambda_3}e_3' \\ \vdots \\ \sqrt{\lambda_p}e_p' \end{bmatrix} \qquad (10)$$

where the pair (λ_i, e_i) is the eigenvalue-eigenvector of Σ with $\lambda_1 \geq \lambda_2 \geq, \ldots, \geq \lambda_p$. This fits the covariance structure for the factor analysis model having many factors as variables (i.e. $m = p$) such that the loading matrix has the j^{th} column given by $\sqrt{\lambda_j}e_j$. Supposed that $p > m$ then, the $\lambda_{m+1}e_{m+1}e_{m+1}' + \cdots + \lambda_p e_p e_p'$ can be neglected and Equation (10) reduces to (11)

$$\begin{bmatrix} \sqrt{\lambda_1}e_1 & \sqrt{\lambda_2}e_2 & \sqrt{\lambda_3}e_3 & \cdots & \sqrt{\lambda_m}e_m \end{bmatrix} \begin{bmatrix} \sqrt{\lambda_1}e_1' \\ \sqrt{\lambda_2}e_2' \\ \sqrt{\lambda_3}e_3' \\ \vdots \\ \sqrt{\lambda_m}e_m' \end{bmatrix} = L_{(p\times m)}L'_{(m\times p)} + \Psi \qquad (11)$$

Equation (11) is gives a consistent estimates of the factor loadings and specific variances based on the principal factor model technique.

In order to examine the contribution of a factor on a variable as well as relative importance of each factor to a model, Ogum (2002) suggested that

$$\begin{bmatrix} proportion\ of\ the\ total \\ sample\ varianve\ due \\ to\ j^{th}\ factor \end{bmatrix} = \left.\begin{bmatrix} \dfrac{\lambda_j}{S_{11} + \cdots + S_{pp}}\ for\ an\ FA\ with\ covariance\ matrix\ S \\ \dfrac{\lambda_j}{p}\ for\ an\ FA\ model\ with\ correlation\ matrix\ \rho \end{bmatrix}\right\} (12)$$

Furthermore, the relative importance of each factor to a model is given as

$$\rho_{ij} = e_{ij}\sqrt{\frac{\lambda_i}{S_{jj}}} \qquad (13)$$

Note that Equations (4)-(13) are obtained from Ogum (2002).

Study Design and Data Presentation

In this section, the factor model to be estimated as well as the samples from four departments is presented.

2.3.1 Study Design

Let Y_1 be the score of students in Operation Research/Quantitative Techniques, (OR/QT) and Y_2 be the score of students in GNS (Use of English). Furthermore, let F_1 be the impact of quantitative ability of students' performance

and F_2 be the impact of verbal ability of students' performance then, the factor analysis model for this design is defined as:

$$\left.\begin{aligned} Y_1 &= \ell_{11}F_1 + \ell_{12}F_2 + e_1 \\ Y_2 &= \ell_{21}F_1 + \ell_{22}F_2 + e_2 \end{aligned}\right\} \qquad (14)$$

where ℓ_{ij} = loadings of the i^{th} variable on the j^{th} factor, F_1 = Quantitative ability, F_2 = Verbal ability, e_i = error term. Equation (14) is a regression-like model where the loadings can be treated as parameters and estimated through the OLS (Ordinary Least Square). However, the F_1 and F_2 are unobservable and hence we have no data on them. This account for why the regression approaches may fail and hence the orthogonal factor model and/or the principal factor method are used. Following from Equation (14), data were collected on the exam score of students in OR/QT and GNS. Convenient sampling method was used to draw two departments from school of technology namely HND I Statistics and HND I Computer Science and two departments from school of business studies namely HND I Accountancy and HND I Business Administration of the Federal Polytechnic Idah, Kogi State, Nigeria. The data are presented in Tables I-IV below.

2.3.2 Data Presentation

Table I: Scores in OR and GNS for Statistics Students

S/NO.	OR Score	GNS Score	S/NO.	OR Score	GNS Score	S/NO.	OR Score	GNS Score
1	53	51	22	81	62	43	42	35
2	62	49	23	67	48	44	71	59
3	74	60	24	50	42	45	61	50
4	62	51	25	56	51	46	51	46
5	69	57	26	86	70	47	57	44
6	15	22	27	48	44	48	51	41
7	52	52	28	87	68	49	50	42
8	68	55	29	58	51	50	59	46
9	46	46	30	54	54	51	56	49
10	68	53	31	54	48	52	63	47
11	76	56	32	95	69	53	53	50
12	60	47	33	45	38	54	52	48
13	81	65	34	71	60	55	50	38
14	50	46	35	77	60	56	70	59
15	66	53	36	71	61	57	65	44
16	70	62	37	70	61	58	65	55
17	49	50	38	72	64	59	67	54
18	42	40	39	59	50	60	55	37
19	62	50	40	58	47	61	55	47
20	59	50	41	48	42	62	60	49
21	88	73	42	52	35	63	64	46

Table II: Scores in OR and GNS for Computer Science Students

S/NO.	OR Score	GNS Score	S/NO.	OR Score	GNS Score	S/NO.	OR Score	GNS Score
1	69	58	25	69	59	49	58	54
2	63	59	26	57	49	50	69	58
3	58	49	27	58	43	51	76	68
4	59	55	28	49	50	52	68	57
5	55	43	29	46	38	53	59	57
6	57	51	30	65	59	54	77	60
7	45	45	31	71	68	55	77	60
8	95	78	32	65	51	56	69	72
9	80	67	33	76	67	57	64	54
10	47	42	34	93	74	58	55	51
11	59	49	35	53	42	59	84	72

12	56	46	36	30	26	60	53	46
13	74	63	37	78	59	61	94	78
14	62	51	38	51	45	62	82	69
15	64	54	39	52	42	63	59	51
16	64	55	40	76	67	64	47	44
17	61	61	41	57	59	65	54	49
18	79	63	42	45	38	66	66	66
19	58	53	43	65	60	67	61	51
20	86	71	44	60	56	68	59	48
21	43	38	45	61	55	69	51	47
22	72	61	46	66	53	70	62	56
23	65	62	47	56	50	71	53	50
24	68	60	48	71	59	72	40	37

Table III: Scores in OR and GNS for Accounting Students

S/NO.	QT Score	GNS Score	S/NO.	QT Score	GNS Score	S/NO.	QT Score	GNS Score
1	49	58	37	66	78	73	59	67
2	87	85	38	58	69	74	51	61
3	40	40	39	68	65	75	47	49
4	66	73	40	55	52	76	57	60
5	63	66	41	69	80	77	73	82
6	56	63	42	54	54	78	59	69
7	64	68	43	78	93	79	54	62
8	55	57	44	71	83	80	50	55
9	53	58	45	50	57	81	41	32
10	47	48	46	43	46	82	65	79
11	48	50	47	56	51	83	42	42
12	69	67	48	60	66	84	74	74
13	60	70	49	55	58	85	75	87
14	54	64	50	56	59	86	37	42
15	65	73	51	51	47	87	64	75
16	81	93	52	55	62	88	58	61
17	53	54	53	47	53	89	41	42
18	31	28	54	46	37	90	39	40
19	63	76	55	52	57	91	38	38
20	57	55	56	63	74	92	53	56
21	51	55	57	46	53	93	48	41
22	65	78	58	59	72	94	65	73
23	54	59	59	40	42	95	83	91
24	46	48	60	59	61	96	46	43
25	65	66	61	68	71	97	66	75
26	58	62	62	62	65	98	58	71
27	61	63	63	57	63	99	59	67
28	60	68	64	50	52	100	45	48
29	54	56	65	47	52	101	57	61
30	62	70	66	44	48	102	60	58
31	58	62	67	64	66	103	72	70
32	59	69	68	45	51	104	62	60
33	64	79	69	61	57	105	67	76
34	71	70	70	55	59	106	48	59
35	55	56	71	79	90	107	68	79
36	64	78	72	48	54	108	47	48

Table IV: Scores in OR and GNS for Marketing Students

S/NO.	QT Score	GNS Score	S/NO.	QT Score	GNS Score	S/NO.	QT Score	GNS Score
1	44	49	29	49	60	57	50	62
2	75	99	30	32	33	58	43	54
3	61	84	31	67	84	59	49	60
4	46	54	32	48	60	60	55	70
5	58	64	33	53	59	61	57	65
6	51	59	34	63	83	62	51	61
7	62	79	35	55	61	63	49	54
8	53	68	36	39	50	64	52	66
9	49	71	37	52	71	65	49	58
10	59	67	38	65	64	66	38	44
11	50	69	38	53	67	67	53	64
12	69	86	40	59	72	68	60	82
13	61	65	41	53	61	69	45	61
14	62	89	42	58	76	70	68	76
15	43	44	43	54	64	71	43	46
16	60	78	44	58	71	72	49	67
17	49	68	45	62	81	73	57	76
18	49	68	46	58	72	74	51	69
19	51	72	47	51	62	75	64	69
20	50	63	48	73	83	76	53	57
21	47	65	49	59	80	77	46	55
22	49	54	50	56	77	78	48	54
23	54	53	51	56	69	79	53	68
24	53	74	52	50	58	80	43	52
25	62	75	53	69	88	81	46	59
26	59	69	54	46	57	82	51	65
27	64	81	55	69	98	83	77	93
28	69	97	56	72	87	84	50	60

3.0 SUMMARY OF RESULTS, DISCUSSION AND CONCLUSSION

Summary of Results

The data in Tables I-IV are used to estimate the model in Equation (14) and the result obtained is presented in Table V.

Table V: Estimates of Factor Loadings, Specific Variances, Proportion, and Relative Importance

Schools			Loadings		Spec.Var.	Prop.	Rel. Importance	
Technology	Departments		ℓ_{i1}	ℓ_{i2}	e_i	p_j	ρ_{i1}	ρ_{i2}
	Computer Sc.	Y_1	12.54	10.03	-96.79	96.37%	0.99	0.97
		Y_2	-1.95	2.43	96.79	3.67%	-0.15	0.24
	Statistics	Y_1	13.02	8.99	-77.36	95.75%	0.99	0.96
		Y_2	-1.89	2.74	77.36	4.25%	-0.14	0.29
Business	Accounting	Y_1	10.25	1.89	-172.90	3.31 %	-0.23	0.14
		Y_2	-2.48	13.38	172.90	96.69%	0.97	0.99
	Marketing	Y_1	8.20	1.89	-152.52	5.02%	-0.34	0.15
		Y_2	-2.92	12.69	152.52	94.98%	0.94	0.98

3.2 Discussion of Results

Following from Table V, the summary of parameter estimates shows the followings:

1. For all the four departments considered, quantitative ability, F_1 has positive impact on OR/QT scores whereas, it has negative impact on GNS scores.

2. For all the four departments considered, verbal ability, F_2 has positive impact on both OR/QT and GNS scores.

3. For school of technology, quantitative ability accounted for the largest proportion ($95.75\% \leq p_j \leq 96.37$) whereas, verbal ability accounted for the least proportion ($3.67\% \leq p_j \leq 4.25\%$). For school of business studies, verbal ability accounted for the largest proportion ($94.98\% \leq p_j \leq 96.69\%$) while quantitative ability accounted for the least proportion ($3.31\% \leq p_j \leq 5.02\%$).

4. The two factors have significant relative importance ($0.96 \leq \rho_{ij} \leq 0.99$) for OR scores of students in school of technology and a week relative importance for GNS score ($-0.15 \leq \rho_{ij} \leq 0.29$). On the other hand, the two factors have significant relative importance ($0.94 \leq \rho_{ij} \leq 0.99$) for GNS scores of students in school of business studies and a week relative importance for QT score ($-0.345 \leq \rho_{ij} \leq 0.15$).

5. In the overall, quantitative ability impact positively on OR/QT scores whereas, verbal ability impact positively on GNS scores.

3.3 Conclusion

In modeling cause-effect relationship, data are required to form prior, likelihood, and posterior distributions. Most often, getting data about priors may not be feasible and hence conventional regression models may fail and factor analysis model becomes the only alternative in such scenario. This study is based on the assumption that quantitative ability has positive impact on student performance in OR/QT whereas verbal ability has positive impact on student performance in GNS. Based on the data collected from the two schools, it can be concluded that students with good quantitative background performs well in Operations Research/Quantitative Techniques while those with good verbal ability background perform well in General Studies courses (Use of English). In the overall, candidate's previous background (prior criteria) should serve as a significant determining factor for his/her course of future studies.

References

Ahamed A. T. M. S, Tanzeem N. M. and Rahman R. M (2017), An intelligent system to predict academic performance based on different factors during adolescence, *Journal of Information and Telecommunication*, **1** (2), 155-175

Anderson, D. R., and Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society* B **36**, 99-102.

Bendangnuksung, C., and Prabu, P (2018) Students Performance Prediction using Deep Neural Net International Journal of Applied Engineering Research **13**, 1171-1176.

Banerjee, U.S. and Gupta, S. (2012). Assessment of significant sources influencing the variation of water quality of the river damodar through factor analysis. *Asian Journal of Water, Environment and Pollution* **9**, 87-94.

Bosse, Y., Despres, J-P., Chagnon, Y. C., Rice, T., Rao, D. C., Bouchard, C., Perusse, L., and Marie-Claude, M-C. (2007). Quantitative trait locus on 15q for a metabolic syndrome variable derived from factor analysis. *Obestity* **15**, 544-550.

Dickinson, D., Goldberg, T. E, Gold, J. M., Elvevag, B., Daniel R., Weinberger, D. R. (2011). Cognitive factor structure and invariance in people with schizophrenia, their unaffected siblings, and controls. *Schizophr. Bull.* **37**, 1157-1167.

Johnson, R. A. and Wichern, D, W. (2001). *Applied Multivariate Statistical Analysis*, 5th edition. Prentice Hall.

Nkiruka, O. E. , Uchenna, O. J. , Ayuba, Y. , Josaphat, A. , and Abdulrasheed, B. (2017). Application of Factor Analysis on Academic Performance of Pupils. *American Journal of Applied Mathematics and Statistics*, *5*(5), 164-168.

Ogum, I. C. (2002). *Introduction to Methods in Multivariate Analysis*, Afri Towers Limited, Abah- Nigeria

Öğütcü, S. (2020) Student Performance Score Prediction Using Artificial Neural Network with the Support of Exploratory Factor Analysis and Clustering, *2020 IEEE Third International Conference on Data Stream Mining & Processing (DSMP)*, 90-95,

Shamsuddeen S., Ahmad L., Umar U., Gulumbe, S. U., Aminu B. M, (2019) Student's Academic Performance Prediction Using Factor Analysis Based Neural Network, *International Journal of Data Science and Analysis*. **5**(4) , 61-66.

Witten, D., Tibshirani, R., Gu, S. G., Fire, A., and Weng-Onn Lui, W-O. (2010). Ultra-high throughput sequencing-based small RNA discovery and discrete statistical biomarker analysis in a collection of cervical tumours and matched controls. *BMC Biology*. **8**, 1741- 7007.