Exploratory Factor Analysis of Demographic Characteristics of Antenatal Clinic Attendees and their Association with HIV Risk

Wilbert Sibanda

Philip D. Pretorius

*School of Information Technology, North-West University, Vaal
Triangle campus, Vanderbijlpark, South Africa, 1900
Email: Wilbert Sibanda@nwu.ac.za*

Doi:10.5901/mjss.2014.v5n20p303

Abstract

This research was conducted to determine the applicability of the exploratory factor analysis (EFA) technique to study the correlation between demographic characteristics and HIV risk amongst pregnant women attending antenatal clinics in South Africa. EFA was therefore used as a factor reduction technique to identify the number of latent constructs and underlying factor structure amongst demographic characteristics with regards to their influence on the risk of HIV infection. An iterated principal axis factor (IPAF) technique with three factors and varimax rotation was used as a method of extraction. IPAF method refines the communalities until they converge and has the advantage of analysing both the correlation and covariances. The findings revealed a high positive correlation between agewoman and agepartner (0.81), agewoman and gravidity (0.69), agewoman and parity (0.67) and gravidity and parity (0.93). However, a negative correlation was observed between parity and educational level of the pregnant woman. Based on a scree plot of eigenvalues against demographic characteristics, three components were selected that accounted for 55 percent of total variance. The interpretation of the principal components was based on determining which demographic characteristics factors were strongly correlated within each component. The first component was highly correlated with time dependent demographic characteristics such as age of pregnant women, male sexual partners, gravidity and parity. The second component was found to be highly correlated with spatially related demographic characteristics such as province. The third component was correlated with sexually transmitted diseases such as HIV and syphilis. However, the study showed that the three extracted components were not at all correlated. In summary, this research demonstrated that it is possible to reduce the annual South African antenatal HIV seroprevalence data from eleven demographic characteristics to three principal components using a factor analysis with a principal component extraction method.

Keywords: *exploratory factor analysis; demographic characteristics; HIV; South Africa*

1. Introduction

The main reason of conducting research in epidemiological and biological sciences is to gather enough data to provide a basis for both short and long-term sound decisions. In South Africa, the annual antenatal HIV survey is the only existing national surveillance for determining HIV prevalence (Sibanda & Pretorius, 2011a). The National Department of Health in South Africa has been conducting antenatal surveys since 1990 (Department of Health, 2010). Antenatal clinic data contains various demographic characteristics for each pregnant woman such as; pregnant woman's age, marital status, race, level of education, gravidity, parity, name of clinic, HIV and syphilis results.

This study aims to use exploratory factor analysis to study the correlations between a pregnant woman's demographic characteristics and the risk of acquiring an HIV infection. This research is preceded by our earlier work that used various data mining techniques such as factorial design, multilayer perceptrons (Sibanda & Pretorius, 2011b) and response surface methodologies (Sibanda & Pretorius, 2012) that demonstrated that the age of the pregnant woman was highly correlated to the risk of acquiring an HIV infection amongst antenatal clinic attendees in South Africa. The above results were confirmed by our recent research that involved the development and validation of an HIV risk scorecard model (Sibanda & Pretorius, 2013), based on 2007 South African annual antenatal HIV seroprevalence data.

2. Literature Review

This research used the 2010 annual South African antenatal HIV seroprevalence data.

2.1 Factor Analysis

Factor analysis is a popular statistical technique to extract a small number of factors from a large set of variables. The relationship of factors to each other is determined by the rotation technique selected during the analysis. The most popular rotation technique is varimax (Costello & Osborne, 2005). A varimax rotation attempts to simplify the columns of the factor matrix achieving the maximum simplification when only ones and zeroes are present in the columns of the matrix. It results in factors that are independent of each other. Exploratory factor analysis (EFA) is therefore a variable reduction technique (Suhr, 2006). EFA hypothesises an underlying construct, a variable not measured directly, and estimates factors which influence responses on desired variables.

Factor analysis is based on the correlation matrix of the variables involved, and correlations usually need a large sample size before they stabilise (Tabachnik, 2007)

2.2 Factor Extraction

In this research, an iterated principal axis factor method with three factors was used as a method of extraction and a varimax rotation technique was used. The iterated principal axis method is the most commonly used technique that inherently improves the communalities with each successive iteration until convergence. Varimax is a method used to explain variance and its advantage is that it can analyse both correlation and covariance.

The determination of the number of factors to extract should be guided by theory, but also informed by running the analysis, extracting a few different factors, and observing which number of factors provides optimal results. The scree plot was used as a graphical display of the variance of each component in the dataset in order to determine how many components should be retained to explain a high percentage of the variation in the data. A scree plot shows the eigenvalues on the y-axis and the number of factors on the x-axis. The variance of each component is calculated using the following formula:

$$Variance = \frac{\mu}{\sum_{n=1}^{n} \mu i} \quad (1)$$

Where μi is the ith eigenvalue and $\sum_{n=1}^{n} \mu i$ is the sum of all eigenvalues

The scree plot shows the variance for the first component and then for the subsequent components and thus shows the additional variance that each component is adding. The purpose of a factor analysis is to reduce the large number of variables that describe a complex concept to a few understandable intrinsic factors. Therefore, the purpose of the study is to end up with a small number of factors that explain the maximum amount variability in the HIV antenatal data.

2.3 Communalities

After the extraction of factors, the researchers studied the communalities, which explained the variance in each of the original variables that was explained by the extracted factors. The aim was to obtain higher communalities. Therefore, variables with communalities below 0.5 were considered ideal for exclusion from the study, as these variables contained below half of the variance in the original variable. This was used as a proportion of each variable's variance that could be explained by the factors.

2.4 Total Variance Explained

Total variance was used as a measure of the total amount of variability of the original variables explained by each factor solution. More factors provide more variance, however additional variables explain little variation. Every factor analysis has the same number of factors as it does variables, and those factors are listed in the order of the variance they explain. It is always easier to have more total variance by keeping more factors in the solution, but later factors explain so little variation. Therefore, if a given number of factors explain the most of the variability in the original variables, then those factors are a good substitute for all the factors. It is therefore logical to remove the rest of the factors without compromising the original variability.

2.5 Rotated Pattern Matrix

The extraction method was the principal component technique. Oblimin rotation method with Kaiser normalisation was utilised. Oblimin rotation is a general form for obtaining oblique rotations used to transform vectors associated with factor analysis to a simpler structure. The rotated pattern matrix also provides information on the factors that are highly correlated with each component.

2.6 Structure Matrix

The structure matrix provides information on the correlation between variables and components.

2.7 Component Correlation Matrix

This was used to determine correlation between components.

3. Findings and Discussion

3.1 Factor Analysis

3.1.1 Descriptive Statistics

Table 1. Descriptive statistics of demographic characteristics

	Mean	Std. Deviation	Analysis N
Prov	4.42	2.443	30327
District	23.67	14.673	30327
Age woman	25.36	6.332	30327
Population Group	1.20	0.609	30327
Education level	1.93	0.510	30327
Marital status	1.22	0.441	30327
Gravida	2.12	1.227	30327
Parity	1.03	1.173	30327
Age partner	29.74	7.530	30327
RPR result	0.01	0.121	30327
HIV result	0.30	0.458	30327

Table 1 shows that the average age of a pregnant woman attending an antenatal clinic for the first time in South Africa in the year 2010 is about 25.36 years old. Her male sexual partner's age is on average 29.74 years old. In addition, on average a pregnant woman attending an antenatal clinic has a primary education. The vast majority of these pregnant women were not married and present to the clinic for their second pregnancy. Due to the fact that this survey is conducted in public health institutions, the majority of women were found to be African.

3.1.2 Correlation matrix

Table 2. Correlation matrix

	Province	District	Agewoman	Race	Education	Marital status	Gravidity	Parity	Agepartner	Syphilis	HIV
Prov	1.00	-0.24	0.02	0.38	-0.03	0.03	0.03	0.02	0.00	-0.18	-0.08
Distr	-0.24	1.00	-0.03	-0.15	-0.03	-0.03	-0.01	0.01	-0.02	-0.01	0.02
Agewoman	0.02	-0.03	1.00	0.02	-0.11	0.33	0.69	0.67	0.81	0.06	0.18
Race	0.38	1.00	0.00	1.00	-0.02	0.07	0.04	0.02	-0.05	0.02	-0.17
Educat	0.03	-0.03	-0.11	-0.02	1.00	-0.03	-0.24	-0.25	-1.17	-0.01	-0.05
Marital	0.03	-0.03	0.33	0.07	-0.03	1.00	0.32	0.31	0.33	-0.02	-0.04
Gravida	0.03	-0.01	0.69	0.04	-0.24	0.32	1.00	0.93	0.62	-0.04	0.11
Parity	0.02	0.01	0.67	0.02	-0.25	0.31	0.93	1.00	0.61	-0.02	0.10
Agepartner	0.01	-0.01	0.81	-0.05	-0.12	0.33	0.62	0.61	1.00	0.06	0.18
Syphilis	-0.02	0.02	0.01	0.02	-0.01	-0.02	-0.00	-0.00	0.01	1.00	0.04
HIV	-0.08		0.18	-0.17	-0.05	-0.04	0.11	0.10	0.18	0.04	1.00

High correlations were observed between agepartner and agewoman (0.81), gravidity and agewoman (0.69), agewoman and parity (0.67), gravidity and agepartner (0.62), gravidity and parity (0.93) and parity and agepartner (0.61).

A negative correlation was observed between parity and educational level of the pregnant woman (-0.25). This might be suggesting that, as the educational level of the pregnant woman increases, there is a decrease in the number of children a woman makes. The same pattern was observed for the number of pregnancies (gravidity). It therefore means that education brings about awareness of the need to reduce family sizes in order to improve the living standard of individuals.

3.1.3 Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy

Table 2. KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.741
	Approx. Chi-Square	127962.896
Bartlett's Test of Sphericity	df	55
	Sig.	0.000

The KMO measure of 0.741 indicated that principal component analysis was an appropriate method for this study. KMO predicts if data are likely to factor well, based on correlation and partial correlation. Therefore, KMO measures sampling adequacy as an index used to examine the appropriateness of factor analysis. In general, a KMO value greater than 0.6 is considered to meet the minimum criteria, while values less than 0.5 imply that factor analysis may not be appropriate.

Bartlett's test of sphericity tests the hypothesis that the correlation matrix is an identity, *inter alia* all diagonal elements are zero, implying that all of the variables are uncorrelated. If the significant value for the test is less than the alpha level, the authors reject the null hypothesis and conclude that there are correlations in the data set that are appropriate for factor analysis. This analysis meets this requirement.

3.1.4 Communalities

Table 3. Communalities

	Initial	Extraction
Prov	1.000	0.579
District	1.000	0.266
Age woman	1.000	0.768
Population Group	1.000	0.568
Education level	1.000	0.072
Marital status	1.000	0.238
Gravida	1.000	0.817
Parity	1.000	0.802
Age partner	1.000	0.706
RPR result	1.000	0.000
HIV result	1.000	0.184

Extraction method: principal component analysis

Communalities provide information on how much of the variance in each of the original variables is explained by the extracted factors. Higher communalities are desirable. If the communality for a variable is less than 0.5, it is a candidate for exclusion from the analysis because the factor solution contains less than half of the variance in the original variable and the explanatory power of that variable might be represented better by the individual variable.

Therefore, extraction communalities are estimates of the variance in each variable accounted for by the components. The communalities after six-factor extraction are high compared to factors less than six, which indicate that the extracted components represent the variables well. If the communalities are very low in a principal components extraction, there is no need to extract another component.

3.1.5 Total Variance Explained

The total column gives the eigenvalue, or amount of variance in the original variables accounted for by each component. The percentage of variance column gives the ratio, expressed as a percentage, of the variance accounted for by each component to the total variance in all of the variables. The cumulative percentage column gives the percentage of variance accounted for by the first number of components.

The second section of Table 4 shows the extracted components. They explain nearly 80.65 percent of the variability in the original eleven variables hence the complexity of the data set can be reduced considerably by using these components, with only a 20 percent loss of information. The rotation maintains the cumulative percentage of variation explained by the extracted components, but that variation is now spread more evenly over the components. The larger changes in the individual totals suggest that the rotated component matrix will be easier to interpret than the unrotated matrix.

Table 4. Total variance explained

Component	Initial Eigenvalues			Extraction sum of loaded squares			Rotation sum of loaded squares
	Total	% variance	Cumulative %	Total	% variance	Cumulative %	Total
1	3.42	31.07	31.07	3.42	31.07	31.07	3.41
2	1.58	14.38	45.45	1.58	14.38	45.45	1.60
3	1.05	9.55	55.00				
4	1.03	9.34	64.34				
5	0.98	8.87	73.21				
6	0.82	7.43	80.65				
7	0.75	6.81	87.46				
8	0.60	5.47	92.93				
9	0.52	4.73	97.65				
10	0.19	1.69	99.34				
11	0.73	0.66	100.00				

3.1.6 Scree Plot

The scree plot helps to determine the optimal number of components. The eigenvalue of each component in the initial solution is plotted in order to extract the components on the steep slope. The components on the shallow slope contribute little to the solution. The last big drop occurs between the second and third components. The first three components, as shown in Figure 1, were used.

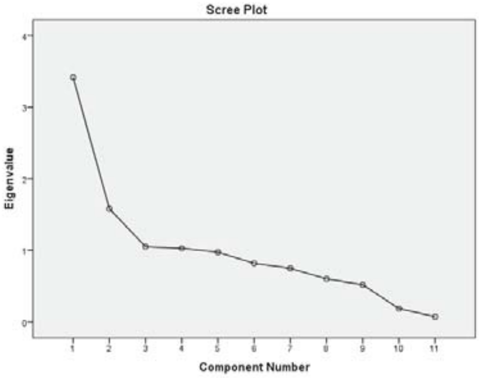

Figure 1. Scree plot

3.1.7 Rotated Component Matrix

The rotated component matrix helps to determine what the components represent. The first component is most highly correlated with gravidity (0.90), parity (0.90), agewoman (0.87) and agepartner (0.83). Therefore, gravidity, parity, agewoman and agepartner were found to be highly correlated and increased together. Gravidity was found to be a better representative of the first component because since it was less correlated with the other two components. The second component was highly correlated with province (0.78), while the third component was found to be highly correlated with HIV result. This suggests that a researcher can focus on gravidity, province and HIV for further studies. As a rule of the thumb, a variable <0.3 is not contributing significantly to that component.

Table 5. Description of components

Component	Demographic characteristics	Description
1	Gravidity	This component contains, ages of the pregnant women and their male sexual partners as well as the number of pregnancies and children born to the woman
	Parity	
	Agewoman	
	Agepartner	
2	Provine	This component represents the geographical location of the antenatal clinic
3	HIV	This component represents sexually transmitted diseases
	Syphilis	

Table 5 shows the distribution of demographic characteristics within the three components as presented by the structure matrix. The structure matrix represents correlation between variable and component. Interestingly, demographic characteristics related to the age of the pregnant woman and her male sexual partner, are found to be highly correlated in component one. These characteristics are also correlated closely to the number of pregnancies and children the woman had. The second component is characterised by the prominence of the geographical area from which the pregnant woman is from, such as her province. Sexually transmitted diseases are prominent in this component.

3.1.8 Component Plot in Rotated Space

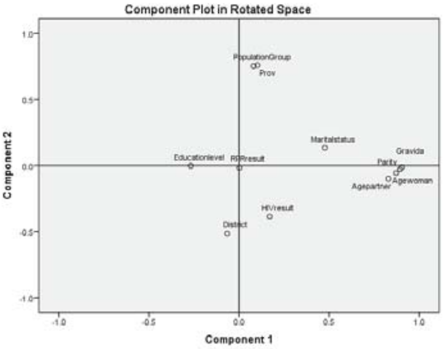

Figure 2. 2-D component plot in rotated space

The 2-D component plot shows that gravidity, parity, agepartner, agewoman, population group (race) and province are highly correlated to single components. Therefore, scoring highly on a single component means that these demographic characteristics are explained overwhelmingly by a single component.

4. Concluding Remarks

This research has shown that it is possible to reduce the size of the data file from eleven demographic characteristics to three components using a factor analysis with a principal component extraction. The age of the pregnant woman, age of her male sexual partner, gravidity and parity were found to be highly correlated, and thus placed in one component. Sexually transmitted diseases were found to be highly correlated and belonged to another component. Lastly, demographic characteristics that related to geographical location of the antenatal clinic such as province were also found to belong to another component.

5. Acknowledgements

Wilbert Sibanda acknowledges funding from the South African Centre for Epidemiological Modeling (SACEMA), Medical Research Council (MRC) and North-West University. This work is based in part on the research supported by the National Research Foundation of South Africa (Grant no. 86946). Any opinion, findings and conclusions or recommendations expressed in this material are those of the author(s) and therefore the NRF does not accept any liability in regard thereto.

References

Department of Health, (2010). *National antenatal sentinel HIV and syphilis prevalence survey in South Africa*.

Costello A. B. & Osborne J. W (2005). Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation,*10 (7).

Sibanda, W. & Pretorius, P. D. (2011a). Application of two-level factorial design to determine and optimize the effect of demographic characteristics on HIV prevalence using the 2006 South African annual antenatal HIV and syphilis seroprevalence data. *International Computer Applications*, 35 (12), 15-20.

Sibanda W. and Pretorius, P. D. (2011b). Novel application of multilayer perceptrons (MLP) neural networks to model HIV in South Africa using seroprevalence data from antenatal clinics. *International Journal of Computer Applications*, 35 (5), 26-31.

Sibanda W. & Pretorius P. D. (2012). Response surface modeling and optimization to elucidate the differential effects of demographic characteristics on HIV prevalence in South Africa. In proceedings of *2012 IEEE/*ACM International Conference on Advances in Social Networks Analysis and Mining, Instabul, pp. 818-826.

Sibanda W. & Pretorius P. D., (2013) Development and validation of an HIV risk scorecard model. In proceedings of IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Niagara, pp. 916-922.

Suhr D. D. (2006). Exploratory or confirmatory factor analysis? In proceedings of the Thirty First Annual SAS users Group International Conference, Cary, NC, SAS Institute, Paper 200-31.

Tabachnik B. G. & Fidell L. S. (2007). Using multivariate statistics. (5th ed.). Boston: Pearson/Allyn and Bacon.