Principal Component Analysis of Factors Affecting Ovulation Interval of Selected Women

Matthew Chukwuma Michael¹*, Eriobu Nkiru Obioma² and Ashinze Mkpuruoma Akudo³

¹²Department of Mathematics and Statistics, Nigeria
³Department of Statistics, Nigeria

Abstract

The factors affecting ovulation interval, like another biological phenomenon are numerous and interdependent. This paper identified the factors to include Age, Height, Weight, Work Time (Stress), Menstrual Duration, Number of Conceptions, Number of Births and Exposure to Sun. It is the aim of this paper to formulate a smaller number of independent random variables such that minimum information is lost. A total of two hundred (200) women in their reproductive age interval were selected for the study. Questionnaires were used to get the relevant information from them. Eighty eight percent (88%) of the total variation in the ovulation interval was accounted for by six (6) principal components compared to the original eight random variables. It was observed that the most important factors are Stress, Height, Menstrual Duration, Weight and Age, and that only Exposure to Sun did not show any significance in affecting Ovulation Interval.

Keywords: Principal Component; Ovulation; Conception; Factors and multivariate

Methodology

A total of 200 women in their reproductive age were selected for the study. The women were mostly teachers in primary and secondary schools, health workers in public and private health facilities and literate women in market places. The data collection process took place in a period of six weeks. Questionnaires were administered to the women and they supplied the necessary information under proper supervision. The major factors considered for analysis are Age, Height, Weight, Work Time (Stress), Menstrual Duration, Number of Conceptions, Number of Births and Exposure to Sun.

The method of data analysis is Principal Component Analysis. It is a multivariate technique used in data reduction and data analysis [4,5]. The aim of principal component analysis is to ascertain if the joint variation in p variables \(x_1, x_2, x_3, ..., x_p\) can be represented approximately in terms of the joint variation of a fewer number, say...
Let \( k < p \), of hypothetical variables without much loss of information. That is, \( X_{\text{new}} \) is replaced by a linear transformation,

\[
y_{\text{new}} = A'_{p \times p} X_{\text{old}}
\]

Where \( Y_{\text{new}} \), \( k < p \) are significant and approximately chosen from \( Y_{\text{old}} \). The \( Y_{\text{new}} \) are interpretable in terms of the original problem and they are independent [5,6].

\[
y_i = \sum_{j=1}^{p} a_{ij} x_j, \quad i = 1, 2, \ldots, p
\]

The Principal Components (PC) if they are uncorrelated and their variances are as large as possible.

\[
\text{var}(Y) = a_i^2 \sum_i a_i^2, \quad \text{cov}(Y, Y) = a_i a_k \sum_i a_i a_k = 0, \quad i, k = 1, 2, 3, \ldots, p \quad (3)
\]

If the mean vector and covariance matrix of \( X \) are given as and respectively, and has eigen value \( \lambda_1 \), \( \lambda_2 \), \( \ldots \), \( \lambda_p \) with \( \lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \leq \lambda_p \geq 0 \), the 1\( \text{st} \) Principal Component Analysis is given by

\[
i = 1, 2, \ldots, p
\]

With

\[
\text{var}(Y) = \lambda_i, \quad \text{cov}(Y, Y) = \sum_i a_i a_k = 0, \quad i \neq k \quad (5)
\]

The total variance is given as

\[
\sum_{i=1}^{p} \text{var}(Y) = \sum_{i=1}^{p} \lambda_i
\]

Thus, the proportional variance explained by the \( k \text{th} \) principal component is

\[
\zeta_k = \frac{\lambda_k}{\sum_{i=1}^{p} \lambda_i}; \quad k = 1, 2, \ldots, p
\]

If \( 80\% \) to \( 90\% \) of the total population variance for large can be attributed to the first one, two or three components, these components can be used to replace the original variables without much loss of information.

When the variables, \( x_1, x_2, x_3, \ldots, x_p \) are in different units, two or more variables are measured on vastly different ranges or standardized random variables are used instead of the original random variables, the correlation matrix, \( x \), is used to calculate Principal Components (PC) instead of \( \sum [4] \).

**Data Analysis**

The eigen values for the eight random variables considered are 1.8989, 1.2740, 1.1409, 0.9970, 0.8840, 0.8225, 0.6101 and 0.3707 with the following respective proportions 0.237, 0.158, 0.143, 0.125, 0.110, 0.103, 0.076 and 0.047 and cumulative proportions 0.237, 0.397, 0.539, 0.664, 0.774, 0.877, 0.953 and 1.000. The independent Principal Components are thus represented as,

\[
y_{1} = -0.441x_{1} - 0.139x_{2} - 0.023x_{3} + 0.032x_{4} - 0.125x_{5} - 0.604x_{6} - 0.593x_{7} - 0.050x_{8}
\]

\[
y_{2} = 0.156x_{1} + 0.518x_{2} + 0.270x_{3} - 0.331x_{4} + 0.323x_{5} - 0.101x_{6} - 0.274x_{7} - 0.579x_{8}
\]

\[
y_{3} = 0.200x_{1} - 0.162x_{2} + 0.368x_{3} - 0.577x_{4} - 0.669x_{5} - 0.069x_{6} + 0.078x_{7} + 0.086x_{8}
\]

\[
y_{4} = -0.451x_{1} + 0.430x_{2} + 0.260x_{3} - 0.274x_{4} + 0.221x_{5} - 0.034x_{6} + 0.056x_{7} + 0.645x_{8}
\]

\[
y_{5} = -0.055x_{1} + 0.207x_{2} - 0.801x_{3} - 0.518x_{4} - 0.084x_{5} - 0.160x_{6} + 0.109x_{7} - 0.0030x_{8}
\]

\[
y_{6} = 0.329x_{1} - 0.650x_{2} + 0.126x_{3} - 0.425x_{4} + 0.447x_{5} + 0.206x_{6} + 0.037x_{7} - 0.173x_{8}
\]

Where \( x_1 = \text{Age}, \quad x_2 = \text{Height}, \quad x_3 = \text{Weight}, \quad x_4 = \text{Work Time}, \quad x_5 = \text{Menstrual Duration}, \quad x_6 = \text{Number of Conceptions}, \quad x_7 = \text{Number of Births} \) and \( x_8 = \text{Exposure to Sun} \).

The six principal components account for 87.7\%, approximately 88\% of the total variation in the ovulation interval of the studied population. Hence, further analysis of the data demanding the condition of independence of data set can be done using the six principal components.

**Conclusion**

The original data which was collected in eight dimensions known to be correlated has been successfully reduced into six principal components which are shown to be uncorrelated. The six principal components account for 88\% of the total variation in ovulation interval of the studied population. Hence, further analysis of the data demanding the condition of independence of data set can be done using the six principal components.

**Acknowledgement**

None.

**Conflict of Interest**

No conflict of interest.

**References**