MS -Business Intelligence & Analytics

Fall 2015

**BIA – 652 C**

*December 1, 2015*

**Mohit Ravi Ghatikar**

**CWID - 10405877**

Multivariate Data Analytics – HW6

**Ethics Statement**

I pledge on my honor that I have not given or received any unauthorized assistance on this assignment /examination. I further pledge that I have not copied any material from a book, article, the Internet or any other source except where I have expressly cited the source.

Signature \_Mohit Ravi Ghatikar\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date: 12/01/2015\_\_\_

**Question:**

**14.1**

For the depression data set described in Appendix A, perform a principal

components analysis on the last seven variables DRINK–CHRONILL (Table3.3). Interpret the results

**Solution:**

Before running Principle component analysis on the seven categorical variables, we run a univariate and co-relation procedure to get a sense of the data. There are a few variables that a degree of co-relation between them. For example, Acute Illness and Beddays have a co-relation of 0.42. The objective of PCA is to reduce the dimensionality and to remove co-relation.

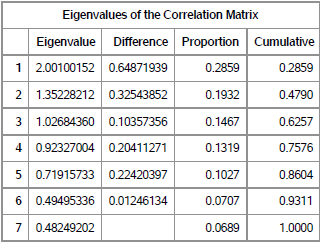
PCA is run of the seven variables by normalizing the data with a mean=0 and Standard deviation=1. The Eigen values are given below. These values explain the degree of variation.

There are three methods to choose eigenvalues:

1) Choose Eigenvalues greater than 1. In this case we would choose the first three Eigenvalues.

2) Set a minimum amount of variance percentage explained by the Eigenvalues. For example, if we set 70% as our limit for variance explained, then the first 4 Eigenvalues would be selected since they explain about 75.6% of the variance.

3) Check the Scree plot and retain those components in the steep curve before the first point that starts the flat line trend.



In our analysis, we choose the first method and select Eigenvalues greater than 1. The coefficients or the loading factors are determined by the Eigenvectors. Therefore, we select three Principal components whose equations are:

a) Principal Component 1 = 0.17 \* Drink + 0.45 \* health – 0.12\* Reg\_Doc + …. + 0.53 \*Chronic Illness.

b) Principal Component 2 = -.28\*Drink + -.27\*Health +……….+ -0.19\*Chronic Illness.

c) Principal Component 3 = 0.67 \* Drink +0.25 \* Health + ……..+ -.08 \* Chronic Illness.

**Question:**

**14.6**

Using the family lung function data described in Appendix A define a new variable RATIO = FEV1/FVC for the fathers. What is the correlation between RATIO and FEV1? Between RATIO and FVC? Perform a principal components analysis on FEV1 and FVC, plotting the results. Perform a principal components analysis on FEV1, FVC, and RATIO. Discuss the results.

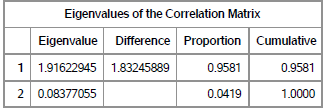
**Solution:**

Firstly, We input a new column in the lung dataset where Ratio = FEV1\_father / FVC\_father.

The co-relation between Ratio and FEV1\_father is **0.18** and the co-relation between Ratio and FVC\_father is **-0.21**.

Before running the PCA we first normalize the data for Ratio, FEV1\_father and FVC\_father with a mean=0 and Standard deviation=1.

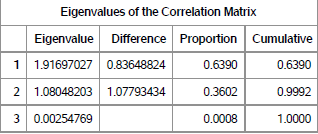
We run a PCA on FVC\_father and FEV1\_father. We will choose only the first Eigenvalue which explains 98% of the variance.



The co-relation between FVC\_father and FEV1\_father is 0.91 which is very high. So after running the PCA we can eliminate one dimension. Therefore, the principal component equation Is:

Principal Component 1 = 0.707 \* FEV1\_father + 0.707 \* FVC\_father.

After this, we run a PCA on ratio, FVC\_father and FEV1\_father. We will choose the first two Eigenvalues since they are greater than 1. They explain about 99% of the variance.



By doing this we eliminate one variable and reduce the dimensionality. The principal component equation Is:

a) Principal Component 1 = 0.703 \*FEV1\_father + 0.709 \* FVC\_father – 0.02 \* Ratio.

b) Principal Component 2 = 0.21\* FEV1\_father – 0.173 \* FVC\_father + 0.96 \* Ratio.

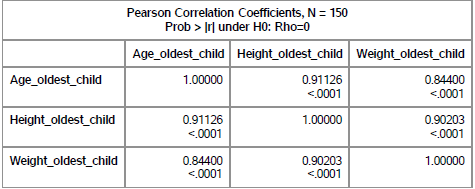
**Question:**

**14.7**

Using the family lung function data, perform a principal components analysis on age, height, and weight for the oldest child.

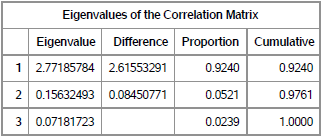
**Solution:**

We first run a co-relation on age, height and weight of the oldest child. The matrix is given below:



As we can see the variables are highly co-related to each other. With the PCA procedure, we can remove the co-relation and reduce the dimensions.

Before running the PCA, we first normalize the data for Age\_oldest\_child, Height\_oldest\_child and Weight\_oldest\_child with a mean=0 and Standard deviation=1.



We will choose only the first Eigenvalue since its greater than 1. It explains 92% of the variance.

The principal component equation is:

Principal Component 1 = 0.57 \* Age\_oldest\_child + 0.58 \* Height\_oldest\_child + 0.57 \* Weight\_oldest\_child.

Here we have reduced the dimensions from three to one.