Diabetes Study Project

Mohammad Anas Khaddam

Department of Mathematics, University of South Carolina Upstate

STATU599-01: Seminar in Statistics

Dr. Bernard Omolo

2022, April 25

Contents

Diabetes Study Project

Introduction:

Background:

You are working as a consulting statistician for a company that has a contract with a medical researcher. She has gathered data on 60 adult female patients for a diabetes study. The variables measured include health and demographic variables for the females. The 7 variables she has are:

Preg = Number of times pregnant

Glu = Plasma glucose concentration (based on an oral glucose tolerance test)

DBP = Diastolic blood pressure (mm Hg)

TST = Triceps skinfold thickness (mm)

Insulin = Two-Hour serum insulin (mu U/ml)

BMI = Body mass index (weight in kg/(height in m^2)

Age = Age of patients (years)

Research Questions and Task:

The medical researcher wanted the following questions to be answered:

1. Are there individual females who are highly unusual (in any way) based on the measured health variables only? If so, identify their numbers.

2. Are there notable associations/relationships between some of the variables? (if so, describe them)

3. Is there a way to graphically represent the raw data for the 60 patients and draw conclusions about the data set from such a graph?

4. Can we find a few indices that describe the variation in the data set using a lesser dimension than the original set of variables? If so, what are those indices? Is there a convenient interpretation of any of the indices?

5. Can we graphically display the data in a low number of dimensions using such indices? What conclusions about the patients (individual patients or groups of patients) can you draw from such a graph?

6. What are any other potentially interesting aspects of the dataset?

You will prepare a roughly 6-page report detailing your analysis of the data and your conclusions. Keep in mind that the report should be written for two audiences: the medical researcher, who has a sense for numbers but is not an expert in statistics, and your own supervisor at the statistical consulting company, who will be judging you and deciding on your possible promotion based on the statistical competency of the report. Your report should be understandable and meaningful to both audiences.

Methods:

Principal Components Analysis: The intellipaat's website states," Principal Component Analysis (PCA) is a statistical technique used for data reduction without losing its properties. Basically, it describes the composition of variances and covariances through several linear combinations of the primary variables, without missing an important part of the original

information. In another term, it is about obtaining a unique set of orthogonal axes where the data has the largest variance. Its main aim is to overcome the dimensionality of the problem. The reduction of dimensionality should be such that when dropping higher dimensions, the loss of data is minimum. Also, the interpretation of principal components can explain associations among variables that are not visible at first glance. It helps analyze the scattering of the observations and recognize the variables responsible for distribution." (What is PCA, 2020)

Mahalanobis Distance: A way to detect multivariate outliers; according to Machine Learning Plus website," Mahalanobis distance is an effective multivariate distance metric that measures the distance between a point (vector) and a distribution. It has excellent applications in multivariate anomaly detection, classification on highly imbalanced datasets and one-class classification and more untapped use cases. Considering its extremely useful applications, this metric is seldom discussed or used in stats or ML workflows." (Prabhakaran 2022)

Correlation between variables: This is used to find the relationships among variables(features). According to the Machine Learning Mastery website ," The statistical relationship between two variables is referred to as their correlation. A correlation could be positive, meaning both variables move in the same direction, or negative, meaning that when one variable's value increases, the other variables' values decrease. Correlation can also be neutral or zero, meaning that the variables are unrelated." (Brownlee, 2020)

Independence of Observations (chi-plot): Chi-plots are used to check whether the variables are independent of one another or somehow related. According to the textbook An Introduction to Applied Multivariate Analysis with R , "Although the scatterplot is a primary data-analytic tool for assessing the relationship between a pair of continuous variables, it is often difficult to judge whether or not the variables are independent–a random scatter of points is hard

for the human eye to judge. Consequently, it is sometimes helpful to augment the scatterplot with an auxiliary display in which independence is itself manifested in a characteristic manner. The chi-plot suggested by Fisher and Switzer (1985, 2001) is designed to address the problem." (Everitt & Hothorn, 2011, pp.34)

Software:

R and Python programming language: PCA, PCA Correlation Circle, Mahalanobis Distance, Correlation, PCA Bi-Plot, and Chi-plots.

Tableau: Profile Plot.

Analysis and Results:

We can reduce the number of health variables into two principal components without losing the essence of the data. In other words, PCA Correlation Circle allows us to plot all the health variables using the first two principal components. Patients that are far away from the circle are considered as multivariate outliers. From **figure 1**, it is apparent that patients 5, 4, 56, 27, and 58 are multivariate outliers; patient 5 is an extreme outlier, and the rest are moderate outliers. However, as the axes of the graph show, the two principal components make about 69.3% variation of the total data (explains Insulin, Glucose, TST, and BMI); there is 30.7% of the total variation missing (which could explain Blood Pressure). For that reason, it is a good idea to substantiate our findings from **figure 1** with a Mahalanobis Distance plot to make sure our findings are accurate and to find out if we missed any potential outlier within the 30.7% of the variation that wasn't plotted in **figure 1**. As shown in figure 2, patient 20 is also considered a multivariate outlier because it is far away from the other clustered points. Therefore, we can conclude that patient 20 is another outlier due to her high diastolic blood pressure since patient

20 was not an outlier in figure 1, which does not explain blood pressure within the 69.3% of the variation.

From **figure 3**, we notice that there is a moderate positive correlation between Age and Number of Pregnancies with a correlation coefficient of 0.617, Age and Plasma Glucose Concentration with a correlation coefficient of 0.559, Age and Two-Hour Serum Insulin with a correlation coefficient of 0.601, and Plasma Glucose Concentration and Two-Hour Serum Insulin with a correlation coefficient of 0.648. We Also notice a high positive correlation between Triceps Skinfold Thickness and Body Mass Index with a correlation coefficient of 0.738.

In order to plot the raw data, a Profile Plot can be used, a line plot for all variables at once, and from it, some patterns could be discovered. We can notice some interesting insights in **figure 4**. Patient 58 had the highest Body Mass Index value of 55 kg/m^2, patient 5 had the highest Insulin level of 846 mu U/ml, and patient 4 had the second-highest Insulin level of 543 mu U/ml, which explains why these three patients were multivariate outliers, as shown in **figure 1**. Also, Patient 20 had the highest diastolic blood pressure (110 mm Hg), confirming what we found in **figure 2**. Patients 8 and 58 both had the same lowest diastolic blood pressure of 30 mm Hg. Patient 38 got pregnant 15 times which is considered the highest number of pregnancies among our data. Furthermore, we also notice that correlated variables tend to follow the same path in the profile plot. We can look carefully at the correlated variables and their direction to confirm this. For example, both Body Mass Index and Triceps Skinfold Thickness were highly correlated, as shown in **figure 3**; therefore, we can see how these two variables follow the same path and direction in **figure 5**. Of course, this also applies to the other correlated variables.

After scaling the data and applying the PCA dimension reduction technique, we can describe the variation in the original set of variables with few indices using a lesser dimension without losing valuable information. Choosing the number of principal components is subjective; the best approach is to choose the smallest number of principle components while maintaining the highest cumulative proportion of variance as possible. The first three principal components are enough because they describe around 77.7% of the total variance as shown in **table 1**. We can backup this decision with something called a Scree Plot which is another way of determining the number of principal components. As shown in **figure 6**, we notice that there is an elbow effect at principal component 3 which aids our previous decision. For those reasons, the first three principal components are sufficient. As shown in **table 2**, Principal Component 1 best explains Glucose, Serum Insulin, and Age, Principal Component 2 primarily describes Body Mass Index and Skinfold thickness, and Principal Component 3 mainly explains Diastolic Blood Pressure and adds to the variability of Serum Insulin. However, it is not recommended to use 3D graphs; therefore, using a PCA Bi-Plot to plot the first two principal components that make up around 63% of the total variation is the best approach. Moreover, grouping the female patients into three different age groups in the PCA Bi-plot will help with finding some interesting insights. The first age group: [20-29] in red color, the second age group: [30-39] in green color, and the last age group 40 and above in blue. From **figure 7**, we notice that most female patients are between 20 and 29 years old. Also, the 20-29 age group patients are plotted very far and in the opposite direction of the Glucose and Insulin arrows, which indicates that this age group has very low glucose and insulin levels compared to the other age groups. Moreover, the patients of age 40 years and above were placed in the exact opposite location of the 20-29 age group, which means that this age group has very high levels of Glucose and Insulin. Also, the second age

group is placed in between the first and third age groups, indicating that this group has moderate levels of Insulin and Glucose. We also notice that the first age group is far away from the Body Mass Index and the Triceps Skinfold Thickness, which means that the young female patients have low weight and low fat. However, patients 58 and 56 are the only two individuals from the first age group with high body mass index and the Triceps Skinfold Thickness values. Also, most of the patients were plotted on the upper half of the PCA Bi-plot, which is in the opposite direction of body mass index and the Triceps Skinfold Thickness arrows; this indicates that most female patients are in shape and have low weights. Furthermore, based on the plot, patient 43 has the lowest Glucose and Insulin among all the patients.

Based on **figure 8**, we notice that the blood pressure variable is independent of all the other demographics and health variables; in other words, the blood pressure variable is not related to the other variables. The lack of dependency between the blood pressure variable and the other variables explains why the blood pressure variable was primarily neglected in the previous analysis since the PCA relies heavily on correlation and the relationship between variables.

Discussion:

Some female patients were highly unusual based on the data of the health variables, especially patient five, who had an extremely high abnormal Insulin level of 846 mu U/ml which should be further investigated because if that was true, patient 5 has a dangerous stage of diabetes which requires urgent treatment. It was noticed that the blood pressure variable did not add any value to the analysis because it was an independent variable, and the analysis focused mainly on PCA which relies on variables that are related to one another. The Age variable was positively

correlated with Insulin levels and Glucose concentrations, and the majority-female patients were in shape.

Conclusions:

From the analysis, most female patients were young, between 20 and 29 years old. The Body Mass Index and Triceps Skinfold thickness were highly correlated, which is reasonable because both variables are indicators of fat and weight. The Age variable was positively correlated with Insulin levels and Glucose concentrations; therefore, as age increases, insulin and glucose also increase, as evident in figure 7. That follows, patients who are 40 and older had the highest glucose and insulin levels.

References:

Brownlee, J. (2020, August 20). How to Calculate Correlation Between Variables in Python. Machine Learning Mastery. https://machinelearningmastery.com/how-to-use-correlation-to-understand-the-relationship-between-variables/#:%7E:text=The%20statistical%20relationship%20between%20two,the%20other%20variables'%20values%20decrease.

Everitt, B., & Hothorn, T. (2011). The chi-plot. An Introduction to Applied Multivariate Analysis with R. (p. 34). Springer Publishing.

Prabhakaran, S. (2022, March 1). Mahalanobis Distance - Understanding the math with examples (python). Machine Learning Plus. https://www.machinelearningplus.com/statistics/mahalanobis-distance/#:%7E:text=Mahalanobis%20distance%20is%20an%20effective%20multivariate%20distance%20metric,and%20one-class%20classification%20and%20more%20untapped%20use%20cases.

Principal component analysis (PCA)| What is PCA? (2020, October 17). Intellipaat Blog. https://intellipaat.com/blog/a-brief-introduction-to-principal-component-analysis/

Appendix:

Graphs:

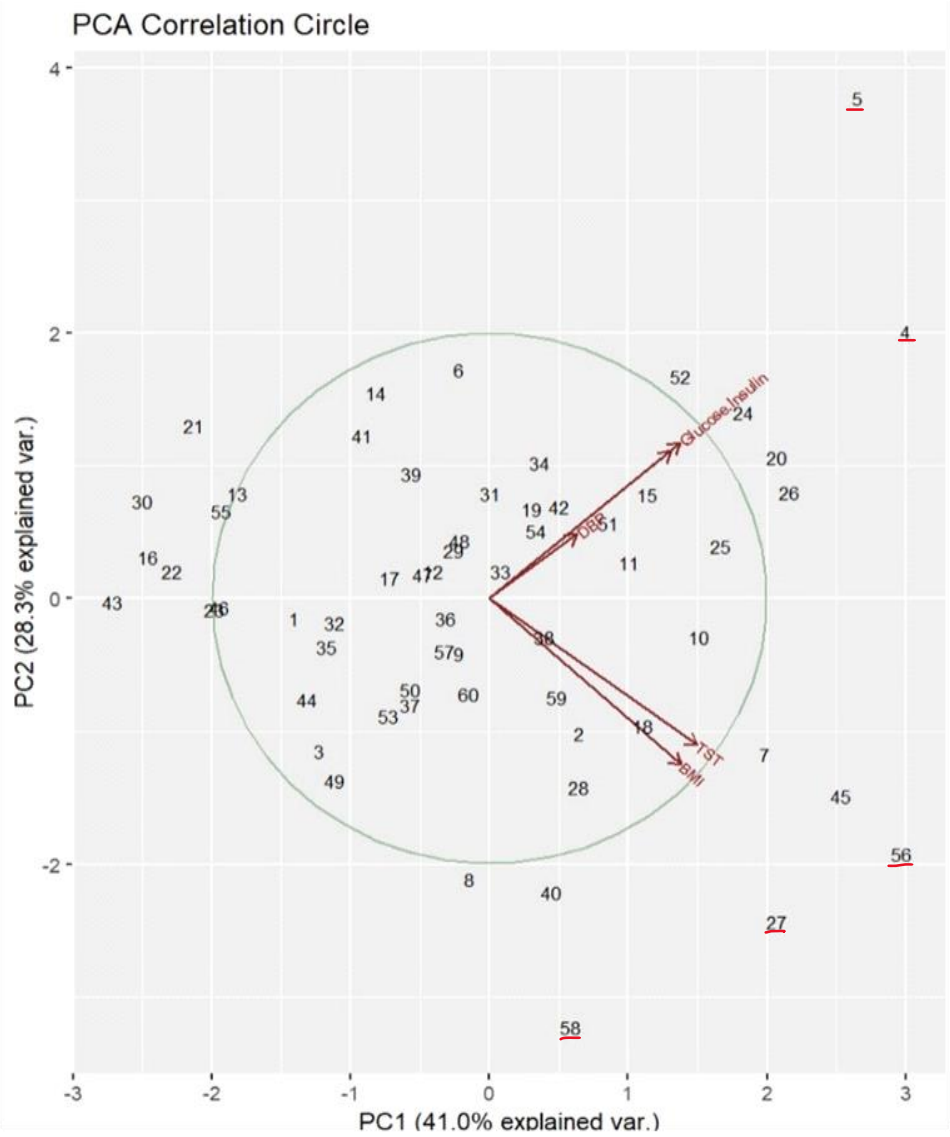

Figure 1: PCA Correlation Circle of Plasma Glucose Concentration through BMI

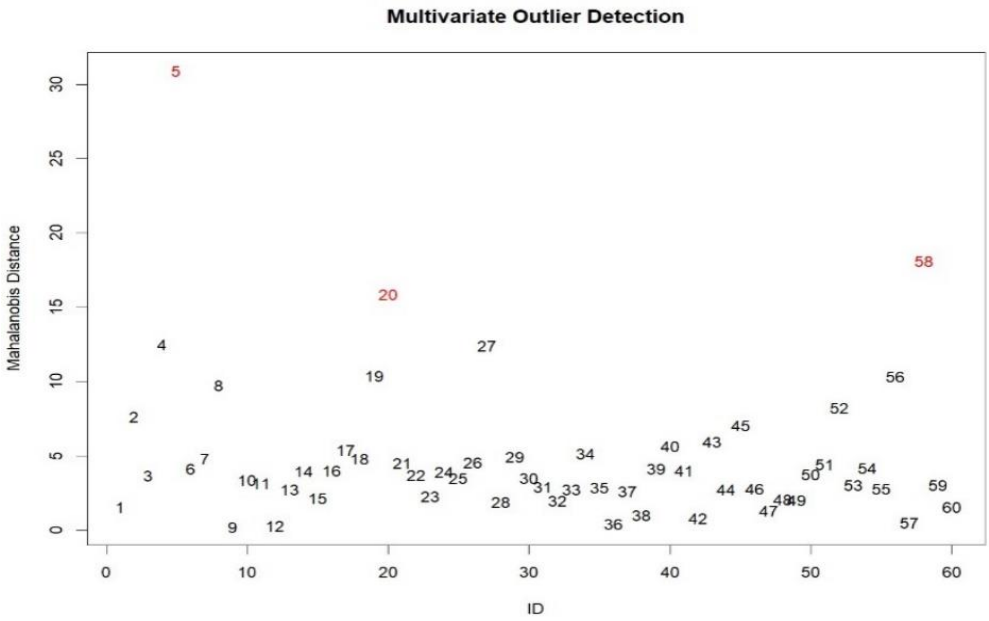

Figure 2: Mahalanobis Distance of Plasma Glucose Concentration through BMI

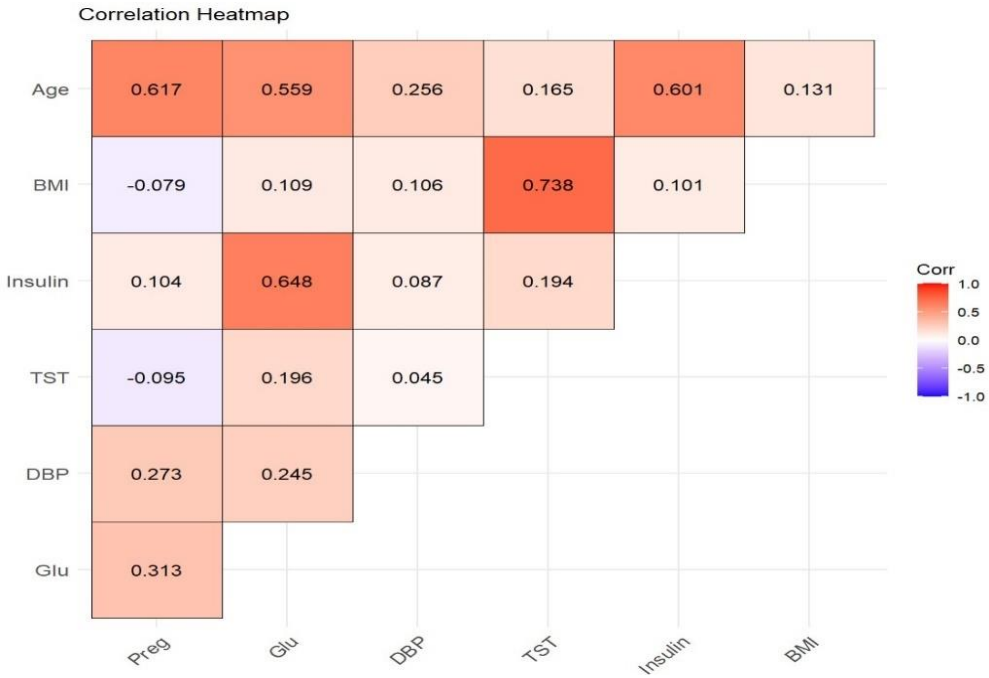

Figure 3: Correlation heatmap between all the variables

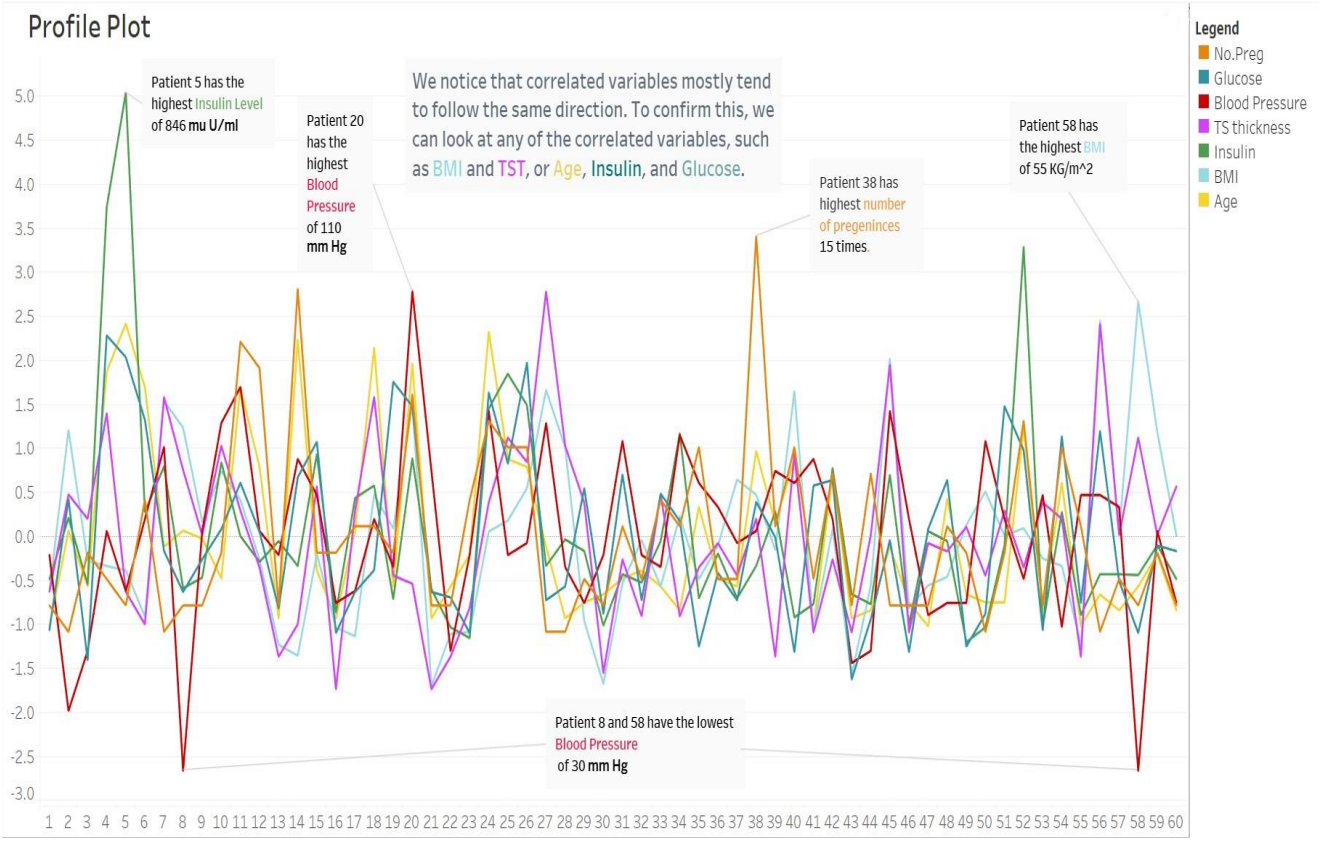

Figure 4: A Profile Plot for all the variables

Figure 5: A Profile Plot highlighting only BMI and TST

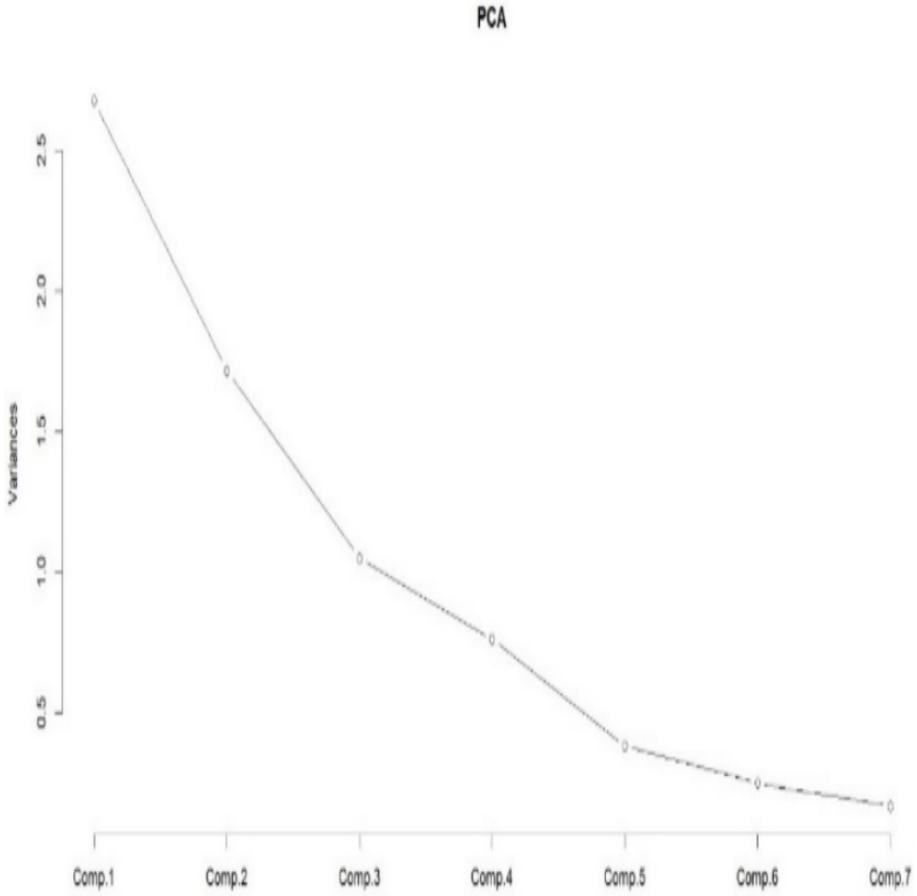

Figure 6: Scree Plot

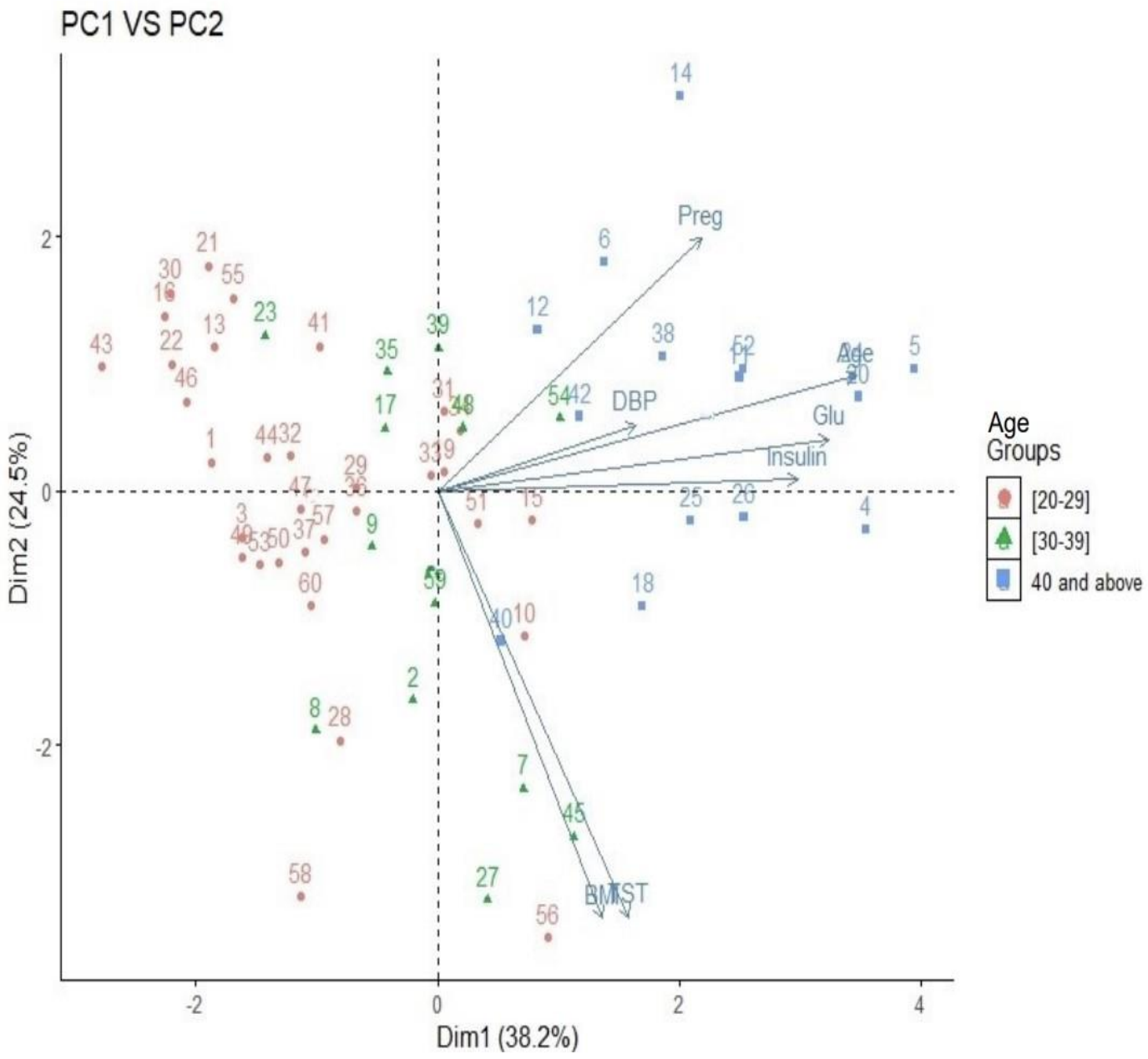

Figure 7: PCA Bi-Plot for PC1 vs PC2

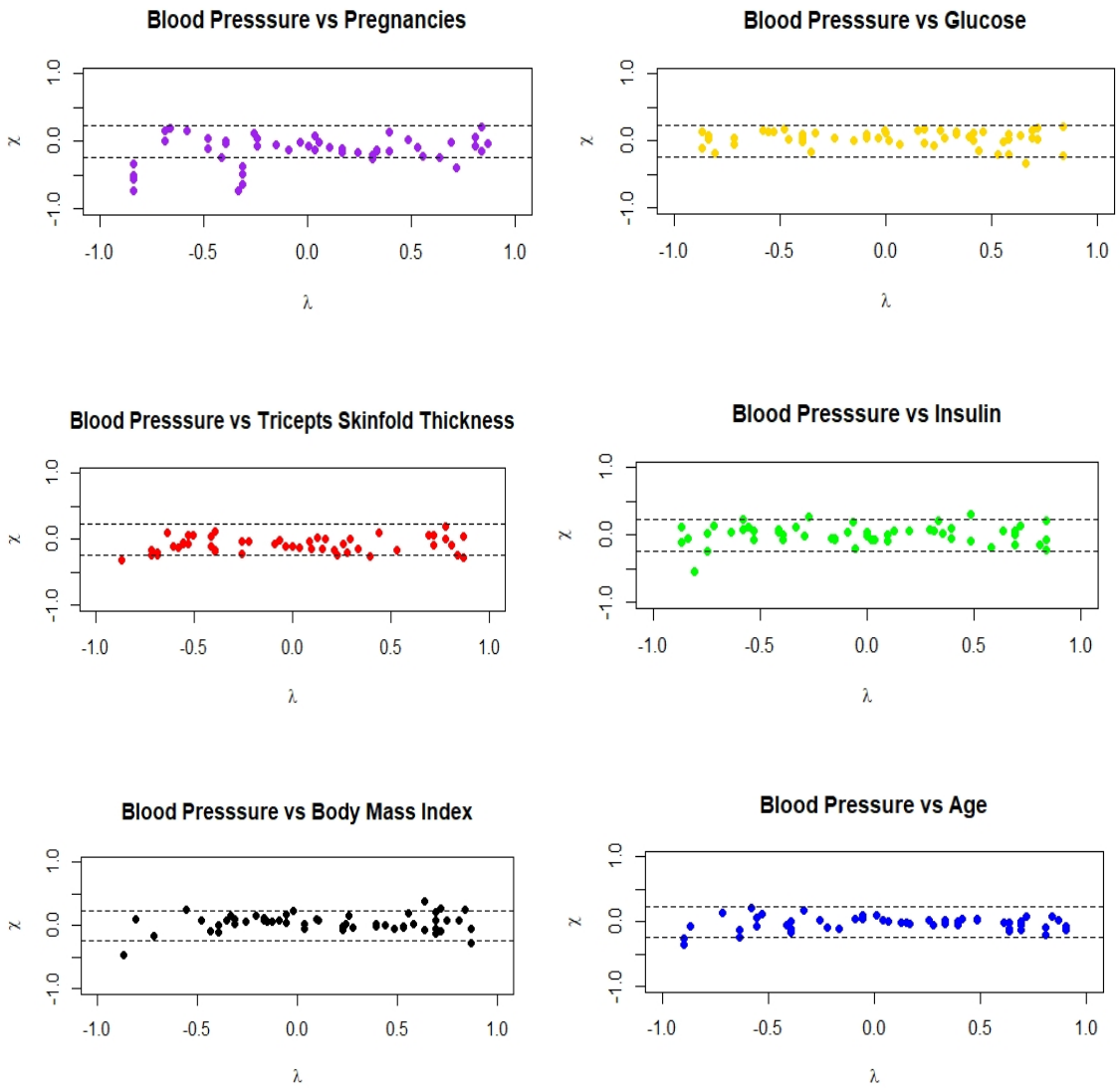

Figure 8: Chi-plot of independence between blood pressure and the other variables

Tables:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.6	Comp.6	Comp.7
Standard deviation	1.6352	1.3098	1.0245	0.8724	0.6177	0.49974	0.41027
Proportion of Variance	0.38197	0.24509	0.14996	0.10874	0.05451	0.035678	0.02405
Cumulative Proportion	0.3819	0.62706	0.777	0.886	0.9403	0.97595	1

Table 1: PCA Importance of Components

	Comp.1	Comp.2	Comp.3
Number of times Pregnant	0.4290448	0.3914202	0.1005621
Glucose	0.4873333	0.0851366	0.2274442
Diastolic Blood Pressure	0.3340075	0.1113054	-0.924376
Triceps Skinfold Thickness	0.2858604	-0.629654	0.0774183
Body Mass Index	0.2689808	-0.629981	-0.068626
Age	0.5592843	0.1838822	0.2701494

Table 2: Correlation Matrix Between the Principal Components and the variables