**Performance Assessment: D212- Dimensionality Reduction Methods**

**A. Research Question**

**1.** For this assessment, the research question is at follows: what are the principal components within the medical dataset?

**2.** The goal of this data analysis is to determine which variables are statistically relevant to one another.

**B. Method Justification**

**1.** PCA works by trying to understand which variables within the data set are important. It does this through variance. “The greater the variance the more the information” (Cheng 2023). What this means is that by analyzing the variance and correlation between data points, we can then understand which variables are significant. The variables that have a low variance can be omitted from the data set. This helps to reduce a large data set into a smaller one because smaller data sets are easier to work with. First PCA cannot work with data that is not standardized so the first thing is to standardize the data and then generate the covariance matrix. The covariance matrix can then be broken down to find the eigenvalues, which show the variance of each of the principal components. By using the eigenvalues or by generating a scree plot you can look at the variance of the variables and determine how many principal components are in your data set. With that information we then expect the outcome of being able to reduce the dimensions of the data set into a smaller number of components which is easier to work with while still retains most of the information from the original data set. The key word is “most” as it will only retain the percentage of variance that the user chooses as their cut off point when determining how to reduce the data set dimensions based on the scree plot.

**2.** One assumption of PCA is that the “The feature set must be correlated and the reduced feature set after applying PCA will represent the original data set but in an effective way with fewer dimensions” (Vadapalli 2020). What this means is that the variables within the data set must be statistically correlated to one another so the PCA can determine which of those variables are the most statistically relevant in order to effectively reduce the number of dimensions.

**C. Data Preparation**

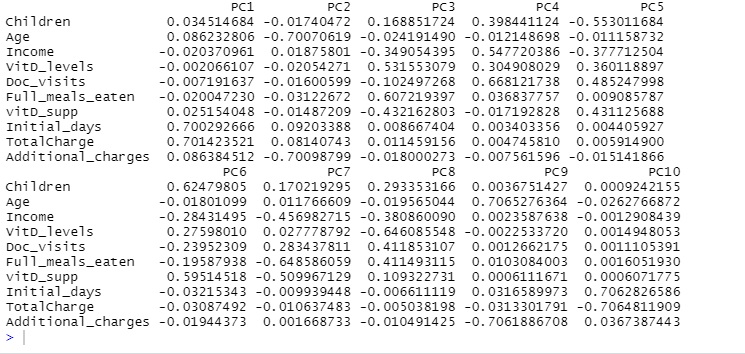
**1.** Before performing the PCA analysis, I will remove the variables from the data set that I will not be using in the analysis, either because they are categorical variables, or the data is irrelevant (i.e.. Latitude, longitude, UID, case number, etc.) This will leave the data set with the following numerical variables:

* Children
* Age
* Income
* VitD\_levels
* Doc\_visits
* Full\_meals\_eaten
* VitD\_supp
* Initial\_days
* TotalCharge
* Additional\_charges

**2.** Using the scale() function in R, these variables have been normalized. The normalized data set has been attached alongside this written assessment.

**D. Analysis**

**1.** The loading matrix for the principal components is shown below:



The loading matrix here tells us how each principal component is affected by each of the variables inside of it. Values above 0 positively affect the component and values below (negative) negatively affect the component. Since these are absolutely values the closer to +-1 then the more positive or negative that particular variable affects that component. For example, TotalCharge in PC1 shows the loadings at 0.7014 which suggests a strong positive impact and VitD\_levels are -0.0021 which shows negative impact, and so on. The full loadings can be seen for each variable and principal component above.

The correlation matrix for the data set is shown below:

A graph showing the amount of charge

Description automatically generated (Zelta 2023)

As you can see, the correlation matrix mostly shows a perfect correlation between the variable with itself as seen from the straight diagonal line. There also exists a perfect correlation between total charges and the initial days as well as a correlation of 0.5 between age and additional charges. This makes sense as it can be assumed that staying in the hospital longer would affect the total charges and we can also assume that older patients are more likely to face additional charges when in the hospital.

**2.** The scree plot for the data set is shown below:

A graph of a number of blue bars

Description automatically generated with medium confidence (Zelta 2023)

The elbow rule states that we observe the scree plot of the data and “select all components just before the line flattens out” (Mangale 2020). By observing the scree plot above, we would select the first two principal components. After the second component, there is a sharp decline in variance as the scree plot “elbows” out.

**3.** The following is the proportion of variance for each of the seven principal components as described in the previous step:

* Comp 1: 32.2% or 0.322
* Comp 2: 21.4% or 0.214

These numbers were obtained from the summary of the PCA as seen below:

A white text with black numbers

Description automatically generated

**4.** By adding up the following percentages, the total variance of the two components is equal to 53.6% as a percentage or 0.536.

**5.** To summarize, a scree plot was created of the principal component analysis, and by using the elbow rule it was determined that the first two PCs within the data should be retained, with a variance of 32.2% and 21.4% respectively. This gives us a total variance of 53.6% for the two PCs. The data should be reduced from 10 to 2 PCs in total, giving us a much easier data set to work with while retaining a little over half of the total variance.

**E. Sources**

Vadapalli, Pavan. “PCA in Machine Learning: Assumptions, Steps to Apply & Applications.” upGrad Blog, PCA in Machine Learning: Assumptions, Steps to Apply & Applications | upGrad blog, 11 Nov. 2020, [www.upgrad.com/blog/pca-in-machine-learning/](http://www.upgrad.com/blog/pca-in-machine-learning/).

Mangale, Sanchita. “Scree Plot.” Medium, Medium, 28 Aug. 2020, sanchitamangale12.medium.com/scree-plot-733ed72c8608#:~:text=The%20scree%20plot%20criterion%20looks,the%20side%20of%20a%20mountain.).

Kelta, Zoumana. “Principal Component Analysis (PCA) in R Tutorial.” DataCamp, DataCamp, 13 Feb. 2023, www.datacamp.com/tutorial/pca-analysis-r.

Cheng, Casey. “Principal Component Analysis (PCA) Explained Visually with Zero Math.” Medium, Towards Data Science, 6 May 2023, towardsdatascience.com/principal-component-analysis-pca-explained-visually-with-zero-math-1cbf392b9e7d#9b5c.