Analysis of HCV blood test result

Chris Taehwan Kim

2021 4 16

Introduction

The human body is more complex than what we think it is. Our body produces millions of new cells every second and proteins in our body interact and react to each others. Each organs takes specific roles for our body to function properly. Veins are the path to deliver essential sources all over the body parts to maintain other systems to function well. Our body is so intelligent that if there is any problem in our body, the body system tries to cure and protect our body from the harmful threats. Clearly, our body system knows more than what we know about ourselves. This means our body react almost immediately to the problems such as diseases, viruses and deficiency of essential sources. From this report, we're going to study about the blood test result of HCV(Hepatitis C Virus) patients, then analyze what signals the body gives to the patients and based on the signals, we're going to perform a few of statistic techniques to explain following phenomena.

Background information of the data

The name of my data set is "HCV data Data Set" and the data set is available from "UCI Machine Learning Repository". The source website is http://archive.ics.uci.edu/ml/datasets/HCV+data. The data set contains total 14 attributes. The first column represents the number of patients ID/No. Other than that, 2 categorical variables such as sex and types of blood donors and the rest of attributes are quantitative variables. There is total of 615 instances are in the data set but there are some missing values. The data set was donated on 2020-06-10. To briefly explain what Hepatitis C is to fully understand the data set and the statistic results, the Hepatitis C is a liver disease that caused by inflammation. There are 3 stages of Hepatitis C such as Hepatitis C to Fibrosis, then Cirrhosis. It is important to know these terms to interpret the data set. The data set contains laboratory values such as type of enzymes in the blood of blood donors and Hepatitis C patients and contains demographic values like age.

Preliminary analysis of the data

Before we apply any statistic techniques, it is important to learn about the data set. When we're analyzing a multivariate data set, it is very important to check how much we can trust the statistic summary we obtained and to have strength to our results, we have to make sure that the data set is in a good shape to analyze and check multivariate normality assumption since most of the statistic techniques require normality. That being said, the data set has some missing values and sometime, samples with missing values can be obstacles to our study. It is important to trim our data set to use only observations with complete values. Therefore, we're going to eliminate observations with missing values. Now, let's import the data set and find the new total rows and columns.

After we eliminate observations with missing values, we have new total of 589 instances(rows) and 14 attributes(columns). That means We trimmed out total of 26 instances with missing values from the original data set. We're going to use trimmed version of data set. Before we obtain statistic summary, we want to check multivariate normality assumption. If the multivariate data set is not normally distributed, we have to transform the data set to be approximately normal before we use any techniques. In order to do so, we're going to use graphical test to test normality. Now, let's plot a QQ plot to check if we have to transform the data set.

These are QQ plots of each variables with numerical values in the data set. In QQ plot, we say the distribution is normal when the points in the QQ plot are following a straight line. However, we reject the normality when the points are breaking away from the straight line. In these QQ plots, there are some significant evidences that the points in the tail are breaking away from red straight line. This could be a sign that the data need transformation. To be more accurate, let's put them all together and interpret the overall multivariate normality.

To give a brief interpretation, we noticed the first half points are following the green straight line but the other half points are slowly breaking away from the line. This means in overall, it gives us enough evidence to reject multivariate normality of the data. Therefore, we can use transformation to make not normally distributed random variables to be approximately normal.

There are many transformation methods we can use such as log transformation, Box-cox transformation and some general power transformation. However, we're going to use the Tukey's power transformation. The Tukey's power transformation is a power transformation on the data set to find the best fit normal distribution as possible. We can estimate the best power of variable lambda to transform the data set. We used a "transformTukey" function from package "rcompanion". Let's find out which variables are causing to reject normality. When we look at the QQ plot of each variables, some QQ plots are extremely breaking away from the red straight line. If we can transform those variables to be approximately normal, then we can fail to reject multivariate normality assumption due to have insufficient evidence. It seems like variable 5,6,7 and 11 are causing large variances. Therefore, let's use Tukey's power transformation to transform these variables.

After we used Tukey's power transformation, we evaluated suitable power of lambda for each variables that I previously selected. In the principle of Tukey's power transformation, if the

evaluated power of lambda is negative, we have to apply the transformation differently. If the lambda is positive, we can simply use transformation equation of DATA^lambda but if the lambda is negative, then we have to additionally multiply -1 which mean the transformation equation is -1*DATA^lambda. The suitable power of lambda for ALT(variable 5), AST(variable 6), BIL(variable 7) and GGT(variable 11) are 0.15, -1.125, -0.325 and -0.45. Therefore, we applied the power of lambda for each variables in the following instruction. After the transformation, I plotted QQ plots for each variable's before and after transformation to compare the differences. As we can see, the points are less likely to breaking away from the red straight line. It is not perfect, but acceptable range of errors. Now, let's interpret the overall multivariate normality.

In overall, the points are more fit to the green straight line than the previous QQ plot. However, there are some points on the right tails are still breaking away from the straight line. It is possible to make those points to be more fit to the straight line if we use other transformation methods that can perform better but since we're studying a real data set, we cannot transform the non-normal data set to be perfectly normal. Therefore, we assume that the transformed data set is approximately normal.

Statistic summary

Now, let's obtain statistic summary.

After we transform the data, We obtained new means and variances for each variables, covariance structure and correlation structure. One thing we can note on this summary is that we can interpret all the correlations by each variables and since covariances are not zero, we can assume that the relationship between each variables exists. At the end of this report, We can explain more about the relationship between each variables.

Principal Component Analysis

In the preliminary analysis, We have transformed the data to be approximately normal and obtained statistic summary. Based on what we've got, we can perform PCA(Principal Component Analysis). The PCA can be use for the purpose of feature extraction. We can preserve the main structure that is present in the feature space without losing any information. Also, we can identify important features in the data. By reducing dimension, we can easily interpret the projections of each features and what characteristics can be described to the features.In the PCA, we're going to find principal components of linear combinations of variables that can explain the overall variance of the data.

As a preliminary analysis for PCA, let's recall the numbers of observations and variables.

```
## [1] 589  11
```

```
##  [1] "Age"  "ALB"  "ALP"  "ALT"  "AST"  "BIL"  "CHE"  "CHOL" "CREA" "GGT"
## [11] "PROT"
```

The total of quantitative variables we're going to use for PCA is 11 variables and 589 observations for each variables. Excluding the age variable, the rest of variables are the names of enzymes in the blood. The fist step of the PCA is to find the vector of coefficients(which is eigenvectors) for each principal components to construct linear combinations. We can use prcomp function to find the eigenvectors.

So we have total of 11 PCs and PC scores for each variables. With this information, we can construct linear combinations for each PCs. For example, When we look at the PC1 scores for each variables, 0.08866846 is the coefficient for Age, -0.42638397 is the coefficient for ALB and so on. Therefore, we can construct a linear combination for PC1 which is PC1 = 0.08866846*Age* - *0.42638397*ALB -0.05435692*ALP* - *0.41132277*ALT - 0.18018153*AST + 0.01030395*BIL - 0.47044882*CHE* - *0.37800407*CHOL + 0.02734955*CREA* - *0.23217314*GGT - 0.43215787*PROT. When we look at the vectors weighted on each variables in PC1, We can interpret which variables are heavily weighted. For PC1, PROT(protein), CHOL(cholesterol), CHE(cholinesterase),ALB(albumin) AND ALT(alanine aminotransferase) are heavily weighted and other variables are relatively less weighted. Hence, we can say that PC1 is roughly corresponds to PROT, CHOL, CHE, ALB and ALT and less likely corresponds to other variables. For PC2, AST(aspartate aminotransferase) and GGT(gamma-glutamyl transferase) are heavily weighted than other variables. So PC2 is roughly corresponds to AST and GGT. Other PC's linear combinations can be constructed and interpret in a similar manner.

Now, let's look at the summary for PCs.

There are 3 categories in the summary. Standard deviation, Proportion of Variance and Cumulative Proportion. Proportion of variance is the ratio of how much proportion of variance for each PC's are taking from total variation. The larger proportion of variance means more information about variance is inherent in the PCs. Usually, the proportion of variance is in order of largest to smallest which mean the PC1 has the largest information about total variation which is 21.46% of information about total variation is contained in PC1. The cumulative proportion is the cumulation of proportion of variances and this allows us to interpret how many PCs can explain the total variation. For example, if we want to explain 90% of total variation, we have to look for the cumulative proportion with approximately 0.9. In this case, we have to select up to first 8 components because PC8's cumulative proportion is 0.88263(closest to 0.9). In general, we retain first few components which can explain around 70~90% of total variation. Now, let's plot a screeplot to display the change of variances for each components and interpret it.

A screeplot is a graphical tool we use in the PCA to visualize the change of variance for each components. Most of the time, a screeplot has a negative slope but the slope is slowly converging to 0 as the change of variance is slowly decreasing. The purpose of using screeplot is to find the points where the screeplot is becoming narrow. To give a short interpretation of the screeplot we obtained, we can see that the PC1 to PC2 has the largest change of variance as we mentioned above and as we go along the next component, the screeplot is slowing decreasing. The screeplot is starting to get narrower when we're at PC7 to PC8. Therefore, we can make a conclusion that we use first 7 or 8 components to explain the total variation because we can explain 70-90% of total variation with these components.

We can also construct a biplot to visualize the characteristic of each variables and to obtain the general picture of the data set. Biplot is one of the main graphical tool used as screeplot in PCA. It is also known as a score plot and the score plot is mixture of a scatterplot of two principle components and a loading plot. Both plots are simultaneously plotted and the purpose of using biplot is to interpret the distance between observations and underlying relationship between observations and variables. We can simply use biplot function to display biplot.

The blue vectors are the loading plot and the red numbers are the scatterplot of two principle components(PC1 and PC2). So let's start with loading plot. In general, a loading plot can show how strongly each variables can influences a principle components by looking at the directions of vectors. When we look at the biplot, every vectors are pinned at the origin and we can easily group vectors into two groups. ALB, CHE, PROT, CHOL and ALT are relatively drawing a same direction to the left and GGT, AST, ALP, BIL, Age and CREA are drawing a downward direction.

For convenience, let's call them group A and B respectively.

For group A, we noticed that the group A's vectors are the variables I mentioned earlier in the summary that they are heavily weighted in PC1. So as we can see in the loading plot, group A's vectors have large negative loadings on PC1 since the vectors are directing to the left. The length of vectors are approximately equal to each other which mean there is no variables weighted exceptionally that influencing PC1 among them. All of group A vectors are equally influencing PC1.

For group B, there are total of 6 vectors drawing a direction to downward and we can assume that these variables have negative loadings on PC2. We also mentioned that GGT and AST are heavily weighted on PC2 so that the length of vectors for GGT and AST are significantly longer than other group variables. This means GGT and AST are the main variables that influences PC2 and other 4 variables also influences PC2 but not as strong as GGT and AST.

We can also note that the relationship between each variables are divided into group A and B. Because in loading plot, the distance/angle between two vectors represent the correlations between them. If the distance/angle between two vectors are small(less than 90 degrees), then they are positively correlated but if they are large(equal or greater than 90 degrees), then they are not likely to be correlated or negatively correlated if it's close to 180 degrees. Suppose that we grouped vectors by vectors with the same directions, we can assume that grouped vectors are positively correlated to each other and if they are not in the same group, they are either not likely to be correlated or positively correlated but not significant.

In our data, we have 2 categorical variables, category and sex. In category, we classified sample of bloods with specific conditions. There are total of 5 categories in the variable. Blood donor's are group of people whose not suffering from Hepatitis C virus, suspect Blood donor's are unknown conditions, and 3 stages of Hepatitis C virus. We're going to plot a biplot again and this time, we're going to classify samples with categorical variable and see if we can find similar characteristic between samples.

Again, this is the same biplot we obtained before but with colours in the scatterplot. The

colour of samples classifies with a categorical variable named "Category" and the description on the right describes which colour represent the type of bloods. Now, let's look at the scatterplot correspond to their category. we can see that samples are most likely clustered together based on their categories and not many outliers are out of their clusters. However, some of categories are mixed up that we cannot separately group them except for Cirrhosis and suspect blood donor. Most of red points(Blood Donor) are clustered around the origin and green(Hepatitis) and blue(Fibrosis) points are also clustered in the same field as red points. This means these 3 categories are sharing the same characteristics that we cannot separate the clusters into one another. But If we look at the pink(Cirrhosis) and brown(suspect Blood Donor) points, they are significantly separated from the other 3 category's clusters which means Cirrhosis and suspect Blood Donor shares the same characteristics. If we give a little bit more interpretation on brown points, brown points are classified as suspect Blood Donor that we suspect them from Hepatitis C virus. Except for ID number 537 and 536, most of brown points are clustered together with pink points. Because other suspect donors are sharing the same blood conditions with Cirrhosis patients, we could rationally suspect them from Cirrhosis.

Next, let's look at the loading plot. We can interpret what shape of structure is underlying between variables and the characteristic of samples. Previously, we grouped vectors into two group A and B. Most of Group A's vectors have direction to the top left corner and Group B's vectors have direction to the downward. We noticed that pink points are clustered around the down right corner which is the opposite side of group A's vectors are pointing and group B's vectors are more closer to the pink points cluster. This means Cirrhosis patients have less ALB, CHE, PROT, CHOL and ALT than regular people and GGT,AST,ALP, BIL, CREA and age are relatively higher than regular people. Since age is in the direction to the Cirrhosis, We can also know that older people have more risk of having Cirrhosis. Other group B vectors have high correlation with liver damage that high values in these variables directly influences liver conditions which is correspond to Cirrhosis.

In overall, we can summarize graphical interpretation in the following way:

- Group A(ALB,CHE,PROT,CHOL,ALT) are heavily weighted on PC1 and have large negative loadings. Regular people have more group A variables in the blood than Cirrhosis patients and less group A variables can be suspicious.
- Group B(GGT,AST,ALP,BIL,CREA,AGE) are heavily weighted PC2 and have large negative loadings. Regular people have less group B variables in the blood than Cirrhosis patients and more group B variables can be suspicious.

Factor Analysis

In the previous study, we've performed PCA method to find the underlying structures between variables and we've grouped vectors into 2 groups with similar characteristics/directions. By using the statistic summary in PCA, we can perform factor analysis to determine the latent variables that helps to understand the group of variables in terms of smaller number of underlying unobservable factors.

There are two methods of estimation. One is principal factor method and other one is maximum likelihood method. For this study, we're going to use principal factor method because we have already performed the simliar steps in PCA so we can recall those summaries to proceed. Principal factor analysis is simply using eigenvalues and eigenvectors of the covariance matrix to estimate common factors that can account for the common variances. The first step of factor analysis is choosing the appropriate number of factors. There are two default methods to choose factor m. One is number of positive eigenvalues of the sample covariance matrix and second one is number of eigenvalues greater one for the sample correlation matrix. In PCA, We kept first 8 PCs to explain approximately 90% of total variation. Let's use the second method to choose number of factors and plot a screeplot to graphically visualize the result.

So we have total of 11 eigenvalues of correlation matrix and a screeplot. Our assumption is choosing number of eigenvalues that are only greater than 1 and we noticed first 4 eigenvalues are greater than 1 and the screeplot shows that the first 4 points are above 1($y > 1$) and the rest of points are below 1($y < 1$). Therefore, it is reasonable to choose 4 factors in this data. Now, we estimated the proper number of factors for the study, so for this time, let's set the number of factors to principle factor method with a factor rotation. The factor rotation helps to minimize the complexity of the factor loadings to make structure more interpretable. We're going to use orthogonal rotation in which applying the constraint to factors to being uncorrelated. The varimax criterion is a common method for orthogonal rotation that helps to have a few large factor loadings and as many as small loadings that are closely to zero. This constraint will help to settle ambiguous factor loadings in principal factor method.

So we obtained a table of factor scores for each variables with 4 factors(4 RCs) and h2 represent communality and u2 represent unique variances. It is important to check the communalities for each variables before we interpret factor scores because when the communalities are below 0.5, we assume that the variance shared with other variables are not significant. Since most of communalities are above 0.5, we can assume that it's significant. In the previous study, we kept 8 principle components to explain approximately 90% of the total variance and we noticed that in factor analysis method with 4 factors, explains 62% of total variance(Cumulative Var at RC4). Usually in factor analysis, when we can explain 60% or higher of total variance with selected number of factors, we assume that we have valid number of factors for the model.

Now let's look at the factor scores for each RCs. Factor scores are reliable than manifest variable values due to ambiguousness and the factor scores are the major components of latent variables. We have selected 4 factors which mean we have 4 latent variables at the end and we choose number of manifest variables contained in latent variables with respect to large factor scores. So we can summarize in the follow way:

- For RC1, CHE, CHOL and ALT have relatively large factor scores. Therefore, CHE, CHOL and ALT are the main manifest variables measures RC1.
- For RC2, AST, BIL and GGT have relatively large factor scores. Therefore, AST, BIL and GGT are the main manifest variables measures RC2.

- For RC3, ALB and PROT have relatively large factor scores. Therefore, ALB and PROT are the main manifest variables measures RC3.
- For RC4, CREA and ALP have relatively large factor scores. Therefore, CREA and ALP are the main manifest variables measures RC4.

In overall, these 4 latent variables can be specified into certain field of subjects in biology based on their combinations of manifest variables. Unfortunately, in order to know the background of these latent variables, we cannot easily define them due to lack of knowledge in biology. Therefore, in this case, we need expert's opinions.

Conclusion

In this report, we performed two major statistical techniques, Principle Component Analysis and Factor Analysis. Both of them have similar processes of analyzing underlying structures in the data and creating new variables but in different perspectives. Some of the results are quite similar to each other but We obtained many different aspects of interpretations of the data throughout these techniques. In PCA, we divided vectors into 2 groups and defined each group's characteristics and the clusters of samples. In FA, we obtained 4 latent variables in terms of manifest variables with large factor scores. To putting these results together, we can conclude that Cirrhosis(stage 3 of Hepetitis C virus) patients are suffering from deficiency of group A vectors(ALB,CHE,PROT,CHOL,ALT) and these manifest variables are the major components of RC1 and RC3. Therefore, RC1 and RC3 are the field of subject that needs to be study in order to understand these deficiencies. High group B vectors(GGT,AST,ALP,BIL,CREA,AGE) can be suspicious and exposed risk to other side effects and except for AGE, these manifest variables are the major components of RC2 and RC4. Therefore, RC2 and RC4 are the field of subject that needs to be study in order to lower the risk of Cirrhosis.

References

- http://archive.ics.uci.edu/ml/datasets/HCV+data
- https://datacookbook.kr/39
- https://rpubs.com/Evan_Jung/pca
- https://blog.bioturing.com/2018/06/18/how-to-read-pca-biplots-and-scree-plots/
- https://rcompanion.org/handbook/I_12.html