Multivariate Analysis on Data from Cancer Patients

Ramos de Pina, Samuel · Mendonça, Cristiano · Ogura, Joana · Pereira, M. Bernardo · Gaehtgens, Thomas

December 11, 2022

Abstract

1. Introduction

1.1. The Problem

The goal of this project is to explore and apply different techniques to a multivariate data set concerning cancer patients and try to find and identify intriguing elements on the data. This ability is fundamental because it is what allows to extract knowledge and information out of the given data. In this project, the specific problem addressed is the prediction of survival of cancer patients doing chemotherapy using molecular profiling. This problem is pertinent because if we can have a more accurate prediction on the survival of these patients compared to the established international medical predictors, we can target research into the study of biochemical processes involved in the most deadly genes and also provide more accurate predictions to the medical staff treating those patients.

1.2. The Data

The dataset of the project regards patients with a particular cancer called *Diffuse Large B-Cell Lymphoma* (DLBCL), which is a kind of blood cancer. These patients took part in a study (1) published in the New England Journal of Medicine back in 2002. The data consists of two separate tables. The main table, `Patients`, stores 12 different attributes about 240 patients with tumor-biopsy samples which were used to extract the micro-array gene expression. The features associated with each patient in this table and their respective types are the following:

`Numerical`	*DLBCL sample (LYM number)*
	Follow-up (years)
	Germinal center B cell signature
	Lymph node signature
	Proliferation signature
	BMP6
	MHC class II signature
	Outcome predictor score
`Categorical`	*Analysis Set*
	Status at follow up
	Subgroup
	IPI Group

Table 1: Types of Features in `Patients`

One very relevant variable in this table is the *DLBCL sample (LYM number)* which enables the connection from this table to the second, called `Microarray`.

In addition to the `Patients` table, the `Microarray` table stores information about the expression of 7291 genes in 293 biopsy samples. Therefore, 53 biopsy samples in this dataset did not come from patients in the previous table, because this one only contained 240.

1.3. Project Structure

At first, exploratory data analysis will be performed in order to interpret the information of the dataset. Afterwards, some multivariate analysis techniques will be applied in order to retrieve insights from the dataset.

2. Exploratory Data Analysis (EDA)

The first procedure applied was the merging of the two tables of information regarding the patients and the gene expressions. Notice that the `Microarray` table had to be transposed in order to have the patients represented by the row corresponding to their LYM number, in accordance with the `Patients` table.

After merging both tables, we get an expanded `Patients` table where, for each patient entry, we also have the expression level for each gene in their corresponding biopsy sample. Since we originally had 7291 genes, after this merge we have a table, `merged_data`, with 240 patients (rows) and 7303 (12 + 7291) features (columns).

In this final merged table, we can conceptually divide the data in two separate groups of features: *patients features* and *genes features*. In the first group of variables, that corresponds to the original `Patients` dataset, there are only missing values in the *IPI group* attribute, with 6.94% of NAs. In the second group, which is much larger, we have an average of 10.13% of NAs in each column (gene). In total, there are 177 352 missing values in this table, which corresponds to 1 in 10 values being missing.

The variable we aim to target is *Outcome predictor score*. em que sentido, target?

For the purpose of visualization, for the categorical variables, namely *Analysis set*, *Status at follow-up*, *Subgroup* and *IPI group*, we obtained the following bar charts.

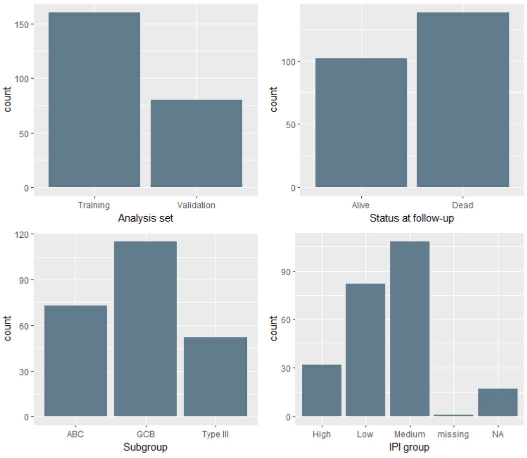

Figure 1: Bar plots of categorical variables

As for the continuous variables, they are distributed as follows.

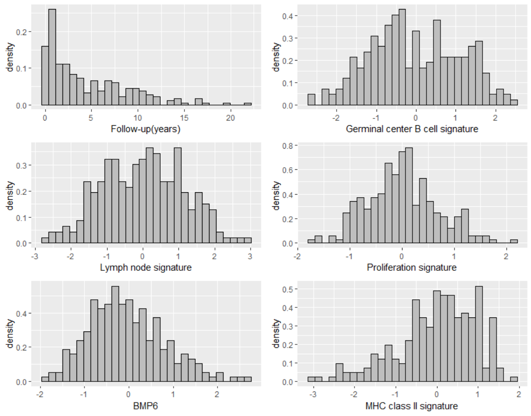

Figure 2: Histograms for numerical variables

Lastly, our target variable, *Outcome predictor score* has the following density.

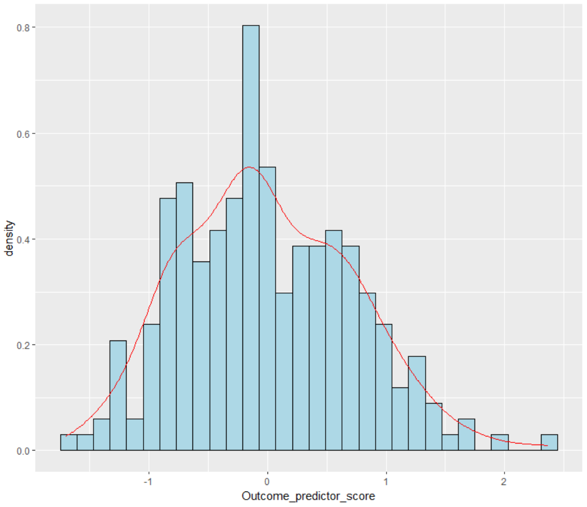

Figure 3: Density of Outcome predictor score

Now we present the graphs where independent continuous variables are in relation to our target variable.

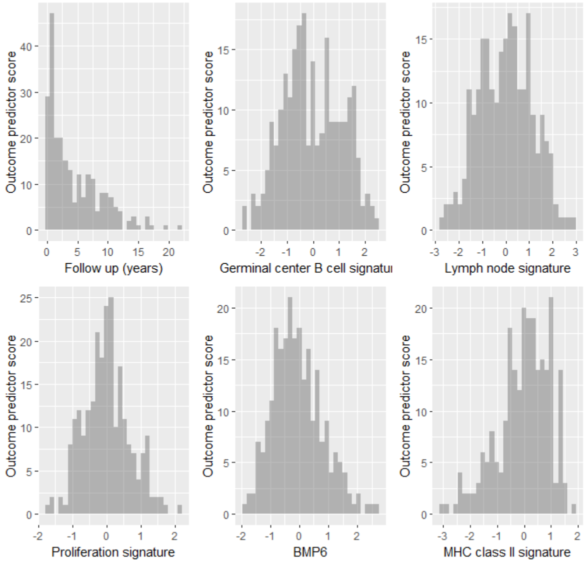

Figure 4: Continuous variables in relation to target variable

For the correlation analysis, we obtained the following matrix.

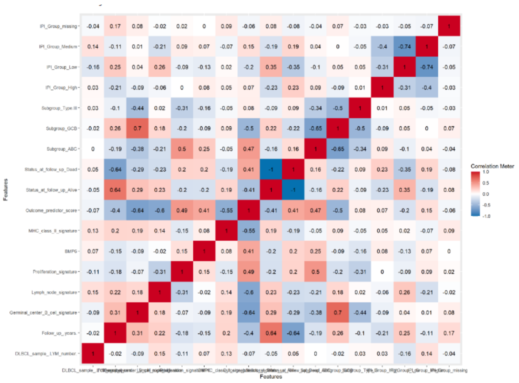

Figure 5: Correlation Matrix

It can be observed that variables such as ... present the highest positive correlation with our target variable whilst ....present the highest negative correlation.

3. Methods

4. Results

5. Conclusion and future work

As concluding remarks ...

References

A. Rosenwald, G. Wright, W. C. Chan, J. M. Connors, E. Campo, R. I. Fisher, R. D. Gascoyne, H. K. Muller-Hermelink, E. B. Smeland, J. M. Giltnane, *et al.*, "The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma," *New England Journal of Medicine*, vol. 346, no. 25, pp. 1937–1947, 2002.