2019-12-08

Factor Analysis of Mixed Data (FAMD) and Multiple Linear Regression in R

Nestor Pereira

Technological University Dublin, D19125953@mytudublin.ie

Using Multiple linear regression based on principal component analysis (Factor analysis of mixed data FAMD) for predict the final score of secondary students from Portugal

PEREIRA LINARES, Nestor
Post-Graduate in Data Science in the
Technological University Dublin,
Ireland
D19125953@mytudublin.ie

Abstract

In the previous projects, it has been worked to statistically analysis of the factors to impact the score of the subjects of Mathematics and Portuguese for several groups of the student from secondary school from Portugal.

In this project will be interested in finding a model, hypothetically multiple linear regression, to predict the final score, dependent variable G3, of the student according to some features divide into two groups.

One group, analyses the features or predictors which impact in the final score more related to the performance of the students, means variables like study time or past failures.

The second group analyses the predictors more relate to a family situation or family relationships.

The approach to constructing the linear model is using the principal component results from the analyses of the principal component instead of the original features or predictors.

The linear model proposal is:

score G3 = a + b1*(PC1) + b2*(PC2) + ... + bk*(PCk)

bi = Coefficients

PCi = principal component, i: 1, 2, …, k dimensions

Keywords *— Principal Component Analyses, Machine Learning, Multiple Linear Regression, Logistic regression, Accuracy.*

I. INTRODUCTION

It will be worked with a dataset from the University of Minho, Portugal which describe the data collected from two public school and show the score of the secondary school during the period 2005 – 2006.[1]

The data include information about two subjects: Math and Portuguese.

I will be interested to predict the final score of the student based in two aspects:

First, it will be studied the impact of the variables like age, study time, past failures, extra school support, extra classes, access to the Internet, interest in higher education, health status, absences, and additional variables which tell the subject (Math, Portuguese), in the final score of the student. These variables are more related with the performance of the student.

Second, it will be studied the impact of the variables more related with the family situation or family life, like age, parent's cohabitation, mother's job and education, father's job and education, student's guardian, absences.

In linear regression have many variables called predictors, introduce noise and redundancy into the data and increase the variance in the predictive model. Also, it demands independence between the predictors mean not collinearity.

In order to avoid those problems, in this project it will be used the method Principal Components Analyses (PCA) not only to reduce the number of predictors, also to avoid the redundancy and multicollinearity between them.

Due that the variables are numeric and categorical, it will be used the extension method called Factor Analysis of Mixed Data (FAMD) to deal with data quantitative and data qualitative. [2]

Finally, it will be constructed two multiple linear regression models for the two aspects or groups describe before.[3] [4] [5]

II. PROBLEM DEFINITION

The principal objective in this project consist of predict the final score of the secondary student based on the result of the reduction variables applied to the original dataset.

A. *Hypothesis 1: Performance*

In order to deal with the problem, it will be defined two hypotheses, the first one is related with variables more about performance,

Ho: there are no significance prediction (or effect) of the student final score G3 by the features related with performance.

Ha: It is possible to predict the student final score G3 by these features using a multiple linear regression.

B. *Hypothesis 2: Family situation*

The second hypothesis is related with variables more about the family life or family situation,

Ho: there are no significance prediction (or effect) of the student final score G3 by the features related with family situation.

Ha: It is possible to predict the student final score G3 by these features using a multiple linear regression.

This project will be finding the linear regression model which fix better for the problem applying the technique of reduction variables called FAMD.

III. OBJECTIVES

That was mention before, the principal aim in this project is predicting the final score of the student building a multiple linear regression.

The final model will be evaluated based on the performance indicators: [7][10][11]

- RMSE, Root Mean Squared, average error, measure how far the observations are from the regression line, lower value is better model.
- RSE, Residual Standard Error, called sigma, is an average error performed by the model in predicting the final score G3.
- Accuracy of the model calculated from the RSE divided by the mean of score G3.
- R-Square (adjust), percentage of the variation in the final score G3 explained by the predictor variables.
- F statistic, if the predictor variables are statistically significantly related to the final score G3.
- Multicollinearity, it just to verify that is no present in the predictor's variable.

The objective-based on the sampling techniques using 80% to train the regression model and 20% for testing the model in order to achieve estimations with no high bias and at the same time, no high variance.

IV. TECHNICAL APPROACH AND MATERIAL

In this section, it is boarded a brief description of the technical approach to tackle the problem.

Firstly, it will prepare the dataset, dealing with messing values, delete outliers using the rule "Tukey Fence", and choosing the variables considered more relevant for the analysis for the two aspects (performance and family situation) in this research.

Secondly, the extract statistics information of the dependent variable, score G3, and previous information about the last part of the project I.

Later, in order to do the correlation analyses, all variables are transformed in variables numeric. It will be used the scale z-score for numeric variables and the "Dummy" technique for categorical variables.

During the correlation analyses also, it is applied to Bartlett's test of Sphericity and check the Kaiser-Meyer-Olkin (KMO) measure and determinant to know the applicability of PCA.

Applying the FAMD for two subsets of the dataset: performance and family situation, and based on the eigenvalues, will be found the best components or dimensions to reduce the subset of the dataset.

Finally, the dimensions founded in the previous step for both subset of the dataset, allow to construct and train (with 80% of the dataset) the linear regression model using the dimensions as the predictor's variables.

The F-statistics results will be allowed to find enough statistical evidence to reject the Ho (no effect) in both cases in favour of the Ha.[2][7]

A. *Material*

The dataset is composed by joint of two datasets, one with score of students in Math, and another with the score in the subject Portuguese.

The original dataset has 1.044 observations with 334 variables.

The dataset, subset from the original, using into the analyses of the Hypothesis 1 (Performance) has 990 observations with 10 variables, and the dataset for the analyses of the hypothesis 2 (Family situation) has 990 observations with 8 variables.

The predictors variables for measure the effect in the hypothesis 1 (performance) are:

- age, variable numeric
- study time, variable categorical
- past failures, variable numeric
- extra school support, variable categorical
- paid extra classes, variable categorical

- access to the Internet, variable categorical
- interest in higher education, variable categorical
- health status, variable categorical
- absences, (variable numeric) and additional
- subject (Math, Portuguese). variable categorical

The predictors variables for measure the effect in the hypothesis 2 (family situation) are:

- age, variable numeric
- parent's cohabitation, variable categorical
- mother's job. variable categorical
- mother's education. variable numeric
- father's job. variable categorical
- father's education. variable numeric
- student's guardian, variable categorical
- absences. variable numeric

After dealing with missing values and outliers, the number of observations is **990**, and the dependent variable, final score G3, look like normal distribution:

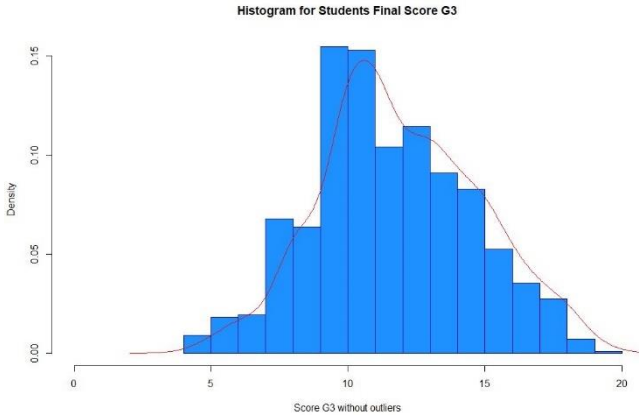

Fig. 1. Final score G3 after outliers

V. DESCRIPTIVE STATISTICS OF THE VARIABLES

The variable "target" final score G3 is the dependent variable and look like as a normal distributed after applying the process to deal with missing values and delete outliers.

The technique using for deleting outliers was the "**Tukey Fence**" in which the observations outside to the range (fence) of [1Q - 1.5 (IQR), 3Q + 1.5 (IQR)] are deleted.

1Q: first quartile

3Q: third quartile

IQR: inter-quartile range

Apply this rule affect specifically to score G3 zero, that it is correct for this analysis because it means that the student probably abandons the course before it finished.

The Tukey fence is showing in the next graphs.

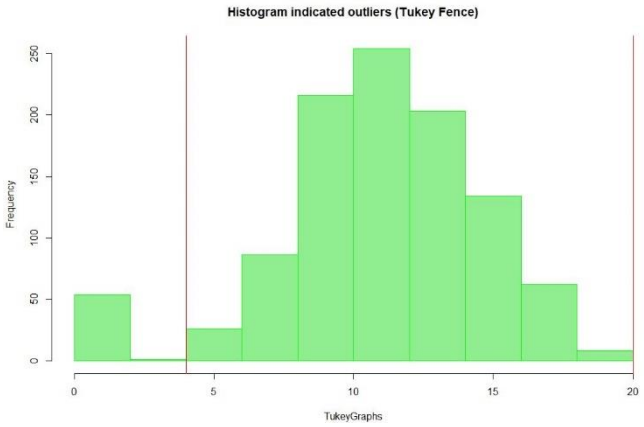

Fig. 2. Final score G3 and tukey fence

The general statistics descriptiion of the dependent variable final score G3 show a homogeneous distribution of the observations around the mean, and also the median closer to the mean.

The skewness is 0.151 considered very low and support the assumption than the distribution is normal.

```
> # ----
> # Statistics descriptive of the final score G3
> # ----
> print("Statistics descriptive for final score: ")
[1] "Statistics descriptive for final score: "
> describe(df$G3, IQR=F, na.rm = T)
df$G3
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     990        0       17    0.988    11.96    3.266        8        8       10       12       14       16       17

Value          4     5     6     7     8     9    10    11    12    13    14    15    16    17    18    19    20
Frequency      1     8    18    19    67    63   153   151   103   113    90    82    52    35    27     7     1
Proportion 0.001 0.008 0.018 0.019 0.068 0.064 0.155 0.153 0.104 0.114 0.091 0.083 0.053 0.035 0.027 0.007 0.001
> summary(df$G3)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   4.00   10.00   12.00   11.96   14.00   20.00
```

Fig. 3. Statistics metrics of the final score G3 (in R)

A. *Normality test for dependent variable score G3*

Even though the graphs and the statistical measures indicate that the dependent variable G3 has a normal distribution, the Kolmogorov-Smirnov test does not provide enough evidence to support it.

```
> # Kolmogorov test
>
> ks.test(df$G3, pnorm)

        One-sample Kolmogorov-Smirnov test

data:  df$G3
D = 0.99997, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Fig. 4. Kolmogorov test: reject Ho (Normal distribution)

Probably because there is some observations outsider showing in the next graphs.

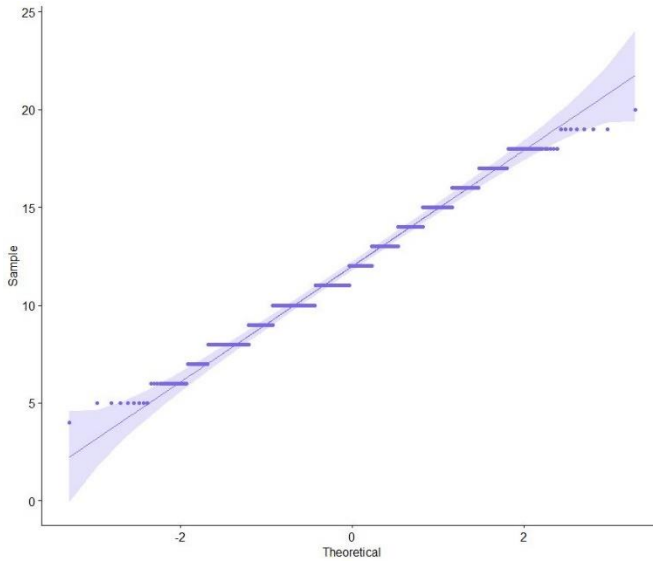

Fig. 5. Plot distribution score G3 vs normal distribution

The same conclusion applying the alternative test Jarque Bera, the result not provides enough evidence to support the assumption of normality in the dependent variable G3.

```
> jarque.bera.test(df$G3)

        Jarque Bera Test

data:  df$G3
X-squared = 8.9086, df = 2, p-value = 0.01163
```

Fig. 6. p-value < 0.05 reject the null hypothesis Ho: normal distribution.

Following the graphical inspection and the statistically describe values, it can be assumed that the variable score G3 has a normal distribution.

B. Statistics information from the previuos project

The previous project show clearly statistical evidence that the variable "study time" has an effect in the final score G3.

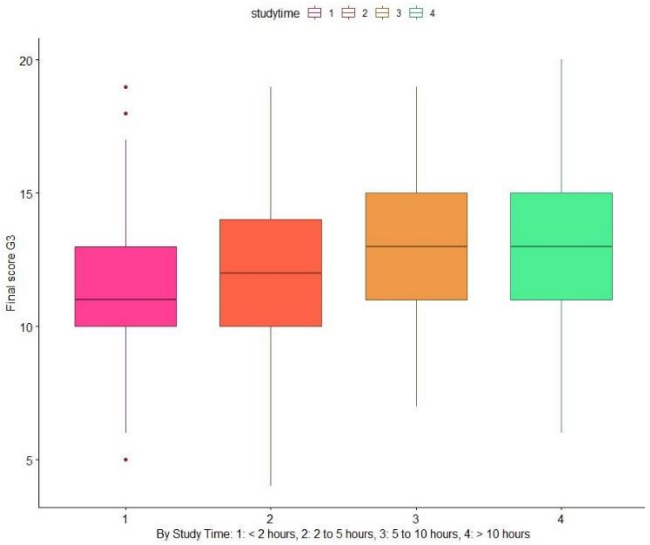

Fig. 7. Plot distribution score G3 by study time

Another essential information from this project was that the score G3 in the subject Math is statistically less than the score G3 in the subject Portuguese.

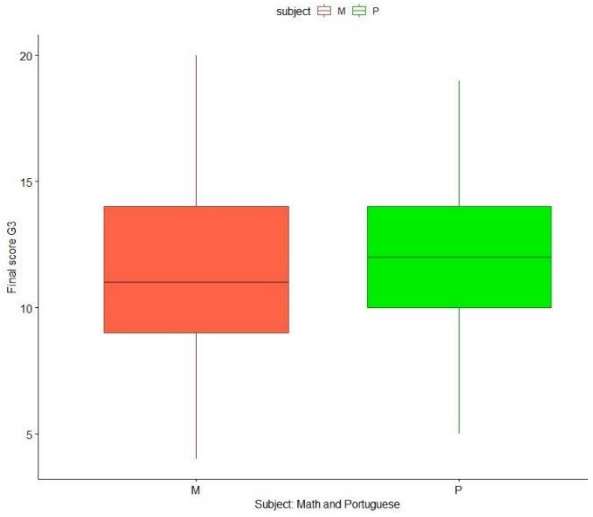

Fig. 8. Plot distribution score G3 by subject (Math vs Portuguese)

VI. PRE-PROCESSING (FOR CORRELATION ANALYSES)

Before apply the correlation and the analyses of multicollinearity it is necessary to transform the data in numeric. To do that it will be used two techniques: scalar the data using z-score transformation and apply "dummy" codes to convert categorical variables in numeric.

This process was made in four steps:

- Apply z-score to numeric variables, standardise the variable.
- Dummy code for categorical variables of 2 levels, binary variables transform to 0 or 1.

- Dummy code for categorical variables of 2 levels but are not numeric, created a new "sub-variable" for each category.
- Dummy code for categorical variables that have three or more levels, created a new variable for each category.

The final dataset looks like this (in R):

```
> ls.str(df.scaled)
absences :  num [1:990] 0.21 -0.108 0.847 -0.426 -0.108 ...
age :  num [1:990] 1.056 0.246 -1.376 -1.376 -0.565 ...
at_home_fjob :  num [1:990] 0 0 0 0 0 0 0 0 0 0 ...
at_home_Mjob :  num [1:990] 1 1 1 0 0 0 0 0 0 0 ...
course_reason :  num [1:990] 1 1 0 0 0 0 0 0 0 0 ...
failures :  num [1:990] -0.375 -0.375 4.507 -0.375 -0.375 ...
father_guardian :  num [1:990] 0 1 0 0 1 0 0 0 0 0 ...
Fedu :  num [1:990] 1.452 -1.274 -1.274 -0.365 0.543 ...
G3 :  num [1:990] -2.06 -2.06 -0.677 1.051 -0.677 ...
health :  num [1:990] -0.374 -0.374 -0.374 1.023 1.023 ...
health_fjob :  num [1:990] 0 0 0 0 0 0 0 0 0 0 ...
health_Mjob :  num [1:990] 0 0 0 1 0 0 0 0 0 0 ...
higher :  num [1:990] 1 1 1 1 1 1 1 1 1 1 ...
home_reason :  num [1:990] 0 0 0 1 1 0 1 1 1 1 ...
internet :  num [1:990] 0 1 1 1 0 1 1 0 1 1 ...
Medu :  num [1:990] 1.226 -1.437 -1.437 1.226 0.338 ...
mother_guardian :  num [1:990] 1 0 1 1 0 1 1 1 1 1 ...
other_fjob :  num [1:990] 0 1 1 0 1 1 1 0 1 1 ...
other_guardian :  num [1:990] 0 0 0 0 0 0 0 0 0 0 ...
other_Mjob :  num [1:990] 0 0 0 0 1 0 1 1 0 1 ...
other_reason :  num [1:990] 0 0 1 0 0 0 0 0 0 0 ...
paid :  num [1:990] 0 0 1 1 1 1 0 0 1 1 ...
Pstatus :  num [1:990, 1:2] 1 0 0 0 0 0 0 0 1 1 0 ...
reputation_reason :  num [1:990] 0 0 0 0 0 1 0 0 0 0 ...
schoolsup :  num [1:990] 1 0 1 0 0 0 0 1 0 0 ...
services_fjob :  num [1:990] 0 0 0 1 0 0 0 0 0 0 ...
services_Mjob :  num [1:990] 0 0 0 0 0 0 1 0 0 1 0 ...
studytime :  num [1:990] 0.0267 0.0267 0.0267 1.226 0.0267 ...
subject :  num [1:990, 1:2] 1 1 1 1 1 1 1 1 1 1 ...
teacher_fjob :  num [1:990] 1 0 0 0 0 0 0 0 1 0 0 ...
teacher_Mjob :  num [1:990] 0 0 0 0 0 0 0 0 0 0 0 ...
```

Fig. 9. Dataset predictor variables after trasformation

Now, all variables are numeric, and it is possible to apply the correlation functions.

VII. CORRELATION ANALYSES AND MULTICOLLINEARITY

The linear regression model assume that the predictor variable is independent and also have a correlation, have effect or impact, into the dependent variable.

It will be analysed the correlation for the groups of predictors related with the **hypothesis 1 (performance)**. It is shown the following results:

	age	studytime	failures	schoolsup	paid	higher	internet	health	absences	subject.M	subject.P	G3
age	1.00	0.00	0.29	-0.20	-0.03	-0.23	-0.02	-0.03	0.18	-0.03	0.03	-0.08
studytime	0.00	1.00	-0.13	0.06	0.10	0.17	0.05	-0.06	-0.09	0.06	-0.06	0.19
failures	0.29	-0.13	1.00	0.03	-0.01	-0.26	-0.08	0.05	0.15	0.05	-0.05	-0.35
schoolsup	-0.20	0.06	0.03	1.00	0.02	0.07	-0.02	0.00	-0.02	0.05	-0.05	-0.16
paid	-0.03	0.10	-0.01	0.02	1.00	0.12	0.11	0.01	0.07	0.50	-0.50	-0.09
higher	-0.23	0.17	-0.26	0.07	0.12	1.00	0.08	0.00	-0.10	0.11	-0.11	0.25
internet	-0.02	0.05	-0.08	-0.02	0.11	0.08	1.00	-0.03	0.09	0.08	-0.08	0.11
health	-0.03	-0.06	0.05	0.00	0.01	0.00	-0.03	1.00	-0.02	0.01	-0.01	-0.09
absences	0.18	-0.09	0.15	-0.02	0.07	-0.10	0.09	-0.02	1.00	0.20	-0.20	-0.22
subject.M	-0.03	0.06	0.05	0.05	0.50	0.11	0.08	0.01	0.20	1.00	-1.00	-0.11
subject.P	0.03	-0.06	-0.05	-0.05	-0.50	-0.11	-0.08	-0.01	-0.20	-1.00	1.00	0.11
G3	-0.08	0.19	-0.35	-0.16	-0.09	0.25	0.11	-0.09	-0.22	-0.11	0.11	1.00

Fig. 10. Correlation values between predictors for hypothesis 1 (performance)

Some variables, study time, features, higher and absences, have a low correlation with the dependent variable G3. Also, some predictor variable shows a correlation between them, for example, age with failures, school sup and higher.

These results indicate that, perhaps, the assumption of independence does not have enough support. Therefore, the final model would be of high variance and unstable.

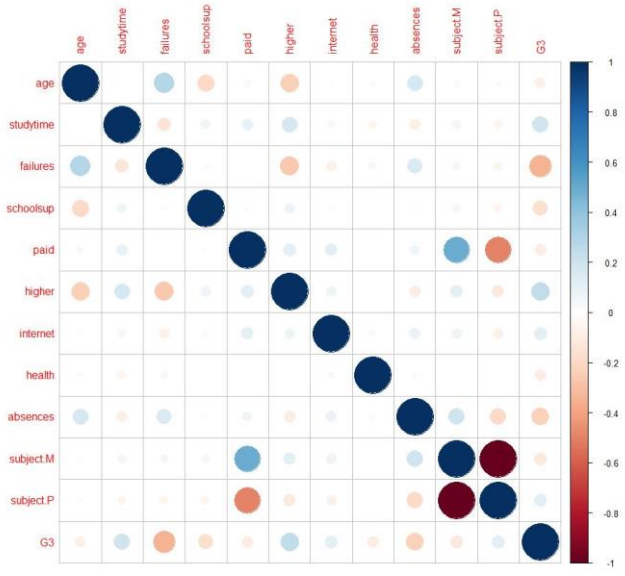

Fig. 11. Correlation between predictors for hypothesis 1 (performance)

Similar results can be observed for the groups of predictors for the **hypothesis 2 (family situation)**, the predictors about the level of education of the parents (Medu, Fedu), absences have a low correlation with the final score G3. Aldo, when the mother and father work as a teacher has an effect in the final score G3. However, predictors like the level of education of the parents and the type of jobs are clearly correlated.

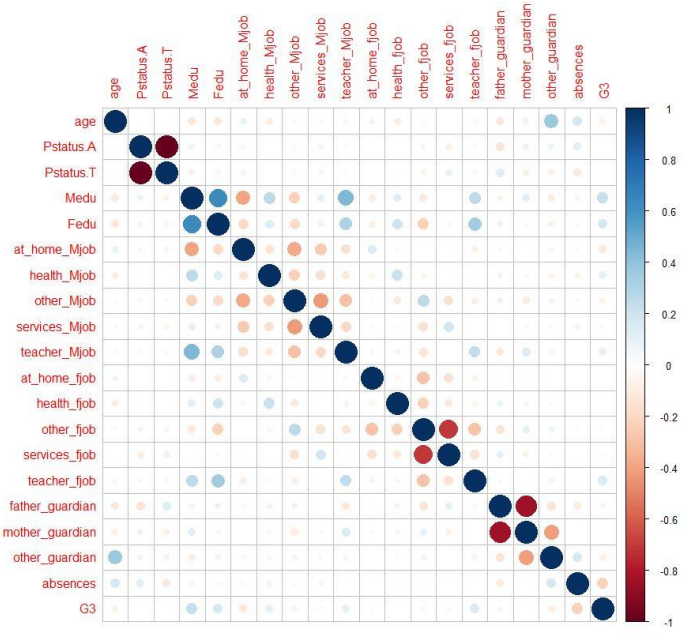

Fig. 12. Correlation between predictors for hypothesis 2 (family situation)

Finally, looking at the multicollinearity, it is shown the correlation between all predictors in the next graphs.

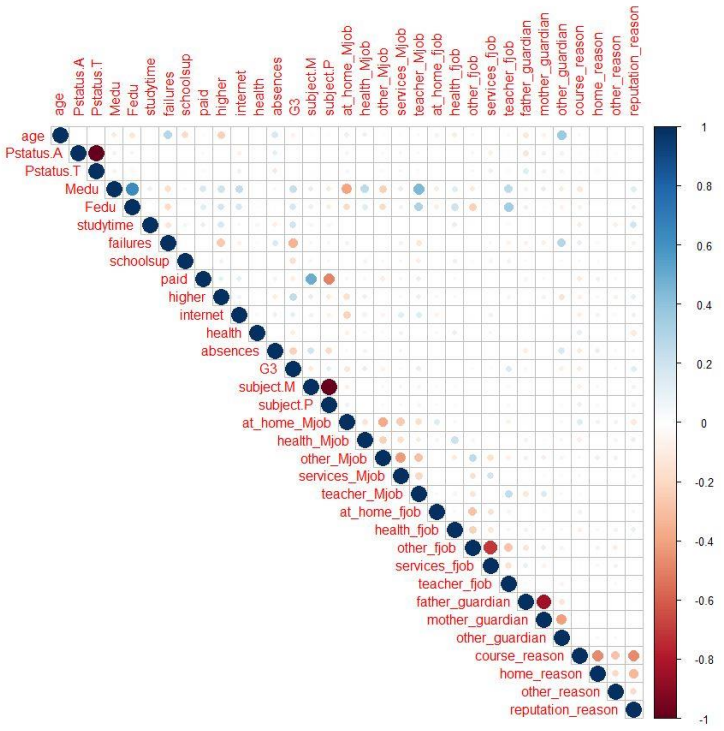

Fig. 13. Correlation between all variables

Unfortunately, it has shown many predictors correlated that means then can be calculated from others, and there are not independent.

The determinant of this correlation matrix is zero (0) means than exist multicollinearity.

As a result, the dataset and the predictor's variables have multicollinearity.

This problem can be tackled using a technique for dimension reduction, principal component analysis, which reorganise the dataset in components or dimensions independents.

A. *Applicability of PCA*

There are two tests or indicators that help to investigate if the PCA technique can be applying in this dataset.

The Bartlett's sphericity test measures if there are significant difference between the correlation matrix and the identify matrix (perfect correlation). In this case, value $p < 0.05$ therefore, PCA is applicable,

```
        Bartlett's Test of Sphericity

Call: bart_spher(x = df.scaled)

     X2 = Inf
     df = 465
p-value < 2.22e-16
```

Fig. 14. Bartlett's test indicate the PCA is applicable.

Another measure, Kaiser-Meyer-Olkin (KMO), indicate how well suited the dataset with the PCA. In this case, the value of the Measure of Sampling Adequacy (MSA) is just 0.5 means that PCA could be useful.

```
Kaiser-Meyer-Olkin factor adequacy
Call: psych::KMO(r = df.scaled)
Overall MSA =  0.5
```

Fig. 15. Indicator MSA Measures sampling adequacy = 0.5

Finally, it is possible to conclude that the dimension reduction method is applicable for this dataset.

VIII. PRINCIPAL COMPONENTS ANALYSIS: FMCA

In this chapter, it will be used the principal component analysis (PCA) to reduce the number of variables (features).

Specifically, it will be used the extension of PCA for mixed variables, numerical and categorical, by using the method called **Factor Analysis of Mixed Data** (FAMD).

Having many predictors introduce noise and redundancy in the data and increase the variance in the model. Therefore, reduce the numbers of predictor would help to find a better linear model to predict the score of the student, and also to avoid the collinearity between them.

A. FAMD for Hypothesis 1: performance

In this case, it will be used the subset of the dataset with the interested variables for the hypothesis 1 more related to performance (df.performance).

```
> str(df.performance)
'data.frame':   990 obs. of  10 variables:
 $ age       : int  18 17 15 15 16 16 16 17 15 15 ...
 $ studytime: int  2 2 2 3 2 2 2 2 2 2 ...
 $ failures : int  0 0 3 0 0 0 0 0 0 0 ...
 $ schoolsup: Factor w/ 2 levels "no","yes": 2 1 2 1 1 1 1 2 1 1 ...
 $ paid     : Factor w/ 2 levels "no","yes": 1 1 2 2 2 2 1 1 2 2 ...
 $ higher   : Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 2 ...
 $ internet : Factor w/ 2 levels "no","yes": 1 2 2 2 1 2 2 1 2 2 ...
 $ health   : int  3 3 3 5 5 5 3 1 1 5 ...
 $ absences : int  6 4 10 2 4 10 0 6 0 0 ...
 $ subject  : Factor w/ 2 levels "M","P": 1 1 1 1 1 1 1 1 1 1 ...
```

Fig. 16. df.performance, subset variables related with performance to test hypothesis 1.

Note that the subset has a mixed variable, the implementation of the method FAMD can deal with those variables and do not need the previous transformation.

```
> eig.val <- get_eigenvalue(res.famd)  # Evaluate the eigenvalues
> head(eig.val,11)
       eigenvalue variance.percent cumulative.variance.percent
Dim.1   1.7590565        17.590565                    17.59056
Dim.2   1.6059071        16.059071                    33.64964
Dim.3   1.1224234        11.224234                    44.87387
Dim.4   1.0158282        10.158282                    55.03215
Dim.5   0.9891901         9.891901                    64.92405
Dim.6   0.9102151         9.102151                    74.02620
Dim.7   0.8133280         8.133280                    82.15948
Dim.8   0.7086895         7.086895                    89.24638
Dim.9   0.5978081         5.978081                    95.22446
Dim.10  0.4775540         4.775540                   100.00000
```

Fig. 17. Eigenvalues for df.performance: indicate a principal dimensions

Applying the FAMD is obtained the four (4) dimensions can be explained the 55% of the variance.

The criteria used to choose those four (4 dimensions) are:

1. Eigenvalues > 1
2. Percentage of explained variances

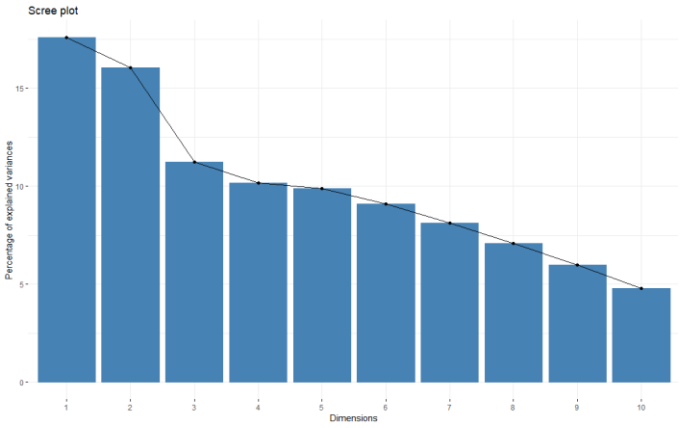

Fig. 18. % of explained variances for df.performance by dimensions.

Those four (4) dimensions or components will be used in the linear regression model.

Under the assumption that exists a linear relationship between the dimensions (or components) and the predictor variables, each dimension can be calculated by the predictor variables and their percentage of contribution.

The following graphs shows the correlation between the predictors and the two principal dimensions.

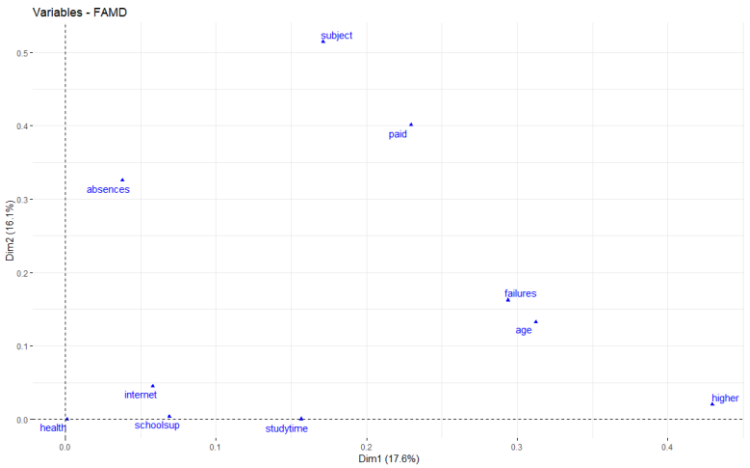

Fig. 19. Correlation between the predictors and the two principal dimensions.

The percentage of contribution of each variable to each dimension is showing in the next table,

```
> head(var$contrib[,1:4]) # % contributions
                Dim.1        Dim.2      Dim.3       Dim.4
age        17.75244618 8.266396e+00 10.4531165  0.87075180
studytime   8.89603754 2.987781e-02 10.8145866 17.01191340
failures   16.69884802 1.011904e+01  6.7116011  5.06730493
health      0.06578895 3.144853e-04 22.1059397 38.55533101
absences    2.15619993 2.026799e+01  0.3731641  0.04808883
schoolsup   3.92731216 2.512326e-01 29.3958835 31.65010268
```

Fig. 20. % contribution of each variable to each dimension.

This contribution is showing in the next graph for the four (4) dimensions together,

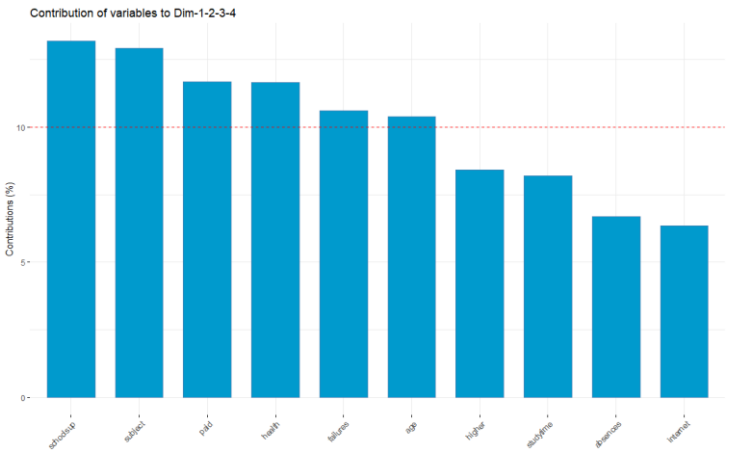

Fig. 21. % contribution of each variable to the four (4) principal dimensions.

Note that the variables “school sup”, “subject”, “paid extra class”, “health”, “failures”, and “age” have the principal contribution to the dimensions. The contributions of the variables are not uniform.

Finally, the function FAMD additionally calculated the results for individuals. Those values are the values of the dimensions and will be used the sample observations to train and test the

linear regression model. Therefore, the **new set of data** for the regression model is:

```
> head (var.ind[,1:4],10)
        Dim.1       Dim.2       Dim.3       Dim.4
1  -0.2489012  0.27223919  1.7962156  2.2515994
2  -0.4743682  0.46564906 -0.6237877 -0.1570285
3  -0.5178644  3.04812082  2.9652786  2.3898830
4  -2.4088725  1.05834467  0.3608534 -0.8533141
5  -1.2116146  1.04094170  1.4959862 -0.6823285
6  -1.5223273  1.88534184  0.4155010 -1.2446614
7  -0.9095429 -0.05406930 -0.3226918 -0.2466660
8  -0.6264744  0.03657842  1.4011458  3.0439634
9  -2.1694043  0.93086011 -0.5396549  0.3808541
10 -2.0976900  0.93581838  0.7749166 -1.3552347
```

Fig. 22. New set de data for training and testing the regression model.

This new set of data will be divided in a set for training (80%) the regression model and a set for test (20%) to evaluate the model.

B. FAMD for Hypothesis 2: family situation

A similar way to the previous analysis, it will be used the subset of the dataset with the interesting variable for the hypothesis 2 more related to the family situation (df.family).

```
> str(df.family)
'data.frame':   990 obs. of  8 variables:
 $ age     : int  18 17 15 15 16 16 16 17 15 15 ...
 $ Pstatus : Factor w/ 2 levels "A","T": 1 2 2 2 2 2 2 1 1 2 ...
 $ Medu    : int  4 1 1 4 3 4 2 4 3 3 ...
 $ Fedu    : int  4 1 1 2 3 3 2 4 2 4 ...
 $ Mjob    : Factor w/ 5 levels "at_home","health",..: 1 1 1 2 3 4 3 3 4
 $ Fjob    : Factor w/ 5 levels "at_home","health",..: 5 3 3 4 3 3 3 5 3
 $ guardian: Factor w/ 3 levels "father","mother",..: 2 1 2 2 1 2 2 2 2 2
 $ absences: int  6 4 10 2 4 10 0 6 0 0 ...
```

Fig. 23. df.family, subset variables related with family situation to test hypothesis 2.

Note again that the subset has a mixed variable, the implementation of the method FAMD can deal with those variables and do not need the previous transformation.

```
> head(eig.val,16)
       eigenvalue variance.percent cumulative.variance.percent
Dim.1   2.4598408        16.398939                    16.39894
Dim.2   1.5315937        10.210625                    26.60956
Dim.3   1.2495145         8.330096                    34.93966
Dim.4   1.2081802         8.054535                    42.99419
Dim.5   1.1458519         7.639013                    50.63321
Dim.6   1.0918387         7.278925                    57.91213
Dim.7   1.0089095         6.726063                    64.63820
Dim.8   0.8663462         5.775641                    70.41384
Dim.9   0.8361070         5.574046                    75.98788
Dim.10  0.7869688         5.246459                    81.23434
Dim.11  0.7747441         5.164961                    86.39930
Dim.12  0.6575385         4.383590                    90.78289
Dim.13  0.6003973         4.002648                    94.78554
Dim.14  0.5335956         3.557304                    98.34285
Dim.15  0.2485732         1.657154                   100.00000
```

Fig. 24. Eigenvalues for df.family: indicate a principal dimensions

In this case, applying the FAMD is obtained the seven (7) dimensions can be explained the 64,63% of the variance.

The criteria used to choose those seven (7) dimensions are:

1. Eigenvalues > 1
2. Percentage of explained variances

Note that in this case, the number of dimensions is too close to the numbers of predictors, with eight (8) variables, mean that only reduce in one dimension. However, the principal argument is that those dimensions are not correlated. Multicollinearity is not present.

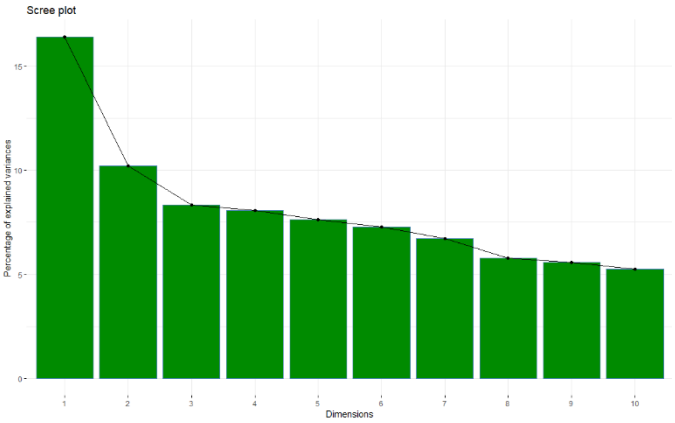

Fig. 25. % of explained variances for df.family by dimensions.

Those seven (7) dimensions or components will be used in the linear regression model.

The following graphs shows the correlation between the predictors and the two principal dimensions.

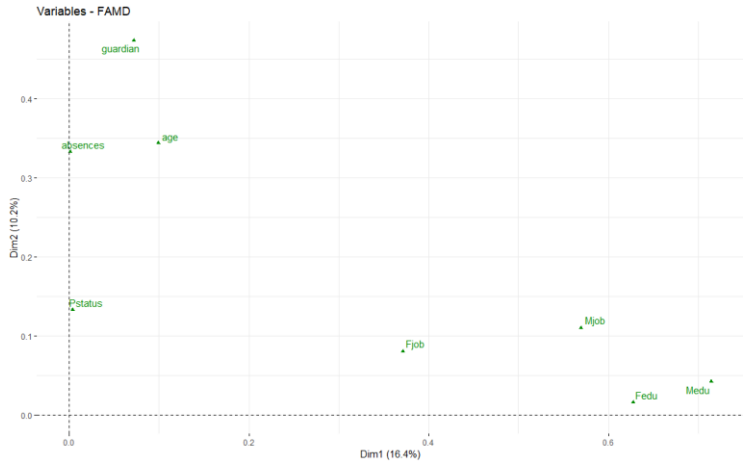

Fig. 26. Correlation between the predictors and the two principal dimensions.

The percentage of contribution of each variable to each dimension is showing in the next table,

```
> head(var$contrib[,1:7]) # % contributions
             Dim.1      Dim.2     Dim.3      Dim.4      Dim.5       Dim.6      Dim.7
age      4.04160084 22.429467 7.3742025  1.3526319  3.63819634  1.576171088  3.0633095
Medu    29.04453565  2.755483 0.2096874  0.2663103  0.02068247  0.002974376  0.6879248
Fedu    25.52029513  1.042269 0.9047598  0.2772884  0.67202932  0.871327587  1.2451926
absences 0.04996702 21.711996 0.7615206  2.8984672  3.04537666  0.104973807  1.2431530
Pstatus  0.15272061  8.672942 7.1053585  1.6488377  6.78522475 10.459285840 21.5204908
Mjob    23.15343714  7.198480 27.3980098 40.9161171 53.67143446 33.500277468 12.3734889
```

Fig. 27. % contribution of each variable to each dimension.

This contribution is showing in the next graphs for the seven (7) dimensions together,

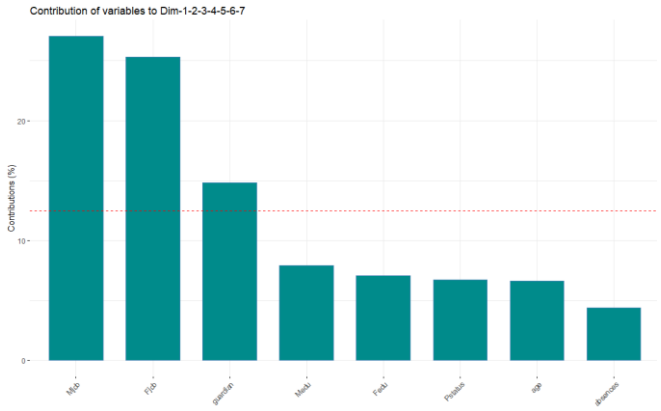

Fig. 28. % contribution of each variable to the seven (7) principal dimensions.

Note that the variables about parent's job: "Mjob", "Fjob", and "guardian" have the principal contribution to the dimensions.

Finally, and alike the previous analyses, the function FAMD calculated the results for individuals. Those values are the values of the dimensions and will be used the sample observations to train and test the linear regression model. Therefore, the **new set of data** for the regression model is:

```
> head (var.ind[,1:7],6)
      Dim.1      Dim.2       Dim.3       Dim.4       Dim.5      Dim.6      Dim.7
1  1.9370879  2.1472357 -0.11349946 -1.21002369  1.10831965  0.7444551  1.9924888
2 -2.1710242 -1.1715788  0.45565733 -0.08705416  1.02588039 -0.2807393  0.1962251
3 -1.8700950 -0.7464521 -0.54623486 -0.75690938  0.39598364  1.2802689 -0.3095289
4  1.4966880 -1.5971541  0.23937062  1.28222742 -0.36026947  1.0397389 -1.4005555
5  0.1572315 -0.8698056 -0.74001544  0.52630086  0.01207562 -1.7662135  0.6692281
6  1.0422553  0.4815382 -0.09684321 -0.20339603 -1.35919909  0.5868873 -0.1576951
```

Fig. 29. New set de data for training and testing the regression model.

This new set of data will be divided in a set for training (80%) the regression model and a set for test (20%) to evaluate the model.

IX. APPLY MULTIPLE LINEAR REGRESSION MODEL FOR THE HYPOTHESES

In this chapter, the linear regression model will be constructed and trained for the new subset to data to verify the hypotheses, using the dimensions have founded in the FAMD analysis.

Each new set of data for each Hypothesis will be split into a training set (80%) and test set (20%), the training set will be used to train the model and calculate the coefficients. The test set will be used to compare with the score predicted by the model in order to evaluate the performance of the model.

A. Apply multiple linear regression model for Hypothesis 1: performance

The first fact to mention in this part of the analysis is the Determinant apply to the correlation matrix for the new set of data is 1, mean that there is no multicollinearity in the dataset.

Applying the multiple linear regression model to the training dataset obtains the coefficients for the model and the F statistic,

$$F(4,790) = 60.65, \text{p-value} < 2.2e\text{-}16$$

Mean that the test is statistically significant for Hypothesis 1, and there is enough evidence to reject the Ho in favour of the Ha. There is an effect of the dimensions in the final score of the student.

```
> model <- lm(G3 ~ Dim.1 + Dim.2 + Dim.3 + Dim.4, data=train.data)
> summary(model)

Call:
lm(formula = G3 ~ Dim.1 + Dim.2 + Dim.3 + Dim.4, data = train.data)

Residuals:
    Min      1Q  Median      3Q     Max
-7.1861 -1.7448 -0.2198  1.6721  7.8196

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.96709    0.09125 131.148   <2e-16 ***
Dim.1       -0.49004    0.06913  -7.089    3e-12 ***
Dim.2       -0.62925    0.07056  -8.918   <2e-16 ***
Dim.3       -0.85173    0.08501 -10.019   <2e-16 ***
Dim.4       -0.21200    0.09168  -2.312    0.021 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.572 on 790 degrees of freedom
Multiple R-squared:  0.2349,    Adjusted R-squared:  0.2311
F-statistic: 60.65 on 4 and 790 DF,  p-value: < 2.2e-16
```

Fig. 30. Results MLR to the set of data Hypothesis 1: performance.

1) Performance of the model.

Applying the model to the test dataset is obtained the following results:

- RMSE: 2.38, measure how far the observations are from the regression line, similar to RSE.

- RSE: 2.57, it will be used to calculate the average prediction error rate.
- Average predictor error rate: 21.5 % is an average error performed by the model in predicting the final score G3.
- Accuracy of the model: 78.49%, it is moderate good.
- R-Square (adjust): 0,23, the model explains a low portion of the variance in the score G3.
- Multicollinearity, the indicator **vif**, variance inflation factor, is less than 2.5, indicate no multicollinearity.

2) Influential outliers

Using the Cook's distance to measures of influential Outliers, show a few impacts.

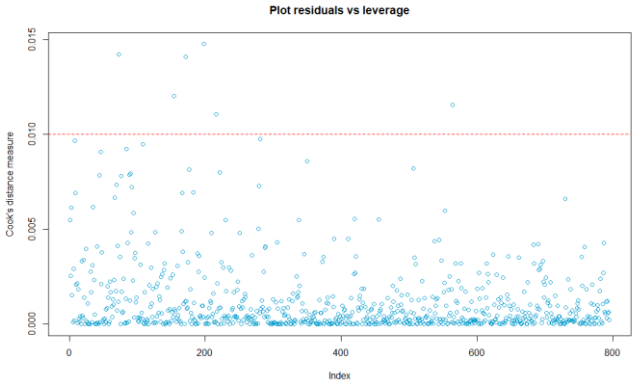

Fig. 31. Cook's distance Hypothesis 1: performance.

3) Evaluate residual and normality of the residual error

The next graphs show than there is no correlation between the dimensions, a good value for the residual error and also normality on this residual error.

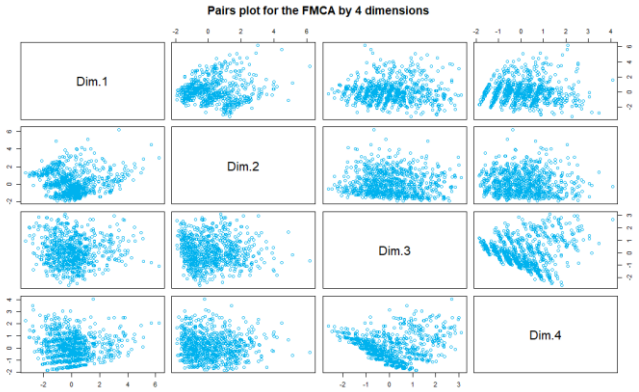

Fig. 32. Correlation between dimensions on Hypothesis 1: performance.

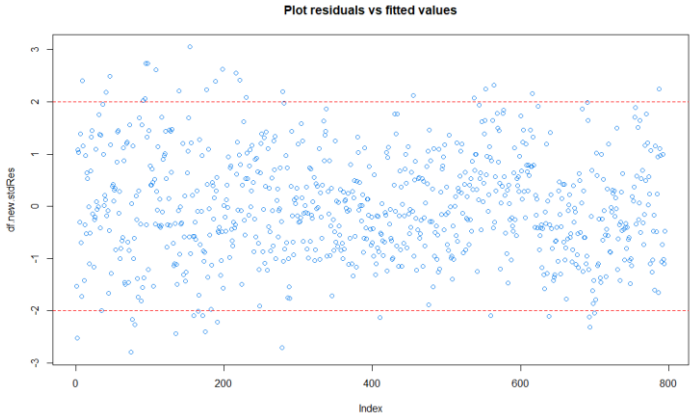

Fig. 33. Residuals error Hypothesis 1: performance.

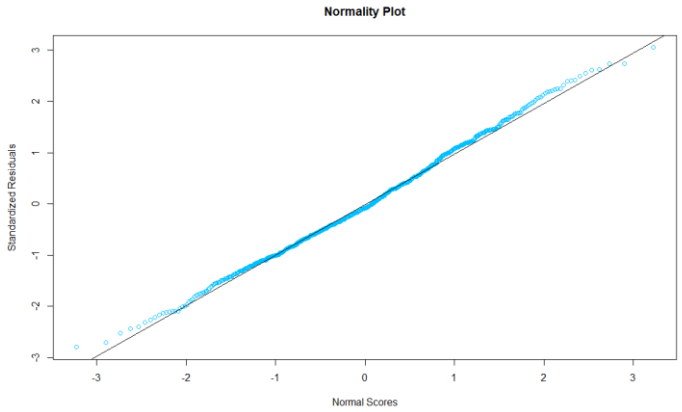

Fig. 34. Normality for the residuals error Hypothesis 1: performance.

B. Apply multiple linear regression model for Hypothesis 2: family situation

A similar to the previous analysis, the Determinant apply to the correlation matrix for the new set of data is 1, mean that there is no multicollinearity in the dataset.

Applying the multiple linear regression model to the training dataset obtains the coefficients for the model and the F statistic,

$$F(7,787) = 8.88, \text{ p-value: } 1.49e\text{-}10 < 0.05$$

Mean that the test is statistically significant for Hypothesis 2, and there is enough evidence to reject the Ho in favour of the Ha. There is an effect of the dimensions in the final score of the student.

```
> model <- lm(G3 ~ Dim.1 + Dim.2 + Dim.3 + Dim.4 + Dim.5 +
+             Dim.6 + Dim.7, data=train.data)
> # linear model with 8 PCA
> summary(model)

Call:
lm(formula = G3 ~ Dim.1 + Dim.2 + Dim.3 + Dim.4 + Dim.5 + Dim.6 +
    Dim.7, data = train.data)

Residuals:
    Min      1Q  Median      3Q     Max
-7.6254 -1.8122 -0.2396  1.9315  7.4517

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.96315    0.10065 118.861  < 2e-16 ***
Dim.1        0.43874    0.06472   6.779 2.37e-11 ***
Dim.2       -0.24971    0.08214  -3.040  0.00244 **
Dim.3       -0.07736    0.08904  -0.869  0.38522
Dim.4       -0.15434    0.09152  -1.686  0.09210 .
Dim.5        0.19269    0.09343   2.062  0.03950 *
Dim.6       -0.06353    0.09600  -0.662  0.50832
Dim.7        0.05527    0.10168   0.544  0.58686
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.836 on 787 degrees of freedom
Multiple R-squared:  0.07322,   Adjusted R-squared:  0.06497
F-statistic: 8.882 on 7 and 787 DF,  p-value: 1.49e-10
```

Fig. 35. Results MLR to the set of data Hypothesis 2: family situation.

1) Performance of the model.

Applying the model to the test dataset is obtained the following results:

- RMSE: 2,61, average error performed by the model similar to RSE.
- RSE: 2.83, it will be used to calculate the average prediction error rate.
- Average predictor error rate: 23.71 % is an average error performed by the model in predicting the final score G3.
- Accuracy of the model: 76.28%, it is low-moderate good.
- R-Square (adjust): 0.06, the model explains a low portion of the variance in the score G3.
- Multicollinearity, the indicator **vif**, variance inflation factor, is less than 2.5, indicate no multicollinearity.

2) Influential outliers

Using the Cook's distance to measures of influential Outliers, show a few impacts.

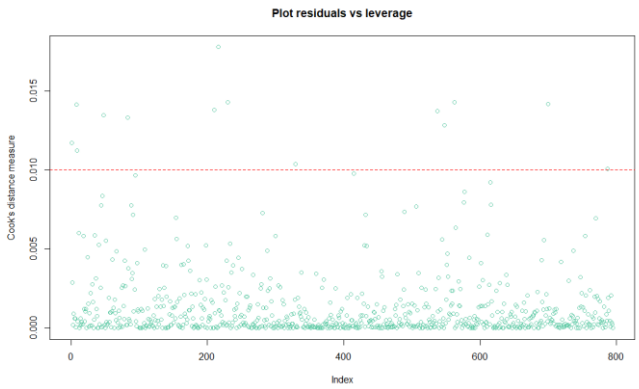

Fig. 36. Cook's distance Hypothesis 2: family situation.

3) Evaluate residual and normality of the residual error

The next graphs show than there is no correlation between the dimensions, a good value for the residual error and also normality on this residual error.

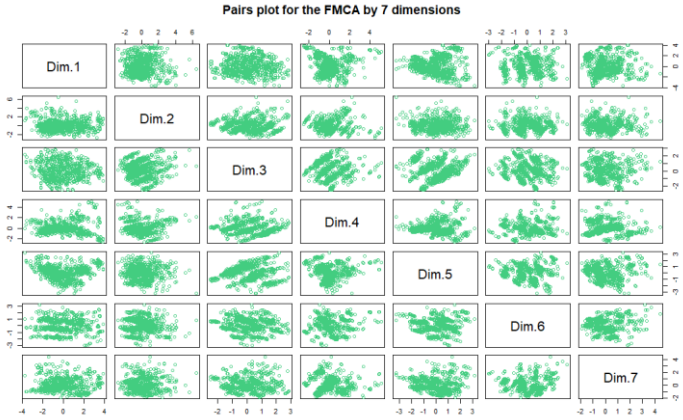

Fig. 37. Correlation between dimensions on Hypothesis 2: family situation.

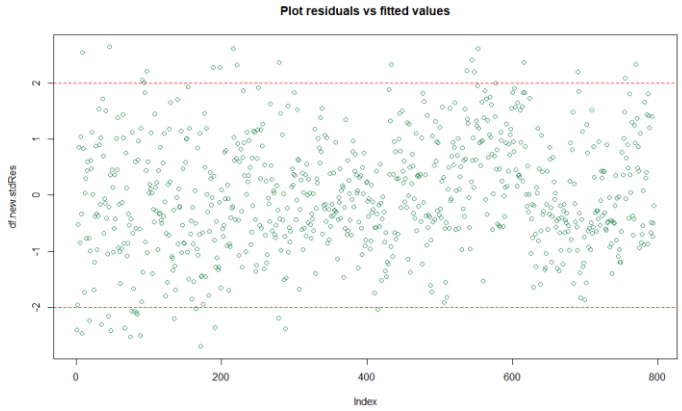

Fig. 38. Residuals error Hypothesis 2: family situation.

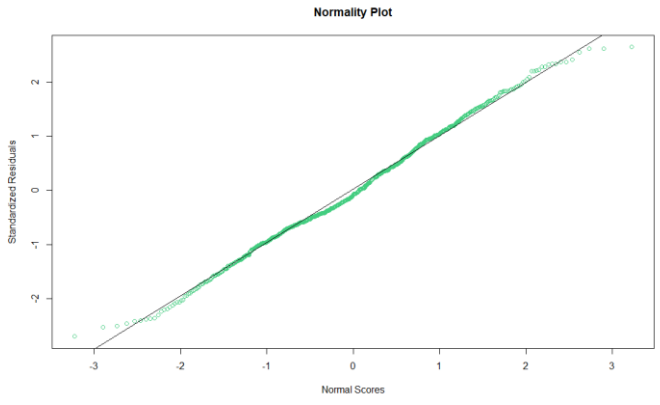

Fig. 39. Normality for the residuals error Hypothesis 2: family situation.

X. FINAL RESULTS AND CHOSE THE BEST MODEL TO FIX THE PROBLEM

Having statistical evidence to support the Hypothesis alternative in both cases, it is founded a linear regression model to fix the problem and predict the score G3 with a moderate level of accuracy, 78.49% for the Hypothesis 1, and 76.28% for the Hypothesis 2, respectively.

In the analyses of Hypothesis 1, about variables related to the performance, it was tested this model:

$$\text{Score G3} = 11.96 + (-0.49)\text{Dim.1} + (-0.62)\text{Dim.2} + (-0.85)\text{Dim. 3} + (-0.21)\text{Dim.4}$$

The four (4) dimensions can be calculated with the percentage of contribution (weights) of the original predictor variables for performance.

The Hypothesis 2, about variables related with the family situation, was tested this model:

$$\text{Score G3} = 11.96 + (0.43)\text{Dim.1} + (-0.24)\text{Dim.2} + (-0.07)\text{Dim.3} + (-0.15)\text{Dim.4} + (0.19)\text{Dim.5} + (-0.06)\text{Dim.6} + (0.05)\text{Dim.7}$$

The seven (7) dimensions can be calculate with the percentage of contribution (weights) of the original predictor variables for family situation.

In this approach, it was found that even though the predictor error in both cases is moderate-low, is a viable solution to tackle the problem with a linear regression model. Otherwise, the multicollinearity of the original variable increases the error and not support the assumption of independence in the predictors.

XI. CONCLUSION AND FUTURE WORK

Using a mixed technique, dimension reduction and linear regression, was possible to build a model to predict the final score G3 for secondary school student in this case. This approach shows the potentiality to mix in a coherent way two different techniques to tackle the regression problem.

Looking more in details the results of the models, perhaps it would be interesting to make more test to optimise the model. In the first case, the model for hypothesis 1, note that the predictor dimension 4 has p-value < 0.05 mean that the contribution to the linear model could be dismissed. A similar analysis would be done for the model in Hypothesis 2, in which some dimensions, 3,4,5,6, and 7, has p-value < 0.05 and could be dismissed too.

Finally, the project has demonstrated the potential to use this approach in regression models.

REFERENCES

[1] Cortez P., Silva A., University of Minho, 2008, https://www.researchgate.net/publication/228780408_Using_data_mining_to_predict_secondary_school_student_performance. Access: November 2019

[2] Kassambara, A. 2017, "Practical Guide to Principal Component Methods in R", STHDA, http://www.sthda.com.

[3] Çamdevýren, H., Demýra N., Kanik A., Keskýn S., 2005, Use of principal component scores in multiple linear regression models for prediction of Chlorophyll-a in reservoirs, Ecological Modelling, Elsevier. Access: November 2019.

[4] Ul-Saufie A.Z, Yahya A., Ramli N., 2011, Improving multiple linear regression model using principal component analysis for predicting PM10 concentration in Seberang Prai, Pulau., International Journal of Environmental Sciences, Volume 2. Access: November 2019.

[5] Mendes M., 2011, Multivariate Multiple Regression Analysis Based on Principal Component Scores to Study Relationships between Some Pre- and Post-slaughter Traits of Broilers. https://www.researchgate.net/publication/298444587_Multivariate_Multiple_Regression_Analysis_Based_on_Principal_Component_Scores_to_Study_Relationships_between_Some_Pre-_and_Post-slaughter_Traits_of_Broilers Access: November 2019.

[6] K. Ping Shung (2018) "Accuracy, Precision, Recall or F1", 2018, https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9 Access: July 2019

[7] Frank, E., Hall, M. A., Pal, C.J. and Witten, I.H., 2017, "Data mining: Practical machine learning tools and techniques". 4th edition. Cambridge, Massachusetts: Elsevier/Morgan Kaufmann. (pp 147).

[8] MarinStatsLectures, 2018, https://statslectures.com/r-stats-datasets Access: June 2019

[9] Jason Brownlee, 2019, "Machine Learning Mastery with Python".

[10] James, Witten, Hastie, Tibshirani, 2013, "An Introduction to Statistical Learning with Applications in R", Springer (pp 25,176)

[11] J. Brownlee, 2018, "How to Reduce Variance in a Final Machine Learning Model", https://machinelearningmastery.com/how-to-reduce-model-variance/ Access: July 2019