STAT 448 Individual Report

Principal Component Analysis of Students' Grade Data

Name: Xinyan Yang(xinyany2) Group number: 6

December 3, 2017

Introduction

Background Information

This dataset comes from Kaggle(https://www.kaggle.com/uciml/student-alcohol-consumption). The data were obtained in a survey of students math and portuguese language courses in secondary school. It contains a lot of interesting social, gender and study information about students.

The original dataset contains two different csv files which collect students' Math grades and Portuguese language grades seperately. Since the Portuguese grades dataset has more observations than the Math grades one, I choose it in order to get a more accurate analysis.

Description of the Dataset

The dataset originally has 649 observations and 33 variables. Each row represents a single student's information which looks like below:

school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian
GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother

traveltime	studytime	failures	schoolsup	famsup	paid	activities	nursery	higher
2	2	0	yes	no	no	no	yes	yes

internet	romantic	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
no	no	4	3	4	1	1	3	4	0	11	11

Here is a simple description of some attributes in the data.

Variable	Description
school	student's school (binary: 'GP' or 'MS')
famsize	family size
Pstatus	parent's cohabitation status
Medu	mother's education
Mjob	mother's job
studytime	weekly study time
failures	number of past class failures
schoolsup	extra educational support
famsup	family educational support
higher	wants to take higher education
Dalc	workday alcohol consumption
Walc	weekend alcohol consumption
G1	first period grade
G2	second period grade
G3	final grade

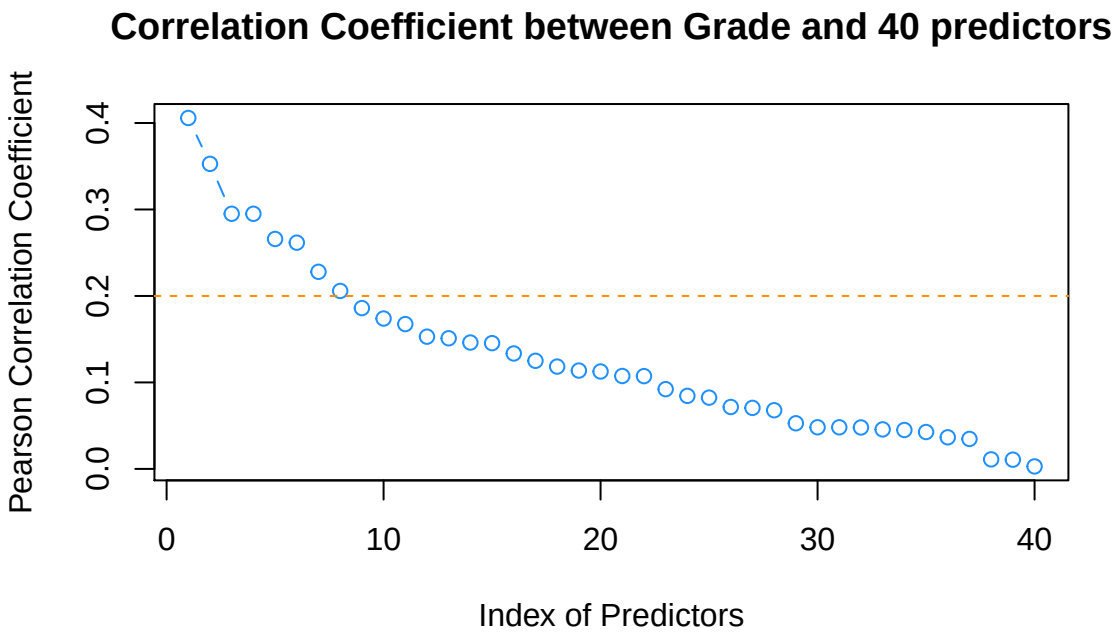
Because we have three variables G1, G2 and G3 which represents to three grades and we are not sure if G3 is the linear combination of G1 and G2, I got the average of these three grades as my target response variable which denoted by y. Besides, since the dataset has some categorical variables, I transformed them into numerical flag values. After all the datapreprocessing, we now have 41 variables including the response y.

The Task and Motivation

Since our dataset has 41 predictors and many of them are correlated with each other(which can be seen from below analysis), I will try to reduce it to fewer dimensions and get rid of the correlation effect using PCA (Principal Component Analysis). The goal is to find out which are the most important factors for predicting the grade.

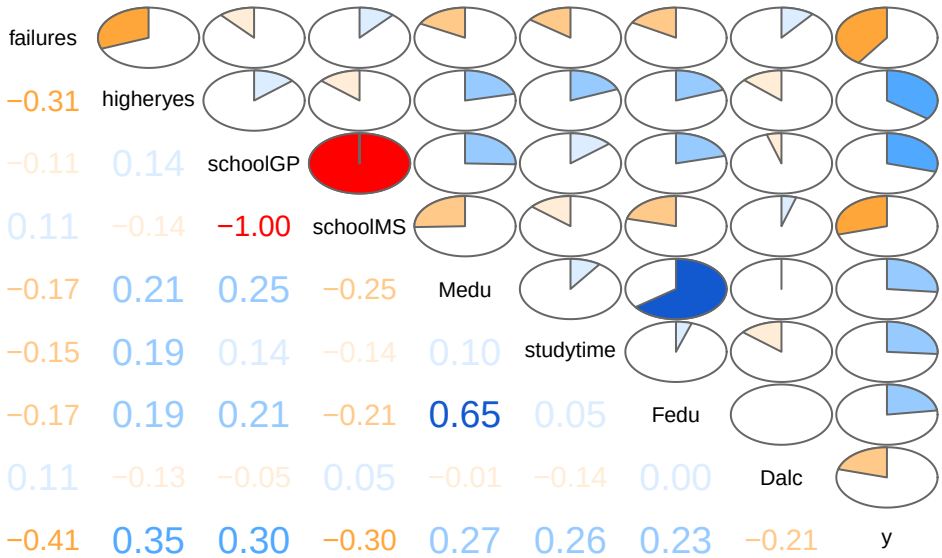
Basic Descriptive Statistics and charts

First we try to get a basic idea of the correlations between the predictors and our response variable grade. So I get the pearson correlation coefficient between X and y, convert them into positive values and plot them in a decreasing order as shown below.



As we can see from the plot, half of the variables have a correlation coefficient less than 0.1. And it turns out that if we set the threshold value to be 0.2, we could pick 8 predictors out of 40, which can greatly reduce the complexity of PCA model without losing too much information. And these 8 predictors are failures, higheryes, schoolGP, schoolMS, Medu, studytime, Fedu and Dalc.

Next by looking at the correlation matrix between these 8 predictors, we can check if there is any multicolinearity issue.



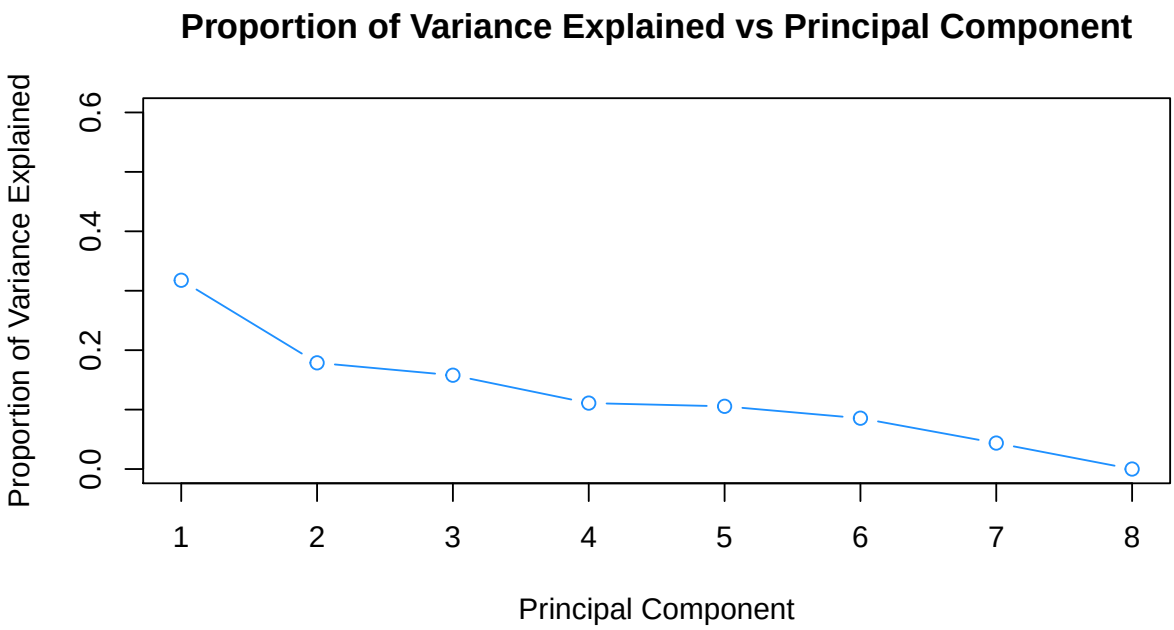
As we can see, all these variables are more or less correlated with each other. For example, schoolMS and schoolGP or Fedu(father education) and Medu(mother education) are highly correlated. If we include all of them in a model to predict y (grades), we would introduce a lot of multicolinearity. That is exactly why we should use PCA- a great tool of reducing dimensionality.

Methods and Results

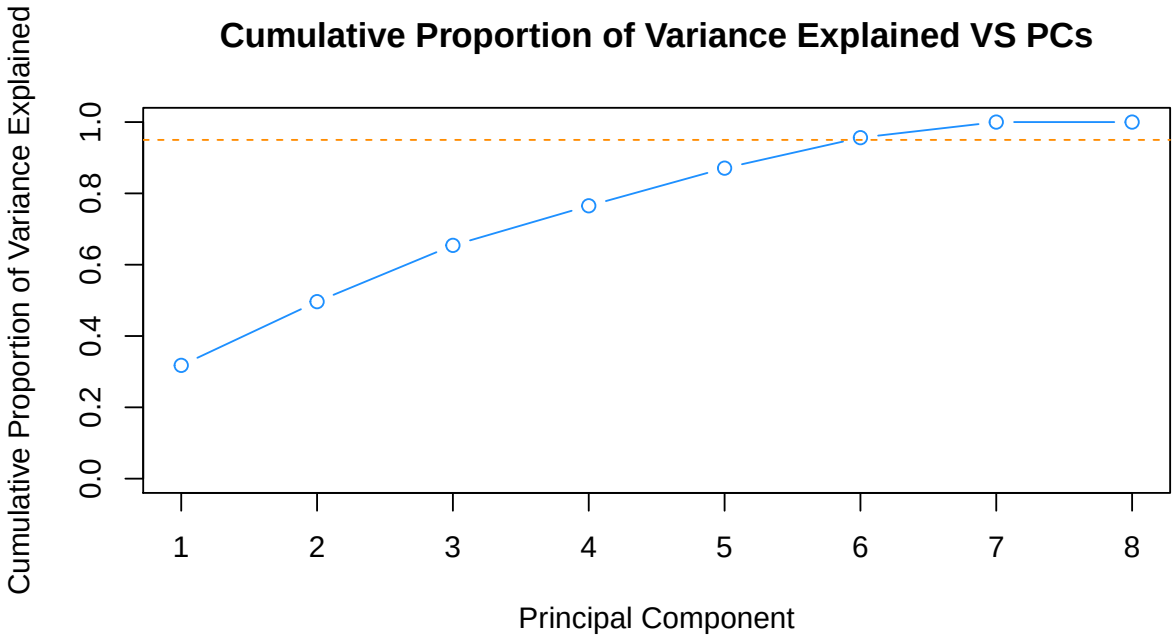
Perform PCA and Choose PCs

Before we do PCA, we need to remove our target variable first.

Frequently we will be interested in the proportion of variance explained by a principal component. The plot below shows us how much variance of the dataset is explained by the 1st, the 2nd ... 8th principle component. Since the 8th PC can explain nearly no varinace, it is kind of unimportant for our dataset.



Since the cumulative proportion can help us decide how many PCs should we keep, which actually helps reduce dimension of the dataset, I then plot the cumulative proportion of variance explained by PCs.



As expected, we see that only 6 PCs can help explain over 95% variance of the data. Therefore I would choose 6 PCs for my future analysis which helps reduce the dimensionality.

Explanation of PCs

The rotation matrix below can help us understand what these PCs composed of and what is the relationship

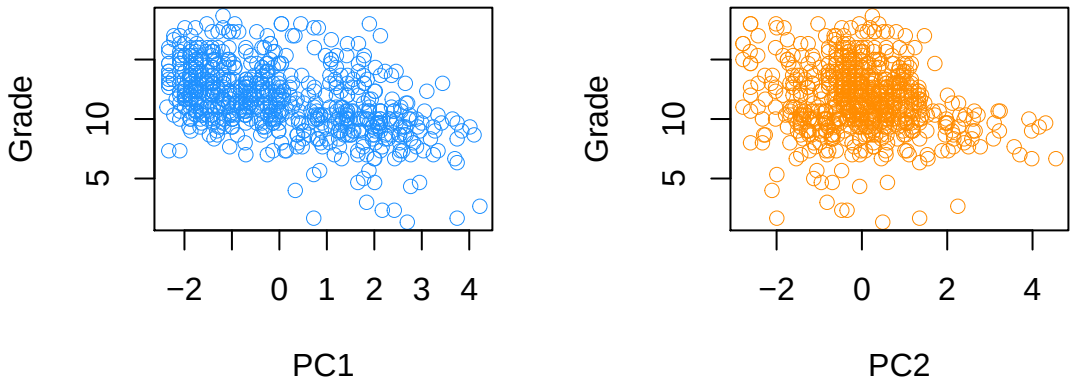
between PCs and the original predictors.

	PC1	PC2	PC3	PC4	PC5	PC6
failures	0.2398590	0.3678625	-0.2754432	-0.4935477	0.3022018	-0.6294728
higheryes	-0.2719702	-0.3821253	0.2833750	0.3161310	-0.1336017	-0.7626610
schoolGP	-0.5103056	0.4787248	0.0663429	0.0396848	-0.0639171	-0.0051981
schoolMS	0.5103056	-0.4787248	-0.0663429	-0.0396848	0.0639171	0.0051981
Medu	-0.4029265	-0.3160939	-0.4193590	-0.2024740	0.0847901	0.0320050
studytime	-0.1930525	-0.1472024	0.4447564	-0.0892246	0.8474169	0.1233739
Fedu	-0.3741094	-0.3386783	-0.4613655	-0.2103699	0.0317574	0.0539387
Dalc	0.0899972	0.1549703	-0.4984963	0.7484049	0.3953993	-0.0538830

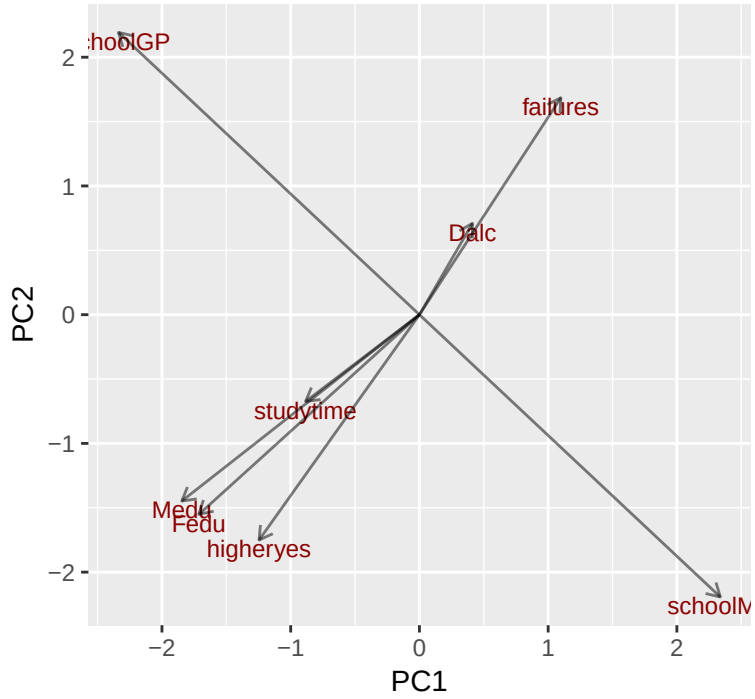
By checking the predictors with high coefficients, we can get a general idea of what these PCs represent.

PC	Predictors with large coefficients	Explanation
1	schoolGP, schoolMS	school
2	failures, higheryes	study motivation
3	Fedu, Medu, studytime	study time
4	Dalc	alcohol consumption

To get a better idea of the relationships between the predictors and PCs, I draw a biplot using PC1 and PC2 as the axes. Before we get into that plot, let's take a look at the relationship between the grade and each PC.



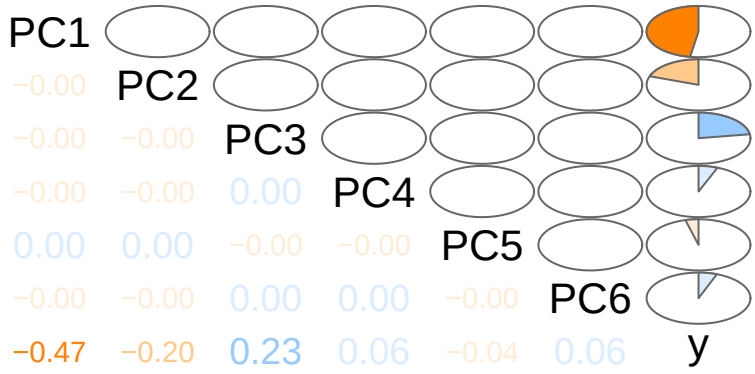
We can see that those who have negative PC1 or PC2 tend to have higher grades than those who have positive PC.



As you can see, the predictors are basically seperated into two groups which point to opposite directions. The group which points to the left, bottom corner has variables like studytime, Medu, Fedu and higheryes. These variables tend to have positive influence(as we talked before) on the response variable grades. The other group which points to the right, top corner has variables like failures and Dalc, which would cause a negative influence on grades.

Check the correlation again

The power of PCA is not only in that it can reduce the dimensionality of data, but also in that it automatically get rid of correlation effect between predictos. For the matrix below I removed PC7 and PC8, as well as added our target variable.



As shown in the plot, there's no correlation effect between PCs any more. In word, PCA does solve the multicolinearity issue of our dataset.

Principal Component Regression

To check if our PCs would get the same prediction power as the previous variables, we can perform a linear regression using PCs as predictors.

```
##
## Call:
## lm(formula = y ~ ., data = pca_corr)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.1534  -1.5366  -0.1405   1.4573   6.4247
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.62506    0.09200 126.360  < 2e-16 ***
## PC1         -0.84263    0.05775 -14.590  < 2e-16 ***
## PC2         -0.46228    0.07700  -6.003 3.23e-09 ***
## PC3          0.56803    0.08193   6.933 1.01e-11 ***
## PC4          0.17897    0.09772   1.832   0.0675 .
## PC5         -0.12537    0.10020  -1.251   0.2113
## PC6          0.19677    0.11131   1.768   0.0776 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.344 on 642 degrees of freedom
## Multiple R-squared:  0.3221, Adjusted R-squared:  0.3158
## F-statistic: 50.84 on 6 and 642 DF,  p-value: < 2.2e-16
```

Let's compare it to the model which use the original predictors.

```
##
## Call:
## lm(formula = y ~ ., data = stu_selected)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.2442  -1.4873  -0.1122   1.4545   6.4352
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.6404     0.4441  19.457  < 2e-16 ***
## failures     -1.3141     0.1658  -7.928 9.95e-15 ***
## higheryes     1.5845     0.3243   4.885 1.31e-06 ***
## schoolGP      1.0868     0.2022   5.374 1.08e-07 ***
## schoolMS          NA         NA      NA       NA
## Medu          0.2309     0.1088   2.123 0.034165 *
## studytime     0.4605     0.1151   4.002 7.02e-05 ***
## Fedu          0.1138     0.1106   1.029 0.304085
## Dalc         -0.3863     0.1014  -3.808 0.000153 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 2.345 on 641 degrees of freedom
## Multiple R-squared:  0.3223, Adjusted R-squared:  0.3149
## F-statistic: 43.56 on 7 and 641 DF,  p-value: < 2.2e-16
```

It turns out that two R-squared barely have any difference, which means that our PCA doesn't lose any prediction power while at the same time reduce the dimensionality.

Conclusion

Our dataset originally has 40 predictors and most of them are correlated with each other. After performing PCA, we both reduce the dimension and get rid of the multicolinearity issue without losing any prediction power. Although sometimes it's hard for us to understand what these principal components represent, some useful conclusions can be drawn from the biplot.

In general, if a student wants to strive for higher education and spends longer time studying, he or she may get higher grades while consuming more alcohol during the week or having more past failures in exams could have a negative influence on the grades, which is kind of reasonable.