Students' Performance in the Portuguese School

Multivariate Data Analysis

Department of Statistics, National Taipei University

Yu-Ting Yeh

November 29, 2022

1 Data Description

The data has 46 students' information. Including students' family status and their studying time, this data has 12 independent metric variables and others. Through principal Component Analysis (PCA) and Factor Analysis (FA), the dimensions will be reduced for bringing a simpler as well as more clear interpretation of data.

The following table shows 10 students with 12 metric independent variables and some description about these variables.

Medu	Fedu	traveltime	studytime	failures	famrel	freetime	goout	Dalc	Walc	health	absences
3	2	2	1	1	2	5	5	5	5	5	10
1	1	3	2	3	5	4	4	3	3	2	8
3	3	2	2	0	4	5	4	2	3	3	2
1	3	1	1	1	4	3	3	2	3	3	7
1	1	3	1	1	4	4	4	3	3	5	4
4	3	2	2	0	4	5	5	1	3	2	4
3	3	1	2	0	5	3	4	1	1	5	0
4	4	2	2	0	4	3	3	1	2	5	4
3	2	2	2	0	1	2	3	1	2	5	2
1	1	2	1	0	3	3	2	1	2	3	4

Figure 1: 10 students' information

Table 1: Metric Variable description

Variable name	Description
Medu	mother' s education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
Fedu	father' s education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
traveltime	home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
studytime	weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
failures	number of past class failures (numeric: n if 1<=n<3, else 4)
famrel	quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
freetime	free time after school (numeric: from 1 - very low to 5 - very high)
goout	going out with friends (numeric: from 1 - very low to 5 - very high)
Dalc	workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
Walc	weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
health	current health status (numeric: from 1 - very bad to 5 - very good)
absences	number of school absences (numeric: from 0 to 93)

2 Principal Component Analysis

Firstly, principal component analysis (PCA) replaces original variables with fewer dimensions which virtually explain for the covariance matrix of independent variables by linear transforming data. The table 2 and table 3 shows the eigenvalues and eigenvectors from covariance matrix. Then, the principal components is $\widehat{y_i} = \widehat{\mathbf{e_i}^\mathbf{T}}\mathbf{x}$, where $i = 1, \ldots, 12$

2.1 Principal Components

Table 2: The first six principal components

Variable	$\widehat{\mathbf{e}_1}$	$\widehat{\mathbf{e}_2}$	$\widehat{\mathbf{e}_3}$	$\widehat{\mathbf{e}_4}$	$\widehat{\mathbf{e}_5}$	$\widehat{\mathbf{e}_6}$
Medu	0.03	-.48	0.5	-.03	0.11	-.21
Fedu	-.03	-.39	0.52	0.01	0.04	-.02
traveltime	-.03	0.18	0.07	-.02	0.08	-.13
studytime	-.04	-.14	0.04	-0.1	0.24	0.21
failures	0.04	0.15	-.16	0.14	0.24	-.24
famrel	-.01	-.05	-.08	0.35	0.75	-0.1
freetime	0.13	-.04	-.04	0.51	0.2	0.43
goout	0.1	-.24	0.01	0.39	-.33	0.45
Dalc	0.14	0.21	0.24	0.31	-.08	-.49
Walc	0.1	0.19	0.16	0.52	-.34	-.21
health	-.05	0.63	0.6	-.12	0.16	0.39
absences	0.97	0.01	-.01	-.22	0.07	0.03
Variance	19.4	2.81	2.09	1.79	1.31	0.83
Cum. Prop.	62%	72%	79%	85%	89%	92%

Table 3: The last six principal components

Variable	$\widehat{\mathbf{e}_7}$	$\widehat{\mathbf{e}_8}$	$\widehat{\mathbf{e}_9}$	$\widehat{\mathbf{e}_{10}}$	$\widehat{\mathbf{e}_{11}}$	$\widehat{\mathbf{e}_{12}}$
Medu	0.46	0.18	-.27	0	0.09	0.37
Fedu	-.45	-0.2	0.35	-.07	-.39	-.22
traveltime	-.08	0.77	0.56	-.01	0.03	0.17
studytime	0.06	-.11	0.24	0.81	0.34	-.13
failures	0.37	-.45	0.47	0.01	-.32	0.39
famrel	-.35	-.03	-.14	-.22	0.32	0.07
freetime	0.28	0.28	-.13	0.12	-.49	-.25
goout	0.09	-.13	0.35	-.31	0.45	0.15
Dalc	0.29	-.06	0.1	-.05	0.26	-.61
Walc	-.37	-.02	-0.2	0.41	-.03	0.38
health	0.06	-.13	-.06	-0.1	0.07	0.11
absences	-.09	0	0.01	0	0.01	0.04
Variance	0.69	0.51	0.38	0.32	0.28	0.24
Cum. Prop.	94%	96%	97%	98%	99%	100%

2.2 Definition of principal Components

This scree plots indicates one dimension should be chosen for the data. Yet, based on Kaiser rule, the rule leaves five dimensions because variance of each of five dimension is more than 1. From my perspective, I prefer to choose five dimensions given that they account for 89% of whole variance.

Define five principal components:

1st PC: Absences

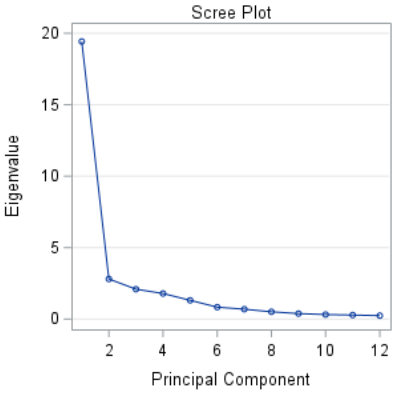
2nd PC: Health

3rd PC: Parents degree (Medu, Fedu)

4th PC: Entertainment (freetime, goout, Dalc, Walc)

5th PC: Family relationship

Figure 2: Scree plot



2.3 Conclusion

- Through PCA, we replace 12 independent metric variables with 5 fewer principal components.
- Now, when adopting this approach, we have index of absences, health, parent degree, entertainment, family relationship.
- On the whole, our new indexes explains for nearly 90% variance, which means PCA shows fabulous performance of this dimension reduction on this data.

3 Factor Analysis

Here, the purpose of FA is to describe the covariance relationship among many variables in terms of a few latent variables. In the meanwhile, these factors can retain the original covariance relationship as possible.

3.1 Appropriate method to Estimation

Two methods, principal components method and maximum likelihood method, can perform estimation of FA. Although maximum likelihood method has the test to check

whether the model is the good fitting, the original independent variables should observe the assumption of normality. Through the exploratory data analysis, the distribution of the variable, absences, does not seem to distribute normally. (Figure 3) We, in turn, adopt the principal components method.

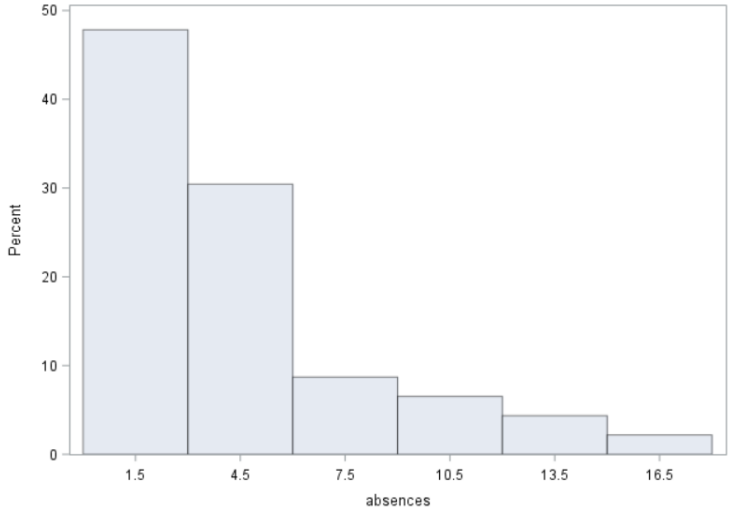


Figure 3: distribution of absences

3.2 Properties of Latent Variables

Table 4: The Result of Factor analysis

Variable	factor1 $\widehat{\mathbf{l}_{i1}}$	factor2 $\widehat{\mathbf{l}_{i2}}$	factor3 $\widehat{\mathbf{l}_{i3}}$	factor4 $\widehat{\mathbf{l}_{i4}}$	factor5 $\widehat{\mathbf{l}_{i5}}$	factor6 $\widehat{\mathbf{l}_{i6}}$	Commun. $\hat{h}_i$	Spec. Var. $\hat{\psi}_i$
Medu	0	0.9	.03	0	-.1	.12	.85	.15
Fedu	-.1	.87	.07	-.1	.03	-.1	.79	.21
traveltime	.13	-.1	-.3	.11	.35	-.1	.24	.76
studytime	-.6	.27	.02	.14	.03	-.1	.45	.55
failures	.24	-.3	-.1	.52	0	.14	.47	.53
famrel	-.1	.07	0.1	.92	0	-.1	.88	.12
freetime	.18	0	.77	0.4	0	.24	.84	.16
goout	0.2	.14	.82	-.2	-.3	.09	.86	.14
Dalc	.82	0.1	.05	.19	.17	.33	.85	.15
Walc	.84	0	.36	.03	0.1	.02	.84	.16
health	.06	0	-.1	-.1	.98	0	.99	.01
absences	.29	0	.26	.02	-.1	.91	1	0
Variance	3.49	2.32	2.96	1.43	2.51	15.52		
Cum. prop.	11%	19%	29%	33%	41%	92%		

3.3 Definition of latent variables

Carry out factor analysis of covariance matrix by the estimation of principal component analysis. The reason of choosing covariance matrix is the desire to both retrieve meaning latent variables and retain the original variance as possible. The following are the definition of latent variables.

factor1 : disadvantage for academic performance (Dalc, Walc, failures, absences)
factor2 : advantage for academic performance (Medu, Fedu, studytime)
factor3 : entertainment(freetime, goout)
factor4 : family relationship
factor5 : health
factor6 : absences

Three factors respectively represent different perspectives of variables. First of all, the first group of variables in factor 1 contains disadvantageous factors including primarily the consumption of alcohol, and part of the degree of attendance and failures in class. In the second factors, the factor 2 cares about their advantageous factors such as parents' education and hard work they pay. Factor 3 focuses on the entertainment. Factor 4 focuses on the family relationship. Factor 5 focuses on the individual health. At last, the attendance of class dominates factor6.

Fortunately, the defined factors work on this data. Those observations whose factor scores are more than 1 have the value above average. For instance, the average score of health is 3.39. However, the observations whose factor 5 scores are more than 1 all achieve score 5 on the health column. This indicates such definitions of factors in this data are feasible. See below tables display the relationship of original variables and factors.

Obs	Variable	Medu	Fedu	traveltime	studytime	failures	famrel	freetime	goout	Dalc	Walc	health	absences
1	Mean	2.35	2.28	1.91	1.83	0.46	3.83	3.33	3.09	1.76	2.52	3.39	3.76
2	StD	1.22	1.13	0.76	0.71	0.81	1.06	1.1	1.05	1.08	1.11	1.45	4.27

Figure 4: simple statistics of variables

Obs	Dalc	Walc	Factor1
1	5	5	2.68117
29	3	4	1.28033
36	4	3	1.12092
42	4	5	1.65077
43	3	4	1.91103
44	3	3	1.02610
45	3	4	1.67387

(a) factor1 > 1

Obs	Medu	Fedu	studytime	Factor2
6	4	3	2	1.21217
8	4	4	2	1.54654
15	4	4	2	1.72103
18	4	4	3	1.68467
21	4	4	2	1.31992
26	4	4	3	1.69361
29	4	4	2	1.49707
32	4	4	2	1.60087
38	4	4	1	1.35278

(b) factor2 > 1

Obs	freetime	goout	Factor3
1	5	5	1.33012
3	5	4	1.44058
6	5	5	1.96693
12	3	4	1.16692
25	5	5	2.00172
42	5	4	1.01052

(c) factor3 > 1

Obs	famrel	Factor4
2	5	1.79522
19	5	1.33202
28	5	1.62361
29	5	1.10451
36	5	1.34523
39	5	1.20925
42	5	1.43894
44	5	2.01512

(a) factor4 > 1

Obs	health	Factor5
5	5	1.10931
7	5	1.09496
8	5	1.16435
12	5	1.27777
13	5	1.14488
18	5	1.07404
21	5	1.05581
33	5	1.11305
38	5	1.28675
39	5	1.31461
45	5	1.11650

(b) factor5 > 1

Obs	absences	Factor6
21	10	2.13140
24	8	1.50681
25	14	2.02247
31	17	3.05830
36	14	2.46618
38	7	1.07462
42	11	1.05926

(c) factor6 > 1

3.4 Conclusion

- Due to increase of loading by rotation, we retrieve more clear indexes than principal component analysis.
- After factor analysis, the data has six perspectives including advantages for studying, disadvantages for studying, health, family relationship, entertainment, absences.
- On the whole, we have reduced dimensions from 12 to 6; Our new indexes can explain for 92% variance as well.