Factor Analysis

Mpelias Michael

27 March 2016

Contents

1 Introduction

1.1 Goal of this lesson

The purposes of Factor Analysis are:

- to find the hidden factors behind observed variables: The hidden factors cannot measured directly, but should be "natural groupings" of observed variables
- The reduction of the number of variables inter-correlated: In this meaning, it resembles the principal component analysis.

The goal of this lesson is :

- To comprehend the intuition behind the Factor Analysis
- To understand the procedure in R
- And to learn from an example how to make a valid Factor Analysis

1.2 Description of the file that will be used

The `GREECS` file contains variables (d1 - d15) that are used for psychological evaluation of patients with acute coronary syndrome (GREECS study). we have 1523 observations and 15 variables. Below, are detailed descriptions of these variables:

"How many days (in the last 90 days) ...." :

- d1: You were not satisfied with your life
- d2: Did you find it difficult to sleep at night
- d3: Did you feel taking more responsibility than you can
- d4: People in your environment appreciated you and respected you
- d5: Were you satisfied with your socializing
- d6: Were you satisfied with the performance at work
- d7: Did you feel free in your work
- d8: Did you feel your work is penetrating your personal life
- d9: Did you have financial security and peace of mind from your work
- d10: Were you satisfied with your income
- d11: Did you have enough time for yourself
- d12: The communication with your partner was good
- d13: Did you feel oppressed in your relationship
- d14: You relied on your family for a problem
- d15: Your family environment had been influencing your decision making

All these variables are categorical with five categories (**0: 0-5, 1: 5-15, 2: 15-45, 3: 45-60, 4:> 60 days aka ordinal**) this categorization was made on (almost) Normally Distributed data. Finally the variable **cvd_event** expresses recurrence cardiovascular event one month after the first episode (1 = yes, 0 = no).

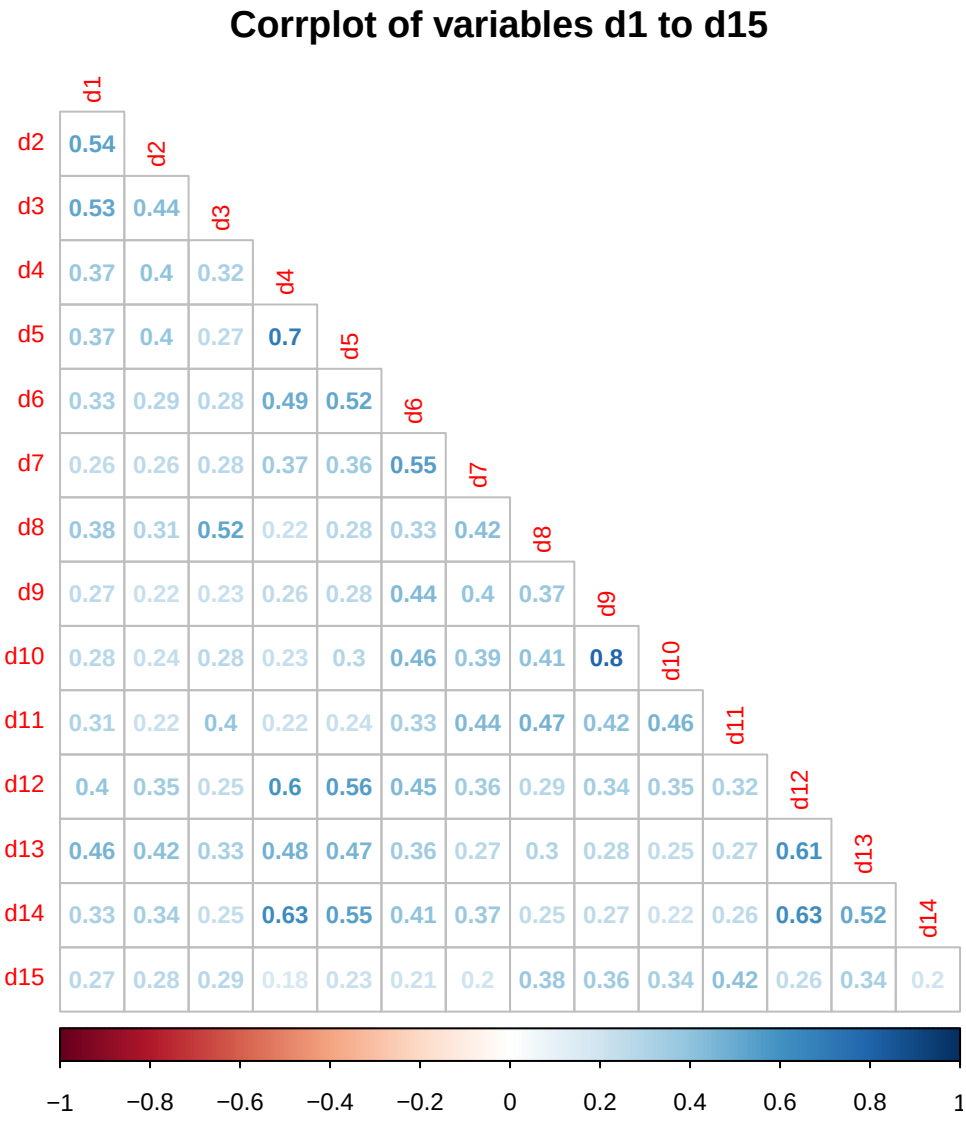
2 Factor Analysis

2.1 Checking if our data are suitable for Factor Analysis

2.1.1 1st Prrequisite: sample size & Correlation adequacy

Similarly to PCA a sample size of 50 observations is very poor, 200 is fair, 300 is good, 500 very good, and more than 1,000 excellent

Our data in order to be suitable for factor analysis they must be correlated enough (about 30-40% of total correlation table elements).



From the results it is concluded that we have a sufficient number of correlations between variables with $|r| > 0.3$ in order to carry out factor analysis that will result in good results for interpretation.

2.1.2 Kaiser-Meyer-Olkin's sampling adequacy criteria

We proceed to the Kaiser-Meyer-Olkin test. Kaiser-Meyer-Olkin's sampling adequacy criterion is a measure of data correlation of the correlations matrix or variance-covariance matrix and takes values between (0,1). The proposed minimum price to be taken in order for the factor analysis to be credible is 0.5, while values> 0.8 are considered to be very good.

The Kaiser-Meyer-Olkin test with MSA (individual measures of sampling adequacy for each item), is not implemented into a package, but it can be easily created with the following function.

```
kmo <- function(x)
{

r <- cor(x,use = "na.or.complete") # Correlation matrix
r2 <- r^2 # Squared correlation coefficients
i <- solve(r) # Inverse matrix of correlation matrix
d <- diag(i) # Diagonal elements of inverse matrix
p2 <- (-i/sqrt(outer(d, d)))^2 # Squared partial correlation coefficients
diag(r2) <- diag(p2) <- 0 # Delete diagonal elements
KMO <- sum(r2)/(sum(r2)+sum(p2))
MSA <- colSums(r2)/(colSums(r2)+colSums(p2))
return(list(KMO=round(KMO,3), MSA=round(MSA,3)))
}
kmo(fa)
```

```
$KMO
[1] 0.883

$MSA
   d1    d2    d3    d4    d5    d6    d7    d8    d9   d10   d11   d12
0.903 0.914 0.852 0.876 0.900 0.921 0.887 0.899 0.790 0.784 0.901 0.910
  d13   d14   d15
0.919 0.914 0.898
```

Variables with MSA being below 0.5 indicate that the item does not belong to a group and should be removed from the analysis.

2.1.3 Sphericity of Bartlett test

Another criterion is sphericity of Bartlett test a statistical test with null hypothesis:

- (H_0): The correlation matrix is the identity matrix (i.e. the diagonal values are equal to 1 and all others 0)
- (H_1): The correlation matrix and the Covariance- Variance matrices are not equal to identity matrices, thus there are correlations between the analysis variables.

The Barlett Sphericity Test also can be created easily with the following function:

```
Bartlett.sphericity.test <- function(x)
{
method <- "Bartlett's test of sphericity"
object.name <- deparse(substitute(x))   ## Save the Name of the Matrix (if any)
x <- subset(x, complete.cases(x)) # Omit missing values
n <- nrow(x)        # count Row length
p <- ncol(x)           # count Column length
chisq <- (1-n+(2*p+5)/6)*log(det(cor(x)))
names(chisq) <- "X-squared"
df <- p*(p-1)/2
names(df) <- "df"

p.value <- pchisq(chisq, df, lower.tail=FALSE)

return(structure(list(statistic=chisq, parameter=df, p.value=p.value,
method=method, data.name=object.name), class="htest"))
}
```

```
    Bartlett's test of sphericity

data:  fa
X-squared = 6373.8, df = 105, p-value < 2.2e-16
```

KMC test had a value of 0.883> 0.5 showing that our data are correlated enough to make a factor analysis. Also the p value Bartlett's test is (p <0.001) which indicates that we accept the null hypothesis at a significance level of a = 0.05, so the correlation matrix is not the identity matrix and it can be concluded that our data are sufficiently correlated with each other.

The presence of multicollinearity (variables with high correlation between them) may cause problems in factor analysis as it is impossible to assess the unique contribution of highly correlated variables with the factors. A first approach is to identify the variables that have a large coefficient of correlation between them (r> 0.9) and remove them from the analysis. A second way to find out if we have multicollinearity is to check the Det of the correlation matrix. A determinant value greater than 0.000001 means that our data do not show multicollinearity and may proceed with the factor analysis. In our case, the value of the determinant is 8.1×10^{-4} so our data do not have multicollinearity.

3 Starting The Factor Analysis

3.1 A First Model

The command for Factor Analysis is `factanal` and we add the following options `factors = 10 ,na.action = na.omit , rotation = "none , factors = 10"`. The default factoring method is Maximum Likelihood ,where the multi-variable assumption of Normality is considered to be `TRUE`.Applying the maximum likelihood method our variables are considered continuous or bivalent. In this example, because we have psychometric test, the ordinal categorical variables with more than 5 categories can be considered to come from a normal distribution, we proceed to the analysis.

Table 1: Factor Analysis / correlation

	Eigenvalue	Difference	Proportion	Cumulative
ML1	5.5258217	3.9925756	0.4862873	0.4862873
ML2	1.5332461	0.4724340	0.1349298	0.6212171
ML3	1.0608122	0.3196259	0.0933543	0.7145715
ML4	0.7411863	0.0119862	0.0652264	0.7797979
ML8	0.7292001	0.1474652	0.0641716	0.8439694
ML6	0.5817349	0.1557014	0.0511942	0.8951637
ML9	0.4260334	0.0631232	0.0374921	0.9326558
ML5	0.3629102	0.1406202	0.0319371	0.9645929
ML7	0.2222900	0.0422386	0.0195621	0.9841550
ML10	0.1800513	NA	0.0158450	1.0000000

```
    LR Test: Indepented vs. Saturated

data:
Chi-squared = 6380.9, Degrees Of Freedom = 105, p-value < 2.2e-16
```

In the first table we take the Eigenvalues, the Difference, the Proportion and the Cumulative proportion. After that an LR-test Independent vs Saturated.

```
library(pander)
loadings = round(a$loadings[1:length(row.names(a$loadings)), 1:length(colnames(a$loadin
Uniqueness= a$uniquenesses
df2 = data.frame(cbind(loadings,Uniqueness))

pander(round(df2 , 4) ,
       caption = "Factor Loadings (pattern matrix) and Unique Variances",
       split.table = 75 )
```

Table 2: Factor Loadings (pattern matrix) and Unique Variances (continued below)

	ML1	ML2	ML3	ML4	ML8	ML6	ML9	ML5
d1	0.6112	0.2506	0.7122	0.09	-0.0042	-0.1786	1e-04	-0.0997
d2	0.5052	0.2147	0.254	0.0275	0.0764	0.0528	0.018	0.0058
d3	0.5068	0.1376	0.4448	0.1542	-0.003	0.6114	-0.0063	0.2915
d4	0.7068	0.52	-0.2281	-0.1622	-0.0062	-0.117	0.0027	0.3569
d5	0.7225	0.4635	-0.2476	-0.2338	-0.0043	0.1224	4e-04	-0.3495
d6	0.6718	0.024	-0.1381	0.1076	0.0176	-0.0013	-0.0655	-0.0344
d7	0.6583	-0.0712	-0.2777	0.6906	-0.0023	-0.0277	-9e-04	-0.0414
d8	0.518	-0.0886	0.1617	0.1833	0.1312	0.2547	0.2044	0.0154
d9	0.7203	-0.5861	-0.0244	-0.179	-0.0028	-0.138	-0.001	0.06
d10	0.7247	-0.5823	0.0091	-0.1823	-0.0014	0.1187	-0.0013	-0.0273
d11	0.5144	-0.1772	0.075	0.1713	0.2194	0.154	0.3509	0.038
d12	0.625	0.2099	-0.0614	-0.0807	0.51	-0.0869	-0.1704	0.0213
d13	0.5383	0.2185	0.1152	-0.0516	0.4559	-0.0363	-0.0564	0.0057
d14	0.5702	0.2993	-0.1359	-0.0223	0.3333	-0.0817	-0.1079	0.0942
d15	0.3894	-0.1064	0.1171	-0.035	0.2803	0.0808	0.461	-0.0106

	ML7	ML10	Uniqueness
d1	-0.039	-0.001	0.005
d2	0.0061	0.2407	0.5664
d3	0.1798	-3e-04	0.0115
d4	-0.0749	1e-04	0.005
d5	0.071	-4e-04	0.005
d6	-0.0282	-0.0224	0.5104
d7	-0.0098	9e-04	0.005
d8	0.0641	-0.0325	0.5349
d9	0.278	-3e-04	0.005
d10	-0.2875	8e-04	0.005
d11	-0.0091	-0.233	0.4182
d12	-0.0429	-0.0844	0.2489
d13	0.0479	0.1748	0.4013
d14	0.0357	-0.0297	0.4259
d15	0.0855	0.1665	0.4893

In the second table the Loadings are presented as evaluated by factor analysis. Thus estimated 10 new parameters whose loadings (which are coefficients expressing the relationship of each factor to each variable) are shown in the ten columns and the last column shows the Uniqueness of each initial variable, which expresses the percentage of volatility of each original variable that is not explained by the factorial model. The communality of the initial variables is the

percentage of variability of each variable expressed by the factorial model and is calculated by Subtracting 1 minus the Uniqueness of each variable. Thus, the following table shows that the Uniqueness of of the first variable is 0.5 % so the Factor is interpreting the (1-0.5 % = 99.5 %) of the d1 and leaves 0.5 % unexplained.

For the second variable of the model the analysis considers that interprets 43.36410199953 of the variation, which is vary small so if we find a small Uniqueness to a variable of interest we may need to add one or more factors in our analysis to the increase that value.

Finally we may check the Degrees of Freedom of the model which are 0, so the model is already Saturated and needs a reduction in the factors. `But how many shall we keep?`

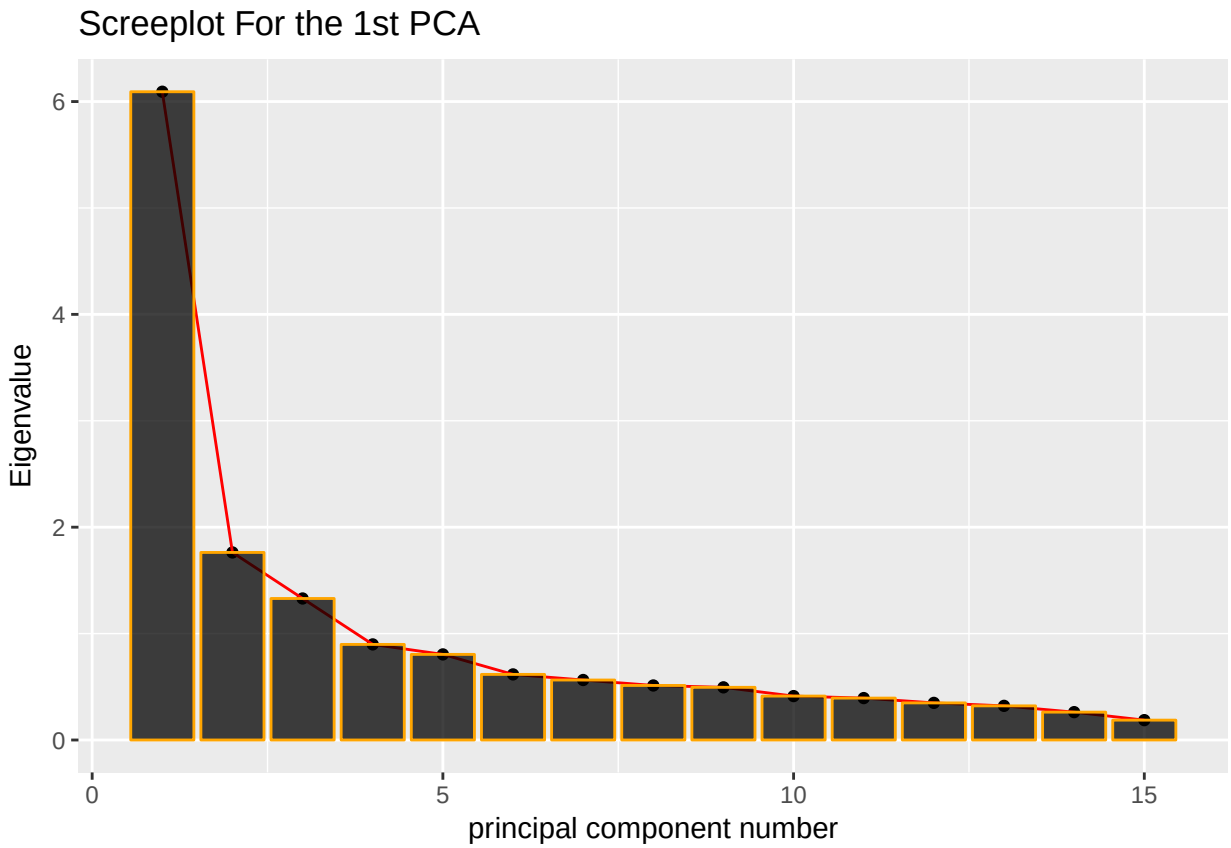
It is proposed to take the same amount of Factors as of Main Components from a PCA. In PCA we pick either by Kaiser's criterion or by Scree-plots.

```
## Warning: In prcomp.default(x, ...) :
##  extra arguments 'scores', 'rotation' will be disregarded
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```



The kaiser criterion 6.0923987, 1.762567, 1.3299129, 0.8983158, 0.804533, 0.6164792, 0.5637711, 0.5125383, 0.4949792, 0.4126762, 0.3937905, 0.3487179, 0.3208857, 0.2617845, 0.1866501 proposes that we should keep 3 variables, while the `Screeplot` 5. We will follow the intuitive path of the scree-plot.

Table 4: Factor Analysis / correlation

	Eigenvalue	Difference	Proportion	Cumulative
ML1	4.5510869	2.1554068	0.5119155	0.5119155
ML2	2.3956801	1.3646826	0.2694709	0.7813864
ML3	1.0309975	0.4827854	0.1159687	0.8973551
ML4	0.5482120	0.1838789	0.0616640	0.9590191
ML5	0.3643332	NA	0.0409809	1.0000000

```
    LR Test: Indepented vs. Saturated

data:  chi2(105) = 6380.89
Chi-squared = 6380.9, Degrees Of Freedom = 105, p-value < 2.2e-16

    LR Test: 5 Factors vs. Saturated
```

```
data:  chi2(40) = 194.57
Chi-squared = 194.57, Degrees Of Freedom = 40, p-value < 2.2e-16
```

Table 5: Factor Loadings (pattern matrix) and Unique Variances

	ML1	ML2	ML3	ML4	ML5	Uniqueness
d1	0.4407	0.3917	0.3641	-0.2087	-0.0915	0.4678
d2	0.3957	0.3993	0.2555	-0.1799	-0.1499	0.5639
d3	0.4164	0.3022	0.5639	-0.0408	-0.1685	0.3872
d4	0.4606	0.6665	-0.2242	0.0211	-0.1985	0.2535
d5	0.4967	0.5684	-0.1977	0.018	-0.1699	0.3619
d6	0.6002	0.2895	-0.0878	0.2756	-0.0807	0.4658
d7	0.527	0.2378	0.0684	0.5026	0.0708	0.4035
d8	0.5174	0.1536	0.4389	0.1594	0.0767	0.4848
d9	0.8043	-0.1804	-0.0474	0.0203	0.0684	0.3132
d10	0.9236	-0.3054	-0.0423	-0.0352	-0.0239	0.05
d11	0.5526	0.0773	0.296	0.1991	0.2025	0.5204
d12	0.5427	0.5371	-0.1803	-0.1151	0.2482	0.3096
d13	0.4491	0.5217	0.0462	-0.2399	0.2386	0.4095
d14	0.4305	0.5974	-0.1799	0.0095	0.1284	0.4089
d15	0.4188	0.1042	0.2493	-0.061	0.1954	0.7097

From the above table we shall write the statistical model of the factors (the model coefficients are listed in Table 2 charges) .

d1= 0.4407*ML1 + 0.3917*ML2 + 0.3641*ML3 + -0.2087*ML4 + -0.0915*ML5

d2= 0.3957*ML1 + 0.3993*ML2 + 0.2555*ML3 + -0.1799*ML4 + -0.1499*ML5

d3= 0.4164*ML1 + 0.3022*ML2 + 0.5639*ML3 + -0.0408*ML4 + -0.1685*ML5

d4= 0.4606*ML1 + 0.6665*ML2 + -0.2242*ML3 + 0.0211*ML4 + -0.1985*ML5

d5= 0.4967*ML1 + 0.5684*ML2 + -0.1977*ML3 + 0.018*ML4 + -0.1699*ML5

we have estimated the parameters of the factorial model without rotation. In order to apply rotation

Table 6: Factor Analysis / correlation

	Eigenvalue	Difference	Proportion	Cumulative
Factor1	4.5505433	2.1543448	0.5118531	0.5118531
Factor2	2.3961985	1.3651849	0.2695286	0.7813817
Factor3	1.0310135	0.4827616	0.1159702	0.8973519
Factor4	0.5482519	0.1839278	0.0616683	0.9590202

	Eigenvalue	Difference	Proportion	Cumulative
Factor5	0.3643241	NA	0.0409798	1.0000000

```
    LR Test: Indepented vs. Saturated

data:  chi2(105) = 6380.89
Chi-squared = 6380.9, Degrees Of Freedom = 105, p-value < 2.2e-16

    LR Test: 5 Factors vs. Saturated

data:  chi2(40) = 194.57
Chi-squared = 194.57, Degrees Of Freedom = 40, p-value < 2.2e-16
```

Table 7: Factor Loadings (pattern matrix) and Unique Variances

	Factor1	Factor2	Factor3	Factor4	Factor5	Uniqueness
d1	0.4407	0.3918	0.3641	-0.2087	-0.0915	0.4678
d2	0.3956	0.3993	0.2555	-0.1799	-0.1499	0.5639
d3	0.4164	0.3023	0.564	-0.0409	-0.1685	0.3871
d4	0.4605	0.6666	-0.2242	0.021	-0.1985	0.2535
d5	0.4966	0.5685	-0.1977	0.018	-0.1699	0.3619
d6	0.6002	0.2896	-0.0878	0.2756	-0.0807	0.4658
d7	0.5269	0.2379	0.0684	0.5027	0.0708	0.4034
d8	0.5174	0.1537	0.4389	0.1594	0.0766	0.4848
d9	0.8043	-0.1803	-0.0474	0.0203	0.0685	0.3132
d10	0.9237	-0.3052	-0.0422	-0.0352	-0.0239	0.05
d11	0.5526	0.0774	0.296	0.1991	0.2024	0.5204
d12	0.5426	0.5372	-0.1803	-0.1151	0.2482	0.3096
d13	0.449	0.5218	0.0462	-0.2399	0.2386	0.4095
d14	0.4304	0.5974	-0.1799	0.0095	0.1284	0.4089
d15	0.4188	0.1043	0.2493	-0.061	0.1954	0.7097

The interpretation of the factors is done by controlling of elements of the matrix after the rotation and the correlation of variables with the greatest burden in each factor. Thus we have :

- The first factor is highly correlated with variables d4, d5, d12 and d14 expressing satisfaction derived by the person from the social environment (family, partner, social relationships) and may be characterized as **positive social environment**
- The second factor is highly correlated with variables d1, d2, d3, d13 and d15 which are

associated with symptoms of anxiety and depression, and we can assume that is the factor of **Anxiety-Depression**

- The third factor is highly correlated with variables d9, d10 and d11, which are related to the work and can be considered as reflecting the **Job satisfaction**
- The fourth factor most associated with job satisfaction d6, d7, d8 , d11 and we believe that is the factor **working environment**
- The fifth factor is highly correlated with d12 and d13 questions and we believe that is the factor **partner relationship satisfaction**

Now that we have our scores we can save them as predicted scores over our dataset. The reason we are doing this is in order to check if these Factor variable are affecting after all the recurrence cardiovascular event or not. So we shall create a new Data.Frame with the predicted values

```
Call:
glm(formula = cvd_event ~ ., family = "binomial", data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7594  -0.4676  -0.4321  -0.3842   2.4840

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.31489    0.13540 -17.097   <2e-16 ***
ML2          0.10838    0.14298   0.758   0.4485
ML3          0.03967    0.16278   0.244   0.8074
ML1         -0.09839    0.14575  -0.675   0.4996
ML4          0.14897    0.17228   0.865   0.3872
ML5         -0.40041    0.19362  -2.068   0.0386 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 437.28  on 699  degrees of freedom
Residual deviance: 431.39  on 694  degrees of freedom
  (202 observations deleted due to missingness)
AIC: 443.39

Number of Fisher Scoring iterations: 5
```