Faculty of Economics and Political Science
English Section
Statistics Major

Multivariate data analysis

How Character Strengths Navigated the COVID-19 Storm

Under the Supervision of

Dr. Sara Osama

T.A Sama Alaa-Eldin

Submitted by

Name	**Code**	**Mail**
Sara Mohammed Hassan	5200300	sarah.abdelghani2020@feps.edu.eg
Mariam Walid Hassan	5200914	maryam.waleed2020@feps.edu.eg
Mawadda Abu-Taleb	5200653	mawada.sameh2020@feps.edu.eg

January 2024

Introduction

The COVID-19 pandemic cast a long shadow over the world, plunging us all into collective isolation. As doors shut, routines unravelled and the world grinded to a halt, the psychological toll grew evident, casting a cloud of anxiety, stress, and a weakened sense of self-efficacy. Yet, amidst this misfortune glimmered a silver of resilience. This study delves into the fascinating potential of character strengths as shields against the pandemic's psychological onslaught.

Drawing on data from 944 Italian individuals navigating the early days of lockdown, this study focuses on the protective role of character strengths in sustaining mental health and self-efficacy during lockdown.

By analysing their responses to surveys assessing key traits like transcendence, interpersonal connection, openness, and restraint, the study aimed to find how certain qualities might have buffered the blows of isolation.

In this report we will compare a sample of 850 observations to the original study after making the sample undergo several techniques of data reduction and description to unveil any hidden patterns, where the data will undergo factor analysis, cluster analysis and multi-variate regression to observe how much of the variance in the dependent variables was explained.

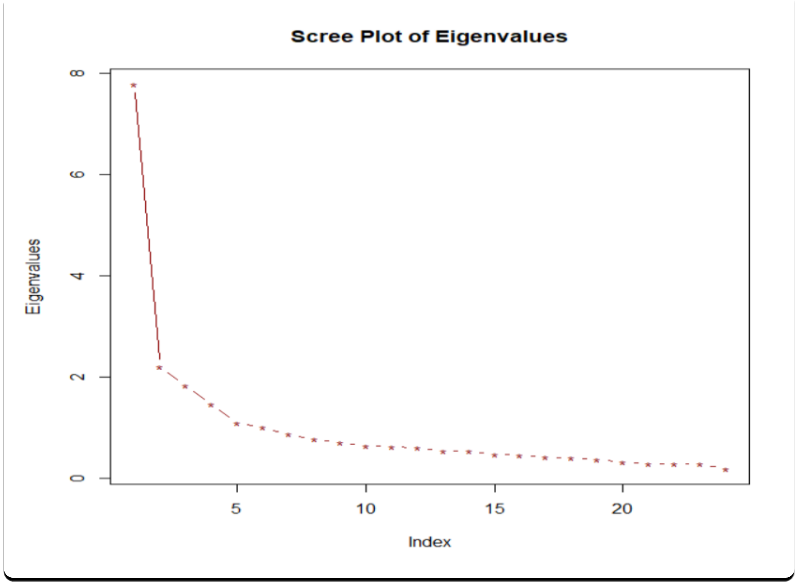
We will be working on the same explanatory variables in the original paper and the same three dependent variables which were: DASS21 (Depression Anxiety and Stress Scale), GHQ12 (General Health Questionnaire), and SEC (Self-efficacy for Covid-19).

Factor analysis

Factor analysis is a statistical technique that reduces a set of variables by extracting all their commonalities into a smaller number of factors. We will be focusing on the **standardized data** as we aim for describing it.

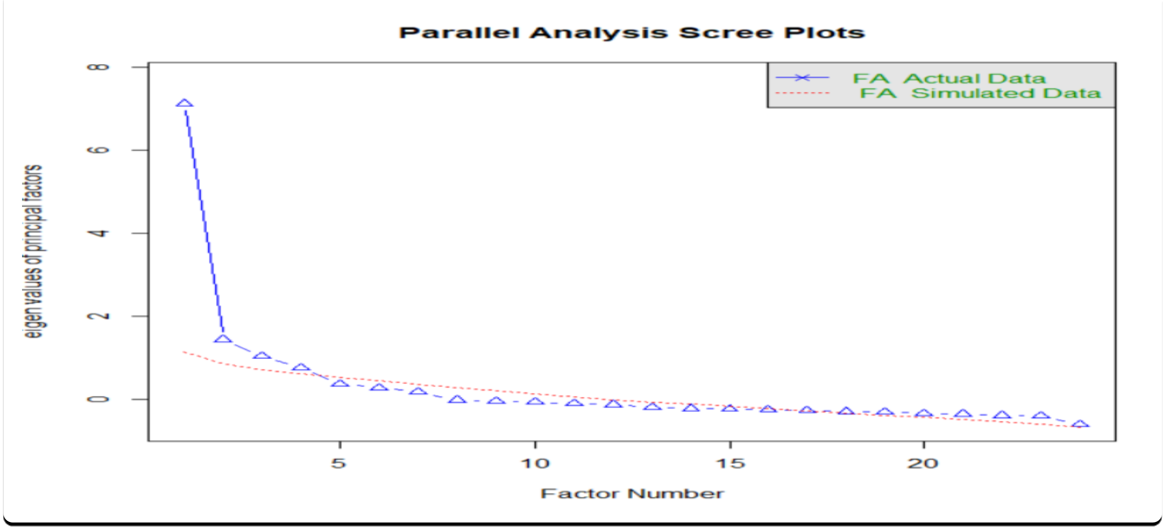
1. Determining the number of common factors to retain

Using the scree plot, as it's the best approach in determining number of common factors to retain in the model, which is a line graph of values of the eigenvalues against the eigenvalues order.



The scree plot of eigenvalues suggests that we should include **four factors** in our model.

For more validation, we used the parallel analysis scree plot, which is a technique designed to help take some of the subjectivity out of interpreting the scree plot.



Also, the parallel analysis scree plot agrees with the results obtained from the scree plot as shown in the graph below, we would retain **four factors** only.

2. Estimating factor loadings

While exploring the best way to estimate the factor loadings, comparing both principal component method and principal factor method, it was found that both give very close results as we have large number of variables and small number of common factors, we will use the **principal component method**, to avoid having communalities to exceed 1.

	MR1	MR2	MR3	MR4
Appreciation_of_beauty	0.5535331	0.037843482	-0.05321665	0.114756810
Bravery	0.4985836	-0.288820843	0.19450621	0.200888922
Creativity	0.5702498	-0.287683817	0.25837486	0.222051969
Curiosity	0.6932406	-0.371221840	0.13519857	0.013919441
Fairness	0.5550784	0.346781485	-0.31930108	0.256639733
Forgiveness	0.3555257	0.166530493	-0.30432980	-0.160982295
Gratitude	0.7000262	-0.061304702	-0.15163833	-0.275223788
Honesty	0.5753644	0.149757515	0.06389708	0.009849324
Hope	0.7328027	-0.267550419	0.01452263	-0.292733975
Humilty	0.2727970	0.463106008	-0.22431762	-0.126671129
Humor	0.4361034	-0.227291415	-0.06951432	0.263087578
Judgment	0.4216323	0.378275748	0.53555366	0.087810277
Kindness	0.6385705	0.160720501	-0.33287153	0.217434013
Leadership	0.6037014	0.143511404	-0.15062596	0.273612697
Love	0.5432624	-0.025510931	-0.10633696	-0.077984676
Love_of_learning	0.4242900	-0.099979920	0.15099005	0.067179560
Perseverance	0.5854254	-0.042400854	0.15647683	-0.248876483
Perspective	0.4779769	0.190735935	0.38887506	0.138714543
Prudence	0.3263478	0.662666060	0.38893674	-0.194197225
Self_regulation	0.4478269	0.149181876	0.15110466	-0.282349211
Social_intelligence	0.6385939	0.007083113	0.03902839	0.197802446
Spirituality	0.5300419	-0.107541972	-0.16925531	-0.270917130
Teamwork	0.5187811	0.259996852	-0.32496202	0.057502101
Zest	0.7778680	-0.393310251	-0.05168710	-0.172734370

As shown in the above table, it can be noticed that loadings overlap, and it would be hard to interpret.

We need to make factor rotation.

3. Factor rotation

Using factor rotation to minimize the complexity of the factor loadings to make the structure simpler to interpret. While keeping the extracted factors orthogonal (uncorrelated), as well ordered in terms of importance.

Comparing the varimax rotation and oblimin rotation, we used the **varimax rotation** which is an orthogonal rotation method that maximizes the variance of the factor loadings on each factor, while keeping the factor orthogonal.

	MR1	MR4	MR2	MR3
Appreciation_of_beauty	0.19377262	0.37115921	0.076225715	0.1346556233
Bravery	0.26391644	0.08703744	0.029765041	0.4676162728
Creativity	0.28561705	0.08796335	0.091289302	0.5286645725
Curiosity	0.59590469	0.03212588	-0.001461356	0.3582399474
Fairness	-0.10945561	0.80859486	0.049400932	-0.0302021800
Forgiveness	0.26226484	0.31949823	-0.030442910	-0.3015802691
Gratitude	0.69537244	0.16426554	0.026419175	-0.1544411444
Honesty	0.24261915	0.27025190	0.262246190	0.0677000015
Hope	0.84567697	-0.04778695	0.029786968	0.0155134131
Humilty	0.01871658	0.39055460	0.204637619	-0.3777483177
Humor	0.13437002	0.32292133	-0.158088754	0.3408254559
Judgment	-0.04292504	0.03575273	0.736527437	0.2492531986
Kindness	0.07831149	0.73797704	-0.060351872	0.0280064065
Leadership	0.01639713	0.62853034	0.054333268	0.1692999537
Love	0.40156603	0.23109802	0.022075391	-0.0175626479
Love_of_learning	0.24023419	0.07376490	0.126460814	0.2444546131
Perseverance	0.59555544	-0.07319908	0.251681096	0.0004414452
Perspective	0.04547193	0.11489164	0.502157922	0.3046408258
Prudence	0.00550651	0.01973369	0.835244437	-0.1896747065
Self_regulation	0.44069549	-0.06352183	0.351091368	-0.1357550402
Social_intelligence	0.18358302	0.38920062	0.131664356	0.2752925826
Spirituality	0.61810097	0.08801465	-0.052543791	-0.1672503919
Teamwork	0.10237891	0.60853036	0.012634931	-0.1564195189
Zest	0.82961560	0.04980846	-0.114640720	0.1378571184

After rotation the overlap issue was solved, and from the above table we can indicate that the first factor is the most important, and the third factor is the least important.

We will define factors with the highest loadings for each factor that **exceed 0.4**, for the first factor it was found that (curiosity, gratitude, hope, love, perseverance, self- regulation, spirituality, zest) are identified as “flourishing “, for the second factor (judgement, perspective, prudence) are identified as “virtue “, for the third factor (bravery, & creativity) are identified as “wisdom “, for the fourth factor (fairness, kindness, leadership, teamwork) are identified as “resilience “.

Computing the communalities

	Communalities
Appreciation_of_beauty	0.3238322
Bravery	0.4101921
Creativity	0.5240114
Curiosity	0.6368606
Fairness	0.5961866
Forgiveness	0.2726628
Gratitude	0.5925372
Honesty	0.3576513
Hope	0.6944872
Humilty	0.3552494
Humor	0.3158949
Judgment	0.6153947
Kindness	0.5916844
Leadership	0.4826030
Love	0.3131740
Love_of_learning	0.2173291
Perseverance	0.4309452
Perspective	0.4353076
Prudence	0.7346136
Self_regulation	0.3253578
Social_intelligence	0.4485013
Spirituality	0.3945531
Teamwork	0.4456390
Zest	0.7922804

Where the uniqueness is (1- communalities).

Interpreting some of the above communalities:

- Hope captures 69.45% of total variation.
- Creativity captures 52.4% of total variation.
- Honesty captures 35.78% of total variation.
- Forgiveness captures 27.27% of total variation.

4. **Correlation matrix for factors**

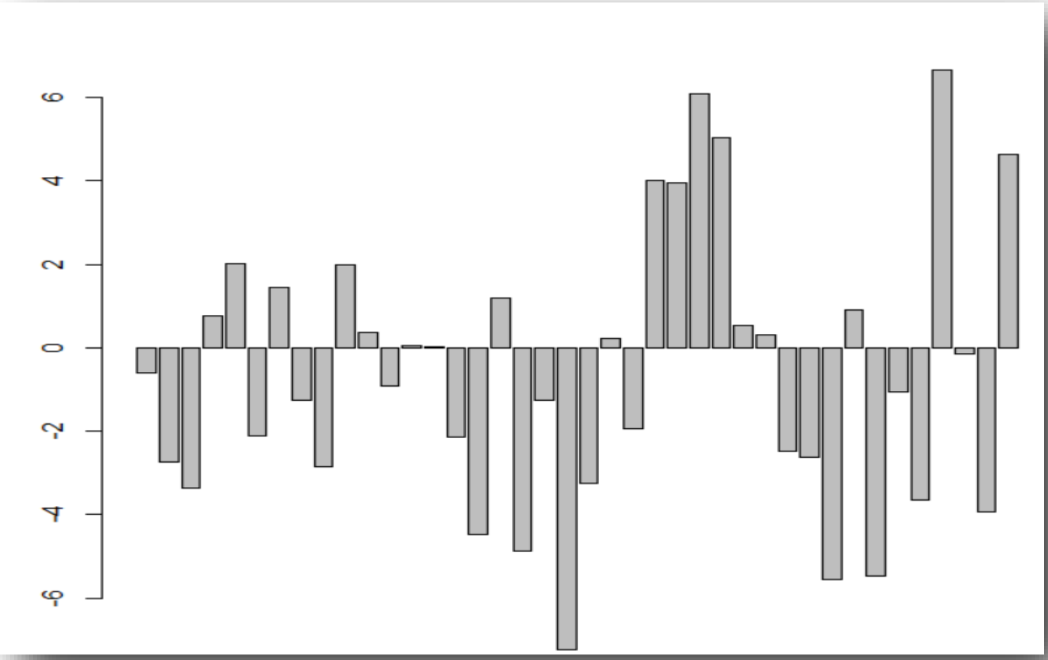
	Transcendence	Interpersonal	Openness	Restraint
Transcendence	1.000000000	0.15734369	0.04496192	0.005230671
Interpersonal	0.157343694	1.00000000	0.06775303	0.013478363
Openness	0.044961917	0.06775303	1.00000000	0.060894964
Restraint	0.005230671	0.01347836	0.06089496	1.000000000

As shown in the above matrix, representing the correlation between the factors into consideration, we can confirm that there is no correlation between the factors.

Cluster analysis

Moving forward into applying cluster analysis, to help us in processing our data into clusters.

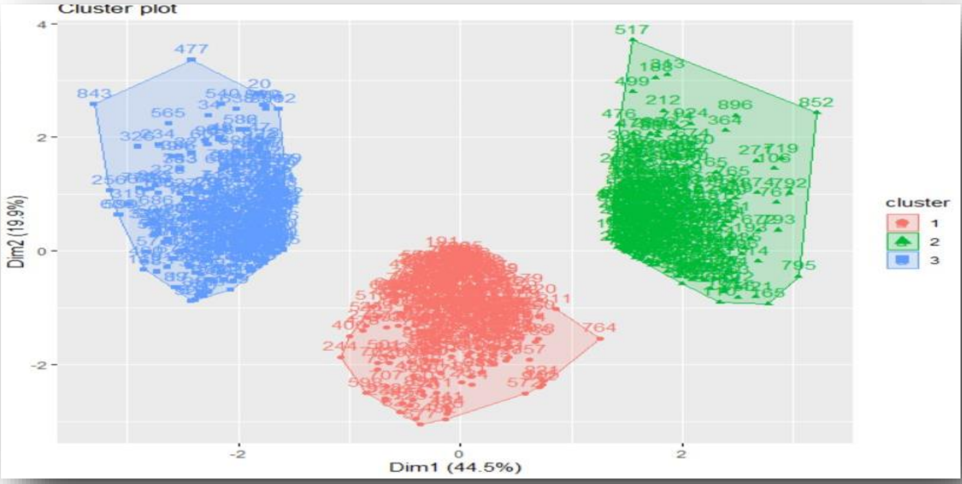
Trying to visualize the number of clusters to be taken through making histogram for the first principal component.



The above histogram suggests either 2 groups (one positive and the other negative) OR that we have 4 groups (1 positive, 1 negative, 1 low negative, 1 low positive).

Using K-mean method

Trying two clusters, three clusters, four clusters, it was found that **three clusters** is our best way of classifying our data.



- The <u>first cluster</u> captures **35.06%** of the data, with **298** observations.
- The <u>second cluster</u> captures **34.43%** of the data, with **293** observations.
- The <u>third cluster</u> captures **30.51%** of the data, with **259** observations.

Mean of clusters

	1	2	3
Transcendence	0.5357035	-0.13898343	-0.4588481
Interpersonal	0.6485657	-0.09215071	-0.6414106
Openness	0.4397214	-0.19392737	-0.2865326
Restraint	0.3132530	-0.90088436	0.6565996

The first cluster has the highest means in variables (transcendence, interpersonal, openness), while the third cluster has the lowest means in (transcendence, interpersonal, openness).

We can identify the cluster as; first cluster as “high unity”, second cluster as “low unity”, third cluster as “moderate unity”

<u>Discriminant analysis</u>

Applying the discriminant analysis was for a validation purpose, to state the level of correctly specifying the data in cluster analysis.

Testing the assumptions of discriminant analysis:

i. Normality assumption

Upon checking if our variables follow multivariate normal, it was found that our variables do not follow multivariate normal, then in order to know which variable caused this, we checked the normality assumption for each variable on the univariate level, it was found that the variables (interpersonal, & openness) are the ones that are not normally distributed.

- ✓ Since we have a large sample size, then the normality assumption is satisfied following the central limit theorem.

ii. Equality of variance covariance matrix

```
        Box's M Test

Chi-Squared Value = 84.2024 , df = 20  and p-value: 7.51e-10
```

p-value< 0.05, we reject H0 at 95% level of significance indicating unequal covariances. The assumption is not satisfied.

iii. Equality of mean vectors

```
> #check equality of means
> #3andna el covs not equal w unknown fa nestakhdm hotelling
> TwoSamplesHT2(factors[,c(1:4)], factors$clusters_3.cluster)$p.value
             [,1]
[1,] 4.278766e-150
```

p-value< 0.05, reject H0 at 95% level of significance, indicating unequal mean vectors. The assumption is satisfied.

- ➢ Based on the results above, the **quadratic discriminant analysis** is more suitable to use, since the assumption of equality of covariances was violated.

Splitting the data & validation

The dataset was split into two subsets: a training set comprising 66.82% of the data and a test set consisting of the remaining 33.18%.

The quadratic model was fitted on the train data of 568 observations, and we used the fitted model to predict the classes for the test data of 282 observations.

Constructing the classification table

```
> xtabs(~predicted$class + test_data$clusters_3.cluster)
               test_data$clusters_3.cluster
predicted$class  1  2  3
              1 95  3  1
              2  2 94  0
              3  0  2 85
```

Based on the above table,

Percentage of correctly classification in **cluster 1**= **97.94%**

Percentage of correctly classification in **cluster 2= 94.95%**

Percentage of correctly classification in **cluster 3**= **98.84%**

Percentage of **overall** correct classification= **97.16%**

So, we can conclude that classifying the data into three clusters gave us a high percentage of correct classification, which supports that we should classify our data into three clusters, which supports the results of the cluster analysis.

Multivariate Regression Model

```
Call:
lm(formula = DASS_21 ~ Age + factor(Gender) + Work + factor(Student)
    Sons + Openness + Restraint + Transcendence + Interpersonal,
    data = mvsample)

Residuals:
    Min      1Q  Median      3Q     Max
-24.525  -6.552  -1.200   4.811  47.006

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            35.98167   10.49881   3.427 0.000636 ***
Age                    -0.04943    0.02745  -1.801 0.072044 .
factor(Gender)Male     -3.17855    0.72573  -4.380 1.32e-05 ***
Work                    0.67960    0.20481   3.318 0.000941 ***
factor(Student)other   -2.44954    9.57925  -0.256 0.798229
factor(Student)Student -1.25708    9.59801  -0.131 0.895825
Sons                   -1.71715    0.69571  -2.468 0.013759 *
Openness                0.10342    0.03030   3.413 0.000670 ***
Restraint               0.02365    0.04654   0.508 0.611478
Transcendence          -0.28774    0.02539 -11.331  < 2e-16 ***
Interpersonal           0.04564    0.03009   1.517 0.129696
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.519 on 933 degrees of freedom
Multiple R-squared:  0.2166,    Adjusted R-squared:  0.2082
F-statistic:  25.8 on 10 and 933 DF,  p-value: < 2.2e-16
```

DASS_21.

We will reject the null hypothesis, since the p-value is less than 0.05 therefore at least one of the coefficients in the model is significant.

The independent variables (Age – Factor (Gender) – Work – Factor (Student) – Sons - Openness – restraint – transcendence – interpersonal) explain 20.82% of the variation in the response variable DASS_21.

On comparing the results to the result in the original report we will find that the percent of variation explained in this model is 8.18% less than the model in the report where the variables in the original model explained 29% of the variation in DASS_21.

```
Response GHQ_12 :

Call:
lm(formula = GHQ_12 ~ Age + factor(Gender) + Work + factor(Student
    Sons + Openness + Restraint + Transcendence + Interpersonal,
    data = mvsample)

Residuals:
     Min       1Q   Median       3Q      Max
-15.0243  -2.9291   0.0798   2.8902  13.4025

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            18.486853   5.077588   3.641 0.000287 ***
Age                     0.030495   0.013275   2.297 0.021828 *
factor(Gender)Male     -0.990245   0.350989  -2.821 0.004884 **
Work                    0.180864   0.099053   1.826 0.068181 .
factor(Student)Other    4.495488   4.632853   0.970 0.332124
factor(Student)Student  5.609594   4.641927   1.208 0.227176
Sons                   -0.041430   0.336472  -0.123 0.902030
Openness               -0.007979   0.014655  -0.544 0.586251
Restraint               0.045803   0.022508   2.035 0.042140 *
Transcendence          -0.104439   0.012282  -8.504  < 2e-16 ***
Interpersonal           0.031504   0.014554   2.165 0.030670 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.604 on 933 degrees of freedom
Multiple R-squared:  0.128,	Adjusted R-squared:  0.1187
F-statistic:  13.7 on 10 and 933 DF,  p-value: < 2.2e-16
```

GHQ_12

We will reject the null hypothesis, since the p-value is less than 0.05 therefore at least one of the coefficients in the model is significant.

The independent variables (Age – Factor (Gender) – Work – Factor (Student) – Sons - Openness – restraint – transcendence – interpersonal) explain 11.87% of the variation in the response variable GHQ_12.

On comparing the results to the result in the original report we will find that the percent of variation explained in this model is 5.3% less than the model in the report where the variables in the original model explained 17% of the variation in GHQ_12.

```
Response SEC :

Call:
lm(formula = SEC ~ Age + factor(Gender) + Work + factor(Student) +
    Sons + Openness + Restraint + Transcendence + Interpersonal,
    data = mvsample)

Residuals:
     Min       1Q   Median       3Q      Max
-13.7453  -2.1069  -0.0497   2.2379  13.3703

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             3.1523817  3.8078893   0.828  0.40796
Age                     0.0069964  0.0099554   0.703  0.48238
factor(Gender)Male      0.9935124  0.2632207   3.774  0.00017 ***
Work                   -0.1518360  0.0742839  -2.044  0.04123 *
factor(Student)Other   -0.5520493  3.4743648  -0.159  0.87379
factor(Student)Student -0.9687592  3.4811699  -0.278  0.78085
Sons                    0.1202903  0.2523337   0.477  0.63368
Openness                0.0130500  0.0109902   1.187  0.23536
Restraint              -0.0005034  0.0168798  -0.030  0.97621
Transcendence           0.1082204  0.0092105  11.750  < 2e-16 ***
Interpersonal          -0.0202995  0.0109146  -1.860  0.06322 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.453 on 933 degrees of freedom
Multiple R-squared:  0.2581,	Adjusted R-squared:  0.2502
F-statistic: 32.46 on 10 and 933 DF,  p-value: < 2.2e-16
```

SEC

We will reject the null hypothesis, since the p-value is less than 0.05 therefore at least one of the coefficients in the model is significant.

The independent variables (Age – Factor (Gender) – Work – Factor (Student) – Sons - Openness – restraint – transcendence – interpersonal) explain 25.02 % of the variation in the response variable SEC.

On comparing the results to the result in the original report we will find that the percent of variation explained in this model is 4.98% less than the model in the report where the variables in the original model explained 30% of the variation in SEC.

Conclusion

It was noticed that initial multi-variate regression model in the original paper demonstrated a higher level of variation explaining (higher R-squared), it can be argued that the multi-variate regression model that was done after a series of data reduction and classification techniques including cluster analysis and factor analysis led to a decline in the percent of variation explained by the model.

This implies that these techniques, while they potentially aided in variable grouping and dimension reduction, they may have inadvertently resulted in the removal or obscuring some of the original variation within the dependent variables.