Using Exploratory Factor Analysis Model (EFA) for Determination the main Factors of Train's Accidents in Egypt (APPLIED STUDY) Marie M. Mahmoud and Maie M. Kamel

Applied Mathematical Sciences, Vol. 4, 2010, no. 38, 1883 - 1897

Using Exploratory Factor Analysis Model (EFA) for Determination the main Factors of Train's Accidents in Egypt (APPLIED STUDY)

Marie M. Mahmoud and Maie M. Kamel

Dept. of Statistics & Mathematics
Faculty of Comm., Tanta University, Egypt
marie_tanta@yahoo.com

Abstract

Factor analysis (FA) is used to summarize the data covariance structure in a few dimensions of the data. FA attempts to identify underlying variables that explain the pattern of correlations within a set of observed variables. It is often used in data reduction to identify a small number of factors that explain most of the variance that is observed in a much larger number of manifest variables. FA can also be used to generate hypotheses regarding causal mechanisms or to screen variables for subsequent analysis (for example, to identify collinearity prior to performing a linear regression analysis).The variables should be quantitative at the interval or ratio level. Data for which Pearson correlation coefficients can sensibly be calculated should be suitable for factor analysis. The data should have a bivariate normal distribution for each pair of variables, and observations should be independent. The computed estimates are based on the assumption that all unique factors are uncorrelated with each other and with the common factors. The present study aims to determination the main Factors of Train's Accidents in Egypt. The statistical packages, MINITAB, SPSS, and FACTOR will be used for the EFA modeling. The present study also makes comparisons between the different methods. Our model leads to: Using Principal Components method with MINITAB and SPSS, there are **3** factors together represent 79.7%, 79.228% of the variability respectively, (<u>the Collision & the Falling</u>, <u>the Parathion & the Exceeded</u>, and <u>the Fires & the Human errors factor</u>). But with FACTOR package,

there are **2** factors together represent 60.268% of the variability (the Collision & Falling, and the Fires, the Parathion & the Exceeded factor). Using Maximum Likelihood method with MINITAB, there are the same**3** factors using PC method, together represent 74.7% of the variability. But with SPSS and FACTOR package, there are two factors together represent 52.259%, 60.268% of the variability respectively (the Collision & Falling, and the Fires, the Parathion & the Exceeded factor). Then the best estimation method is the P C with MINITAB.

Keywords: Train's Accidents, Multivariate analysis, Factor Analysis, Principal Components, and Maximum Likelihood

(I)- Introduction

Multivariate analysis procedures (**MA**) are used to analyze the data when we have made multiple measurements on items or subjects. You can choose a method depending on whether you want to analyze the data covariance structure for the sake of understanding it or to reduce the data dimension, and assign observations to groups. Analyzing the data covariance structure and assigning observations to groups are characterized by their non-inferential nature, that is, tests of significance are not computed. There may be no single answer but what may work best for your data may require knowledge of the situation.

1 - Factor Analysis (FA) is used to summarize the data covariance structure in a smaller number of dimensions. The emphasis in factor analysis, however, is the identification of underlying "factors" that might explain the dimensions associated with large data variability. The assumptions of factor analysis are variable normality distributed, no outliers among cases, Linearity, no variable uncorrelated with any others, and also no variable correlated one with each others. Note that, the absolute number of the determinate of the correlation matrix must be not equal to zero.

Sample size considerations: Sample size calculation for **Exploratory Factor Analysis Model (EFA)** according to Stevens rule, the factor is reliable if it has:

3 or more variables with loadings of 0.8 and any n
4 or more variables with loadings of 0.6 and any n
10 or more variables with loadings of 0.4 and n ≥ 150
Factors with only a few loadings require n ≥ 300

The Model: The following assumes that the ***p*** observed variables (the X_i) that of the ***n*** subjects have been measured for each

$$
\begin{aligned}
X_1 &= u_1 + a_{11}F_1 + a_{12}F_2 + \ldots\ldots + a_{1m}F_m + e_1 \\
X_2 &= u_2 + a_{21}F_1 + a_{22}F_2 + \ldots\ldots + a_{2m}F_m + e_2 \\
&. \\
&. \\
X_p &= u_p + a_{p1}F_1 + a_{p2}F_2 + \ldots\ldots + a_{pm}F_m + e_p
\end{aligned}
\qquad ..(1)
$$

in the matrix form this can be written as

$$X_{p*1} = u_{p*1} + A_{p*m}F_{m*1} + \varepsilon_{p*1} \qquad \ldots\ldots\ldots\ldots..(2)$$

where i = 1,2,...., p

- The F are the m common factors, The F are assumed to be independent and have mean zero and standard deviation one $E(F) = 0_{m*1}$, $\mathrm{Cov(F)} = \mathrm{E(F'F)} = \mathrm{I}_{\mathrm{m*m}}$
- The ***A*** are the factor ***pxm*** factor loadings
- The ε are the independent *p* specific errors,

$E(\varepsilon) = 0_{p*1}$

, and also $Cov(\varepsilon) = E(\varepsilon'\varepsilon) = \psi_{p*p}$

are independent of each other. εF and $Cov(\varepsilon, F) = E(\varepsilon, F') = 0_{p*m}$

using the standardized score with a mean zero and unit variance

The orthogonal factor model is

$$X_i - u_i = a_{i1}F_1 + a_{i2}F_2 + \ldots.. + a_{im}F_m + e_i \ , \qquad i = 1,2,...., p \qquad (3)$$

in the matrix form this can be written as

$$\underline{Z}_{p*1} = A_{p*m}F_{m*1} + e_{p*1} \qquad \ldots (4)$$

Using the above assumption, leads us to the correlation matrix of X

$$\sum_{p*p} = AA' + \psi \qquad \ldots\ldots (5)$$

Where $\boldsymbol{\sum p*p}$ is the correlation matrix of $\mathbf{X}_{px1}$.

This implies that:

$$Var(X_i) = \sum_{j=1}^{m} a_{ij}^2 + \psi \ \ldots\ldots\ldots. \ (6)$$

The Variance of **X** can be decided to two parts:

a) - Part one is related to the common factors which is called Communality (The sum of ***Xi*'s** squared factor loadings)

b) - Part two is unrelated to the common factors which are specificity of $\boldsymbol{X_i}$.

2- Principal Components Analysis (PC). A factor extraction method used to form uncorrelated linear combinations of the observed variables. The first component has maximum variance. Successive components explain progressively smaller portions of the variance and are all uncorrelated with each other. Principal components analysis is used to obtain the initial factor solution. It can be used when a correlation matrix is singular. Principal components analysis is computationally expensive if the number of dependent variables is large. Principal component analysis is used to understand the underlying data structure (the covariance structure in the original variables) and/or to create a smaller number of uncorrelated variables using this structure.

<u>The procedure for factor analysis</u>:

(2-1) Determined the Factor Loading Matrix Lp*m. Exploratory factor analysis (EFA) is a multi-step process. So the principal components analysis (PCA) is the prefect way to factor analysis .Components analysis is only a data reduction

method. PCA used the sample correlation matrix or covariance matrix to specified in terms of the pairs of the Eigenvalus-eigenvectors (**λ_i , e_i**), the ith PC is:

$$PC_i = \sum_{j=1}^{p} e'_{ij} X_j \qquad i = 1,2,..., \ p \tag{7}$$

with following profanities:

$* \, e'e = ee' = I$

$* \, e'_i e_j = 0 \quad \forall \, i \neq j$

$* \sum_{j=1}^{p} e_{ij}^2 = 1 \quad \forall \, PC$

$* \, e_{11}e_{21} + e_{12}e_{22} + ... + e_{1p}e_{2p} = 0$

for the PC_1 & PC_2 *to be* orthogonal

the proportion of total variance explained by the i^{th} Pc is equal to $\dfrac{\lambda_i}{\sum_{i=1}^{p} \lambda_i}$

according to Kaiser's criterion, for every PC_m has λ_m **>1** must be in the analysis.

(2-2) Factor Rotation can be orthogonal or oblique (with orthogonal rotation the new factors are uncorrelated, while with oblique rotation the new factors are correlated). Whichever type of rotation is used the factor loading for the new factors must be either close to zero or very different from zero. ***Varimax rotation*** (by Kaiser 1958) is one method of orthogonal rotation that is often used where it maximizes the sum of the squared factor loadings across the columns. This tends to force each variable to load highly on as few factors as possible. Ideally it will cause each variable to load on only one factor if the Factor Loading Matrix L_{p*m}

then, $\hat{L}^* = LT \qquad \text{where } TT' = T'T = I$ …….. (8)

$\hat{L}^*{}_{p*m}$ is the rotated loading matrix

(2-3) Determined the Factor Scores

3 - The maximum likelihood method (ML) estimates the factor loadings, assuming the data follows a multivariate normal distribution. As its name implies, this method finds a solution by maximizing the likelihood function. Equivalently, this is done by minimizing an expression involving the variances of the residuals. The algorithm iterates until a minimum is found or until the maximum specified number of iterations (the default is 25) is reached. Suppose we have **p** variables and want to fit a model with **k** factors. Let **R** be the (p x p) correlation matrix of the variables, **L** be the (p x k) matrix of factor loadings, and **Y** be the vector of length **p** containing the unique variances. Then we need to find values for **L** and **Y** that maximize the likelihood function, f (L, Y). This involves a two step procedure, first finding a value for **Y**, then for **L**. You can specify the initial value of **Y**. Then calculate **Y** as (1 - communalities).

For a fixed value of **Y**, we maximize f (**L, Y**) with respect to **L**. This is a simple matrix calculation. The value of **L** is then substituted into f (**L, Y**). Now **f** can be

viewed as a function of **Y**. A simple transformation of this function gives, where g1 < g2 < ... gp are eigenvalues of **YR**-1Y**. We then minimize **m(Y).** This gives an estimate of **Y**, which is then put into the likelihood **f (L, Y).** Then the likelihood is again maximized with respect to **L**. Then a new value for **m(Y)** is computed, and so on. Iterations continue for up to **K1** steps, where **K1** is the value given in Max iterations or **25** if you do not enter a value.

Convergence is reached at step n, if either of the following is true:

1- The function m(Y) does not change very much. Specifically, if

| [m(Y) at step n] - [m(Y) at step (n - 1)] | < 10**-6…… (9)

2- None of the unique variances change very much. Specifically, if

| log (Yi at step n) – log (Yi at step n - 1) | < K2……… (10)

For all **i = 1**..., p, where **Yi** is the unique variance corresponding to variable **i**. The value of K2 can be given in Convergence. By default it is **0.005**.

A matrix of second derivatives is used in the minimization of m(**Y**). When minimizing a function like m(**Y**), it is possible to find values of Y that are 0 or negative. Once the algorithm converges, a final check is done on the unique variances. If any are less than **K2**, they are set equal to zero. The corresponding communality is then equal to one. This is called a Heywood case.

4 - Number of factors: The choice of the number of factors is often based upon the proportion of variance explained by the factors, subject matter knowledge, and reasonableness of the solution. Initially, try using the principal components extraction method without specifying the number of components. Examine the proportion of variability explained by different factors and narrow down your choice of how many factors to use. An eigenvalue plot may be useful here in visually assessing the importance of factors. Once you have narrowed this choice, examine the fits of the different factor analyses. Communality values, the proportion of variability of each variable explained by the factors, may be especially useful in comparing fits. You may decide to add a factor if it contributes to the fit of certain variables. Try the maximum likelihood method of extraction as well.

5 - Rotation: Once you have selected the number of factors, you will probably want to try different rotations. Johnson and Wichern suggest the varimax rotation. A similar result from different methods can lend credence to the solution you have selected. At this point you may wish to interpret the factors using your knowledge of the data.

(II) – Applied study

The factor analysis model is: $X = \mu + L F + e$ ……….. (11)

Where **X** is the (p x 1) vector of measurements, **μ** is the (p x 1) vector of means, **L** is a (p x m) matrix of loadings, **F** is a (m x 1) vector of common factors, and **e** is a (p x 1) vector of residuals. Here, **p** represents the number of measurements on a subject or item and **m** represents the number of common factors. **F** and **e** are assumed to be independent and the individual **F's** are independent of each other. The mean of **F** and **e** are **0**, Cov (F) = I, the identity matrix,

and Cov (e) = γ, a diagonal matrix. The assumptions about independence of the **F's** make this an orthogonal factor model. Under the factor analysis model, the (p x p) covariance matrix of the data, **X**, is: $\text{Cov}(X) = L L' + \gamma$……… **(12)**
Where γ is a (p x p) matrix of variances of residuals. The i^{th} diagonal element of **L L'**, the sum of the squared loadings, is called the i^{th} communality. The communality values can be judged as the percent of variability explained by the common factors. The i^{th} diagonal element of is **Y** called the i^{th} specific variance. The specific variance is that portion of variability not explained by the common factors. The sizes of the communalities and/or the specific variances can be used to judge the goodness of fit.

The tables 1 and 2 represent the data of this study, according to the data of the Central Agency for Public Mobilization And Statistics (CAPMAS) of Egypt.
In this paper, PC, ML methods using three Statistical packages, MINITAB, SPSS, and FACTOR will be used for the Exploratory Factor Analysis modeling. The present study also makes comparisons between the different methods.

Table (1): The main variables of the Train's Accidents in Egypt

No	Case	Code	No.	Case	Code
1	Falling to for non stores	**A**	8	These Parathion of trains on all lines	**H**
2	Vehicles	**B**	9	Exceeded Lights all types	**I**
3	Open doors of the trains	**C**	10	Violators of the work of railway	**J**
4	The collision of rail Vehicles at the outlets	**D**	11	The collision of rail gate outlets	**K**
5	The collision of rail Vehicles on the non outlets	**E**	12	Objection to the Lines	**L**
6	Fires trains	**F**	13	Exceeded Station Pavements	**M**
7	The collision of the trains each other or about the collision	**G**	14	Direction Switching overload	**N**

Table (2): Data of the Train's Accidents in Egypt

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
2000	324	33	4	63	69	33	4	68	14	3	475	84	111	19
2001	226	26	5	74	70	31	1	57	14	0	441	120	35	11
2002	186	22	2	62	64	47	6	45	15	2	392	102	25	14
2003	1057	20	2	52	68	42	3	62	10	5	342	93	23	13
2004	200	28	2	63	83	34	4	38	8	1	405	91	15	6
2005	227	23	5	66	74	33	3	76	15	1	409	74	120	19
2006	262	16	2	36	72	126	5	45	14	1	433	70	23	10
2007	245	12	3	66	79	200	2	60	24	0	426	80	27	16
2008	227	25	0	51	72	198	4	47	45	0	450	127	36	22

Depend on the correlation matrix, we can condensate some of these variables.
Table (3): The condensate variables of the Train's Accidents in Egypt

No.	Case	Code
1	Falling to for non stores	**A**
2	Vehicles	**B**
3	Fires trains	**F**
4	These Parathion of trains on all lines	**H**
5	Open doors of the trains, Violators of the work of railway, Direction Switching overload, and Objection to the Lines	**CJNL**
6	The collision of rail Vehicles at the outlets, On the non outlets, The collision of the trains each other or about the collision, and The collision of rail gate outlets.	**DEGK**
7	Exceeded Station Pavements, and Exceeded Lights all types	**IM**

(III)-Results (Parameter Estimation).

<u>1- Factor Analysis, Using Principal Components</u>

To investigate what "factors" might explain most of the variability. Use the Principal components extraction method and examine an eigenvalues plot in order to decide upon the number of factors.

a) - *MINITAB (v.15)* results:

Rotated Factor Loadings and Communalities (Varimax Rotation)

Variable	Factor1	Factor2	Factor3	Communality
CJNL	0.215	0.215	-0.558	0.405
DEGK	0.887	-0.259	-0.211	0.898
IM	0.333	-0.879	-0.182	0.916
A	-0.879	-0.078	0.042	0.781
B	0.153	-0.205	-0.931	0.932
F	0.486	0.258	0.652	0.728
H	-0.161	-0.944	0.053	0.921
Variance	2.0024	1.8926	1.6851	5.5802
% Var	0.286	0.270	0.241	0.797

From the rotated Factor matrix, there are 3 eigenvalues > 1.Then according to Kaiser's criterion there are three main factors represent 79.7% of the variability.
The following matrix represents the degree of correlation between all variables and all factors. It is named the `Sorted Rotated Factor Loadings and Communalities` matrix, where Sorting is done by the maximum absolute loading for any factor. Variables that have their highest absolute loading on factor 1 are printed first, in sorted order. Variables with their highest absolute loadings on factor 2 are printed next, in sorted order, and so on. We will eliminate the Factor Loadings < absolute 0.5.

Sorted Rotated Factor Loadings and Communalities

Variable	Factor1	Factor2	Factor3	Communality
DEGK	0.887	0.000	0.000	0.898
A	-0.879	0.000	0.000	0.781
H	0.000	-0.944	0.000	0.921
IM	0.000	-0.879	0.000	0.916
B	0.000	0.000	-0.931	0.932
F	0.000	0.000	0.652	0.728
CJNL	0.000	0.000	-0.558	0.405

```
Variance   2.0024   1.8926   1.6851       5.5802
% Var       0.286    0.270    0.241        0.797
```

We can definition the factor1 using the variables A, DEGK because it have large Factor Loadings with the factor1, which we can named it the Collision and the Falling factor, also, we can definition the factor2 using the variables H, IM and named it the Parathion and the Exceeded factor, and definition the factor3 using the variables B, F, CJNL and named it the Fires and the Human errors factor

Finally from the Factor Score Coefficients matrix, the 3 factors can be written as follow:

$$F_1 = 0.08cjnl + 0.43\deg k + 0.132im - 0.448A + 0.011B + 0.298F - 0.107H$$
$$F_2 = 0.198cjnl - 0.091\deg k - 0.457im - 0.075A + 0.009B + 0.075F - 0.539H$$
$$F_3 = -0.367cjnl - 0.043\deg k + 0.019im - 0.021A - 0.553B + 0.411F - 0.144H$$

Factor Score Coefficients

Variable	Factor1	Factor2	Factor3
CJNL	0.080	0.198	-0.367
DEGK	0.430	-0.091	-0.043
IM	0.132	-0.457	0.019
A	-0.448	-0.075	-0.021
B	0.011	0.009	-0.553
F	0.298	0.075	0.411
H	-0.107	-0.539	0.144

b) SPSS *(v.16)*, Results

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.766	39.514	39.514	2.766	39.514	39.514	2.200	31.431	31.431
2	1.582	22.596	62.110	1.582	22.596	62.110	1.737	24.819	56.251
3	1.198	17.118	79.228	1.198	17.118	79.228	1.608	22.977	79.228
4	.707	10.106	89.334						
5	.448	6.401	95.735						
6	.272	3.888	99.623						
7	2.640E-02	.377	100.000						

Extraction Method: Principal Component Analysis.

Rotated Component Matrix [a]

	Component		
	1	2	3
A			.608
B	.657		
CJNL		.839	
DEGK			.827
F	-.889		
H	.785		
IM		.786	

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 10 iterations.

Component Score Coefficient Matrix

	Component		
	1	2	3
A	.246	-.287	.413
B	.238	.126	.233
CJNL	.027	.545	-.289
DEGK	-.123	-.031	.543
F	-.464	.164	.149
H	.328	.133	-.044
IM	-.206	.470	.203

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

There are three main factors represent 79.228% of the variability. From the Rotated Components Matrix, we can definition the factor1 using the variables B, F, H and we can named it the Fires and Exceeded, also, we can definition the factor2 using the variables CJNL, IM and named it the Human errors factor, and definition the factor3 using the variables A, DEGK and named it the Collision and Falling factor .

Finally from the component Score Coefficients matrix, the 3 factors can be written as follow:

$$F_1 = 0.027cjnl - 0.123\deg k - 0.206im + 0.246A + 0.238B - 0.464F + 0.328H$$

$$F_2 = 0.545cjnl - 0.031\deg k + 0.470im - 0.287A + 0.126B + 0.164F + 0.133H$$

$$F_3 = -0.289cjnl + 0.543\deg k + 0.203im + 0.413A + 0.233B + 0.149F - 0.044H$$

c) ***Factor Results:***

variable	1	2	3	4	5	6	7
code	A	B	F	H	DEGK	CJNL	IM

ADEQUACY OF THE CORRELATION MATRIX
Determinant of the matrix = 0.001521920902680
(Bartlett's statistic = 31.4 (df = 21; P = 0.071641)
Kaiser-Meyer-Olkin (KMO) test = 0.13448 (unacceptable)

DETAILS OF ANALYSIS--

Participants' scores data file : E:\ Train's Accidents. Data

Number of participants : 9

Number of variables : 7

Variables included in the analysis : ALL

Number of components : 2

Procedure for determining the number of dimensions : Parallel Analysis using marginally bootstrapped samples (PA - MBS)

Dispersion matrix: Pearson Correlations

Method for components extraction: Principal Components Analysis

Rotation to achieve factor simplicity : Raw Varimax

Clever rotation start : Weighted Varimax

Maximum number of iterations : 100

EXPLAINED VARIANCE BASED ON EIGENVALUES

Variable	Eigenvalue	Proportion of Variance	Cumulative Proportion of Variance
1	2.39583	0.34226	0.34226
2	1.82293	0.26042	0.60268
3	1.36140	0.19449	
4	0.97627	0.13947	
5	0.25308	0.03615	
6	0.18489	0.02641	
7	0.00560	0.00080	

ROTATED LOADING MATRIX (loadings lower than absolute 0.500 omitted)

variable	A	B	F	H	DEGK	CGNL	IM
C1	0.871				0.911		
C2		0.665	-0.625	0.770			0.791

The Factor Package reduced the factors to only two factors represent 60.268% of the variability, with eliminated the Human errors factor, this is illogic.

From the Rotated Loading Matrix, we can definition the factor1 using the variables A, DEGK and we can named it <u>the Collision and Falling factor</u>, also, we can definition the factor2 using the variables B, F, H, IM and named it <u>the Fires, the Parathion and the Exceeded factor</u>.

DISTRIBUTION OF RESIDUALS

Summary Statistics for Standardized Residuals

Smallest Standardized Residual = -1.11

Median Standardized Residual = 0.06

Largest Standardized Residual = 0.88

Mean Standardized Residual = 0.04

Root Mean Square of Residuals (RMSR) = 0.1836

Expected mean value of RMSR for an acceptable model = 0.3536 (Kelly's criterion)

<u>2- Factor Analysis, Using Maximum Likelihood and a Rotation</u>

We perform a maximum likelihood extraction, varimax rotation, and interpret the factors.

a) MINITAB results

Sorted Rotated Factor Loadings and Communalities

Variable	Factor1	Factor2	Factor3	Communality
DEGK	0.936	0.000	0.000	1.000
A	-0.812	0.000	0.000	0.676
H	0.000	0.982	0.000	1.000
IM	0.000	0.817	0.000	0.774
B	0.000	0.000	0.972	1.000
F	0.000	0.000	-0.626	0.611
CJNL	0.000	0.000	0.000	0.167
Variance	1.8234	1.8030	1.6015	5.2279
% Var	0.260	0.258	0.229	0.747

From the sorted rotated Factor matrix (using Varimax Rotation).There are 3 factors (the same factors using PC method) represent 74.7% of the variability, after eliminating the eigenvalues < one and the Factor Loadings < absolute 0.5.

Factor Score Coefficients

Variable	Factor1	Factor2	Factor3
CJNL	-0.000	0.000	-0.000
IM	0.040	-0.250	0.057
DEGK	0.072	-0.448	0.102
H	0.965	0.264	0.099
F	0.011	-0.066	0.015
B	0.042	0.092	-1.045
A	-0.071	0.441	-0.101

From the Factor Score Coefficients matrix, the 3 factors can be written as follow:

$$F_1 = -0.000cjnl + 0.072\deg k + 0.040im \_ 0.071A + 0.042B + 0.011F - 0.965H$$
$$F_2 = 0.000cjnl - 0.448\deg k - 0.250im + 0.441A + 0.092B - 0.066F + 0.264H$$
$$F_3 = -0.000cjnl - 0.102\deg k - 0.057im - 0.101A - 1.045B + 0.015F + 0.099H$$

Because the Coefficients of the cjnl variable equal zero for the three previous factors, we will choose 2 factors only.

Factor Analysis: CJNL; IM; DEGK; **H; F; B; A**

Maximum Likelihood Factor Analysis of the Correlation Matrix

Factor Score Coefficients

Variable	Factor1	Factor2
CJNL	0.016	0.003
IM	0.199	0.036
DEGK	0.604	0.110
H	-0.312	0.960
F	0.018	0.003
B	0.044	0.008
A	-0.269	-0.049

Then:
$$F_1 = 0.016cjnl + 0.604\deg k + 0.199im - 0.269A + 0.044B + 0.018F - 0.312H$$
$$F_2 = 0.003cjnl + 0.110\deg k + 0.036im - 0.049A + 0.008B - 0.003F + 0.960H$$

Here, two factors represent only 52.3% of the variability. The factor1 using the variables A, DEGK (<u>the Collision and Falling factor</u>), and the factor2 using the variables F, H, IM (<u>the Fires, the Parathion and the Exceeded factor</u>). With eliminated the Human errors factor, this is illogic.

b) SPSS results

Total Variance Explained

Factor	Initial Eigenvalues			traction Sums of Squared Loadir			otation Sums of Squared Loadin		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.396	34.226	34.226	1.706	24.370	24.370	1.876	26.800	26.800
2	1.823	26.042	60.268	1.952	27.889	52.259	1.782	25.458	52.259
3	1.361	19.449	79.717						
4	.976	13.947	93.663						
5	.253	3.615	97.279						
6	.185	2.641	99.920						
7	603E-03	.005E-02	100.000						

Extraction Method: Maximum Likelihood.

There are two factors represent only 52.259% of the variability. The factor1 using the variables A, DEGK (the Collision and Falling factor), and the factor2 using the variables H, IM (the Parathion and the Exceeded factor), and also with eliminated the Human errors factor, this is illogic.

Factor Score Coefficient Matrix

	Factor	
	1	2
A	-.264	-.047
B	.043	.008
F	.018	.003
CJNL	.016	.003
DEGK	.612	.112
IM	.193	.038
H	-.310	.957

Extraction Method: Maximum Likelihood.
Rotation Method: Varimax with Kaiser Normalization.

$$F_1 = 0.016cjnl + 0.612\deg k + 0.193im - 0.264A + 0.043B + 0.018F - 0.310H$$

$$F_2 = 0.003cjnl + 0.112\deg k - 0.038im - 0.047A + 0.008B - 0.003F + 0.957H$$

c) Factor results

Unrestricted Factor Analysis

Method for components extraction: Exploratory Maximum Likelihood.

EIGENVALUES OF THE REDUCED CORRELATION MATRIX

Variable	1	2	3	4	5	6	7
Eigenvalue	2.1209	1.3895	0.9809	0.1649	-0.0173	-0.3619	-0.4558

EXPLAINED VARIANCE BASED ON EIGENVALUES

Variable	Eigenvalue	Proportion of Variance	Cumulative Proportion of Variance
1	2.39583	0.34226	0.34226
2	1.82293	0.26042	0.60268
3	1.36140	0.19449	
4	0.97627	0.13947	
5	0.25308	0.03615	
6	0.18489	0.02641	
7	0.00560	0.00080	

ROTATED LOADING MATRIX (loadings lower than absolute 0.500 omitted)

variable	V1	V2	V3	V4	V5	V6	V7
F1	-0.544	0.897			0.695		
F2				0.990			0.812

Here, two factors represent only 60.8% of the variability, also with eliminated the Human errors factor. So, we will reanalysis with eliminated loadings lower than absolute 0.3 rather than 0.5. The results as the following rotated loading matrix shows that there are two factors, the factor1 using the variables B, A, DEGK, CJNL, IM (the Collision, Human errors, and Falling factor), and the factor2 using the variables H, IM (the Parathion and the Exceeded factor).

ROTATED LOADING MATRIX (loadings lower than absolute 0.300 omitted)

variable	V1	V2	V3	V4	V5	V6	V7
F1	-0.544	0.897	-0.352		0.695	0.348	0.431
F2				0.990			0.812

EXPLAINED VARIANCE AND RELIABILITY OF ROTATED FACTORS

Factor	Variance	Reliability estimate
1	2.021	0.887
2	1.800	0.986

DISTRIBUTION OF RESIDUALS

Summary Statistics for Standardized Residuals

Smallest Standardized Residual = -1.07

Median Standardized Residual = -0.04

Largest Standardized Residual = 1.33

Mean Standardized Residual = 0.01

Root Mean Square of Residuals (RMSR) = 0.2055

GOODNESS OF FIT STATISTICS

Root Mean Square Error of Approximation (RMSEA) = 0.36

Test of Approximate Fit : H0: RMSEA < 0.05; P = 0.037

(Chi-Square with 8 degrees of freedom = 16.299 (P =0.039916)

Chi-Square for independence model with 21 degrees of freedom = 31.358

Note that: the residuals of Maximum Likelihood have an approximation normal distribution if it is compared with the residuals of Principal Components

(V) – Conclusions

The goal of factor analysis is to find a small number of factors, or unobservable variables that explains most of the data variability and yet makes contextual sense. We need to decide how many factors to use, and find loadings that make the most sense for our data. In this paper seven factors describe the data perfectly, but the goal is to reduce the number of factors needed to explain the variability in the data. Examine the eigenvalues and loadings factor plot.

Using Principal Components method with MINITAB and SPSS, there are three factors together represent 79.7%, 79.228% of the variability respectively. (the Collision and the Falling factor (A, DEGK), the Parathion and the Exceeded factor (H, IM), and the Fires and the Human errors (B, F, and CJNL)). But with FACTOR package, there are two factors only together represent 60.268% of the variability (the Collision and Falling factor (A, DEGK), and the Fires, the Parathion and the Exceeded factor (B, F, H, and IM).

Using Maximum Likelihood method, the results indicate that this is a Heywood case. Using Maximum Likelihood method with MINITAB, there are three factors (the same factors using PC method) together represent 74.7% of the variability, after eliminating the eigenvalues less than one and the Factor Loadings less than absolute 0.5. But with SPSS and FACTOR package, there are two factors only together represent 52.259%, 60.268% of the variability respectively (the Collision and Falling factor (A, DEGK) and the Fires, the Parathion and the Exceeded factor (B, F, H, and IM).

From the previous Conclusions, then the best estimation method is the Principal Components with MINITAB package.

References

[1] B. Muthin, and D. Kaplan. A comparison of some methodologies of the factor analysis of non-normal Likert variables. British Journal of Mathematical and Statistical Psychology, (1992), 45.

[2] C. Chatfield and A. J. Collins, Introduction to Multivariate Analysis. Chapman and Hall. London. (1980).

[3] D.E. Johnson, Applied Multivariate Methods for Data Analysis Pacific Grove, CA: Brooks/Cole Publishing. (1998).

[4]- H. Harmon, Modern Factor Analysis, Third Edition, University of Chicago Press. (1976).

[5] J.F. Jr. Hair, R.E. Anderson, R.L. Tatham, and W.C. Black, Multivariate Data Analysis, (5th Edition). Upper Saddle River, Prentice Hall. N.J. (1998).

[6] J. Stevens, Applied Multivariate Statistics for the Social Sciences "4th Edition. Mahwah, Lawrence Erlbaum Associates. N.J. (2002).

[7] K.V. Mardia, J.T. Kent, and J.M. Bibby, Multivariate Analysis. Academic Press. New York: (1979).

[8] R.A. Johnson, and D.W. Wichern, Applied Multivariate Statistical Analyses (5th Edition). Upper Saddle River, NJ: Prentice Hall. (2002).

[9] S.E. Fienberg, The Analysis of Cross-Classified Categorical Data. The MIT Press. (1987).

[10] S. Sharma, Applied Multivariate Techniques. United States: John Wiley & Sons. (1996).

[11] T.W. Anderson, An Introduction to Multivariate Statistical Analysis, Second Edition, John Wiley & Sons. (1984).

[12] W. Dillon, and M. Goldstein, Multivariate Analysis, Methods and Applications, John Wiley & Sons. (1984).

Received: November, 2009