**FACTOR ANALYSIS**

Factor Analysis

It is a dimension reduction technique.

Definition: A class of procedures preliminarily used for data reduction and summarization.

In marketing research, there may be a large number of variables, most of which are correlated and which must be reduced to manageable level. Relationship among sets of many interrelated variables are examined and represented in terms of a few underlying factors.

In ANOVA, multiple regression and discriminant analysis, one variable is considered as dependent or criterion variable, and the other as independent or predictor variable. However, no such distinction is made in factor analysis. Rather, factor analysis is am interdependence technique, in that an entire set of interdependent relarionship is examined.

Interdependence Technique: Multivariate statistical technique in which the whole set of interdependent relationship is examined.

Factor: An underlying dimension that explains the correlations among a set of variable.

Applications of Factor Analysis in Marketing Research:

1. It can be used in marketing segments for identifying the underlying variables on which to group the customers. New car buyers might be grouped based on the relative emphasis they place on economy, convenience, performance, comfort and luxury. This might result in five segments:
   * Economy seekers
   * Convenience seekers
   * Performance seekers
   * Comfort seekers
   * Luxury seekers.
2. In product research, a factor analysis can be employed to determine the brand attributes that influence the customer choice. Toothpaste brand might be evaluated in terms of protection against cavities, whiteness of teeth, taste, fresh breath and price.
3. In advertising studies, factor analysis can be used to understand the media consumption habits of the target market. The users of the frozen food may be heavy viewers of cable TV, see lots of movies and listen to country music.
4. in pricing studies, it can be used to identify the characteristic of price sensitive customers. E.g. these customers might be methodical, economy minded, and home centered.

Factor Analysis Model:

The unique factors are uncorrelated with each other and with the common factors. The common factors themselves can be expressed as k



Fi = Wi1X1+Wi2X2+……+WikXk. where ∑Wij =1.

And  J=1

Fi : estimate of ith factor;

Wi ; weight or factor score co-efficient ;

k : number of variables.

It is possible to select weights or factor score co-efficient so that the first factor explains the largest portion of the total variance. Then a second set of weights can be selected, so that the second factor accounts for most of the residual variance, subject to being uncorrelated with the first factor. The same principle could be applied to selecting additional weights for the additional factors.

Thus, the factors can be estimated so that their factor scores, unlike the values of original variables, are not correlated.

Max. no. of factors = Total no. of variables under study.

**Definitions:**

* correlation matrix: A correlation matrix is a lower triangular matrix showing the simple correlations r, between all possible pairs of variables included in the analysis. The diagonal elements, which are 1 are usually omitted.
* Communality: it is the amount of variance, a variable shares with all other variables being considered. This is also the proportion of variance explained by the common factors.
* Eigen value: The eigen value represents the total variance explained by each factor.
* Factor loadings: Factor loadings are simple correlations between the variables and the factors.
* Factor loading plot: A factor loading plot is a plot of the original variables using the factor loadings as co-ordinates.
* Factor matrix: A factor matrix contains the factor loadings of all the variables on the entire factor exacted.
* Factor scores: Factor scored are composite scores estimated for each respondent on the derived factors.
* Percentage variance: It is a percentage of total variance attributed to each variable.
* Residuals: Residuals are the difference between observed correlations as given in input correlation matrix and reproduced correlations as estimated from factor matrix.
* Scree plot: A scree plot is a plot of eigen values against the number of factors in order of extraction.

Stochastic associated with Factor Analysis:

* Bartlett’s test of Sphericity:

It is a test statistics used to examine the hypothesis that the variables are uncorrelated in the population. In other words, the population correlation matrix is an identity matrix.

* Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy:

The KMO measure of sampling adequacy is an index to examine the appropriateness of factor analysis. High values (0.5 to 1) indicate factor analysis is appropriate. Values below 0.5 imply that factor analysis is not appropriate.

**Summary:**

To check appropriateness of factor analysis

* + Data should be (continuous) scale.
  + The number of observations must be greater than 4 to 5 times of the number of variables.
  + The value of KMO must be greater than or equal to 0.5.
  + There should be high correlation among the set of variables.
  + The Bartlett’s test of Sphericity must be significant.

(i.e. H0: R=I should be rejected).

**Factor Analysis if Data is Given:**

Data file= Purchase of toothpaste

The researcher wants to determine the underlying benefits, consumers seeks from the purchase of the toothpaste. A sample of 30 respondents was interviewed using moll intercept interviewing. The respondents were asked to indicate their degree of agreement with the following statements using a 7point scale (1 : Strongly disagree, 7 : Strongly Agree) :

V1 : It is important to buy a toothpaste that prevents cavities.

V2 : I like a toothpaste that gives shiny teeth.

V3 : A toothpaste should strengthens your gums.

V4 : I prefer a toothpaste that frashens breath.

V5 : Prevention of tooth decay is not an important benefit offered by a toothpaste.

V6 : The most inmportant consideration in buying a toothpaste is attractive teeth.

**Interpretation:**

From correlation matrix (R) we can observe correlations between any 2 variables. This leads us to conclusion that variable-1 is highly correlated with the variable-3 i.e. a thoothpaste that prevents the cavities should also be able to strengthen the gums.

If we consider only one factor then p-value is 2.42e-05, which is significant. So, we reject the null hypothesis i.e. test of the hypothesis that 1 factor is sufficient. Now we go for 3 factor analysis.

Futher, we consider only three factors then p-value is 0.022, which is significant. So, we reject the null hypothesis i.e. test of the hypothesis i.e. test of the hypothesis that 2 factor is sufficient.

If we consider only TWO factor then p-value is 0.266, which is insignificant. So, we do not reject the null hypothesis i.e. test of the hypothesis that 2 factor is sufficient.

Further the eigen values correspond to factor 1 & factor 2 are 2.542 & 1.892 resp. which are more than 1 indicating that two factors are appropriate.

Further the PCA suggests that there are 2-principal components which play major part in explaining total variation. The total variation explained by them is:

0.604 + 0.0544 = 0.6584 i.e. 65% of the total variation can be explained by these 2-PCs only.

So if we are ready to lose the remaining information regarding variation of the data then by considering only 2-PCs we have reduced dimension of the data. The selection of PCs is done by observing the scree plot & eigenvalues which are greater than 1.

Also the factor loadings which are nothing but the simple correlations between the variables and the factors suggest the high importance of all variables except the 5th one. This means that the preferences towards the toothpaste which prevents the tooth decay is less as compare to rest of the other variables.

One could summariza data by stating that a consumer appears to seek two major kinds of benefits from toothpaste:

* + 1. Health benefit
    2. Social Benefit

Factors.

We can make factor 1, has high coefficient for variable 1 (Prevention of cavities), variable 3 (strong gum) & negative coefficient for variable 5( prevention of tooth decay is not important) as health benefit factor & remaining as factor 2 i.e., highly related with variable 2 (shiny teeth), variable 4 (fresh teeth), variable 6 (attractive teeth) may be labeled as social benefit factor. So we can minimize the dimensions of data.

We extracted two factors from data namely health benefit factor & social beefit factor. To study particular factor, we only look at the particular set of variables which represents these factors.