MARKING RESEARCH
&
ANALYTICS

Notes by Professor Dr. Mohammad Azam Khan,
Department of Statistics & Operations Research,
Aligarh Muslim University, AMU

MOHAMMAD WASIQ

Marketing Research & Analytics

DSM - 3003

Unit-I

Reducing Data Complexity : consumer brand rating data, principal component analysis and perceptual maps, exploratory factor analysis, multidimensional scaling. Linear modeling: handling highly correlated variables, linear models for binary outcomes: logistic regression, hierarchical models, Bayesian hierarchical linear models, quick comparison of the effects.

Unit II

Confirmatory Factor Analysis and Structural Equation Modelling : the motivation for structural models, scale assessment: confirmatory factor analysis (cfa), general models: structural equation models, the partial least squares (pls) alternative.

Unit III

Segmentation : Clustering and Classification : segmentation philosophy, segmentation data, clustering, classification, prediction: identifying potential customers. association rules for market basket analysis: the basics of association rules, retail transaction data: market baskets, finding and visualizing association rules, rules in non-transactional data: exploring segments again.

Unit IV

Choice Modelling : choice-based conjoint analysis surveys, simulating choice data, fitting a choice model, adding consumer heterogeneity to choice models, hierarchical bayes choice models, design of choice-based conjoint surveys. behavior sequences: web log data, basic event statistics, identifying sequences (sessions), markov chains for behavior transitions.

DSM-3003 Marketing Research & Analytics

Mohammad Wasiq

Table of Contents

This course is taught by **Dr. Mohammad Azam Khan**

1 Syllabus

1.0.1 Unit — I

Reducing Data Complexity: consumer brand rating data, principal component analysis and perceptual maps, exploratory factor analysis, multidimensional scaling. Linear modeling: handling highly correlated variables, linear models for binary outcomes: logistic regression, hierarchical models, Bayesian hierarchical linear models, quick comparison of the effects.

1.0.2 Unit — II

Confirmatory Factor Analysis and Structural Equation Modeling: the motivation for structural models, scale assessment: confirmatory factor analysis (CFA, general models: structural equation models, the partial least squares (pls) alternative.

1.0.3 Unit — III

Segmentation: clustering and classification: segmentation philosophy, segmentation data, clustering, classification, prediction: identifying potential customers. association rules for market basket analysis: the basics of association rules, retail transaction data: market baskets, finding and visualizing association rules, rules in non-transactional data: exploring segments again.

1.0.4 Unit — IV

Choice Modeling: choice-based conjoint analysis surveys, simulating choice data, fitting a choice model, adding consumer heterogeneity to choice models, hierarchical bayes choice models, design of choice-based conjoint surveys. behavior sequences: web log data, basic event statistics, identifying sequences (sessions), Markovchains for behavior transitions.

2 Unit - I

Reducing Data Complexity: consumer brand rating data, principal component analysis and perceptual maps, exploratory factor analysis, multidimensional scaling.

Linear modeling: handling highly correlated variables, linear models for binary outcomes: logistic regression, hierarchical models, Bayesian hierarchical linear models, quick comparison of the effects.

2.1 Principal Factor Analysis (PFA)

- **Factor Analysis involving both PCA and CFA** (It is a statistical approach that can be used to analyze inter-relationship among a large number of variables and to explain these variables in terms of their common underlying dimension (factor).)

- The objective is to find a way of conducting the information in several original variables into a small set of variates (factor) with a minimal loss of information.

- A researcher can use factor analysis :

For example to better a understand the relationship between customer's rating of a fast food restaurant. Suppose you ask customer to rate the restaurant on the following six variables :

```
-   Food taste
-   Food temperature
-   Freshness
-   Waiting time
-   Cleanliness
-   Behavior of employees
```

The analyst would like to combine these six variables into a smaller number.

By analysis would customer responses the analysis might find that the variables *Food taste, Food temperature* and *Freshness* combined together to form a single factor of *Food quality* & *Waiting time, Cleanliness* and *Behavior of employees* combine to form another single variables *Service quality*.

2.1.1 Common Factor Analysis VS Principal Component Analysis

1. PCA considers the total variance and derives the factors that contain small proportion of unique variables and in some instances error variance.

- CFA in contrast considers only the common or "Shared variance" assuming that both the unique and error variance are not of interest in defining the structure of variables.

`Shared variance` can mean several things, depending on what field you're in. In statistics, shared variance generally refers to covariance. However, the term is sometimes used in textbooks as part of the formal definition for the correlation coefficient (shared variance divided by combined variance).

`Error variance` is the component of variance in a distribution from the error variable, or influences other than what a scientist aims to manipulate. For instance, if one studies the influences of an independent variable (say sleep) on a dependent variable (say running speed) then one may have errors in measurements. There may also be extraneous variables influencing running speed, such as caffeine intake. The error variance measures the statistical spread of these sources other than the independent variable.

---

2. PCA is most appropriate when data reduction is a primary concern focusing on the minimum number of factors needed to account for the maximum proportion of the total variance represented in the original set of variables.

- Prior knowledge suggest that specific and error variance represent a relatively small proportion of the variance.

3. CFA is most appropriate when the primary object is to identify the latent dimensions or construct represented in the original variable.

- The researcher has little knowledge about the amount of the specific and error variance and therefore they wish to eliminate this variance.

2.1.2 Stopping Rule (Criteria for the number of factors to extract)

The following stopping rule for the number of factors to extract is currently being utilized :

2.1.2.1 Priori Criteria

A stopping rule for determine the number of factors. This rule is determined solely on the researcher judgement and experience. The researcher may know the desired structure or other conceptually based considerations so that the number of factors can be predetermined.

2.1.2.2 Latent Root Criteria

The most commonly used technique is Latent Root Criteria. Latent Root Criteria also known as **Kaisar Rule**.

- The eigenvalue of a single factor is simply the sum of the square loadings of the variables on that factors.
- The simple rule is do not retain any factor which account for less variance than a single variable.
- Thus only factors having the latent root or eigenvalue is greater than 1 are significant. All factors with latent roots less than 1 are consider insignificant and discarded.
- The rule is most applicable to PCA where the diagonal value representing the amount of variance for each value is 1.
- In CFA the diagonal value is replaced with the communality (explained variance).
- Latent Root Criteria is most reliable when the number of variables is between 20 and 50 and commonality above 0.4.

2.1.2.3 Parallel Analysis

A Stopping rule based on comparing the factor eigen values to a set of eigen values generated from a random data.

- The basic premise is to retain factor that has eigen value exceeding those which would be generated by random data.

2.1.2.4 Scree Plot

The scree test is used to identify the optimum number of factors that can be extracted before the amount of unique variance begins to dominate the common variance structure.

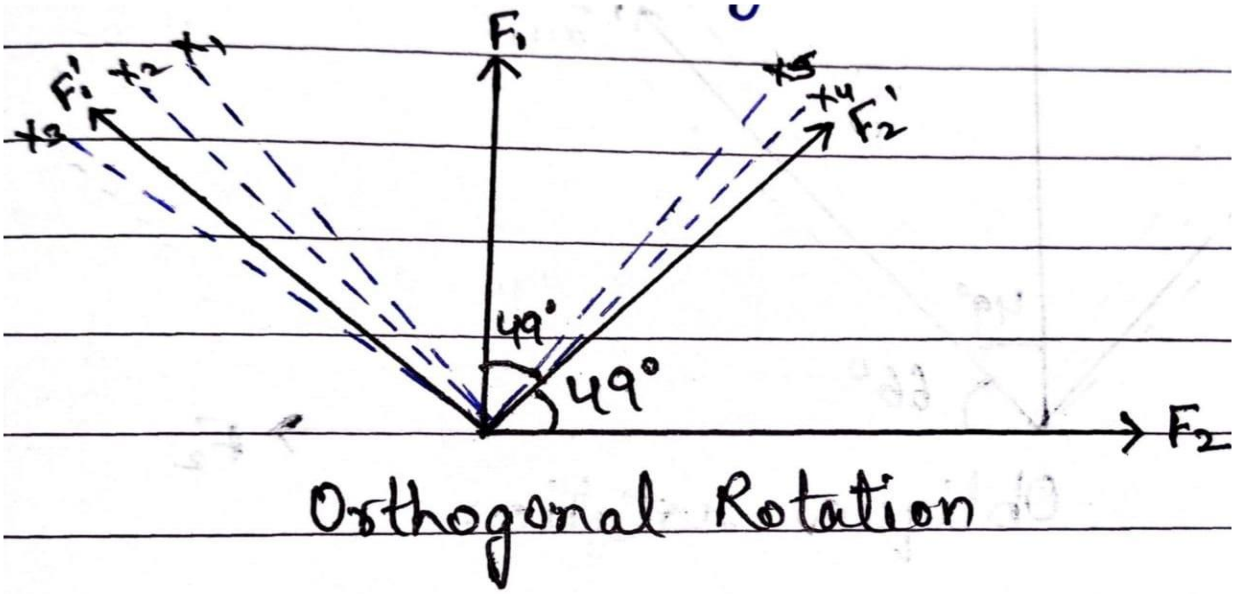
- The scree test is derived by plotting the latent root against the number of factors in their order of extraction and the shape of resulting curve is used to evaluate the cut off point.
- Starting with the first factor, the scree plot slops steeply downwards initially and then slowly and approximately horizontal line and inflection point termed by many as the elbow. Most of researcher do not include the elbow but rather retain all the preceding factors. Some researcher adequate including the elbow.

2.1.2.5 Percentage of Variance Criteria

A stopping rule for the numbers of vectors to retain which is the based on the amount of total variance accounted for in a set of vector or communality of each of the variable. The threshold value is specified by the researcher based on the object of the research and judgements about the quality of data being analysed.

2.1.3 Interpret the Factor Solution

We can make use of factor solution to facilitate the factor interpretation. We do not have to rotate the factor solution but it will facilitate interpretating the findings particularly if we have a reasonably large number of items (say 6 or more).



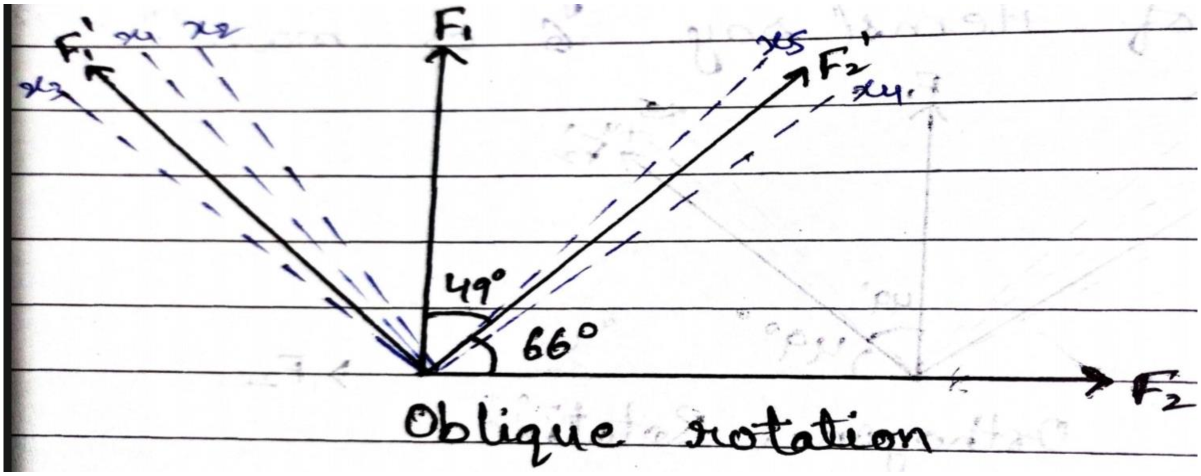
We see that both factors are orthogonally rotated 49^o that means 90^o is maintained between the factors during the rotation procedure. Consequently the factor remains uncorrelated. Which is inline with the PCA initial objective. By rotating the first factor F_1 to $F_1{}'$ is now strongly related to the variables x_1, x_2, x_3 but weakly related to x_4, x_5. Conversely by rotating the second factor F_2 to $F_2{}'$ is now strongly related to the variables x_4, x_5 but weakly related to x_1, x_2, x_3.

Various orthogonal rotation methods exists all of which differ with regard to their treatment of the loading structure.

The **Varimax** rotation is the best known and this procedure aims that maximizing the dispersion of loadings with factors.

Which means a few variables will high loadings while the remaining variables loadings will be considerably small.

We can choose several **oblique** rotation technique. In oblique rotation, the 90^o between the factor is not maintain during the rotation and the resulting factors are therefore correlated.



Promax rotation is commonly used oblique technique. The promax rotation allows for setting an exponent (reffered to as **Kappa** that need to be greater than 1. A kappa value of 3 works well for most application.)

Direct Oblique rotation allows specifing the maximum degree of obliqueness. This degree of delta, which determine the level of the correlation allowed between the factors.

2.1.4 Evaluating the Goodness of Fit of the Factor Solution

A way to check the solution goodness of fit by evaluating how much of each variable, variance of the factor reproduce. If several communality exhibits low values we should consider removing these variables. Considering the varible specific **MSA** measure could help us make this decision.

2.1.5 Compute the Factor Score

After the rotation and the interpretation of the factors we can compute the factors score. Factor score are linear combination of the items and can be used as separate variables in subsequent analysis.

The simplest way to compute factors scores for each observation is to sum all the scores of the items assigned to a factor.

2.1.6 Bartlett Test of Sphericity

A statistical test for the overall significance of all correlations within a correlation matrix.

2.1.7 Measure of Sample Adequacy (MSA)

Calculate both for the entire correlation matrix and each individual variable. MSA values above 0.5 for either the entire matrix or an individual variable indicate appropriateness for performing factor analysis on the overall set of variables.

2.2 Multi-Dimensional Scaling

Multi-dimensional scaling is also known as **perceptual mapping**.

Perceptual mapping is a visual representation of a respondent perception of an object of two or more dim

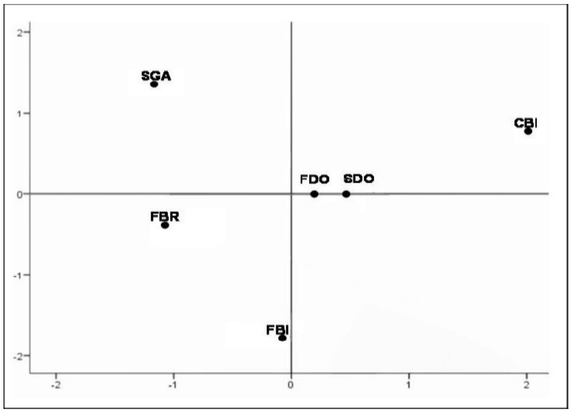
MDS is a procedure that enables a researcher to determine the percieved images of set of objects. The purpose of MDS is to transform consumer judgement over all similarity or preference(for store or brands) into distances represented in multi-dimensional scaling.

2.2.1 Comparing Objects :

MDS is based on the comparision of object. To perform MDS analysis, the researcher perform three basic steps :-

1. Gather major of similarity or preference across the entire set of object to be analyzed.
2. Use MDS technique to estimate the relative position of each object in MDS.
3. Identify and interpret the axis of the dimensional space in terms of perceptual and/or objective attributes.

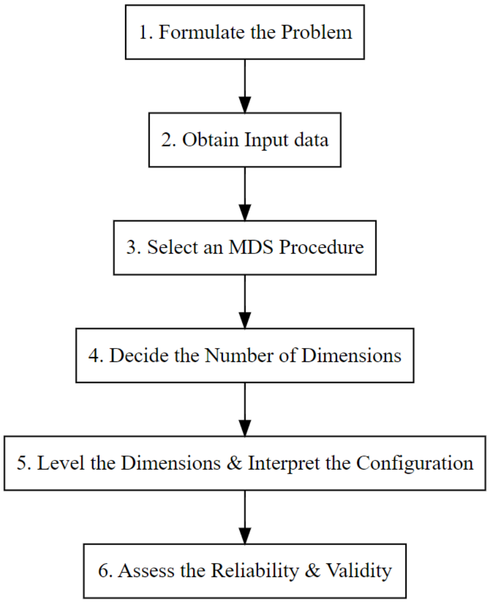
Example : Assume that object A and B judge by respondent to be most similar compared with all other possible pairs of objects.



(AC, BC, AD, and so on), MDS technique will position object AB, so that the distance between them in Multi-dimensional space is similar than the distance between any other two pairs of objects.

The resulting perceptual map is also known as spatial map shows the relative positioning of all the objects.

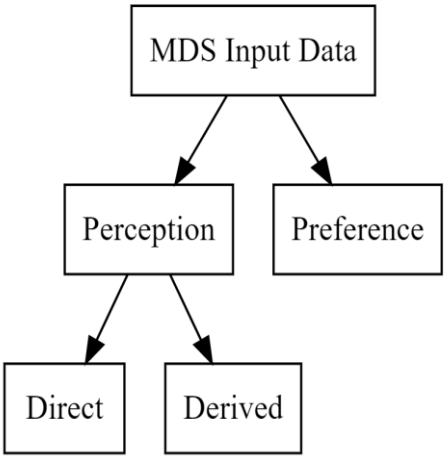
2.2.2 Conducting MDS Procedure


1. Formulate the Problem
2. Obtain Input data
3. Select an MDS Procedure
4. Decide the Number of Dimensions
5. Level the Dimensions & Interpret the Configuration
6. Assess the Reliability & Validity

1. **Formulate the Problem :** Formulating the problems requires that the researcher specify the purpose for which the MDS result would be used and select the brand or other stimuli to be included in the analysis at a minimum 8 brands or stimuli should be included to obtain a well defined spatial map. Including more than 25 brands is likely to be cumbersome. Suppose a researcher is interested in obtaining the consumers perception of automobiles.

If luxury automobiles are not included in the stimulus set. This dimension may not merge in the result. The choice of the numbers and specific brands or stimuli to be included should be based on the statement of the marketing research problem, theory and judgement of the researcher.

2. **Obtain Input Data :**



Input data obtain from the respondents may be related to the perceptions or preference.

2.1. **Perception Data :**

- 2.1.1 Perception Data (Direct Approach) : In direct approaches to gathering perception data respondents are expect to judge how similar or dissimilar. The various brands or stimuli are using there own criteria. Respondents are often required to rate all possible pairs of brands or stimuli in terms of similarity on **likert scale**. These data are referred to as a similarity judgement.
- 2.1.2 Perception Data (Derived Approach) : Derived approaches to collecting perception data are based upon attribute, requiring the respondents to rate the brands or stimuli on the identified attributes using `sematic differntial` or `likert scale`. The respondents are asked to evaluate their hypothetical ideal brand on the same set of attributes. If attributes rating are obtain a similarity measure such as `Eucledian Distance` is derived for each pair of brand.
- **Direct VS Derived Approaches :** Direct approaches has the advantage that the researcher does not have to identify a set of salient feature/attributes. Respondents makes similarity judgements using their criteria. The disadvantages are that the criteria are influced by the brands or stimuli being evaluated are in the same price range then price will not emerge as an important factor. The advantage of attribute based approach is that it is easy to identify respondents with homogeneous perception.

2.2. **Preference Data :** Preference data order the brands or stimuli in terms of respondents preference for some property. A common way in which such data are obtained is preference ranking. Respondents are require to rate the brands from most prefered to the least prefered.

3 **Select on MDS Procedure :**

Selection of specific MDS procedure depends on weather perception or preference data are being scaled or weather the analysis requires both kind of data.

- Non-metric MDS procedure assume that the input data are obtained but they result in metric output. The distances in the resulting spatial map may be assumed to be interval scale.
- In matric MDS method that the output is also metric.
- The metric and non-metric methods produce similar results.

4. **Decide on the Number of Dimensions :**

The objective in MDS is to obtain a spatial map that best fit the input data in the smallest number of dimension.

- **Stress :** The fit of an MDS solution is commonly assesst by the stress, major stress is a lack of fit. Higher values of stress indicates the poorer fits.

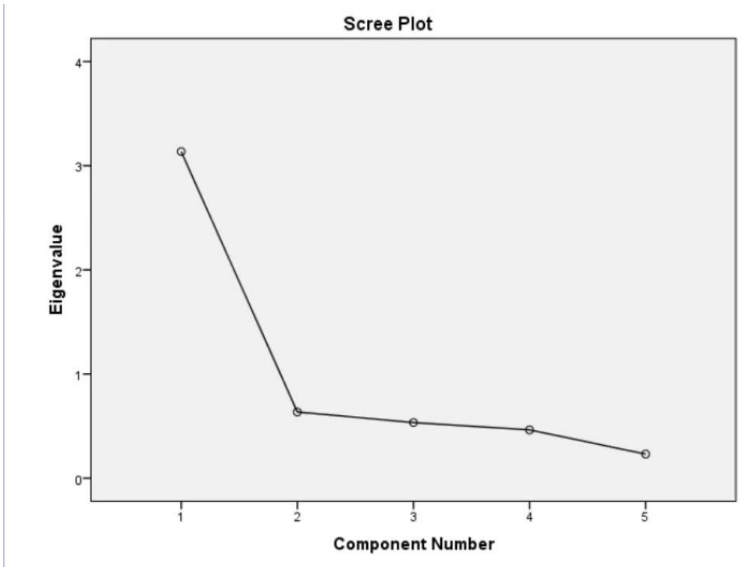
Stress (%)	Goodness of Fit
20	Poor
10	Fair
5	Good
2.5	Excellent
0	Perfect

- **Tucker's coefficient of Congruence Measures the Good Fit :** Higher the value of coefficient indicates the good fit.

- **Following Guidelines are suggested for determining the number of Dimension :**

i. **Prior Knowledge :** Theory of past research may suggest a particular number of distances.

ii. **Elbow Criteria :** A plot of `Stress` VS `Dimensionality` should be examine. The points in this usually form a convex pattern. The point at which an elbow or a sharp band occurs indicates an appropriate number of dimensions.



iii. **Ease of Use :** It is generally easier to work with two dimensional map or configuration than those involving more dimension.

5. **Level the Dimensions & Interpret the Configuration :**

Once a spatial map is developed the dimensions must be label and the configuration interpret labeling the dimensions requires subjective judgement on the part of the researcher.

i. Even if direct similarity judgement are obtained.

ii. Ratings of the brands on researcher supplied attributes may still be collected.

iii. After providing direct similarity or preference data the respondents may be ask to indicate the criteria they used in making their evaluation. These criteria may then be subjectively related to the spatial map to label the dimension.

iv. If possible the respondent can be shown their spatial map and asked to label the dimensions by excepting the configuration.

6. **Assess the Reliability & Validity :**

The input data and consequently the MDS solution are invariably subject to substantial random variablity. Hence it is necessary that some assessment may be made of the reliability and validity of MDS solution.

i. Stress values are indicative of the quality of MDS solution.

ii. Stress measure the badness of fit or the proportion of variance of the optimality scaled data taht is not accounted by the MDS model.

iii. The index of fit or R^2 is a measure of goodness of fit.

iv. If an aggregate label analysis has been done the original data should be split into two or more parts.

Aggregate Analysis : Approach to MDS in which a perceptual map is generated for a group of respondents evaluated of object.

v. MDS Analysis should be conducted separately on each part and the result compared.

2.2.3 Correspondence Analysis

Correspondence Analysis is an MDS technique for scaling the qualitative data in marketing research. The input data are in the form of contingency table indicating a qualitative association between rows and columns.

Correspondence Analysis scale the rows and columns in corresponding units so that each can be displayed graphically in the some low dimensional scale. These spatial maps provides insights into :

i. Similarities and differences within the rows categories with respect to the columns categories.

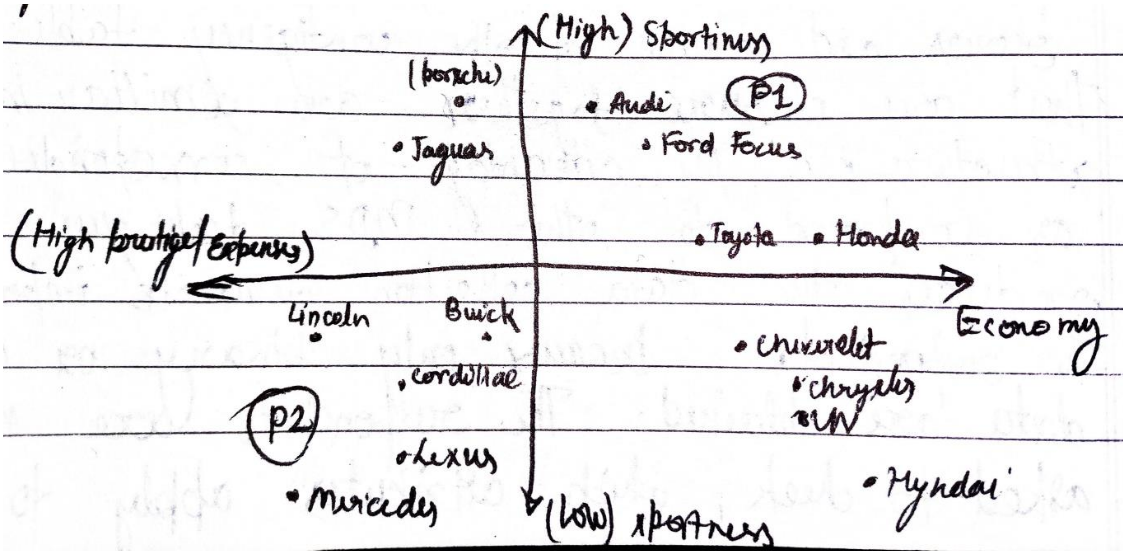
ii. Similarities and differences within the columns categories with respect to the rows categories.

iii. Relationship between rows and columns.

The interpretation of results in **Correspondence Analysis** is similar to that in **PCA**. The correspondence analysis results in the grouping of categories (activities, brands, etc.) found within the contingency table just as PCA involves the grouping of variables.

- The results are interpretation in terms of proxamities among the rows and columns of the contingency table. Categories that are closer together are similar in *structure*.
- The advantage of **CA** as compare to other MDS techniques that it reduces the data collection demand imposed on the respondence because only binary or categorical data are obtain.
- The respondent are marely ask to check which attribute apply to each of several brands. The input data are the number of yes responses for each brand on each attributes.
- The brands and the attributes are then displayed in the same multi-dimensional space.
- MDS including CA is not the only procedure for obtaining perceptual maps. Two other techniques **Discriminant Analysis** and **Factor Analysis** can also be used for this purpose.

Example : A study conducted in 2017 examine consumer perceptions of auto-mobiles by using MDS. Subject rated several auto-mobile attributes and the affect those attributes had on final product choice. Ratings were conducting using a 5 points scale and each subject responses were some across each dimension. The five highest scoring attributes overall were price, fuel, economy, net, horsepower, breaking and acceleration. The use of MDS can help auto-maker to better understand what attributes are most important to consumer.



Joint Space configuration of Automobile consumer performance

MDS map of selected auto-mobile brand drived from similarity data is shown.

In this spatial representation each brand is identify by its distance from the other brand. The closer two brands are for example— **VM** and **Chryeler** the more similar they are percieved to be. The further apart two brands are for example— **VM** and **Mercedes** the similar are percieved to be.

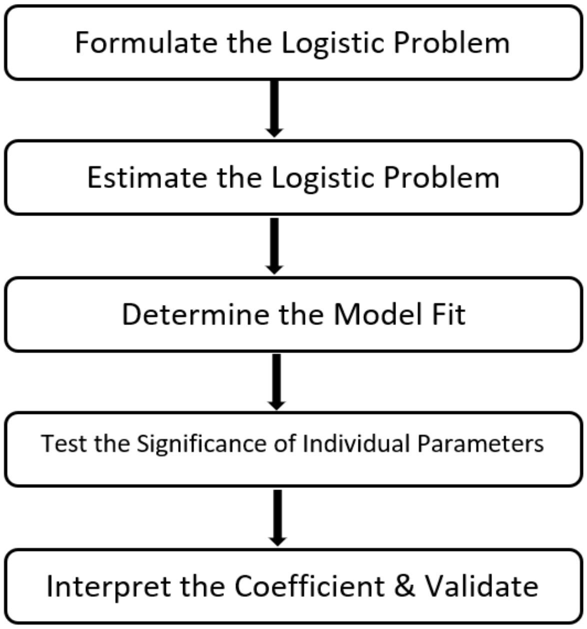
Similar distance that is similarity may also indicate competition. Two illustrate **Honda** competies closely with **Toyota** but not with **Mercedes** or **Poreche**. The dimensions can be interpretated as economy/prestigies and sportiness/non-sportiness. The preference consisted of simple rank order of the brands according to consumer preference. Respondent 1 denoted by $P1$ prefers the sporty cars **Poreche, Jaguar** and **Audi**. The respondent 2 denoted by $P2$ prefers luxury cars **Lincon, Mercedes, Cadillac** and **Lexus**.

Such analysis can be done at the individual respondent level inabiling the researcher to segment the market according to similarities in the respondents ideal points.

2.3 Logistic Regression

Logistic Regression is a specialised form of regression that is formulated to predict and explain a binary(two group/class) categorical variable rather than a metric dependent measure. Logistic Regression has the advantage of being less affected than disciminant analysis when the basic assumption underlying statistical inference, particularly the normality of the variable and the inherent heteroscedasticity introduced by binary dependent measure are not met. Logistic Regression also can accomodate non metric independent variables through dummy variable coding just as regression.

2.3.1 Conducting the Binary Logit Analysis


Formulate the Logistic Problem
Estimate the Logistic Problem
Determine the Model Fit
Test the Significance of Individual Parameters
Interpret the Coefficient & Validate

2.3.2 1. Formulate the Logit Problem

The binary logit model commonly deals with the issue of how likely on observations is to belong to each group. It estimates the probability of an observations belonging to a particular group. We can estimate the probability of a binary event taking place using the binary logit model also called **logistic model**.

Consider an event has two outcomes success and failure. The probability of success may be modeled using the logit model as

$$\boldsymbol{log\left(\frac{p}{1-p}\right) = a_0 + a_1X_1 + a_2X_2 + \cdots + a_kX_k = a_0 + \sum_{i=1}^{k} a_i\,X_i}$$

$$\boldsymbol{p = \frac{e^{a_0+\sum_{i=1}^{k} a_iX_i}}{1+\left(e^{a_0+\sum_{i=1}^{k} a_iX_i}\right)}}$$

where,

- p is the probability of success.
- X_i be the independent variable.
- a_i be the parameters to be estimated.

It can be seen from above equation that although X_i may very from $-\infty$ to $+\infty$, p is constraint to lie between 0 and 1. When X_i approaches to $-\infty$, p approaches to 0 and when X_i approaches to $+\infty$, p approaches to 1. It is desirable because p is the probability and lies between 0 and 1. When OLS regression is used p is not constraint to lie between 0 and 1. It is possible to obtain the estimated values of p are less than 0 and greater than 1.

2.3.3 2. Estimate the Logit Problem

In OLS regression the parameters are estimated to minimize the sum of squared error of prediction. The error term in regression can take on any values and are assume to follow normal distribution when conducting the statistical test. In binary logit model each error can assume only two values. If $Y = 0$ the error is p and if $Y = 1$ the error is $1 - p$. Therefore we would like to estimates the parameters in such a way would like to estimates the parameters in such a way that the estimated values of p would be close to $zero^{(0)}$, when $Y = 0$ and close to 1 when $Y = 1$. The procedure to achieve this and estimates the parameters of the binary model is called **Maximum Likelihood Estimator**.

2.3.4 3. Determine the Model Fit

In multiple regression the model fit is measure by the square of the multiple coefficient correlation R^2, which is also called the **Coefficient of Determination**. In Logistic Regression commonly used measure of model fit are based on the likelihood function and **Cox & Snell** R^2 and **Negelkar** R^2. Both measure are similar to R^2 in multiple linear regression. **Cox-Snell** R^2 is constant in such a way that it can not equal to 1 even if the model perfectly fits the data. The limitation is overcomed by the **Negelkar** R^2.

2.3.5 4. Test the Significance of Individual Parameters

The significance of the estimated coefficients is based on **Wald's Statistic**. This statistic is a test of significance of the logistic regression coefficient based on the asymptotic normality property of Maximum Likelihood Estimates and is estimated as

$$\boldsymbol{Wald} = \left(\frac{\boldsymbol{a_i}}{\boldsymbol{SE_{a_i}}}\right)^2$$

a_i is the logistic coefficient for that predictor variable and SE_{a_i} is standard error of coefficient.

The Wald's statistic is χ^2 distribution with 1 degree of freedom if the variable is metric and (number of categories - 1), if the variables are non-metric.

2.3.6 5. Interpretation of the Coefficients anf Validation

The interpretation of the coefficients as estimated parameters is similar to that in multiple regression taking into account that the nature of the dependent variable is different.

In logistic regression the **log odds** i.e. $\boldsymbol{log}\left(\frac{\boldsymbol{p}}{\boldsymbol{1-p}}\right)$ are a linear function of the estimated parameter. Thus if X_i is increased by one unit the log odds will increase by a_i units. When the effect of other other independent variable is constant. Thus a_i is the size of increase in the log odds of the dependent variable event, when the corresponding independent variable X_i increased by 1 unit and the effect of other independent variables held constant. The sign of a_i will determine weather the probability increases or decreases this amount.

The analysis sample is used for estimating the model coefficients. The validation sample is used for developing the classification matrix.

Example : The table gives the data for 30 respondents, 15 of whom are brand loyal indicated by 1 and 15 of whom are not indicated by 0. We also measure attitute towards the brands, attitute towards the product category and attitute towards shopping. All on $1(non - favourable)$ to $7(favourable)$ scale. The objective is to estimate the probability of a consumer being brand loyal as a function of attitute towards the brands, product and shopping.

3 Unit - II

Confirmatory Factor Analysis and Structural Equation Modeling : The motivation for structural models

Scale Assessment : confirmatory factor analysis (CFA)

General Models : structural equation models, the partial least squares (PLA) alternative.

3.1 Exploratory Factor Analysis (EFA)

- EFA explore the data and provides the researcher with information about how many factors are needed to best represent the data with EFA. All measured variables are related to every factor by a factor loading estimate.
- EFA is conducted without knowing how many factors really exists and which variable belongs to which constructs.
- When EFA is applied the researcher uses stablished guideline to determine which variable load on a particular factor and how many factors are appropriate. The factors that imerge can only be named after the factor analysis is performed.
- The term factors and construct are used interchangeable.

3.2 Confirmatory Factor Analysis (CFA)

With CFA the researcher must specify both the number of factors that exist for a set of variables and which factor, each variable will load on before result can be computed.

- CFA statistics tells us how well our theoretical specification of the factors matches the actual data.
- CFA is a tool that enables us to either confirm or reject the pre-concieved theory.
- CFA is used to provide a confirmatory test of our measurement theory.
- A measurement theory specified how measured variable logically and systematically represent constructs/factors involved in a theoretical model. So measurement theory requires that a factor first be define unlike **EFA** with **CFA** a researcher uses measurement theory to specify a priori the number of factors as well as which variables load on those factors.
- CFA cann't be conducted properly without a measurement theory.

3.2.1 A Simple Example of CFA and SEM

Consider a situation where a researcher is interested in studying factors that impact employee job satisfaction. After reviewing the relevent theory the researcher conclude that two factors have the largest impact supervisor support and work environment. The measured variables for both factors are evaluated using a seven point agree likert scale. The factor supervisor support can be defined as what workers think about the management capabilities of their immediate supervisor.

It can be represented by the following four items :

i. My supervisor recognizes my potential.
ii. My supervisor helps me resolve problem at work.
iii. My supervisor understands the challenges of balancing work and home demands.
iv. My supervisor supports me when I have a problem.

The construct work environment can be defined as the aspect of the environment where people work that impact their productivity. So it can be represented by the four major variables :

i. Supervisors and the workers has similar values and ideas.
ii. My organization provides the equipment needed to perform my job well.
iii. Temperature of my office and other working areas comfortable.
iv. The physical arrangement of work areas at my organization helps me to manage my time on the job well.

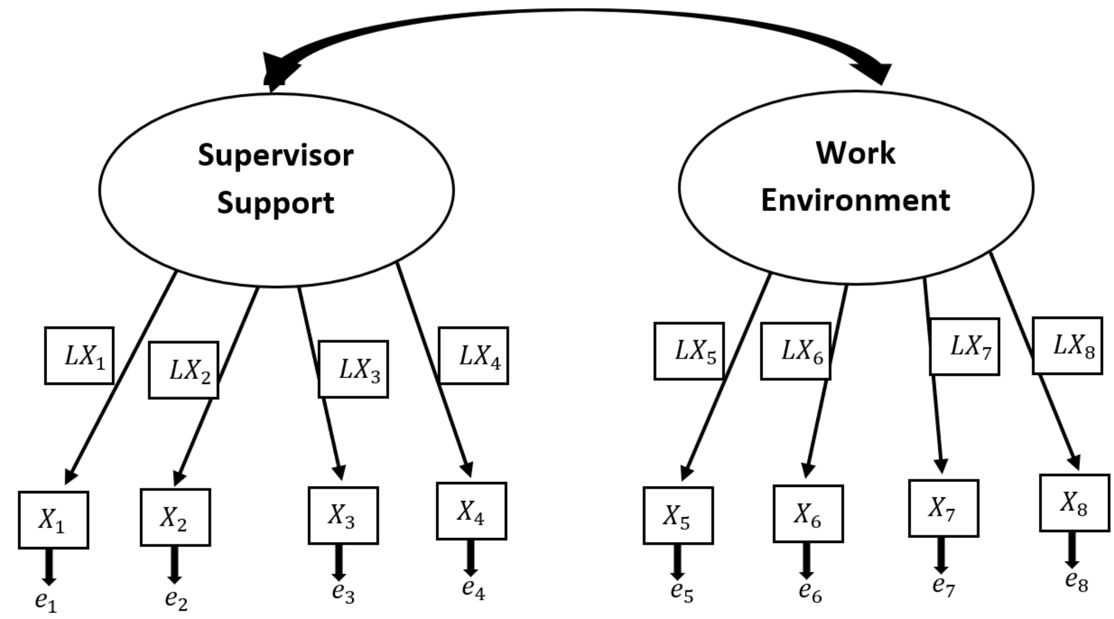
3.2.1.1 Visual Diagram

Measurement theories often are represented using visual diagram called **path diagram**.

The path diagram shows the linkage between specific measured variables and their associated constructs along with the relationship among constructs. **"Paths"** from the latent constructs to measured item are based on the measurement theory.

In **CFA** we must specify the five elements :

1. Latent Construct
2. Measured Variables
3. Items Loadings on the Specific Construct.
4. Relationship among Construct.
5. Error term for each indicator.



First latent constructs are drawn as ellipses and the measured variables are represented by rectangles.

Indicator variables are denoted by X i.e. $X_1, X_2, \dots, X_8$. The relationship between latent constructs respective measured variables (called factor loadings as in EFA) are represented by **arrows** from the construct to the measured variables.

- Final each measured indicator variable has an error term as shown in diagram represented by e.
- From the figure we can say that there are two latent constructs i.e. Supervisor Support and Work Environment. The X_1 — X_8 represent the measured indicator variables and LX_1 — LX_8 are the relationship between the latent constructs and respective measured items (i.e.Factor Loadings). The curved arrows between the two constructs denotes a correlationship between them.
- Finally e_1 — e_8 represent the error associated with measured items.

3.3 Structure Equation Modelling (SEM)

SEM is generally used to understand complex linkage among several independent and dependent variables.

- In SEM, the dependent variables are also called **endogenous** variables whereas independent variable are called **exogenous** variables.

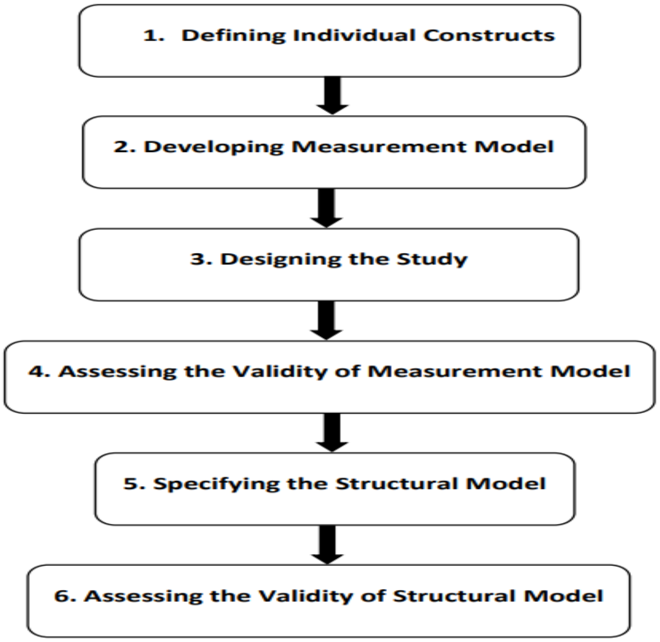
3.3.1 Assumption in SEM

The assumptions recommended for SEM primarily include the following :

i. Observations should be independent of each other.
ii. Relationship between variables should be linear.

iii. Respondents should be selected randomly.
iv. Data should follow a normal distribution.
v. Minimum sample size can be 15 – 20 data points for each independent variable in the regression.

3.3.2 Process of Structure Equation Modelling


1. Defining Individual Constructs
2. Developing Measurement Model
3. Designing the Study
4. Assessing the Validity of Measurement Model
5. Specifying the Structural Model
6. Assessing the Validity of Structural Model

3.3.3 1. Defining Individual Constructs

The constructs need to be theoreticaly defined and operationalized based on relevent literature.

3.3.4 2. Developing Measurement Model

In this step it is required to specify the measurement model based on the latent constructs and their observed indicators. Each latent needs to be identified with its indicator or observed variables.

3.3.5 3. Designing the Study with an appropriate sample size

A sample size within the range of 150 – 400 is best suited for SEM analysis.

Maximum likelihood is popularly used in SEM as its satisfies the Normality assumption and the input matrix is positive definite.

Considering advantage of covariance matrix over correlation matrix in terms of feasibility and better statistical properties. Generally covariance matrix is used as an input in SEM.

3.3.6 4. Checking the Validity of Measurement

The model describe relationship between latent variables and its indicator variables. The validity of the model is commonly evaluated using goodness of fit indicators.

Each of these indicators has a cutoff value which indicates on acceptable fit of the model.

The convergent validity is captured using average variance extracted **(AVE)**. AVE of value 0.5 and above specifies adequate convergence.

Discriminant validity assess based on the AVE and squared correlation among the constructs. When the AVE value from the two standardized constructs is more than the squared correlation between the construct then this shows discriminant validity.

Construct reliability basically checks uni-dimensionality using a composite reliability **CR** score which stablishes weather the set of measures for a given latent construct represent only one construct and has a cutoff of 0.70.

3.3.7 5. Specifying the Structural Model

The step specifying the relationship between the several construct based on the conceptual framework proposed in the study. Note that if a model successfully passes all the stages of step 4. Only then does it become eligible to be tested further using the structural model. In otherwords achieving a good field for the measurement model is a pre-requisite for the structural model.

3.3.8 6. Assessing the Validity of Structural Model

Here we check the model fit using goodness of fit indicator as mention in the below table :

Cutoff Values for fit Indices

Fit Indices	**Cutoff Values**
χ^2/degree of freedom	2 - 5
Root Mean Square Error Approximation (RMSEA)	<0.05
Goodness of Fit Indices (GFI)	>0.95
Standardized Root Mean Square Residual (SRMR)	<0.08

4 Unit - III

Segmentation : Clustering and classification: segmentation philosophy, segmentation data, clustering, classification, prediction: identifying potential customers. association rules for market basket analysis: the basics of association rules, retail transaction data: market baskets, finding and visualizing association rules, rules in non-transactional data: exploring segments again.

4.1 Market Segmentation

Market segmentation is one of the most fundamental strategy of market. To successfully match products and services to costumer needs. Companies have divided markets into groups(segments) of consumers. Costumers and clients with similar needs and wants. Firms can than target each of these segments by positioning themself in a unique segment. Market segmentating is essential for marketing success(the most successful forms segments their market carefully).

4.2 Cluster Analysis

Cluster Analysis is the method for segmentation and identifies homogeneous group of objects(or cases, observations). These objects can individual costumer, groups of costumers, companies or entire countries.

Objects in a certain cluster should be as similar as possible to each other but as distinct as possible from object in other clusters.

Suppose you are interested in segmentating costumers A to G then in order to better target them through for example pricing strategies. The first step is to decide on the characteristic that you will use segment your costumer A to G. In other words you have to decide which clustering variable will be included in the analysis.

For Example : You may want to segment the market based on price consciousness(X) and brand loyalty(Y). These two variables can measure on a scale 0 to 100 with higher values denoting a higher degree of price consciousness and brand loyalty.

The aim of cluster Analysis is to identify group of objects(costumers) that are very similar regarding their price consciousness and brand loyalty and then assign to cluster. After having decided on the clustering variable(price consciousness and brand loyalty). We need to decide on the clustering procedure to form of our group of objects.

These are many different approaches and little guidance on which one to use. The most popular approach in the marketing research including.

1. Hierarchical Method
2. Partioning Method specially K-Means
3. Two Step Clustering

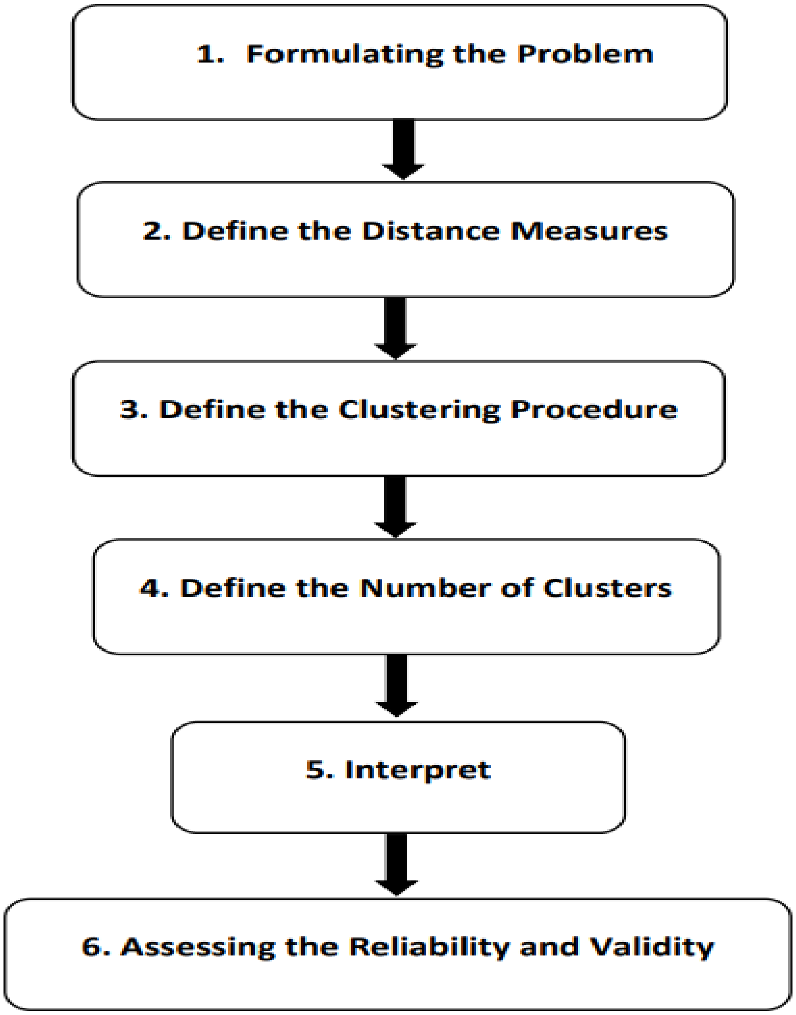
The basic of these procedure is the same namely grouping similar objects into clusters, they take different roots and important consideration before starting grouping is to determine how similarity should be measure. Most methods calculate measures of similarity or dis-similarity by estimating the distance between pairs of objects. Objects with smaller distance objects between one another are consider more similar whereas larger distances objects consider dis-similar.

Example : Megabus is a hugely successful line in US. They completely rethought the nature of their customers and concentration of three specific segments of the market

i. College Kids
ii. Women Travelling in Groups
iii. Active Seniors

To meet customers segments needs Megabus requires the entire driving experience by developing double decker buses with roof and big window and equiped with fast wifi. Megabus success of segmenting and targeting efferts has let to practitioners talk about the Megabus effect.

4.2.1 Conducting the Cluster Analysis


1. Formulating the Problem
2. Define the Distance Measures
3. Define the Clustering Procedure
4. Define the Number of Clusters
5. Interpret
6. Assessing the Reliability and Validity

4.2.2 1. Formulate the Problem

The most important part of formulating the problem is selecting the variables on which the clustering is based. Basically the set of variables selected should describe the similarity between objects in items that are the relevent to the marketing research. The variables should be selected on past reseach, theory or consideration of hypothesis being developed or tested.

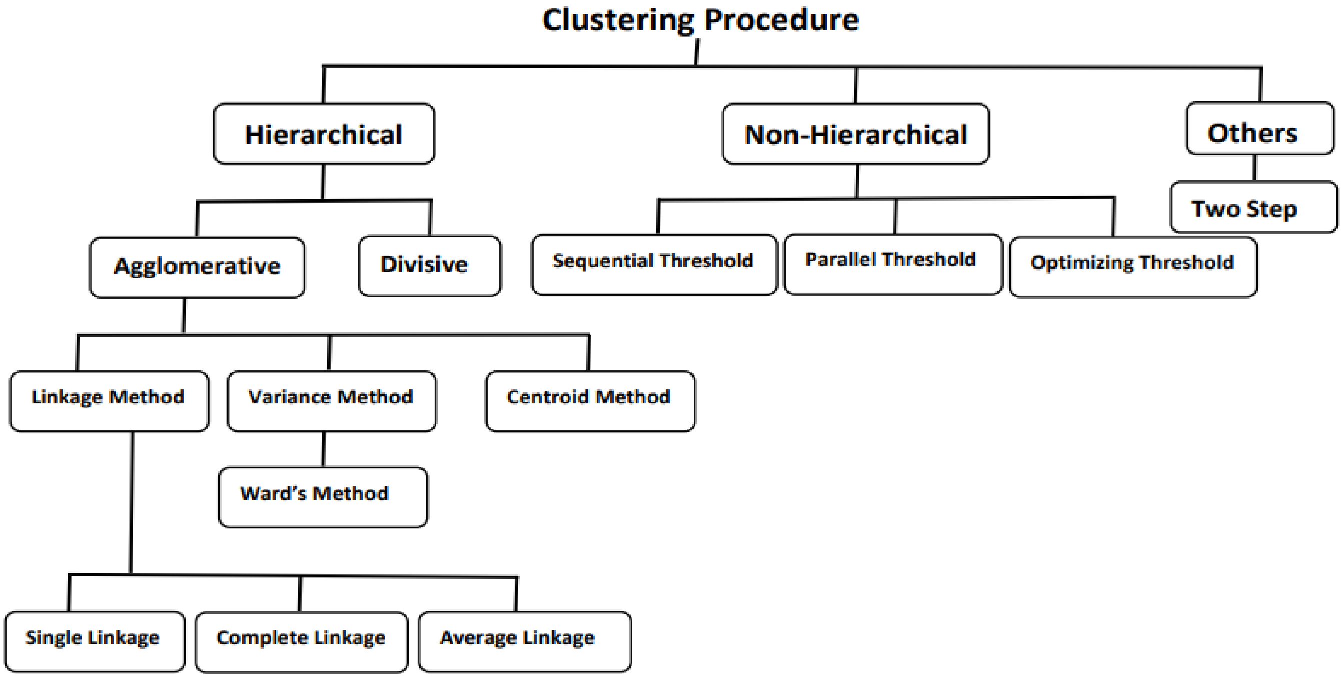
Example : We consider clustering of customers based on attitude towards shopping based on past research. Six(6) attitudinal variables where identified as being the most relevent to the market research problem. Consumers were asked to express their degree of argument with the following statement on a 7 point scale. (1- disagree, 7- agree).

- v_1 : Shopping is fun
- v_2 : Shopping is bad for your budget
- v_3 : I combine shopping with …..
- v_4 : I try to get the best buys while shopping
- v_5 : I don't care about shopping
- v_6 : You can save a lot of money by comparing prices

4.2.3 2. Select a Distance Measure

- Because the objective of clustering is to group similar objects together. Some measure is needed to assess how similar or different, the object are the most common approach is to measure similarity i terms of distance between pair of objects.
- Objects with smaller distances between them are more similar to each other then are those at larger distances.
- The most commonly used measure of similarity is the **Euclidean Distance** or **its Square**. The Euclidean Distance is the square root of the sum of the squared difference in values for each variable.
- Other distance measure are also available like the **City Block** or **Manhattan distance** and the **Chebychev's distance**.
- Use of different distance measures may lead to different clustering result.

4.2.4 3. Select a Clustering Procedure

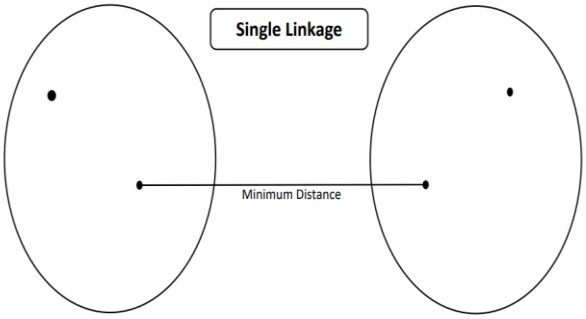


- Clustering procedure can be **Hierarchical** or **Non-Hierarchical** or **Other** Procedure.
- Hierarchical clustering is characterised by the development of hierarchy or tree like structure hierarchical method can be **Agglomerative** or **Divisible**.
- **Agglomerative** clustering starts with each objects in a separate cluster. Clusters are formed by grouping objects into bigger and bigger clusters. This process is continued until all objects are members of a single cluster.
- **Divisive** clustering starts with all the objects group in a single cluster.

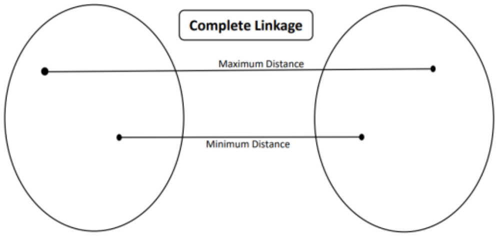
Agglomerative methods are commonly used marketing research. They consists of **linkage method**, **variance method** and **centroid method**.

linkage method include **single linkage**, **complete linkage** and **average linkage**.

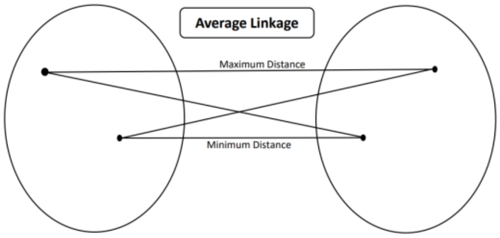
Single linkage is based on minimum distance or the nearest neighbor. The first two objects cluster are those that have the smallest distance between them. The single linkage method doesn't work well when the clusters are poorly defined.



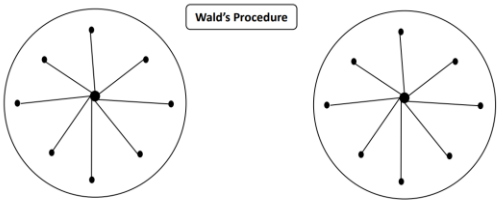
Complete linkage method is similar to single linkage except that it is based on the maximum distance.



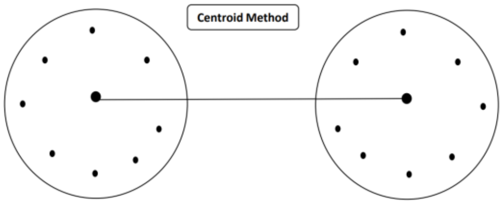
Average Linkage method works similarly, in this method the distance between cluster is defined as the average of the distance between all pairs of objects. The average linkage method uses information on all pairs of distances. Not marely the minimum or maximum distances for this reason it is usually prefer to the single or completely linkage method.



- The **variance method** attempt to generate clusters to minimize the within cluster variance. A commonly used variance method is **Wald's procedure**. For each cluster the means for all variance are computed.



- In the **centroid method** the distance between their centroid(means) of all variables.



- In **Non- Hierarchical** methods which frequently referred to as **K-Means Clustering**. We first assign or determine a cluster centers or group of all objects within a pre-specified threshold value from the center.

Sequential threshold method in which a cluster center is selected and all objects within a pre-specified threshold value from the center are group together.

In **parallel sequential** method we specifies several clusters samples at once. All objects that are within a pre-specified threshold value from the center are group together.

Optimizing Partitioning method, in this method differs from the two threshold procedure in that object can later be re-assign to cluster. To optimize an overall criteria such as average within cluster distance for a given number of clusters.

4.2.5 4. Decide the numbers of Clusters

A major issue in cluster analysis is deciding the number of cluster although there are no hard and fast rules. Some guidelines are available :

i. Theoretical, conceptual or practical consideration may suggest a certain number of clusters.
ii. In hierarchical clustering, the distances at which clusters are combined can be used as criteria.

This information can be obtained from the agglomeration schedule or from the dendogram.

iii. In non-hierarchical clustering, the ratio of total within group variance to between group variance can be plotted against the number of clusters.

The point at which an elbow or a sharp bend occurs indicates an appropriate number of clusters.

4.2.6 5. Interpret and Profile Clusters

Interpret and profiling clusters involves examining the clusters centroids. The centroids represent the mean values of the objects contained in the cluster on each of the variables. The centroids enables us to describe each clusters by assigning it a name or label. If the clustering program does not print this information, it may be obtain through discriminant analysis.

4.3 Market Basket Analysis

Market basket analysis determines customers purchasing patterns by finding significant relationship among the items which they select in their shopping carts. Market basket analysis aids the procedure as well as expense target strategies in numerous business association.

4.3.1 Association Rules

Association rule is the most important data mining technique used in Market basket analysis. All fruits are sorted in the aisle in a supermarket, all dairy products are placed together under another aisle. Hence spending time and intentionally investing resources to place the necessary items in an organized way not only reduces the shopping time of the customer but also help customer to purchase the most appropriate item. One might be keen in clubbing in their market basket.

Association rule is related to the statement **"What goes with what"**. The purchase of products by customers at supermarket are termed as **transaction**. The magnitude of an associative rule can be derived in the existence of three parameter namely : 1. **Support** 2. **Confidence** 3. **Lift**

1. **Support** : Support of an item or set of item is that the fraction of transactions in our dataset that contains the number of that particular item product to the total number of that transaction.

Support gives an idea of how many times an item set has occurs in the overall transaction.

$$\textbf{Support A} = \frac{\textbf{Number of Transactions that contain A}}{\textbf{Total Transaction}}$$

Example : There are 6 transactions in total with various purchases that happen in your cafeteria

Item Transaction	Item Basket
T_1	Cookie, Pasty, Ice-Cream, Cake
T_2	Cookie, Pasty, Maggy
T_3	Cookie, Pasty, Cake
T_4	Bread, Egg, Soap
T_5	Maggy, Berger
T_6	Shake, Berger

We can utilized there core majors that are used in association rule, which are *Support, Confidence* and *Lift*.

Suppose I am interested in to know the **Support Cookie**

$$Supprt(Cookie) = \frac{3}{6} = \frac{1}{2}$$

that is 3 out of 6 transaction purchases containing Cookies have occured 3 times or 50%.

Support can be implemented on multiple items at the same time as well. The Support for Cookie and Cake

$$Supprt(Cookie \,\&\, Cake) = \frac{2}{6} = \frac{1}{3}$$

2. **Confidence :** The Confidence of a consiquent event given an anticident event can be described by using conditional probability. So, it is the probability of event A happening given that event B has already happened. This can be used to describe the probability of an item being purchase when another item is already in the basket. It is measured by the dividing the proportion of transaction with item X and Y over the proportion of transaction with Y.

$$\boldsymbol{Confidence(A \Rightarrow B) = P(A/B) = \frac{P(A \cap B)}{P(A)} = \frac{P(AB)}{P(A)}}$$

$$\boldsymbol{Confidence(A \Rightarrow B)} = \frac{\textbf{Number of transaction that contain A \& B}}{\textbf{Number of transaction that contain A}}$$

Example :

$$Confidence(Cookie \Rightarrow Cake) = \frac{2}{3}$$

We arrive at solution of 2/3. We can understand of the intuition of confidence, it we were to look only at transaction 1 and transaction 3.

3. **Lift :** Lift is the observed to the expected ratio. Lift measures how likely an item is purchased when another item is purchased, while controlling for how popular both items are.

$$\boldsymbol{Lift(A \Rightarrow B) = \frac{Support(A,B)}{Support(A)Support(B)}}$$

A lift of 1 means that both of items are actually independent and without any association. For any value higher than 1, lift shows that actually there is an association. The higher value is higher association.

$$Lift(Cookie \Rightarrow Cake) = \frac{2/6}{3/6 \times 2/6} = \frac{6}{3} = 2$$

The lift of Cookie and Cake is 2 which implies that there is an actually an association between Cookie and Cake.

Fundamental rules for association grant the phenomenon of one item and the inclusion of another. Through the use of data analytics the process is trying to uncover association rules include the following steps :

Step 1 : Set up the data in the transaction presentation. An association algorithm demand input information to be arranged in the transaction format.

Step 2 : Shortlist collection of item that often occur items sets are object aggregation. An association algorithm consider the most commonly occured items, make increasingly relevant the ultimate role that will be pulled to the next level.

Step 3 : Generate the association rule applicable from the item set.

4.3.2 Frequent Item Set Generation

For an association analysis of 2^{n-1} item sets can be discovered apart from the null item set. As when the items increases, the number of items sets also increases exponentially. Therefore, it is necessary to pick and fix a minimum support threshold to error item sets that takes placeless frequently in the transactions.

4.3.3 Extracting Rule From Frequent Item Sets

A brute force solution is to determine the rules for the mining association with Support and Confidence for each rule

$$\boldsymbol{R = 3n - 2^{n+1} + 1}$$

In a dataset of n items, R rule can be found, this process extract all the rules with confidence higher than a minimum confidence threshold.

There are few algorithmic ways to measure the frequent item sets efficiently. The algorithm **Apriori** and **Frequent pattern** F_p growth are two of the most common algorithms for the analysis of association rules.

4.3.4 Use of Market Basket Analysis

1. **Product Placement :** The application of Market Basket Analysis helps retailers and marketers find associations between diverse products. By assessing customers purchasing and locating patterns. Business can identify the most likely product to be purchase together in the store.

Based on the information they arrange or place the product in class proximities to encourage them to purchase both the commodities in the stores.

2. **Physical Shelf Sorting & Exhibition :** An alternate use for product placement in the store is to separate the commodities that are typically bouquet together or at the same time. This step is taken to encourage the customers to wander around the store make impulses purchases while buying what they intent to purchase.

For Example : Shampoos and Conditioners could be placed in different sections in a store so that customers would move around the store looking for both of these items and impulsively by other products as well.

3. **Exploiting Process Cross Sell, Upsell & Bundling Opportunities :** The affinity grouping related to a diverse range of product could be used by business to indicate that customers might be pre-disposed to purchase the group items at the same time. It enables the presentation of the commodities for the cross sell purpose.

4. **Anticipating Customers Purchase Behavior :** The approach is to arrive that the right kinds of incentive that can create value for the customers and the business can retain them. The analytical process can helps to derive information from past purchase decision and and anticipate their future buying behavior. So that their need can be fulfilled by the business entity.

5 Unit — IV

Choice Modeling : choice-based conjoint analysis surveys, simulating choice data, fitting a choice model, adding consumer heterogeneity to choice models, hierarchical bayes choice models, design of choice-based conjoint surveys. behavior sequences: web log data, basic event statistics, identifying sequences (sessions), markov chains for behavior transitions.

5.1 Conjoint Analysis

Conjoint Analysis attempts to determine the relative importance that consumers attach to salient attributes and the utilities they attach to the levels of attributes.

This information is derived from consumers evaluation of brands or from brand profile composed of these attributes and their labels. The participants are represented with stimuli that consist of combination of attribute labels. They are asked to evaluated these stimuli in terms of their desirability.

Like MDS, Conjoint Analysis reliase on participant subject evaluation.

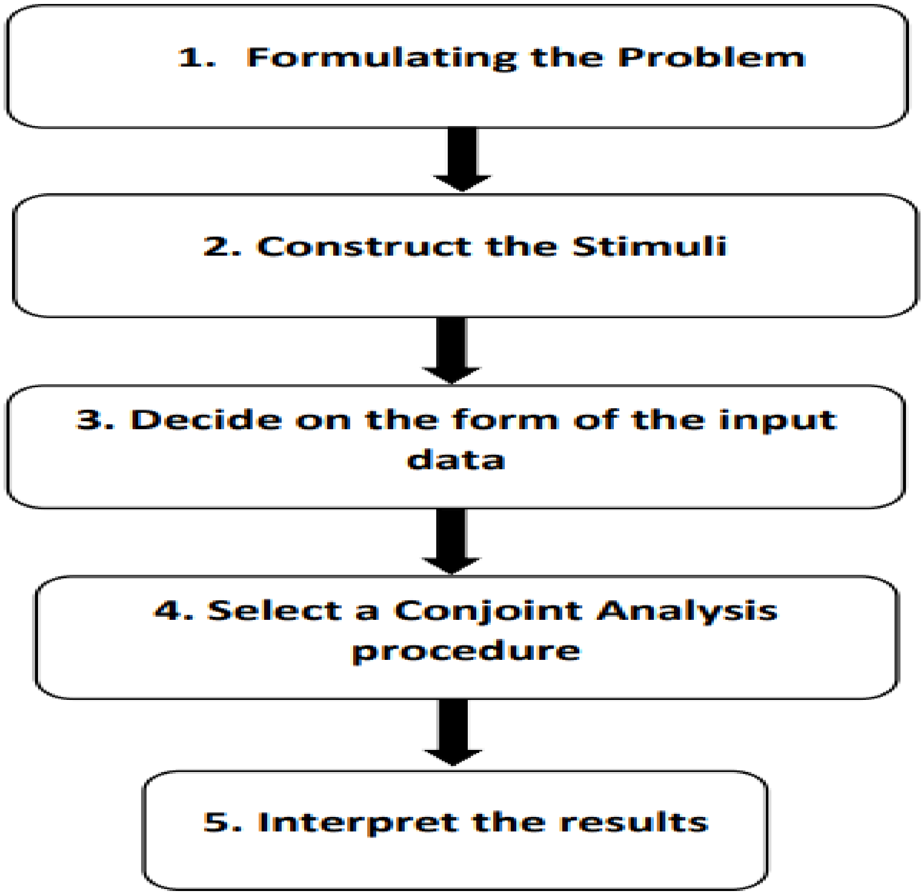
- In MDS however the stimuli are products or brands.

In Conjoint Analysis the stimuli are combinations of attributes labels determined by the researcher.

- The goal in MDS is to develop the Spatial Map depicting the stimuli in a multi-dimensional perceptual or preference space.

Conjoint Analysis on the other hand seeks to develop the part worth or utility functions, describing the utility that consumers attach to the levels of each attributes. The two techniques are complimentary.

5.1.1 Conducting Conjoint Analysis


1. Formulating the Problem
2. Construct the Stimuli
3. Decide on the form of the input data
4. Select a Conjoint Analysis procedure
5. Interpret the results

5.1.2 1. Formulating the Problem

Formulating the conjoint problem the researcher must identify the attributes and attributes levels to be used in constructing the stimuli. Attribute levels denotes the value assumed by the attributes from a theoretical stand point the attribute selected should be salient in influencing consumer preference and choice.

For Example : In the choice of car the price efficiency and fuel efficiency, interior space and so on, should be included from a managerical perspective the attribute and their levels should be characteristics that management can change and take action upon.

To tell a manager that consumer prefers sporty car to one that is conservative(traditional) looking is not helpful unless spotiness and conservativeness are define in terms of attributes over which a manager have control. The attributes can be identify through discussion with management and industry experts, analysis of secondary data, qualitative research and pilot surveys. A typical conjoint analysis study may involves 6 – 7 attributes. Once the salient attribute have been identify their appropriate level should be selected. The number of attributes levels determine the number of parameters that will be estimated and also influences the number of stimuli that will be evaluated by the participants.

5.1.3 2. Construct the Stimuli

To illustrate the conjoint methodology by considering the problem of how students evaluate boots.

For Example : Brands such as Dr. Martins, Timber Land, Belly and Caterpillar.

Qualitative researcher identified three(3) attributes as the material used for the **Upper**, **Country** or region in which boots were designed and manufactured and the **Price.** Each was defined in terms of three levels.

Boot Attributes & Their Levels

Attributes	Number	Description
Upper	3	Lather
	2	Swede
	1	Imitation Letter
Country	3	Italy
	2	USA
	1	ForEast
Price	3	50
	2	125
	1	200

Two brand approaches are available for constructing conjoint analysis stimuli.

1. **Pairwise Approach**
2. **Full Profile Approach**

In the **pairwise approach** also known as *two factor evaluation*, participants evaluate two attributes at a time until all possible pairs of attributes have been evaluated. For each pair participants evaluate all the combination of levels of both the attributes which are presented in a matrix.

Country

		Italy	USA	ForEast
Upper	Lather			
	Swede			
	Imitation Lather			

		Country		
		Italy	USA	ForEast
Price (€)	50			
	125			
	200			

		Price (€)		
		25	125	200
Upper	Lather			
	Swede			
	Imitation Lather			

In **full profile approach** also called *multiple factor evaluation*. In complete profile of brands are constructed for all attributes. Typically each profile is described on a separate index card.

	Product Profile
Upper	Made of Lather
Country	Designed & Made
Price	Costing € 200

It is not necessary to evaluate all the possible combinations, nor is it feasible in all cases. In the **pairwise approach**, it is possible to reduce the number of paired comparisons by using cyclical designs. Likewise, in the **full-profile approach**, the number of stimulus profiles can be greatly reduced by means of fractional factorial designs. A special class fractional designs, *orthogonal arrays*, allows for the efficient estimation of all main effects. Orthogonal arrays permit the measurement of all main effects of interest on an uncorrelated basis. These designs assume that all interactions are negligible. Generally, two sets of data are obtained. One, the estimation set, is used to calculate the part-worth functions for the attribute levels. The other, the holdout set, is used to assess reliability and validity.

The advantage of the **pairwise approach** is that it is easier for the participants to provide these judgements. Its relative disadvantage, however, is that it requires more evaluations than the full-profile approach. Studies comparing the two approaches indicate that both methods yield comparable utilities is more commonly used.

The boots example follows the full-profile approach. Given three attributes. defined at three levels each, a total of $3 \times 3 \times 3 = 27$ profiles can be constructed. To reduce the participant evaluation task a fractional factorial design was employed and a set of nine profiles was constructed to constitute the estimation stimuli set. Another set of nine stimuli was constructed for validation purposes. Input data were obtained for both the estimation and validation stimuli. Before the data could be obtained, however, it was necessary to decide on the form of the input data.

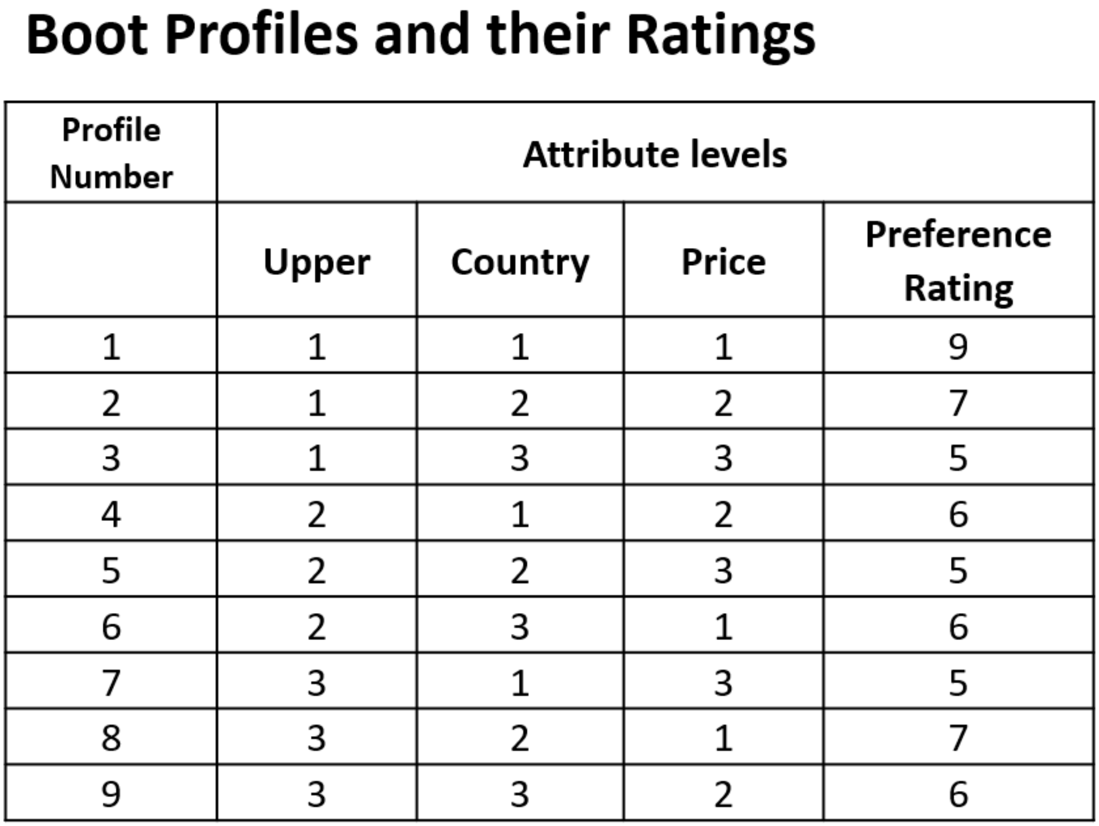
5.1.4 3. Decide on the form of Input Data

As in the case of MDS, conjoint analysis input data can be either non-metric or metric. For non-metric data, participants are typically required to provide rank order evaluations. For the pairwise approach, participants rank all the cells of each matrix in terms of their desirability. For the full-profile approach, they rank all the stimulus profiles. Rankings involve relative evaluations of the attribute levels. Proponents of ranking data believe that such data accurately reflect the behaviour of consumers in the marketplace.

In the metric form, participants provide ratings, rather than rankings. In this case, the judgements are typically made independently. Advocates of rating data believe they are more convenient for the participants and easier to analyse than rankings. In recent years, the use of ratings has become increasingly common.

In conjoint analysis, the dependent variable is usually preference or intention to buy. In other words, participants provide ratings or rankings in terms of their preference or intentions to buy. The conjoint methodology, however, is flexible and can accommodate a range of other dependent variables, including actual purchase or choice.

In evaluating boot profiles, participants were required to provide preference ratings for the boots described by the nine profiles in the estimation set. These ratings were obtained using a nine-point **Likert scale**(1 not preferred, 9 greatly preferred).



Boot Profiles and their Ratings

Profile Number	Attribute levels			
	Upper	Country	Price	Preference Rating
1	1	1	1	9
2	1	2	2	7
3	1	3	3	5
4	2	1	2	6
5	2	2	3	5
6	2	3	1	6
7	3	1	3	5
8	3	2	1	7
9	3	3	2	6

5.1.5 4. Select a Conjoint Analysis Procedure

The basic **conjoint analysis model may be represented by the following formula :

$$\boldsymbol{U(X) = \sum_{i=1}^{m} \sum_{j=1}^{k_j} \alpha_{ij}\, x_{ij}}$$

Where,

- $U(X)$ = overall utility of an alternative -the
- α_{ij} = part-worth contribution or utility associated with the j^{th} level $\left(j = 1,2,\ldots,k_j\right)$ of the i^{th} attribute $(1,2,\ldots,m)$
- k_j = number of levels of attribute i
- m = number of attributes
- $x_{ij} = I$ if the j^{th} level of the i^{th} attribute is present, 0 otherwise.

The importance of an attribute, I_j, is defined in terms of the range of the part-worth α_{ij} across the levels of that attribute :

$$\boldsymbol{I_j = \{max(\alpha_{ij}) - min(\alpha_{ij})\}} \quad \text{for each j}$$

The attribute's importance is normalized to ascertain its importance relative to other attributes, W_i :

$$\boldsymbol{W_i = \frac{I_j}{\sum_{i=1}^{m} I_j}}$$

so that

$$\boldsymbol{\sum_{i=1}^{m} W_i = 1}$$

Several different procedures are available for estimating the basic model. The simplest, and one which is gaining in popularity, is dummy variable regression. In this case, the predictor variables consist of dummy variables for the attribute levels. If an attribute has k_i, levels, it is coded in terms of $k_i - 1$ dummy variables. If metric data are obtained, the ratings, assumed to be interval scaled, form the dependent variable. If the data are non-metric, the rankings may be converted to 0 or 1 by making paired comparisons between brands. In this case, the predictor variables represent the differences in the attribute levels of the brands being compared. Other procedures that are appropriate for non- metric data include *LINMAP, MONANOVA* and the *LOGIT* model.

The researcher must also decide whether the data will be analyzed at the individual-participant or the aggregate level. At the individual level, the data of each participant are analysed separately. If an aggregate-level analysis is to be conducted, some procedure for grouping the participants must be devised. One common approach is to estimate individual-level part-worth or utility functions first. Participants are then clustered on the basis of the similarity of their part-worth functions. Aggregate analysis is then conducted for each cluster. An appropriate model for estimating the parameters should be specified.

The data reported were analysed using ordinary least squares (OLS) regression with dummy variables. The dependent variable was the preference ratings. The independent variables or predictors were six dummy variables, two for each variable. Since the data pertain to a single participant, an individual-level analysis was conducted. The part-worth or utility functions estimated for each attribute as well as the relative importance of the attributes.

Boot Data Coded for Dummy Variable Regression

Preference Rating	Attribute					
	Upper		Country		Price	
Y	X_1	X_2	X_3	X_4	X_5	X_6
9	1	0	1	0	1	0
7	1	0	0	1	0	1
5	1	0	0	0	0	0
6	0	1	1	0	0	1
5	0	1	0	1	0	0
6	0	1	0	0	1	0
5	0	0	1	0	0	0
7	0	0	0	1	1	0
6	0	0	0	0	0	1

Result of Conjoint Analysis

Attribute	Level			
	Number	Description	Utility	Importance
Upper	3	Leather	0.778	0.268
	2	Swede	-0.556	
	1	Imitation Leather	-0.222	
Country	3	Italy	0.445	0.214
	2	USA	0.111	
	1	Far East	-0.556	
Price	3	€ 50	1.111	0.500
	2	€ 125	0.111	
	1	€ 200	-1.222	

The model estimated may be represented as :

$$\boldsymbol{U = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + b_6X_6}$$

where, $X_1, X_2 =$ Dummy variables representing uppers.

$X_3, X_4 =$ Dummy variables representing country.

$X_5, X_6 =$ Dummy variables representing price.

For uppers, the attribute levels were coded as follows :

The levels of the other attributes were coded similarly. The parameters were estimated as follows :

$$b_0 = 4.222$$

$$b_1 = 1.000$$

$$b_2 = -0.333$$

$$b_3 = 1.000$$

$$b_4 = 0.667$$

$$b_5 = 2.333$$

$$b_6 = 1.333$$

Given the dummy variable coding, in which level 3 is the base level, the coefficients may be related to the part-worth. Each dummy variable coefficient represents the difference in the part-worth for that level minus the part-worth for the base level. For uppers, we have the following:

$$\alpha_{11} - \alpha_{13} = b_1$$

$$\alpha_{12} - \alpha_{13} = b_2$$

To solve for the part-worth, an additional constraint is necessary. The part-worths are estimated on an interval scale, so the origin is arbitrary. Therefore, the additional constraint imposed is of the form :

$$\alpha_{11} + \alpha_{12} + \alpha_{13} = 0$$

These equations for the first attribute, uppers, are:

$$\alpha_{11} - \alpha_{13} = 1.000$$

$$\alpha_{12} - \alpha_{13} = -0.333$$

$$\alpha_{11} + \alpha_{12} + \alpha_{13} = 0$$

Solving these equations, we get

$$\alpha_{11} = 0.778$$

$$\alpha_{12} = -0.556$$

$$\alpha_{13} = -0.222$$

The part-worths for other attributes reported can be estimated similarly. For country, we have:

$$\alpha_{21} - \alpha_{23} = b_3$$

$$\alpha_{22} - \alpha_{23} = b_4$$

$$\alpha_{21} + \alpha_{22} + \alpha_{23} = 0$$

For the third attribute, price, we have:

$$\alpha_{31} - \alpha_{33} = b_5$$

$$\alpha_{32} - \alpha_{33} = b_6$$

$$\alpha_{31} + \alpha_{32} + \alpha_{33} = 0$$

The relative importance weights were calculated based on ranges of part-worths, as follows:

Sum of ranges of part-worths= [0.778- (-0.556)]+ [0.445-(-0.556)]+ [1.111-(-1.222)]= 4.668

$$\text{Relative importance of uppers} = \frac{[0.778 - (-0.556)]}{4.668} = \frac{1.334}{4.668} = 0.286$$

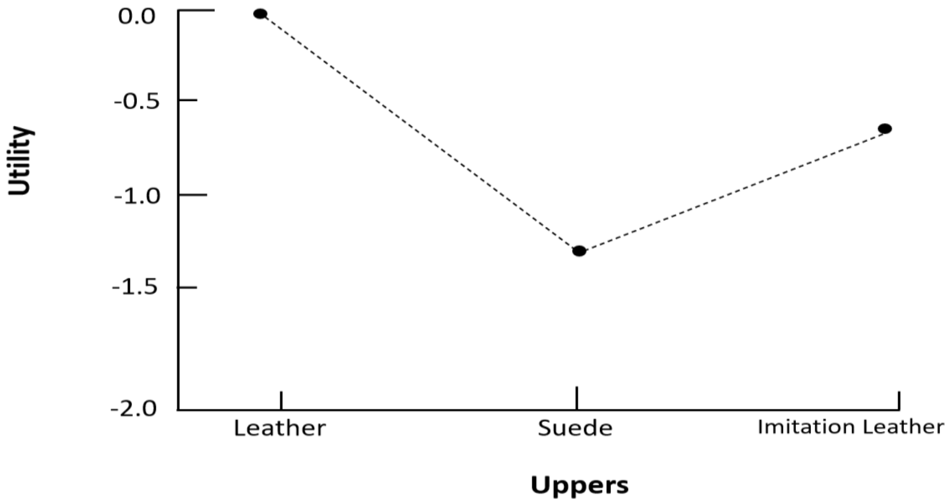
$$\text{Relative importance of country} = \frac{[0.445 - (-0.556)]}{4.668} = \frac{1.001}{4.668} = 0.214$$

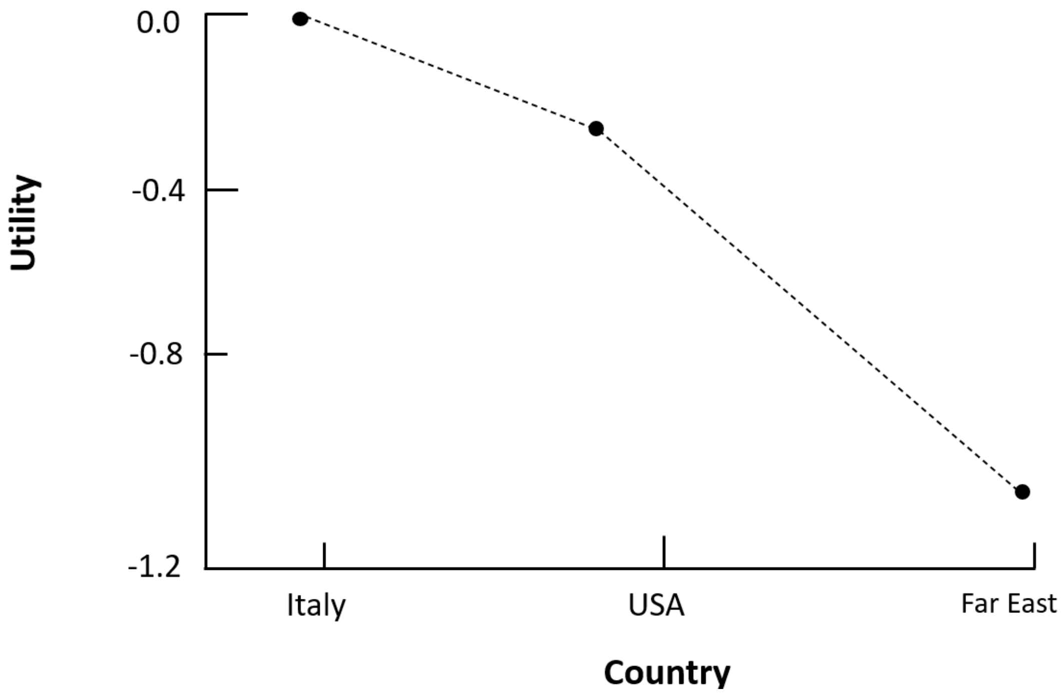
$$\text{Relative importance of price} = \frac{[1.111 - (-1.222)]}{4.668} = \frac{2.333}{4.668} = 0.500$$

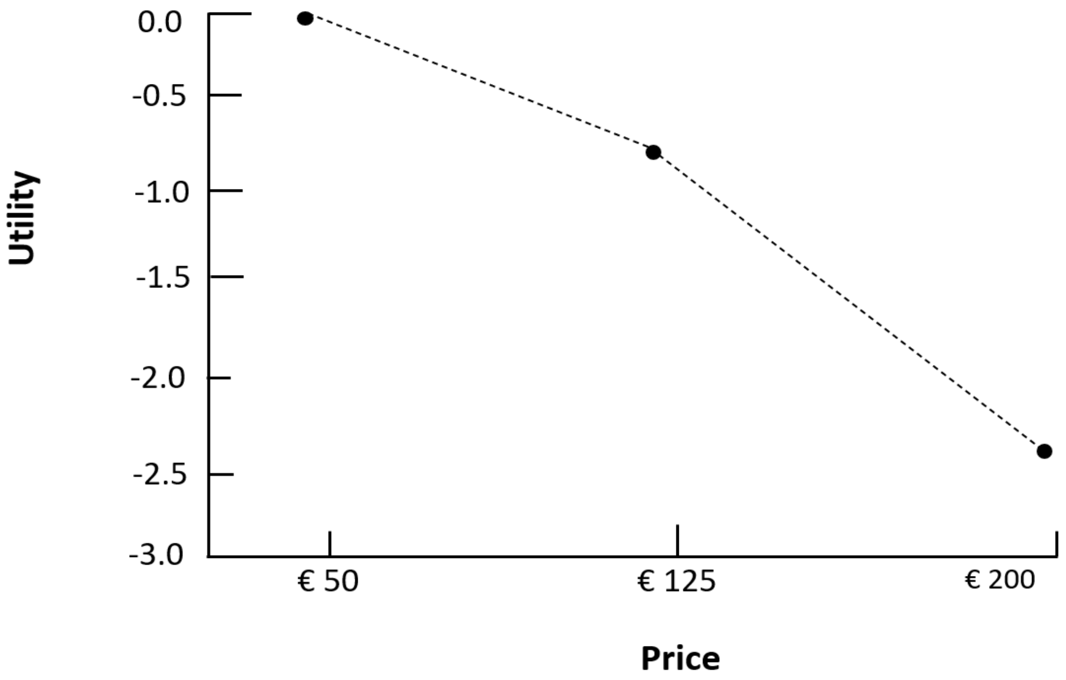
The estimation of the part-worths and the relative importance weights provides the basis for interpreting the results

5.1.6 5. Interpret the Results

For interpreting the results. it is helpful to plot the part-worth functions. The part-worth function values for each attribute. This participant has the greatest preference for leather uppers when evaluating boots, Second preference is for imitation leather uppers, and suede uppers are least preferred An Italian boot is most preferred, followed by American boots and boots from the Far East. As may be expected, a price of 50.00 has the highest utility and a price of 200.00 the lowest. The utility values reported 26.6 have only interval scale properties, and their origin is arbitrary. In terms of relative importance of the attributes, we see that price is number one Second-most important is uppers, followed closely by country. Because price is by far the most important attribute for this participant, this person could be labelled as **“price sensitive”**.




Utility
0.0
-0.4
-0.8
-1.2
Italy
USA
Far East
Country


Utility
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-3.0
€ 50
€ 125
€ 200
Price