The higher ē eigen value ē more important ē variables.

Quantitative - Metric ; Ratio →
Qualitative - None-metric ; Nominal
; Ordinal

Course Description

- Real life examples of Multivariate data.
- The data matrix.
- Calculation of summary statistics, mean vectors, covariance & correlation matrices.
- Test for mean vector, one-sample and Hotelling T^2 tests based on union intersection approach
- Simultaneous confidence intervals for detecting important components.
- Testing equality of two pplns means
- Basic assumptions and application of principal components, discriminant functions and canonical correlations.
- Distributions of linear functions of a random vector and of quadratic forms.
- Multivariate normal regression and correlation analysis.
- Elements of multivariate of variance.
- Use of computer packages.

$= \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0.556 \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

EXAM - Definitions

Introduction.

- The multivariate analysis methods are all statistical techniques that simultaneously analyse multiple measurements on individuals or objects under investigation, that is, any simultaneous analysis of more than two variables can be loosely considered multivariate analysis.

Multivariate Techniques and Real Life Multivariate Data Applications.

- The multivariate analysis techniques available for selection in various fields of application include the following techniques discussed below;

a) Principal Components. This is a statistical approach that can be used to analyse interrelationships among a large no. of variables and to explain these variables in terms of their underlying dimensions or factors.

- The objective is to find a way of condensing the information contained in a no. of original variables into a smaller set of variates (factors) with a minimal loss of info (Hair, Black, Babin & Anderson, 2010).

Example 1.1

An investigator may use factor analysis to help understand relationships between customers' ratings of a fast-food restaurant by asking the customer to take the restaurant on the following 6 variables; food taste, food temperature, freshness, waiting time, cleanliness and friendliness of employees.

The investigator would like to combine these 6 variables, into a smaller number, by analysing the customer responses and the investigator might find that the variables food taste, temparatures, and freshness combine together to form a single factor of food quality, whereas the variables waiting time, cleaniliness, and friendliness of employees combine to form another single variable, service quality.

b) Multiple Linear Regression Analysis.

- The analysis method is used when the research problem involves a single metric dependent variable presumed to be related to two or more metric independent variables. The objective of multiple regression is to predict the changes in the dependent variable in response to

If $x_1 = 2 \cdot 414$ either Let $x_2 = 1$, $x_1 = \frac{1}{2 \cdot 414} = 0 \cdot 414$ $e_1 = \begin{bmatrix} 0.414 \\ 1.000 \end{bmatrix}$ $e_1 = \begin{bmatrix} 0.414 \\ 1.000 \end{bmatrix} \begin{bmatrix} \quad \\ \sqrt{0.414^2 + 1.000^2} \end{bmatrix} \begin{bmatrix} \quad \\ 0.9240 \end{bmatrix}$

$x_2 = 0.414$ or $x_2 = 1$, $x_1 = 0.414$

changes in the independent variables.

Example 1.2

The investigator may be interested in predicting the amount or size of the dependent variable for example, monthly expenditure on leisure (dependent variable) might be predicted from information regarding a household's income, its size, and the age of the head of the house hold (independent variables).

Similarly, the investigator might attempt to predict a company's sales from info on its expenditure for advertising, the no. of sales people, and the no. of stores carrying its products.

(c) Multiple Discriminant Analysis.

- The multiple discriminant " technique is ē appropriate multivariate technique if the single dependent variable is dichotomous (for example male-female) or multichotomous (for example high-medium-low) and therefore non-metric. As with multiple regressions, the independent variables are assumed to be metric.
- Discriminant analysis is applicable in situations in which the total sample can be divided into groups based on a non-metric dependent variable characterizing several classes. The primary objectives of multiple discriminant analysis are to understand group differences and to predict the likelihood that an entity (individual or object) will belong to a particular class or group based on several metric independent variables.
- Logistic regression models, often referred to as logistic analysis, are a combination of multiple regression and multiple discriminant analysis.

Example 1.3

An investigator wishes to distingaish innovators from non-innovators according to their demographic & pyschographic profiles or distinguish heavy product users from light users or distinguish males from females or distinguish national brand buyers from private-label buyers or distinguish good credit risks from poor credit risks.

Example 1.4

The revenue authority uses discriminant analysis to compare selected federal tax returns with a composite, hypothetical, normal tax payer's return (at diff. income levels) to identify the most promising returns and areas of audit.

d) Logistic Regression.

- Logistic " models are distinguished from discriminant analysis primarily in that they accomodate all types of independent variables (metric & non-metric & do not require the assumption of multivariate normality. (independent variables are assumed to be metric)
- However, in many instances particularly with more than two levels of dependent variables, discriminant analysis is the more appropriate analysis.

Example 1.5

- The logistic regression model is used to identify the financial & mangerial data that best differentiated btwn ē successful & unsuccessful firms in order to select the best candidates for investment in the future by reviewing past records of financial management and placed firms data into one of two classes, successful over a five-yr period, and unsucceful after 5 yrs.

e) Canonical Correlation.

- In " analysis the objective is to correlate simultaneously several metric dependent variables and several metric independent variables and is viewed as a logical extension of multiple regression analysis.
- Whereas multiple regressions involve a single dependent variable, canonical correlation involves multiple dependent variables. The underlying principle is to develop a linear combination of each set of variables (both independent and dependent) in a manner that maximizes the correlation between the two sets.

Example 1.6

- The canonical correlation could be used to compare the perception of the world-class companies over the 20 questions with the perception of the company such that the investigator could then conclude whether the perception of the company are correlated with those of world-class companies. The technique provides info on the overall correlation of perception as well as the correlation btwn each of ē 20 qstns.

f) Multivariate Analysis of Covariance and Variance.

- This is a statistical technique that can be used to simultaneously ~~explain~~ explore the rltnship btwn several categorical independent variables (treatments) and two or more metric dependent variables.
- Multivariate analysis of covariance can be used in conjuction with multivariate analysis of variance

to remove (after experiment) the effect on any uncontrolled metric independent variables (covariates) on the dependent variables.

Example 1.7

An investigator may wish to know if a humorous ad will be more effective with its customers than a non-humorous ad. The investigator develop two ads - one humorous and one non-humorous ad then shows a group of customers the two ads. After seeing the ads, the customers would be asked to rate the company and its products on several dimensions, such as modern vs traditional or high quality vs low quality.

The multivariate analysis of variance would be the technique to use to determine the extent of any statistical difference btwn the perception of customers who saw the humorous ad vs those who saw the non-humorous one.

g) Conjoint Analysis.

- This is a dependence technique that brings new sophistication to the evaluation of objects, such as new products, services or ideas. The most direct application is in new product or service dev., allowing for the evaluation of complex products while maintaining a realistic decision context for the respondent.
- The investigator is able to assess the importance of attributes as well as the levels of each attribute while consumers evaluate only a few product profiles, which are combination of product levels.

Example 1.8

Consider a developed product concept that has 3 attributes (price, quality, and colour that have three, two & three levels respectively for example colour has possible levels of for example red, yellow and blue). Instead of having to evaluate all 18 (3x2x3) possible combinations, a subset (9 or more) can be evaluated for 3 attractions to customers, and the investigator knows not only how important each attribute is but also the importance of each level.

h) Cluster Analysis

- This is an analytical technique for developing meaningful subgroups of individuals or objects. The objective is to classify a sample of entities (individuals or objects) into a small number of -

mutually exclusive groups based on ē similarities among ē entities.
- In cluster analysis, unlike discriminant analysis, the groups are not predefined. The technique is used to identify the groups.

Example 1.9
A restaurant owner wishes to know whether customers are patronizing the restaurant for diff. reasons. The data could be collected on perceptions of pricing and food quality. Cluster analysis could be used to determine whether some subgroups (clusters) are highly motivated by low prices vs those who are much less motivated to come to the restaurant based on the price considerations.

(i) Perception Mapping
- In " " (multidimensional scaling), the objective is to transform consumer judgements of similarity or preference (for example preference for stores or brands) into distances represented in multidimensional space. If objects A & B are judged by respondents as being the most similar compared with all other possible pairs of objects, perception mapping techniques will position objects A and B in such a way that the distance btwn them in multidimensional space is smaller than the distance btwn any other pairs of objects.

Example 1.10
The owner of a Burger King franchise wants to know whether the strongest competitor is MacDonald's or Wendy's. A sample of customers is given a survey and asked to rate the pairs of restaurants from most similar to least similar. The results show that the Burger King is most similar to Wendy's, so the owners know that the strongest competitor is the Wendy's restaurant because it is thought to be the most similar.

(ii) Correspondence Analysis.
- This is the technique that facilitates the perceptual mapping of objects (for example products, persons) on a set of non-metric attributes. Investigators are constantly faced with the need to "quality the qualitative data" found in normal variables. Correspondence analysis differs from the interdependence technique discussed earlier in its ability to accomodate both non-metric data and non-linear rltnships.

Example 1-11

The respondents' brand preferences can be cross tabulated on demographic variables (for example gender, income, categories, occupation) by indicating how many people preferring each brand fall into each category of the demographic variables.

Through correspondence analysis, the association, or "correspondence", of brands and the distinguishing xtics of those preferring each brand are then shown in two- or three dimensional map of both brands and respondent xtics.

Brands perceived as similar are located close to one another. Likewise, the distinguishing xtic of respondents preferring each brand is also determined by ē proximity of ē demographic variable categories to the brand's position.

k) Structured Equation Modelling and Confirmatory Factor Analysis.

- This is a technique that allows separate rltnships for each of a set of dependent variables. It provides the appropriate and most efficient estimation technique for a series of separate multiple regression equations estimated simultaneously.
- The multivariate analysis of variance and canonical correlation are not applicable in this situation because they allow only a single rltnship btwn dependent & independent variables. In confirmatory factor analysis the researcher can assess the contribution of each scale item as well as incorporate how well the scale measures the concept (reliability).

Example 1-12

A study by management consultants identifies several factors that affect worker satisfaction supervisor support, work environment, and job performance.

2. DESCRIPTIVE STATISTICS.

The data matrix.

- The measurements collected on several variables (characteristics) are frequently arranged and displayed in various ways that may be, for example, graphs or tabular arrangements.

Arrays:

- The multivariate data arise whenever an investigator seeks to understand a social or physical phenomenon and selects a number $p \geq 1$ of variables or characteristics to record. The values of the variable recorded for each distinct item, individual or experiment unit for the p variables may be recorded in tabular form as shown below:

1	2	...	i	...	p
X_{11}	X_{21}	...	X_{i1}	...	X_{p1}
X_{12}	X_{22}	...	X_{i2}	...	X_{p2}
⋮	⋮	...	⋮	...	⋮
X_{1j}	X_{2j}	...	X_{ij}	...	X_{pj}
⋮	⋮	...	⋮	...	⋮
X_{1n}	X_{2n}	...	X_{in}	...	X_{pn}

Where X_{ij} is the value of the ith variable that is observed on the j^{th} item or trial. The data can alternatively be displayed as rectangular array of n rows and p columns denoted by X as shown below which is commonly referred to as data matrix.

$$X = \begin{bmatrix} X_{11} & X_{21} & \dots & X_{p1} \\ X_{12} & X_{22} & \dots & X_{p2} \\ \vdots & \vdots & \dots & \vdots \\ X_{1n} & X_{2n} & \dots & X_{pn} \end{bmatrix}$$

Mean Vector, Variance - Covariance and Correlation Matrices:

- The info contained in the data can be assessed by calculating certain summary numbers commonly known as descriptive statistics that may be the measure of location or dispersion as described below:

(a) Arithmetic Mean. ◇

- The measure of central location (the arithmetic mean) of the n measurements $X_{11}, X_{12} \dots + X_{1n}$ of the first variable is given as:

$$\bar{X}_1 = \frac{X_{11} + X_{12} + \dots + X_{1n}}{n} = \frac{1}{n}\sum_{j=1}^{n} X_{1j}$$

The general form of the arithmetic mean of the k^{th} variable is written as:

$$\bar{X}_k = \frac{X_{k1} + X_{k2} + \dots + X_{kn}}{n} = \frac{1}{n}\sum_{j=1}^{n} X_{kj} \quad \text{where } k = 1, 2, \dots, p$$

If the n measurements represent a subset of the full set of measurements that might have been obtained, then, $\bar{X}_k$, is the sample mean of the k^{th} variable.

(b) Variance.

The measure of spread is provided by the sample variance defined for n measurements on the first variable given as:

$$S_1^2 = S_{11} = \frac{1}{n-1}\sum_{j=1}^{n} (X_{1j} - \bar{X}_1)^2$$

Where $\bar{X}_1$ is the sample mean of the x_{1j}'s, generally, for p variables, we have

$$S_k^2 = S_{kk} = \frac{1}{n-1}\sum_{j=1}^{n} (X_{kj} - \bar{X}_k)^2 \quad \text{where } k = 1, 2 \dots p$$

12014 99976

c) Standard deviation.

The square root of the sample variance is known as the sample standard deviation.

$$S_k = \sqrt{S_k^2} = \sqrt{S_{kk}}$$

d) Covariance.

The measure of the linear association btwn measurements of 2 variables X_1 & X_2 is given as

$$S_{12} = \frac{1}{n-1}\sum_{j=1}^{n} (X_{1j} - \bar{X}_1)(X_{2j} - \bar{X}_2)$$

The general association of the i^{th} variable and the k^{th} variable is given as:

$$S_{ik} = \frac{1}{n-1}\sum_{j=1}^{n} (X_{ij} - \bar{X}_i)(X_{kj} - \bar{X}_k) \quad \begin{cases} i = 1, 2, \dots p \\ j = 1, 2, \dots p \end{cases}$$

e) Correlation Coefficient (Pearson's Product moment correlation coefficient)

- This is the measure of the linear association btn 2 variables that does not depend on the units of measurement. The sample correlation coefficient for the i^{th} & k^{th} variables is defined as:

$$r_{ik} = \frac{\sum_{j=1}^{n} (x_{ij} - \bar{x}_i)(x_{kj} - \bar{x}_k)}{\sqrt{\sum_{j=1}^{n} (x_{ij} - \bar{x}_i)^2} \sqrt{\sum_{j=1}^{n} (x_{kj} - \bar{x}_k)^2}} \quad \begin{cases} i = 1, 2, \ldots, p \\ k = 1, 2, \ldots, p \end{cases} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}}$$

- The sample correlation coefficient r has the following properties:
- The value of r must be between -1 & $+1$ inclusive, i.e. $-1 \le r \le 1$
- The r measures the strength of ē. linear association.
  - If $r = 0$, this indicates a lack of linear association btn ē variables
  - If $r < 0$, this indicates that there is a tendency for one value in the pair to be larger than ~~the~~ its average when the other is smaller than its average.
  - If $r > 0$, this indicates that there is a tendency for one value in the pair to be larger than the average when the other value is larger than the average and also for both values to be smaller than the average together.
- The value of r_{ik} remains unchanged if the measurements of the i^{th} variable are changed to $y_{ij} = ax_{ij} + b$, $j = 1, 2, \ldots, n$ and the values of the k^{th} variable are changed to $y_{kj} = cx_{ij} + d$, $j = 1, 2, \ldots, n$ provided that the constants a and c have the same sign.
- The quantities s_{ik} & r_{ik} do not, in general, convey all there is to know about the association btwn 2 variables. Non linear associations can exist that are not revealed by these descriptive statistics as they provide measures of linear association (association along a line)

Example 2-1

- The data of variables X_1, X_2 and X_3 were collected and recorded as shown in the table below. Determine the sample mean vector $\bar{x}$, variance-covariance matrix, S_n, and correlation coefficient, r_n for X_1, X_2 and X_3 variables

X_1	4·0	5·0	4·0	3·0
X_2	4·2	5·2	4·5	5·8
X_3	2·0	4·0	3·0	5·0

Solution.

a) Mean vector: $\bar{x}_i = \frac{1}{n} \sum_{j=1}^{n} x_{ij}$

$\bar{x}_1 = \frac{1}{4}(4 + 5 + 4 + 3) = 4$

$\bar{x}_2 = \frac{1}{4}(4{\cdot}2 + 5{\cdot}2 + 4{\cdot}5 + 5{\cdot}8) = 5{\cdot}0$

$\bar{x}_3 = \frac{1}{4}(2 + 4 + 3 + 5) = 3{\cdot}5$

$$\bar{x} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \bar{x}_3 \end{bmatrix} = \begin{bmatrix} 4 \\ 5{\cdot}0 \\ 3{\cdot}5 \end{bmatrix}$$

b) Variance - Covariance Matrix:

$$S_{ik} = \frac{1}{n-1}\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)(x_{kj}-\bar{x}_k) = \frac{1}{n-1}\left\{\sum_{j=1}^{n} x_{ij}x_{kj} - n\bar{x}_i\bar{x}_j\right\}$$

$S_{11} = \frac{1}{3}[(4-4)^2 + \dots + (3-4)^2] = 0.67$

$S_{12} = \frac{1}{3}[(4-4)(4.2-5) + \dots + (3-4)(5.8-5] = -0.20$

$S_{22} = \frac{1}{3}[(4.2-5.0)^2 + \dots + (5.8-5.0)^2] = 0.45$

$S_{13} = \frac{1}{3}[(4-4)(2-3.5) + \dots + (3-4)(5-3.5)] = -0.33$

$S_{23} = \frac{1}{3}[(4.2-5)(2-3.5) + \dots + (5.8-5)(5-3.5)] = 0.87$

$S_{33} = \frac{1}{3}[(2-3.5)^2 + \dots + (5-3.5)^2] = 1.67$

$$S_n = \begin{bmatrix} S_{11} & S_{21} & S_{31} \\ S_{21} & S_{22} & S_{32} \\ S_{31} & S_{23} & S_{33} \end{bmatrix} = \begin{bmatrix} 0.667 & -0.200 & -0.333 \\ -0.200 & 0.453 & 0.867 \\ -0.333 & 0.867 & 1.667 \end{bmatrix}$$

c) Correlation matrix.

$$r_{ik} = \frac{S_{ik}}{\sqrt{S_{ii}}\sqrt{S_{kk}}}$$

$r_{11} = \frac{S_{11}}{\sqrt{S_{11}}\sqrt{S_{11}}} = \frac{0.67}{\sqrt{0.67}\sqrt{0.67}} = 1.000$, $r_{22} = \frac{S_{22}}{\sqrt{S_{22}}\sqrt{S_{22}}} = 1.000$, $r_{33} = \frac{1.67}{\sqrt{1.67}\sqrt{1.67}} = 1.000$

$r_{12} = \frac{S_{12}}{\sqrt{S_{11}}\sqrt{S_{22}}} = \frac{-0.20}{\sqrt{0.67}\sqrt{0.45}} = -0.364$, $r_{13} = \frac{S_{13}}{\sqrt{S_{11}}\sqrt{S_{33}}} = \frac{-0.33}{\sqrt{0.67}\sqrt{1.67}} = -0.316$

$r_{23} = \frac{S_{23}}{\sqrt{S_{22}}\sqrt{S_{33}}} = \frac{0.87}{\sqrt{0.45}\sqrt{1.67}} = 0.997$

$$r_n = \begin{bmatrix} r_{11} & r_{21} & r_{31} \\ r_{12} & r_{22} & r_{32} \\ r_{13} & r_{23} & r_{33} \end{bmatrix} = \begin{bmatrix} 1.000 & -0.364 & -0.316 \\ -0.364 & 1.000 & 0.997 \\ -0.316 & 0.997 & 1.000 \end{bmatrix}$$

3×6

Alternative_1.

X_1	X_2	X_3	X_1^2	X_2^2	X_3^2	X_1X_2	X_1X_3	X_2X_3	$(X_1-\bar{X}_1)$	$X_2-\bar{X}_2$
4	4.2	2	16	17.64	4	16.8	8	8.4	0	—
5	5.2	4	25	27.04	16	26.0	20	20.8	1	—
4	4.8	3	16	23.04	9	19.2	12	14.4	0	—
3	5.8	5	9	33.64	25	17.4	15	29.0	-1	—
16	20	14	66	101.36	54	79.4	55	72.6	0	6

① determin
② calcula

$$\frac{\Sigma(X_1-\bar{X}_1)(X_2-\bar{X}_2)}{\sqrt{\Sigma(X_1-\bar{X}_1)^2}\sqrt{\Sigma(X_2-\bar{X}_2)^2}}$$

a) Mean vector.

$$\bar{X}_i = \frac{1}{n}\sum_{j=1}^{n} X_{ij}$$

$\bar{X}_1 = 1/4 \cdot 16 = 4$, $\bar{X}_2 = 20/4 = 5$, $\bar{X}_3 = 14/4 = 3.5$

$$\bar{X} = \begin{bmatrix} \bar{X}_1 \\ \bar{X}_2 \\ \bar{X}_3 \end{bmatrix} = \begin{bmatrix} 4 \\ 5.0 \\ 3.5 \end{bmatrix}$$

b) Variance - covariance matrix.

$$S_n = \frac{1}{n-1}\left\{\sum_{j=1}^{n} X_{ij} X_{kj} - n\bar{X}_i \bar{X}_k\right\}$$

$S_{11} = 1/3\left[66.0 - 4\left(16/4 \times 16/4\right)\right] = 1/3\left[66.0 - 64\right] = 0.667$

$S_{12} = 1/3\left[79.4 - 4\left(16/4 \times 20/4\right)\right] = 1/3\left[79.4 - 80\right] = -0.200$

$S_{22} = 1/3\left[101.36 - 4\left(20/4 \times 20/4\right)\right] = 1/3\left[101.36 - 100\right] = 0.453$

$S_{13} = 1/3\left[55.0 - 4\left(16/4 \times 14/4\right)\right] = 1/3\left[55.0 - 56\right] = -0.333$

$S_{23} = 1/3\left[72.6 - 4\left(20/4 \times 14/4\right)\right] = 1/3\left[72.6 - 70\right] = 0.867$

$S_{33} = 1/3\left[54.0 - 4\left(14/4 \times 14/4\right)\right] = 1/3\left[54.0 - 49\right] = 1.667$

$$S_n = \begin{bmatrix} S_{11} & S_{12} & S_{13} \\ S_{21} & S_{22} & S_{32} \\ S_{31} & S_{32} & S_{33} \end{bmatrix} = \begin{bmatrix} 0.667 & -0.200 & -0.333 \\ -0.200 & 0.453 & 0.867 \\ -0.333 & 0.867 & 1.667 \end{bmatrix}$$

c) Correlation Matrix.

$$r_{ik} = \frac{S_{ik}}{\sqrt{S_{ii}}\sqrt{S_{kk}}}$$

As done in the previous solution

$$r_n = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} = \begin{bmatrix} 1.000 & -0.364 & -0.316 \\ -0.364 & 1.000 & 0.997 \\ -0.316 & 0.997 & 1.000 \end{bmatrix}$$

Example 2.2

The data for xtics X_1, X_2 & X_3 were collected and recorded in the table as shown below Determine sample mean vector $\bar{x}$, variance covariance matrix, S_n, and correlation coefficient matrix r_n for X_1, X_2 and X_3 variables.

Solution.

X_1	X_2	X_3	X_1^2	X_2^2	X_3^2	X_1X_2	X_1X_3	X_2X_3
0.80	21.4	7	0.64	457.96	49	17.120	5.6	149.8
0.82	268	7	0.6724	718.24	49	21.976	5.74	187.6
0.84	29.2	7.8	0.7056	852.64	60.84	24.528	6.552	227.76
0.82	31.8	7.5	0.6724	1011.24	56.25	26.076	6.15	238.5
0.80	36.0	7	0.64	1296	49	28.600	5.6	252
0.70	31.5	3.8	0.49	992.25	14.44	22.050	2.66	119.7
0.82	268	7	0.6724	718.24	49	21.976	5.74	187.6
0.80	15.2	7.2	0.64	231.04	51.84	12.160	5.76	109.44
0.82	28.0	8	0.6724	784	64	22.960	6.56	224
0.80	26.0	7.6	0.64	676	57.76	20.80	6.08	197.6
0.82	18.4	7	0.6724	338.56	49	15.088	5.74	128.8
0.80	14.2	7.2	0.64	201.64	51.84	11.360	5.76	102.24
0.80	20.4	6.8	0.64	416.16	46.24	16.320	5.44	138.72
0.80	15.8	6.8	0.64	249.64	46.24	12.640	5.44	107.44
0.82	18.5	7.2	0.6724	342.25	51.84	15.170	5.904	133.2
12.06	360.0	104.9	9.710	9285.86	746.29	289.02	84.726	2504.40

a) Mean Vector

$\bar{X}_1 = \frac{12.06}{15} = 0.80$, $\bar{X}_2 = \frac{360}{15} = 24.00$, $\bar{X}_3 = \frac{104.9}{15} = 6.99$

$$\bar{X} = \begin{bmatrix} 0.80 \\ 24.00 \\ 6.99 \end{bmatrix}$$

b) Variance - Covariance Matrix: $X_{ik} = \frac{1}{n-1}\left\{\sum_{j=1}^{n} X_{ij}X_{kj} - n\bar{X}_i\bar{X}_k\right\}$

$S_{11} = \frac{1}{14}\left[9.71 - 15\left(\frac{12.06}{15} \cdot \frac{12.06}{15}\right)\right] = \frac{1}{14}[9.71 - 9.70] = 0.001$

$S_{12} = \frac{1}{14}[289.02 - 15 \times 0.8 \times 24)] = \frac{1}{14}[289.02 - 289.44] = -0.030$

$S_{22} = \frac{1}{14}[9285.86 - 15 \times 24 \times 24)] = \frac{1}{14}[9285.86 - 8640] = 46.133$

$S_{13} = \frac{1}{14}[84.726 - 15 \times 0.8 \times 6.99)] = \frac{1}{14}[84.726 - 84.34] = 0.028$

$S_{23} = \frac{1}{14}[2504.4 - 15 \times 24 \times 6.99] = \frac{1}{14}[2504.4 - 2517.60] = -0.943$

$S_{33} = \frac{1}{14}[746.29 - 15 \times 6.99 \times 6.99] = \frac{1}{14}[746.29 - 733.66] = 0.906$

$$S_n = \begin{bmatrix} S_{11} & S_{12} & S_{13} \\ S_{21} & S_{22} & S_{23} \\ S_{31} & S_{32} & S_{33} \end{bmatrix} = \begin{bmatrix} 0.001 & -0.030 & 0.028 \\ -0.030 & 46.133 & -0.943 \\ 0.028 & -0.943 & 0.906 \end{bmatrix}$$

c) Correlation Matrix.

$$r_{11} = \frac{S_{11}}{\sqrt{S_{11}}\sqrt{S_{11}}} = \frac{0.001}{\sqrt{0.001}\sqrt{0.001}} = 1.00, \quad r_{22} = 1.00, \quad r_{33} = 1.00$$

$$r_{12} = \frac{S_{12}}{\sqrt{S_{11}}\sqrt{S_{22}}} = \frac{-0.030}{\sqrt{0.001}\sqrt{46.133}} = -0.140$$

$$r_{13} = \frac{S_{13}}{\sqrt{S_{11}}\sqrt{S_{33}}} = \frac{0.028}{\sqrt{0.001}\sqrt{0.906}} = 0.925$$

$$r_{23} = \frac{S_{23}}{\sqrt{S_{22}}\sqrt{S_{33}}} = \frac{-0.9043}{\sqrt{46.133}\sqrt{0.906}} = -0.146$$

$$r = \begin{bmatrix} 1.000 & -0.140 & 0.925 \\ -0.140 & 1.000 & -0.146 \\ 0.925 & -0.146 & 1.000 \end{bmatrix}$$

3. TEST FOR MEAN VECTOR AND HOTELLING T^2 TESTS.

Test for Mean Vector.

Consider testing the hypothesis in the univariate case given as:

Hypothesis (two-sided) $H_0: \mu = \mu_0$ against $H_0: \mu = \mu_0$

$H_a: \mu \neq \mu_0$ at α l.o.s $H_0: \mu \neq \mu_0$

- If $X_1, X_2, \ldots, X_n$ are ē observations from a normal distribution, the appropriate test statistic is given as:

$$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \quad \text{where: } \bar{X} = \frac{1}{n}\sum_{j=1}^{n} X_j \text{ and } S^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X})^2 \quad \ldots (3.1)$$

$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$ $\quad t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$

- The test statistic has a student's t distribution with n-1 degrees of freedom. When testing ē hypothesis, we make ē decision of rejecting H_0, if ē observed $|t|$ exceeds a specified point of a t-distribution with n-1 d.f. Rejecting H_0 when $|t|$ is large is equivalent to rejecting H_0 for a large square given as

$$t^2 = \frac{(\bar{X} - \mu_0)^2}{S^2/n} = n(\bar{X} - \mu_0)^2 (S^2)^{-1} (\bar{X} - \mu_0) \quad \ldots (3.2)$$

$t^2 = \frac{(\bar{x} - \mu_0)^2}{s^2/n}$ $\quad t^2 = \frac{(\bar{x} - \mu_0)^2}{s^2/n} = n(\bar{x} - \mu_0)^2 (s^2)^{-1} (\bar{x} - \mu_0)$

- The variance S^2 is the square of the distance from the sample mean $\bar{X}$ to ē test value μ_0. The units of distance are expressed in terms of $S/\sqrt{n}$ or estimated s.d of $\bar{X}$, once $\bar{X}$ & S^2 are observed. The test becomes, reject H_0 in favour of H_a at α l.o.s if:

$$n(\bar{X} - \mu_0)(S_0^2)^{-1}(\bar{X} - \mu_0) > t_{\alpha}^2/2\,(n-1) \quad \ldots (3.3)$$

Where $t^2_{\alpha/n}(n-1)$ denotes the upper $100(1-\alpha/2)^{th}$ percentile of the t dist with n-1 degrees d.f.

- If Ho is not rejected, i.e. $H_0: \mu = \mu_0$ at α l.o.s, then,

$$\frac{|\bar{X} - \mu_0|}{s/\sqrt{n}} \leq t_{\alpha/2}(n-1) \quad \text{---} (3.4)$$

(μ_0 lies in the $100(1-\alpha)$ percent confidence interval

$$\bar{X} \pm t_{\alpha/2}(n-1)\, s/\sqrt{n} \quad \text{or} \quad \bar{X} - t_{\alpha/2}(n-1)\, s/\sqrt{n} \leq \mu_0 \leq \bar{X} + t_{\alpha/2}(n-1)\, s/\sqrt{n} \quad \text{---} (3.5)$$

- In multivariate analysis we consider ē problem of determining whether a given $p \times 1$ vector where, μ_0 is the mean of a multivariate normal dist, i.e, test ē hypothesis.

- Hypothesis: $H_0: \mu_1 = \mu_{10}, \mu_2 = \mu_{20} \ldots \mu_p = \mu_{p0}$ against

$H_a: \mu_i \neq \mu_{i0}$ for at least one value of i at α l.o.s

- The hypothesis is rewritten as:

$$H_0: \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} = \begin{bmatrix} \mu_{10} \\ \mu_{20} \\ \vdots \\ \mu_{p0} \end{bmatrix}$$

against $H_a: \mu_i \neq \mu_{i0}$ for at least one value of i $(i = 1, 2, \ldots, p)$ at α l.o.s

--- (3.6)

- A natural generalization of ē squared distance in eqtn (3.2) is its multivariate analog

$$T^2 = (\bar{X} - \mu_0)' \left(\tfrac{1}{n}\Sigma\right)^{-1} (\bar{X} - \mu_0) = n(\bar{X} - \mu_0)' \Sigma^{-1} (\bar{X} - \mu_0)$$

Where;

$$\underset{(p\times 1)}{\bar{X}} = \begin{bmatrix} \bar{X}_1 \\ \bar{X}_2 \\ \vdots \\ \bar{X}_p \end{bmatrix} = \bar{X}_i = \frac{1}{n}\sum_{i=1}^{n} X_{ij} \quad \underset{(p\times p)}{S} = \frac{1}{n-1}\sum_{j=1}^{n} (X_{ij} - \bar{X}_i)'(X_{ij} - \bar{X}_i) \text{ and } \underset{(p\times 1)}{\mu_0} = \begin{bmatrix} \mu_{10} \\ \mu_{20} \\ \vdots \\ \mu_{p0} \end{bmatrix}$$

Exam: Derive. How to calculate it

- The T^2 is called Hotelling's T^2 in honour of Harold Hotelling, a pioneer in multivariate analysis who first obtained its sampling dist. If the observed statistical distance T^2 is too large, that is, if $\bar{X}$ is "too far" from μ_0, the hypothesis

$$H_0: \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} = \begin{bmatrix} \mu_{10} \\ \mu_{20} \\ \vdots \\ \mu_{p0} \end{bmatrix}$$ is rejected. There are no special tables of T^2 percentage points

random variable (r.v)

required for formal tests of hypothesis because T^2 is distributed as $\frac{(n-1)p}{n-p} F_\alpha(p, n_0 - p)$ where $F(p, n_0 - p)$ denote a r.v with p & $n-p$ d.f from an F- dist.

$\frac{(n-1)p}{n-p} F_\alpha(p, n-p)$

Exq

Example 3.1

The data matrix for a random sample of size $n=3$ from a bivariate normal ppltn is given as $X = \begin{bmatrix} 7 & 9 \\ 9 & 6 \\ 8 & 3 \end{bmatrix}$. State the sampling dist of T^2 and determine whether $\bar{e}$ sample is drawn from a multivariate normal dist with mean, $\mu = \begin{bmatrix} 9 \\ 5 \end{bmatrix}$ at $\alpha = 0.01$ $1.0.5$

Solution

Hypothesis: $H_0: \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} \mu_{10} \\ \mu_{20} \end{bmatrix}$ against $H_a: \mu_i \neq \mu_{i0}$ for at least 1 value of $i (i=1,2,\dots p)$ at $\alpha = 0.01$ $1.0.5$

Test statistic;

X_1	X_2	X_1^2	X_2^2	$X_1 X_2$
7	9	49	81	63
9	6	81	36	54
8	3	64	9	24
24	18	194	126	141

$\bar{X}_1 = {}^{24}/_{3} = 8$, $\bar{X}_2 = {}^{18}/_{3} = 6$

$\bar{X} = \begin{bmatrix} 8 \\ 6 \end{bmatrix}$

$S_{ik} = {}^{1}/_{n-1} \left\{ \sum_{j=1}^{n} x_{ij} x_{kj} - n \bar{x}_i \bar{x}_k \right\}$

$S_{11} = {}^{1}/_{2} \left[194 - 3(8)^2 \right] = 1.0$

$S_{12} = {}^{1}/_{2} \left[141 - 3(8)(6) \right] = -1.5$

$S_{22} = {}^{1}/_{2} \left[126 - 3(6)^2 \right] = 9.0$

$S_n = \begin{bmatrix} 1.0 & -1.5 \\ -1.5 & 9.0 \end{bmatrix}$

$T^2 = n(\bar{x} - \mu_0)' \Sigma^{-1} (\bar{x} - \mu_0) = 3 \begin{bmatrix} -1 & 1 \end{bmatrix} \frac{1}{6.75} \begin{bmatrix} 9.0 & 1.5 \\ 1.5 & 1.0 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \end{bmatrix}$

$= 3 \begin{bmatrix} -1 & 1 \end{bmatrix} {}^{1}/_{6.75} \begin{bmatrix} -7.5 \\ -0.5 \end{bmatrix} = {}^{28}/_{9}$

$= 3.111$, that has approximately F distribution.

Tabulated

$T^2 = \frac{(n-1)p}{n-p} F_\alpha (p, n-p) = \frac{2(2)}{3-2} F_{0.01}(2,1)$

$= 4 \frac{4(4999.5)}{4(4999)} = \frac{19996}{19996}$ 19998.

Decision:

Since $T^2 < 4 F_{0.01}(2,1)$, fail to reject H_0 : $\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} \mu_{10} \\ \mu_{20} \end{bmatrix}$ and conclude that the sample is from a multivariate normal dist with mean vector.

Exq. 3.2

The data of the variables X_1, X_2 & X_3 were collected & recorded. Determine whether the sample is drawn from multivariate normal dist with mean vector μ_0' = [3 6 3] at $\alpha = 0.01$ L.O.S

Solution

Hypothesis: $H_0: \begin{bmatrix}\mu_1\\ \mu_2\\ \mu_3\end{bmatrix} = \begin{bmatrix}\mu_{10}\\ \mu_{20}\\ \mu_{30}\end{bmatrix}$ against $H_a: \mu_i \neq \mu_{io}$ at $\alpha = 0.01$ L.O.S

Test statistic

X_1	X_2	X_3	X_1^2	X_1X_2	X_1X_3	X_2^2	X_2X_3	X_3^2
4	4.2	2	16	16.8	8	17.64	8.4	4
5	5.2	4	25	26	20	27.04	20.8	16
4	4.8	3	16	19.2	12	23.04	14.4	9
3	5.8	7	9	17.4	21	33.64	40.6	49
16	20	16	66	79.4	61	101.36	84.2	78

$\bar{X}_1 = {}^{16}/_4 = 4$, $\bar{X}_2 = {}^{20}/_4 = 5$, $\bar{X}_3 = {}^{16}/_4 = 4$

$\bar{X} = \begin{bmatrix}4\\5\\4\end{bmatrix}$

$S_{ik} =$

$S_{11} = {}^1/_{n-1}\left\{\sum_{j=1}^{n} X_{1j}^2 - n\bar{x}_1^2\right\} = {}^1/_3\left[66 - 4(4)^2\right] = 0.67$

$S_{12} = {}^1/_3\left[79.4 - 4(4\times5)\right] = -0.20$

$S_{22} = {}^1/_3\left[101.36 - 4(5)^2\right] = 0.45$

$S_{13} = {}^1/_3\left[61 - 4(4\times4)\right] = -1.00$

$S_{23} = {}^1/_3\left[84.2 - 4(5\times4)\right] = 1.40$

$S_{33} = {}^1/_3\left[78 - 4(4)^2\right] = 4.67$

$S_n = \begin{bmatrix}0.67 & -0.20 & -1.00\\ -0.20 & 0.45 & 1.40\\ -1.00 & 1.40 & 4.67\end{bmatrix}$

Determine the inverse of S_n

Determinant,

$D = 0.667\{0.453(4.667) - 1.40(1.40)\} + 0.20\{-0.20(4.667) + 1.00(1.40)\}$
$-1.00\{-0.20(1.40) + 1.00(0.453)\}$
$= 0.024$

Using calc to get cofactors. Input a matrix

Step 4 Det (Mat A) x Trsp (Mat A)$^{-1}$

cofactor of element a_{ij} $(-1)^{i+j}$ times minor of a_{ij}

$$\text{Cofactors} = \begin{bmatrix} 0.156 & -0.467 & 0.173 \\ -0.467 & 2.111 & -0.733 \\ 0.173 & -0.733 & 0.262 \end{bmatrix}$$

$$S_n^{-1} = \frac{1}{0.024}\begin{bmatrix} 0.156 & -0.467 & 0.173 \\ -0.467 & 2.111 & -0.733 \\ 0.173 & -0.733 & 0.262 \end{bmatrix} = \begin{bmatrix} 6.6 & -19.7 & 7.3 \\ -19.7 & 89.1 & -30.9 \\ 7.3 & -30.9 & 11.1 \end{bmatrix}$$

$$T^2 = n(\bar{X} - \mu_0)' \Sigma^{-1} (\bar{X} - \mu_0)$$

$$= 4\begin{bmatrix} 1 & -1 & 1 \end{bmatrix}\begin{bmatrix} 6.6 & -19.7 & 7.3 \\ -19.7 & 89.1 & -30.9 \\ 7.3 & -30.9 & 11.1 \end{bmatrix}\begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} = 4\begin{bmatrix} 33.6 & -139.7 & 49.3 \end{bmatrix}\begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}$$

$$= 890.25$$

Tabulated,

$$T^2 = \frac{(n-1)p}{n-p} F_\alpha(p, n-p) = \frac{3(3)}{4-3} F_{0.01}(3,1)$$

$$= 9(5403) \qquad = 48627$$

9(5403.4) 48630.6

Decision;

Since $T^2 < 9.0$ ~~(5403)~~ $F_{0.01}(3,1)$, fail to reject H_0: $\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = \begin{bmatrix} \mu_{10} \\ \mu_{20} \\ \mu_{30} \end{bmatrix}$ and conclude that the sample is from a multivariate normal dist with $\mu_0 = [3\ 6\ 3]$ at $\alpha = 0.01$ l.o.s

Example 3.3

The data of the variable X_1, X_2 & X_3 were collected and recorded in a table as shown below. Determine whether the sample is drawn from a multivariate normal dist with mean vector $\mu_0 = [4\ 50\ 10]$ at $\alpha = 0.01$ l.o.s

Solution:

Hypothesis: H_0: $\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = \begin{bmatrix} \mu_{10} \\ \mu_{20} \\ \mu_{30} \end{bmatrix}$ against H_a: $\mu_i \neq \mu_{i0}$ at $\alpha = 0.01$ l.o.s

Test statistics;

X_1	X_2	X_3	X_1^2	X_1X_2	X_1X_3	X_2^2	X_2X_3	X_3^2
4	36	12	16	144	48	1296	432	144
5	54	12	25	270	60	2916	648	144
4	26	10	16	104	40	676	260	100
5	40	10	25	200	50	1600	400	100
2	20	10	4	40	20	400	200	100
8	54	14	64	432	112	2916	756	196
5	48	16	25	240	80	2304	768	256
7	46	12	49	322	84	2116	552	144
4	36	12	16	144	48	2296	432	144
6	40	12	36	240	72	1600	480	144
50	400	120	276	2136	614	17120	4928	1472

$\bar{X}_1 = {}^{50}/_{10} = 5.0$, $\bar{X}_2 = {}^{400}/_{10} = 40$, $\bar{X}_3 = {}^{120}/_{10} = 12.0$

$\bar{X} = \begin{bmatrix} 5 \\ 40 \\ 12 \end{bmatrix}$

$S_{11} = {}^1/_{n-1}\left[\sum_{j=1}^{n} X_{1j}^2 - n\bar{X}_1^2\right] = {}^1/_9\left[276 - 10\left({}^{50}/_{10}\right)^2\right] = 2.9$

$S_{12} = {}^1/_{n-1}\left[\sum_{j=1}^{n} X_{1j}X_{2j} - n\bar{X}_1\bar{X}_2\right] = {}^1/_9\left[2136 - 10(5)(40)\right] = 15.1$

$S_{13} = {}^1/_9\left[614 - 10(5)(12)\right] = 1.6$

$S_{22} = {}^1/_9\left[17120 - 10(40)^2\right] = 124.4$

$S_{23} = {}^1/_9\left[4928 - 10(40)(12)\right] = 14.2$

$S_{33} = {}^1/_9\left[1472 - 10(12)^2\right] = 3.6$

Determine the inverse of S_n

Determinant, $D = 2.89\{124.44(3.56) - 14.22(14.22)\} - 15.1\{(15.1(3.56) - 1.56(14.22)\}$
$-1.56\{15.11(14.22) - 124.44(1.56)\}$

$= 249.35$

Cofactors, $C = \begin{bmatrix} 240.2 & -31.6 & 21.33 \\ -31.6 & 7.85 & -17.6 \\ 21.33 & -17.6 & 131.2 \end{bmatrix}$

$S_n^{-1} = \frac{1}{249.5}\begin{bmatrix} 240.2 & -31.6 & 21.33 \\ -31.6 & 7.85 & -17.6 \\ 21.33 & -17.6 & 131.2 \end{bmatrix} = \begin{bmatrix} 0.963 & -0.127 & 0.086 \\ -0.127 & 0.031 & -0.070 \\ 0.086 & -0.070 & 0.526 \end{bmatrix}$

$\bar{x} - \mu_0 = \begin{bmatrix} 40 \\ 12 \end{bmatrix} - \begin{bmatrix} 50 \\ 10 \end{bmatrix} = \begin{matrix} -10 \\ 2 \end{matrix}$

$T^2 = n(\bar{x} - \mu_0)' \Sigma^{-1} (\bar{x} - \mu_0)$

$= 10 \begin{bmatrix} 1 & -10 & 2 \end{bmatrix} \begin{bmatrix} 0.963 & -0.127 & 0.086 \\ -0.127 & 0.031 & -0.070 \\ 0.086 & -0.070 & 0.526 \end{bmatrix} \begin{bmatrix} 1 \\ -10 \\ 2 \end{bmatrix}$

$= \begin{bmatrix} 24 & -5.8 & 18.4 \end{bmatrix} \begin{bmatrix} 1 \\ -10 \\ 2 \end{bmatrix}$

$= 119.06$

Tabulated,

$T^2 = \frac{(n-1)p}{n-p} F_\alpha(p, n-p) = \frac{9(3)}{10-3} F_{0.01}(3, 7)$

$= 3.86(8.45) = 32.62$

Decision:

Since $T^2 > 3.86 F_{0.01}(3,7)$, reject $H_0 : \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = \begin{bmatrix} \mu_{10} \\ \mu_{20} \\ \mu_{30} \end{bmatrix}$ and conclude that the sample is not from a multivariate normal dist with $\mu_0 = [4 \ 50 \ 10]$ at $\alpha = 0.01$ i.e. 0.5

NOTE: The H_0 will be rejected if one or some combination of means differs too much from ẽ hypothesized values [4 50 10]. At this stage of analysis if we reject H_0 we would not know which of the hypothesized values would be supported by ẽ data.

Simultaneous Confidence Intervals:

- In order to obtain ẽ method of making inferences from sample, there is need to extend ẽ concept of univariate interval to multivariate confidence region. The rejection region for the mean of μ of p-dimensional normal population before ẽ sample is selected from the previous section is given as:

$p\left[n(\bar{x} - \mu_0)' S^{-1} (\bar{x} - \mu_0) > \frac{(n-1)p}{n-p} F_{\alpha/2}(p, n-p) \right] = \alpha/2$.

- The confidence region for ẽ mean μ of a p-dimensional normal pplτn is then given as:

$p\left[n(\bar{x} - \mu_0)' S^{-1} (\bar{x} - \mu_0) \leq \frac{(n-1)p}{n-p} F_{\alpha/2}(p, n-p) \right] = 1 - \alpha/2$.

- Whatever the values of ẽ unknown μ and Σ in other words $\bar{x}$ will be within

$\left[\frac{(n-1)p}{n-p} F_\alpha(p, n-p) \right]$

- We write the intervals as:

$$\bar{X}_1 - \sqrt{\frac{(n-1)p}{n-p} F_\alpha(p, n-p)} \sqrt{\frac{S_{11}}{n}} \leq \mu_1 \leq \bar{X}_1 + \sqrt{\frac{(n-1)p}{n-p} F_\alpha(p, n-p)} \sqrt{\frac{S_{11}}{n}}$$

$$\bar{X}_2 - \sqrt{\frac{(n-1)p}{n-p} \cdot F_\alpha(p, n-p)} \sqrt{\frac{S_{22}}{n}} \leq \mu_2 \leq \bar{X}_2 + \sqrt{\frac{(n-1)p}{n-p} F_\alpha(p, n-p)} \sqrt{\frac{S_{22}}{n}}$$

$$\bar{X}_p - \sqrt{\frac{(n-1)p}{n-p} F_\alpha(p, n-p)} \sqrt{\frac{S_{pp}}{n}} \leq \mu_p \leq \bar{X}_p + \sqrt{\frac{(n-1)p}{n-p} F_\alpha(p, n-p)} \sqrt{\frac{S_{pp}}{n}}$$

All hold simultaneously with confidence coefficient $1-\alpha$.

Exa 3.4

The data of the variables X_1 & X_2 for $n = 42$ were collected and statistics recorded;

$$\bar{X} = \begin{bmatrix} 0.564 \\ 0.603 \end{bmatrix}, \quad S = \begin{bmatrix} 0.0144 & 0.0117 \\ 0.0117 & 0.0146 \end{bmatrix}, \quad S^{-1} = \begin{bmatrix} 203.018 & -163.391 \\ -163.391 & 200.225 \end{bmatrix}$$

Determine the following,

a) Whether ē sample is drawn from multivariate normal dist with mean vector $\mu = \begin{bmatrix} 0.562 \\ 0.589 \end{bmatrix}$ at $\alpha = 0.05$ l·o·s

b) The 95% confidence interval for the mean.

Solution:

(a) Hypothesis: $H_0: \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} \mu_{10} \\ \mu_{20} \end{bmatrix}$ against $H_a: \mu_i \neq \mu_{i0}$ for at least 1 value of i at $\alpha = 0.05$ l·o·s

Test statistic;

$$T^2 = n(\bar{X} - \mu_0)' \Sigma^{-1} (\bar{X} - \mu_0)$$

$$= 42 \begin{bmatrix} 0.564 - 0.562, 0.603 - 0.589 \end{bmatrix} \begin{bmatrix} 203.018 & -163.391 \\ -163.391 & 200.225 \end{bmatrix} \begin{bmatrix} 0.002 \\ 0.014 \end{bmatrix}$$

$$= 1.27$$

Tabulated

$$T^2 = \frac{(n-1)p}{n-p} F_\alpha(p, n-p) = \frac{41(2)}{40} F_{0.05}(2, 40)$$

$$= \frac{82(3.23)}{40} = 6.62$$

Conclusion: Since test statistic T^2 < tabulated value, then $\bar{X}' = [0.564 \ 0.603]$ is in the acceptance region, therefore we fail to reject H_0 and conclude that the sample data is from the multivariate normal dist with mean, $\mu' = [0.562 \ 0.589]$ at $\alpha = 0.05$ l·o·s

95% C.I of the mean.

b) $\bar{X}_1 - \sqrt{\frac{(n-1)p}{n-p} F_\alpha(p, n-p)} \sqrt{\frac{S_{11}}{n}} \le \mu_1 \le \bar{X}_1 + \sqrt{\frac{(n-1)p}{n-p} F_\alpha(p, n-p)} \sqrt{\frac{S_{11}}{n}}$

$= 0.564 - \sqrt{\frac{41(2)(3.23)}{40}} \sqrt{\frac{0.0144}{42}} \le \mu_1 \le 0.564 + \sqrt{\frac{41(2)(3.23)}{40}} \sqrt{\frac{0.0144}{42}}$

$= (0.516, 0.612)$

$\bar{X}_2 - \sqrt{\frac{(n-1)p}{n-p} F_\alpha(p, n-p)} \sqrt{\frac{S_{22}}{n}} \le \mu_2 \le \bar{X}_2 + \sqrt{\frac{(n-1)p}{n-p} F_\alpha(p, n-p)} \sqrt{\frac{S_{22}}{n}}$

$= 0.603 - \sqrt{\frac{41(2)(3.23)}{40}} \sqrt{\frac{0.0146}{42}} \le \mu_2 \le 5.5 + 0.603 - \sqrt{\frac{41(2)(3.23)}{40}} \sqrt{\frac{0.0146}{42}}$

$= (0.555, 0.651)$

Exa 3.5

The data of the variables X_1, X_2 & X_3 for $n = 40$ were collected and recorded as below

$$\bar{X} = \begin{bmatrix} 5.4 \\ 5.5 \\ 2.5 \end{bmatrix}, \quad S = \begin{bmatrix} 5.6 & 6.0 & 2.2 \\ 6.0 & 7.6 & 2.4 \\ 2.2 & 2.4 & 2.4 \end{bmatrix}, \quad S^{-1} = \begin{bmatrix} 2.251 & -1.860 & -0.204 \\ -1.860 & 1.754 & -0.049 \\ -0.204 & -0.049 & 0.653 \end{bmatrix}$$

Determine the following:

a) Whether the sample is drawn from multivariate normal dist with mean vector $\mu = \begin{bmatrix} 5 \\ 5 \\ 2 \end{bmatrix}$ at $\alpha = 0.05$ & 0.05

b) The 95% C.I for the mean.

Solution

(a) Hypothesis: $H_0: \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = \begin{bmatrix} \mu_{10} \\ \mu_{20} \\ \mu_{30} \end{bmatrix}$ against $H_a: \mu_i \neq \mu_{i0}$ for at least 1 value of $i (i = 1, 2, 3)$ at $\alpha = 0.05$

Test-statistic

$T^2 = \frac{(n-1)p}{n-p} F_\alpha(p, n-p) = \frac{39(3)}{37} F_{0.05}(3, 37)$

$= \frac{117(2.86)}{37} = 9.04$

Conclusion:

The mean vector $\bar{X}' = [5.4 \ 5.5 \ 2.5]$ is in the acceptance region, therefore the sample data is from multivariate normal dist with mean vector $\mu' = [5 \ 5 \ 2]$ at $\alpha = 0.05$ & 0.05

(b) 95% C.I of the mean

$\bar{X}_1 - \sqrt{\frac{(n-1)p}{n-p} F_\alpha(p, n-p)} \sqrt{\frac{S_{11}}{n}} \le \mu_1 \le \bar{X}_1 + \sqrt{\frac{(n-1)p}{n-p} F_\alpha(p, n-p)} \sqrt{\frac{S_{11}}{n}}$

$$= 5.4 - \sqrt{\frac{39(3)(2.86)}{37}}\sqrt{\frac{5.6}{40}} \le \mu_1 \le 5.4 + \sqrt{\frac{39(3)(2.86)}{37}}\sqrt{\frac{5.6}{40}}$$

$$= (4.27, 6.63)$$

$$\bar{X}_2 : = 5.5 - \sqrt{\frac{39(3)(2.86)}{37}}\sqrt{\frac{7.0}{40}} \le \mu_2 \le 5.5 + \sqrt{\frac{39(3)(2.86)}{37}}\sqrt{\frac{7.0}{40}}$$

$$= (4.24, 6.76)$$

$$\bar{X}_3 : 2.5 - \sqrt{\frac{39(3)(2.86)}{37}}\sqrt{\frac{2.4}{40}} \le \mu_3 \le 2.5 + \sqrt{\frac{39(3)(2.86)}{37}}\sqrt{\frac{2.4}{40}}$$

$$= (1.76, 3.24)$$

Testing Equity of the Population Means.

A T^2-statistic for testing ē equality of vector means from 2 multivariate ppltns can be developed by analogy with ē univariate procedures. The T^2-statistic is appropriate for comparing responses from one set of experimental settings (ppltn 1) with independent responses from another set of experimental settings (ppltn 2). If possible ē experimental units should be randomly assigned to ē sets of experimental conditions. Randomization will, to some extent, mitigate ē effect of unit-to-unit variability in a subsequent comparison of treatments. Although, some precision is lost relative to paired comparisons, the inferences in ē two-ppltn cases are, ordinarily, applicable to a more general collection of experimental units simply because unit homogeneity is not required.

Table: The arrangement of observations on p (p = 1, 2, ...) variables.

Population	Sample	Mean	Variance
1	$X_{11}, X_{12}, \ldots, X_{1n_1}$	$\bar{X}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} X_{1j}$	$S_1 = \frac{1}{n_1 - 1}(X_{1j} - \bar{X}_1)'(X_{1j} - \bar{X}_1)$
2	$X_{21}, X_{22}, \ldots X_{2n_1}$	$\bar{X}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} X_{2j}$	$S_2 = \frac{1}{n_2 - 1}(X_{2j} - \bar{X}_2)'(X_{2j} - \bar{X}_2)$
⋮			
q	$X_{q1}, X_{q2}, \ldots X_{qn_2}$	$\bar{X}_q = \frac{1}{n_q}\sum_{j=1}^{n_q} X_{qj}$	$S_q = \frac{1}{n_q - 1}(X_{qj} - \bar{X}_2)'(X_{qj} - \bar{X}_q)$

- We wish to make inference about (mean vector of ppltn 1) − (mean vector of ppltn 2) = $\mu_1 - \mu_2$.

Hypothesis: $H_0 : \mu_1 = \mu_2$ against $H_a : \mu_1 \neq \mu_2$ or $H_0 : \mu_1 - \mu_2 = 0$ against $H_a : \mu_1 - \mu_2 \neq 0$

Assumptions

a) The sample $X_{11}, X_{12}, \ldots, X_{1n_1}$ is a random sample of size n_1 from a p-variate pplτn with mean μ_1 and covariance matrix Σ_1.

b) The sample $X_{21}, X_{22}, \ldots, X_{2n_2}$ is a random sample of size n_2 from a p-variate pplτn with mean μ_2 and covariance matrix Σ_2.

c) $X_{11}, X_{12}, \ldots, X_{1n_1}$ are independent $X_{21}, X_{22}, \ldots, X_{2n_2}$ when n_1 & n_2 are small-

d) Both pplτns are multivariate normal.

e) $\Sigma_1 = \Sigma_2$ (same variance - covariance matrix).

- The assumption that $\Sigma_1 = \Sigma_2$, is much stronger than its univariate counterpart. Here we are assuming that several pairs of variance & covariance are nearly equal. When $\Sigma_1 = \Sigma_2 = \Sigma$, $(X_{1j} - \bar{X}_1)'(X_{1j} - \bar{X}_1)$ is an estimate of $(n_1 - 1)\Sigma$ for pplτn 1 and $(X_{2j} - \bar{X}_2)'(X_{2j} - \bar{X}_2)$ is an estimate $(n_2 - 1)\Sigma$ for pplτn 2. Consequently we pool ē information in both samples in order to estimate ē common variance, Σ.

∴ We set,

$$S_{pooled} = \frac{1}{n_1 + n_2 - 2}\left\{(X_{1j} - \bar{X}_1)'(X_{1j} - \bar{X}_1) + (X_{2j} - \bar{X}_2)'(X_{2j} - \bar{X}_2)\right\}$$

$$= \frac{n_1 - 1}{n_1 + n_2 - 2} S_1 + \frac{n_2 - 1}{n_1 + n_2 - 2} S_2$$

- Since $(X_{1j} - \bar{X}_1)'(X_{1j} - \bar{X}_1)$ from pplτn 1 has $n_1 - 1$ d.f and $(X_{2j} - \bar{X}_2)'(X_{2j} - \bar{X}_2)$ from pplτn 2 has $n_2 - 1$ d.f, the divisor $n_1 + n_2 - 2$ is obtained by combining ē 2 components' d.f. If $X_{11}, X_{12}, \ldots, X_{1n_1}$ is a random sample of size n_1 from $N_p(\mu_1, \Sigma)$ and $X_{21}, X_{22}, \ldots, X_{2n_2}$ is an independent random sample of size n_2 from $N_p(\mu_2, \Sigma)$, then

$$T^2 = [\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)]' \left[\left(\tfrac{1}{n_1} + \tfrac{1}{n_2}\right) S_{pooled}\right]^{-1} [\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)]$$ is distributed as

$$\frac{(n_1 + n_2 - 2)p}{(n_1 + n_2 - p - 1)} F_{\alpha(p, n_1 + n_2 - p - 1)}$$ (Johnson & Wichern, 2005)

Confidence Interval for Difference in Means.

- The C.I for difference in means is given as:

$$\mu_1 - \mu_2 = \bar{X}_1 - \bar{X}_2 = \sqrt{\frac{(n_1 - n_2 - 2)p}{(n_1 + n_2 - p - 1)} F_{\alpha(p, n_1 + n_2 - p - 1)}} \left[\sqrt{\left(\tfrac{1}{n_1} + \tfrac{1}{n_2}\right) S_{11\,pooled}} \quad \sqrt{\left(\tfrac{1}{n_1} + \tfrac{1}{n_2}\right) S_{22\,pooled}}\right]$$

Where,

$$S_{pooled} = \frac{n_1 - 1}{n_1 + n_2 - 2} S_1 + \frac{n_2 - 1}{n_1 + n_2 - 2} S_2$$

$\left(\frac{1}{50} + \frac{1}{50}\right)^{-1} = \left(\frac{1}{25}\right)^{-1} = \frac{25}{9} = 9$

Exam

Exa 3.6

The data of samples from 2 ppltns were collected and recorded below for xtics X_1 & X_2.

$n_1 = 50, \quad \bar{X}_1 = \begin{bmatrix} 8.3 \\ 4.1 \end{bmatrix}, \quad S_1 = \begin{bmatrix} 2 & 1 \\ 1 & 6 \end{bmatrix}$

$n_2 = 50, \quad \bar{X}_2 = \begin{bmatrix} 10.2 \\ 3.9 \end{bmatrix}, \quad S_2 = \begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix}$

$\frac{1}{9}\begin{bmatrix} = \\ = \end{bmatrix}$

Determine the following:

a) Whether ē ppltn have ē same means at $\alpha = 0.05$ l·o·s

b) The 95% C.I for difference in means.

Solution:

Hypothesis: $H_0: \mu_1 = \mu_2$ against $H_a: \mu_1 \neq \mu_2$ or $H_0: \mu_1 - \mu_2 = 0$ against $H_a: \mu_1 - \mu_2 \neq 0$ at $\alpha = 0.05$

Test statistic;

$= \begin{matrix} 2 & 1 \\ 1 & 5 \end{matrix}$

$\bar{X}_1 - \bar{X}_2 = \begin{bmatrix} -1.9 \\ 0.2 \end{bmatrix}$

$\begin{bmatrix} 1 & 0.5 \\ 0.5 & 3 \end{bmatrix} + \begin{bmatrix} 1 & 0.5 \\ 0.5 & 2 \end{bmatrix}$

$S_{pooled} = \frac{n_1 - 1}{n_1 + n_2 - 2} S_1 + \frac{n_2 - 1}{n_1 + n_2 - 2} S_2 = \frac{49}{98}\begin{bmatrix} 2 & 1 \\ 1 & 6 \end{bmatrix} + \frac{49}{98}\begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix} = \begin{matrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{matrix}\begin{bmatrix} 2 & 1 \\ 1 & 5 \end{bmatrix}$

$T^2 = \left[\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)\right]' \left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S_{pooled}\right]^{-1} \left[\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)\right]$

$= \begin{bmatrix} -1.9 & 0.2 \end{bmatrix} \left[\left(\frac{1}{50} + \frac{1}{50}\right)\begin{bmatrix} 2 & 1 \\ 1 & 5 \end{bmatrix}\right]^{-1} \begin{bmatrix} -1.9 \\ 0.2 \end{bmatrix}$

$= \frac{25}{9} \begin{bmatrix} -1.9 & 0.2 \end{bmatrix} \begin{bmatrix} 5 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} -1.9 \\ 0.2 \end{bmatrix}$ (determinant)

$= \frac{25}{9} \begin{bmatrix} -9.7 & 2.3 \end{bmatrix} \begin{bmatrix} -1.9 \\ 0.2 \end{bmatrix}$

$= 52.47$

Tabulated value

$T^2 = \frac{(n_1 + n_2 - 2)p}{(n_1 + n_2 - p - 1)} F_\alpha (p, n_1 + n_2 - p - 1)$

$= \frac{98(2)}{97} F_{0.05}(2, 97) = \frac{98(2)(3.09)}{97} = 6.24$

Conclusion:

Since $T^2 > 6.26$, reject H_0 and conclude that ē ppltn do not have ē same means at $\alpha = 0.05$ l·o·s

b) 95% C.I

$$\mu_1 - \mu_2 = \bar{X}_1 - \bar{X}_2 = \sqrt{\frac{(n_1+n_2-2)p}{(n_1+n_2-p-1)} F_\alpha(p, n_1+n_2-p-1)} \left[\sqrt{(1/n_1 + 1/n_2) S_{11\,pooled}} \right]$$

$$= \begin{bmatrix} -1.9 \\ 0.2 \end{bmatrix} \pm \sqrt{\frac{98(2)}{97} F_{0.05}(2,97)} \left[\sqrt{2(1/50 + 1/50)} \right] = \begin{bmatrix} -1.9 \\ 0.2 \end{bmatrix} \pm \left[\sqrt{6.24(2)(2/50)} \right] 0.707$$

$$\sqrt{6.24(5)(2/50)} \; 1.117$$

$-1.9 - 0.707$ $\quad -1.9 +$ $\quad -1.9 + 0.707 =$

$$= \begin{bmatrix} -2.606 \le \mu_1 - \mu_2 \le -1.193 \\ -0.917 \le \mu_1 - \mu_2 \le 1.317 \end{bmatrix}$$

*Exercise.

4. PRINCIPAL COMPONENTS, DISCRIMINANT FUNCTIONS AND CANONICAL CORRELATION.

Principal Components.

- A principal component analysis is concerned with explaining the variance - covariance structure of variables through a few linear combination of these variables. The general objectives include ē following:

a) Data reduction

b) Interpretation.

- Although p components are required to reproduce ē total system variability, often much of this variability can be accounted for by a small number, k, of ē principal components. If so, there is almost as much information in the k components as there is in ē original p variables. The k principal components can then replace the initial p variables, and the original data set, consisting of n measurements on p variables, is reduced to a data set containing of n measurements on k principal components.

Population Principal Components.

Algebraically, principal components are particular linear combinations of the p r.vs $X_1, X_2, \ldots, X_p$. Geometrically, these linear combinations represent ē selection of a new coordinate system obtained by rotating ē original system with $X_1, X_2, \ldots, X_p$ as ē coordinate axes. The new axes represent the directions with maximum variability and provide a simpler and more parsimonious description of the covariate structure. The development does not require a multivariate normal assumption.

Let the random vector $X' = [X_1, X_2, \ldots, X_p]$ have ē variance - covariance matrix Σ with Eigen values $\lambda_1 \geq \lambda_2 \geq \ldots, \lambda_p \geq 0$. Consider ē linear combination

$Y_1 = a_1'X = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p$

$Y_2 = a_2'X = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p$

$\vdots$

$Y_p = a_p'X = a_{p1}X_1 + a_{p2}X_2 + \ldots + a_{pp}X_p$

We obtain;

$V(Y_i) = a_i'\Sigma a_i \quad , i = 1, 2, \ldots, p \qquad V(Y_i) = a_i'\Sigma a_i$

$Cov(Y_i, Y_k) = a_i'\Sigma a_k \quad , i, k = 1, 2, \ldots, p \qquad Cov(Y_i, Y_k) = a_i'\Sigma a_k$

- The principal components are uncorrelated linear combinations $Y_1, Y_2, \ldots, Y_p$ whose variances are large as possible.

First principal components = linear combination $a_1'X$ that maximizes $V(a_1'X)$ subject to $a_1'a_1 = 1$

Second " " = linear combination $a_2'X$ that maximizes $V(a_2'X)$ subject to $a_2'a_2 = 1$

and $Cov(a_1'X, a_2'X) = Cov(Y_1, Y_2) = 0$

i^{th} " " = linear combination $a_i'X$ that maximizes $V(a_i'X)$ subject to $a_i'a_i = 1$

and $Cov(a_i'X, a_k'X) = Cov(Y_i, Y_k) = 0$ for $k < i$

Exa 4.1

Determine ē principal components if ē random variables X_1, X_2 & X_3 have ē variance - covariance matrix = $\Sigma = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix}$

Solution:

Determine ē eigen values using the relation $A = \lambda_1$ such that $|A - \lambda_1| = 0$

$$\begin{vmatrix} 1-\lambda & -2 & 0 \\ -2 & 5-\lambda & 0 \\ 0 & 0 & 2-\lambda \end{vmatrix} = (1-\lambda)\{(5-\lambda)(2-\lambda)\} - 2(2)(2-\lambda) = 0$$

$= 2\lambda^2 - 12\lambda + 2 - \lambda^3 + 6\lambda^2 - \lambda = -\lambda^3$

$2(\lambda^2 - 6\lambda + 1) - \lambda(\lambda^2 - 6\lambda + 1)$

$2(\lambda^2 - 6\lambda + 1) - \lambda(\lambda^2 - 6) + 1)$

$(2-\lambda)(5 - 6\lambda + \lambda^2 - 4) = (2-\lambda)(\lambda^2 - 6\lambda + 1) = 0$ $\quad 2\lambda^2 - 12\lambda + 2 - \lambda^3 + 6\lambda^2 - \lambda$

$\lambda_1 = 2$ and, $\quad 2\lambda^2 - 12\lambda + 2 - \lambda^3 + 6\lambda^2 - \lambda$

$-\lambda^3 + 8\lambda^2 - 13\lambda + 2$

$-\lambda^3 + 6\lambda^2 - 12\lambda + 2$

$\lambda = \dfrac{6 \pm \sqrt{36-4}}{2} = 3 \pm \sqrt{8} = 3 \pm 2.828$

such that,

$\lambda_2 = 5.828$, $\lambda_3 = 0.172$

Eigen Vectors

$Ax = \lambda x$

a) $\lambda_1 = 5.828$

$$\begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = 5.828 \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}$$

$1X_1 - 2X_2 = 5.828 X_1$ such that $4.828X_1 = -2X_2$ ... (i)

$-2X_1 + 5X_2 = 5.828 X_2$ such that $2X_1 = 0.828X_2$ ... (ii)

$2X_3 = 5.828 X_3$ such that $X_3 = 0$ ... (iii)

consider eqtn (i)

$1X_1 = \dfrac{-2X_2}{4.828}$ $X_1 = -0.414 X_2$ ... (iv)

Let $X_2 = 1$ such that $X_1 = -0.414$ then,

$$e_1 = \begin{bmatrix} -0.414 \\ 1.000 \\ 0.000 \end{bmatrix}$$ such that $$e_1 = \begin{bmatrix} -0.383 \\ 0.924 \\ 0.000 \end{bmatrix}$$ is unit vector.

$e = \dfrac{X}{\sqrt{X'X}} = \dfrac{-0.414}{\sqrt{0.414^2 + 1^2}} = \dfrac{1}{\sqrt{0.414^2 + 1^2}}$

b) $\lambda_2 = 2$

$$\begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = 2 \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}$$

$1X_1 - 2X_2 = 2X_1$ such that $X_1 = -2X_2$ ... (i)

$-2X_1 + 5X_2 = 2X_2$ such that $2X_1 = 3X_2$ ... (ii)

$2X_3 = 2X_3$ such that $X_3 = 0$ ... (iii)

consider eqtn (iii)

Let $X_3 = 1$

consider eqtn (ii)

$X_1 = 3/2 X_2$ ... (iv)

Let $X_2 = 0$ such that $X_1 = 0$ then, $e_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$

c) $\lambda_3 = 0.172$

$$\begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = 0.172 \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}$$

$1X_1 - 2X_2 = 0.172 X_1$ such that $0.828 X_1 = 2 X_2$ ---ii)

$-2X_1 + 5X_2 = 0.172 X_2$ " " $2 X_1 = 4.828 X_2$ ---(iii)

$2X_3 = 0.172 X_3$ " " $X_3 = 0$ ---(iii)

consider eqtn (i)

$$1X_2 = \frac{0.828 X_1}{2} = 0.414 X_1 \quad \text{--- (iv)}$$

Let $X_1 = 1$ such that $X_2 = 0.414$ then,

$C_3 = \begin{bmatrix} 1.000 \\ 0.414 \\ 0.000 \end{bmatrix}$ such that $e_3 = \begin{bmatrix} 0.924 \\ 0.383 \\ 0.000 \end{bmatrix}$ is unit vector.

Therefore, the principal components for $\lambda_1 \geq \lambda_2 \geq \lambda_3$ are

$Y_1 = e_1' X = -0.383 X_1 + 0.924 X_2$ $Y_1 = e_1' X$

$Y_2 = e_2' X = X_3$ $Y_2 = e_2' X$

$Y_3 = e_3' X = 0.924 X_1 + 0.383 X_2$ $Y_3 = e_3' X$

The variable X_3 is one of the principal components because it is uncorrelated with the other two variables.

$V(Y_1) = V(-0.383 X_1 + 0.924 X_2) = V\left\{[-0.383 \ 0.924] \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}\right\}$

$a' \Sigma a$

$= [-0.383 \ 0.924] \begin{bmatrix} 1 & -2 \\ -2 & 5 \end{bmatrix} \begin{bmatrix} -0.383 \\ 0.924 \end{bmatrix} = [-0.383 \ 0.924] \begin{bmatrix} -2.231 \\ 5.386 \end{bmatrix}$

$= 5.83$

$V(Y_2) = V(X_3) = 2.00$

$Cov(Y_1, Y_2) = Cov(-0.383 X_1 + 0.924 X_2, X_3) = Cov\left\{[-0.383 \ 0.924] \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}, X_3\right\}$

$Cov(Y_i, Y_k)$

$a_i \Sigma a_k$

$= [-0.383 \ 0.924] \begin{bmatrix} 0 \\ 0 \end{bmatrix} = 0$

$V(Y_3) = V(0.924 X_1 + 0.383 X_2) = V\left\{[0.924 \ 0.383] \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}\right\}$

$= [0.924 \ 0.383] \begin{bmatrix} 1 & -2 \\ -2 & 5 \end{bmatrix} \begin{bmatrix} 0.924 \\ 0.383 \end{bmatrix} = [0.924 \ 0.383] \begin{bmatrix} 0.158 \\ 0.067 \end{bmatrix}$

$= 0.172.$

It is also readily apparent that,

$\sigma_{11} + \sigma_{22} + \sigma_{33} = 1 + 5 + 2 = 5.828 + 2 + 0.172$

$= 8$

The proportion of total variances accounted for by,

The first principal components is $\frac{\lambda_1}{\lambda_1+\lambda_2+\lambda_3} = \frac{5.828}{5.828+2+0.172} = 0.73$

The first 2 principal components is $\frac{\lambda_1+\lambda_2}{\lambda_1+\lambda_2+\lambda_3} = \frac{5.83+2}{5.828+2+0.172} = 0.98$ of the population variances.

In this case, the components Y_1 and Y_2 could replace the original 3 variables with little loss of information.

$$\rho_{Y_1 X_1} = \frac{c_{11}\sqrt{\lambda_1}}{\sqrt{\sigma_{11}}} = \frac{-0.383\sqrt{5.83}}{\sqrt{1}} = -0.925$$

$$\rho_{Y_1 X_2} = \frac{c_{12}\sqrt{\lambda_1}}{\sqrt{\sigma_{22}}} = \frac{0.924\sqrt{5.83}}{\sqrt{5}} = 0.998$$

Remarks:

- The variable X_2, with coefficient 0.998, records the greatest weight in component Y_1. It also has the largest correlations (in absolute value) with Y_1.
- The correlations of X_1 with Y_1, -0.925 is almost as large as that for X_2, indicating that the variables are about equally important to the principal components. The relative sizes of the coefficients of X_1 & X_2 suggest, however, that X_2 contributes more to the determination of Y_1 than does X_1. Since, in this case both coefficients are reasonably large and they have opposite signs, one would argue that both variables aid in the interpretation of Y_2.
- Finally $\rho_{Y_2 X_1} = \frac{c_{21}\sqrt{\lambda_2}}{\sqrt{\sigma_{11}}} = 0$, $\rho_{Y_2 X_2} = \frac{c_{22}\sqrt{\lambda_2}}{\sigma_{22}} = 0$ & $\rho_{Y_2 X_3} = \frac{c_{23}\sqrt{\lambda_2}}{\sigma_{33}} = \frac{\sqrt{2}}{\sqrt{2}} = 1$ (as it should)

- The remaining correlations can be neglected since the 3rd component is unimportant.

- Variables should probably be standardized if they are measured on scales with widely differing ranges or if the units of measurement are not commensurate. If variables are standardized their subsequent magnitudes will be of the same order.
- The principal components derived from Σ are different from those derived from R. Furthermore, one set of principal components is not a simple function of the other. This suggests that standardization is not inconsequential.

Exa 4.2

The data of the variables X_1, X_2 & X_3 were collected & recorded.

Determine the principal components, loadings, scores and interpret the results.

Solution:

X_1	X_2	X_3	X_1^2	X_1X_2	X_1X_3	X_2^2	X_2X_3	X_3^2
7	4	2	49	28	14	16	8	4
5	6	1	25	30	5	36	6	1
4	4	2	16	16	8	16	8	4
4	4	2	16	16	8	16	8	4
5	2	3	25	10	15	4	6	9
25	20	10	131	100	50	88	36	22

$\bar{X}_1 = 25/5 = 5.0$, $\bar{X}_2 = 20/5 = 4.0$, $\bar{X}_3 = 10/5 = 2.0$

$\bar{X} = \begin{bmatrix} 5 \\ 4 \\ 2 \end{bmatrix}$

Covariance; $S_{ik} = \frac{1}{n-1} \left\{ \sum_{j=1}^{n} X_{ij} X_{kj} - n \bar{X}_i \bar{X}_k \right\}$

$S_{11} = \frac{1}{4} \{ 131 - 5(5)^2 \} = 1.5$, $S_{12} = \frac{1}{4} \{ 100 - 5(5)(4) \} = 0$

$S_{13} = \frac{1}{4} \{ 50 - 5(5)(2) \} = 0$, $S_{22} = \frac{1}{4} \{ 88 - 5(4)^2 \} = 2$

$S_{23} = \frac{1}{4} \{ 36 - 5(4)(2) \} = -1$, $S_{33} = \frac{1}{4} \{ 22 - 5(2)^2 \} = 0.5$

$$S_n = \begin{bmatrix} 1.5 & 0 & 0 \\ 0 & 2 & -1 \\ 0 & -1 & 0.5 \end{bmatrix}$$

Eigen Values

Solve for λ in ē equation $|\lambda - \lambda I| = 0$.

$$\begin{bmatrix} 1.5-\lambda & 0 & 0 \\ 0 & 2-\lambda & -1 \\ 0 & -1 & 0.5-\lambda \end{bmatrix} = 0$$

$0(0-(-1)) \quad (0-1.5)$

$(1.5-\lambda)\{(2.0-\lambda)(0.5-\lambda) - 1.0\} = 0$

$(1.5-\lambda)\{(\lambda^2 - 2.5\lambda)\} = 0$

$\lambda (1.5-\lambda)(\lambda - 2.5) = 0$

$\lambda_1 = 2.5, \lambda_2 = 1.5, \lambda_3 = 0$

Eigen Vectors

Solve for $\hat{x}$ in ē relation $A\hat{x} = \lambda \hat{x}$

$\lambda_1 = 2.5$

$$\begin{bmatrix} 1.5 & 0 & 0 \\ 0 & 2 & -1 \\ 0 & -1 & 0.5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = 2.5 \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

$1.5X_1 + 0X_2 + 0X_3 = 2.5X_1$ or $X_1 = 0$ ---(i)

$0X_1 + 2X_2 - 1X_3 = 2.5X_2$ " $0.5X_2 = -1X_3$ --- (ii)

$0X_1 - 1X_2 + 0.5X_3 = 2.5X_3$ " $1X_2 = 2X_3$ --- (iii)

From eqtn (i), $X_1 = 0$

From eqtn (ii)

Let $X_3 = -1$ such that $X_2 = 2$ then, $C_1 = \begin{bmatrix} 0 \\ 2 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ 2/\sqrt{5} \\ -1/\sqrt{5} \end{bmatrix} = \begin{bmatrix} 0 \\ 0.894 \\ -0.447 \end{bmatrix}$ $e = \frac{X}{\sqrt{X'X}}$

$\lambda_2 = 1.5$

$$\begin{bmatrix} 1.5 & 0 & 0 \\ 0 & 2 & -1 \\ 0 & -1 & 0.5 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = 1.5 \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}$$

$1.5X_1 + 0X_2 + 0.0X_3 = 1.5X_1$ or $0X_1 = 0$ ---(i)

$0X_1 + 2X_2 - 1X_3 = 1.5X_2$ or $0.5X_2 = 1X_3$ ---(ii)

$0X_1 - 1X_2 + 0.5X_3 = 1.5X_3$ or $1X_2 = -1X_3$ --- (iii)

From eqtn (i), Let $X_1 = 1$

From eqtn (ii)

Let $X_2 = 0$ then $X_3 = 0$ such that $C_2 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$

$\lambda_3 = 0$

$$\begin{bmatrix} 1.5 & 0 & 0 \\ 0 & 2 & -1 \\ 0 & -1 & 0.5 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = 0 \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}$$

$1.5X_1 = 0$ or $1.5X_1 = 0$ ---(i)

$2X_2 - X_3 = 0$ or $2X_2 = X_3$ ---(ii)

$-X_2 + 0.5X_3 = 0$ or $X_2 = 0.5X_3$ --- (iii)

From eqtn (i), $X_1 = 0$

From eqtn (ii)

Let $X_2 = 1$ such that $X_3 = 2$ then $C_3 = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1/\sqrt{5} \\ 2/\sqrt{5} \end{bmatrix} = \begin{bmatrix} 0 \\ 0.447 \\ 0.894 \end{bmatrix}$

Principal components

$Y_1 = C_1'X = 0.894X_2 - 0.447X_3$

$Y_2 = C_2'X = X_1$

$Y_3 = C_3'X = 0.447X_2 + 0.894X_3$

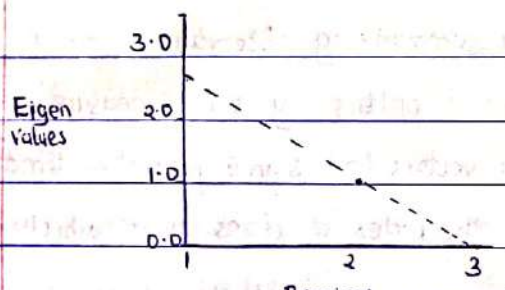
components

The principal loadings are the coefficients of X in the principal components or simply Eigen vectors that are given as:

PC_1	PC_2	PC_3
0.000	1.000	0.000
0.894	0.000	0.447
-0.447	0.000	0.894

- The principal component one indicates that there is a factor that affects X_2 positively while affecting X_3 negatively whereas principal component Y_2 indicates that there is a factor that is affecting by X_1 positively and does not consider any measure of X_2 & X_3 therefore Y_2 measures the scores of X_1.
- Considering ē ~~score~~ scree diagram below the Eigen values of $\lambda_1 = 5/2$ and $\lambda_2 = 3/2$ have values more than 1, then ē principal components of λ_1 & λ_2 are considered sufficient enough in providing enough info. Alternatively ē contribution of ē variances can be used by considering ē relative cumulative variances.

Scree Plot



λ	Cumulative Variance	% Cumulative Variance
2.5	2.5	0.625
1.5	4.0	1.000
0	4.0	1.000

The first principal components contributes 62.5% of ē variations and this is considered lower and therefore more than one principal components are required in ē analysis in order to have at least 90% of ē variance explained by ē principal components. The first 2 principal components have 100% of ē variance explained by ē " " that is more than minimum of 90% and therefore ē first 2 principal components are considered sufficient.

- The principal components scores are computed by first transforming ē initial data of ē variables by substracting the means from ē initial values then multiplying by the principal component loadings, that is,

Principal component score, $PC_{1i} = [X_{1j} - \bar{X}_1^5 \quad X_{2j} - \bar{X}_2^4 \quad X_{3j} - \bar{X}_3^2][e_1]$

" " " , $PC_{2i} = [X_{1j} - \bar{X}_1^5 \quad X_{2j} - \bar{X}_2^4 \quad X_{3j} - \bar{X}_3^2][e_2]$

" " " , $PC_{3i} = [X_{1j} - \bar{X}_1^5 \quad X_{2j} - \bar{X}_2^4 \quad X_{3j} - \bar{X}_3^2][e_3]$

Deviations from ē mean			Loadings			Scores		
X_1	X_2	X_3	C_1	C_2	C_3	PC_1	PC_2	PC_3
2.0	0.0	0.0	0.000	1.000	0.000	0.000	2.000	0.000
0.0	2.0	-1.0	-0.894	0.000	0.447	-2.236	0.000	0.000
-1.0	0.0	0.0	0.447	0.000	0.894	0.000	-1.000	0.000
-1.0	0.0	0.0				0.000	-1.000	0.000
0.0	-2.0	1.0				2.236	0.000	0.000

Exam how to do it →

Discriminant Functions.

- Consider random samples of N_1 & N_2 observation vectors drawn independently from respective p-dimensional multi-normal ppltns with mean vectors μ_1 & μ_2, and a common variance-covariance matrix Σ. Rather than test the usual hypothesis of equal mean vectors we wish to construct a linear compound ~~of~~ or index of summarizing observations from ē groups on a one-dimensional scale that discriminates btwn ē ppltns by some measure of maximal separation. If $\bar{X}_1$ and $\bar{X}_2$ are ē sample mean vectors an S in ē pooled estimate of Σ, we shall determine ē coefficient vector a of the index a x as that which gives ē greatest squared critical ratio.

$$t^2_{(a)} = \frac{[a'(\bar{X}_1 - \bar{X}_2)]^2 N_1 N_2 / (N_1 + N_2)}{a' S a} \quad \text{---} \quad (4.1)$$

Equivalently, that which maximizes ē absolute difference $|a'(\bar{X}_1 - \bar{X}_2)|$ in ē ave. values of the index of ē two groups subject to ē constraint $a'Sa = 1$. As in ē union intersection construction of ē single-sample T^2, the coefficient vector a is given by ē homogeneous system of eqtns

$$\frac{N_1 N_2}{N_1 + N_2} \left[(\bar{X}_1 - \bar{X}_2)(\bar{X}_1 - \bar{X}_2)' - \lambda S \right] a = 0 \quad \text{---} \quad (4.2)$$

Such that

$$\lambda = \max t^2_{(a)} = \frac{N_1 N_2}{N_1 + N_2} (\bar{X}_1 - \bar{X}_2)' S^{-1} (\bar{X}_1 - \bar{X}_2) = T^2 \quad \text{---} \quad (4.3)$$

The rank of ē coefficient matrix is p-1, so that ē system has only ē single solution

$$a = S^{-1}(\bar{X}_1 - \bar{X}_2) \quad \text{---} \quad (4.4)$$

The linear ~~distribution~~ discriminate fnon is given as:

$$y = (\bar{X}_1 - \bar{X}_2)' S^{-1} x \quad \text{---} \quad (4.5)$$

If ē variances of ē responses are nearly equal, the elements give ē relative importance of ē contribution of each response to ē T^2 statistic.

In order to use ē linear discriminant function for classifying an observation of an unknown ppltn we begin by comparing ē mean values of ē scores of ē 2 samples.

$\bar{y}_1 = (\bar{x}_1 - \bar{x}_2)' S^{-1} \bar{x}_1$ and $\bar{y}_2 = (\bar{x}_1 - \bar{x}_2)' S^{-1} \bar{x}_2$

The midpoint of ēse 2 means on ē discriminant fncn scale is $\frac{1}{2}(\bar{x}_1 - \bar{x}_2)' S^{-1} (\bar{x}_1 + \bar{x}_2)$ and we might adopt the classification rule, assign ē individual with ē observation X to ppltn one (1) if $(\bar{x}_1 - \bar{x}_2)' S^{-1} x > \frac{1}{2}(\bar{x}_1 - \bar{x}_2)' S^{-1} (\bar{x}_1 + \bar{x}_2)$ and to ppltn two (2) if $(\bar{x}_1 - \bar{x}_2)' S^{-1} x \leq \frac{1}{2}(\bar{x}_1 - \bar{x}_2)' S^{-1} (\bar{x}_1 + \bar{x}_2)$

- That is, sampling units are assigned to ē group with ē closer discriminant score. Because ē mean point is a value of a random var. it would seem more appropriate to state that rule in terms of ē single statistic.

$W = x' S^{-1} (\bar{x}_1 - \bar{x}_2) - \frac{1}{2}(\bar{x}_1 - \bar{x}_2)' S^{-1} (\bar{x}_1 + \bar{x}_2)$

That is, assign x to ppltn one (1) if $W > 0$, otherwise assign X to ppltn two (2), W is called the Wald-Anderson classification statistic.

Exa 4.3

49 women participating in study of human aging were classified into diagnostic categories "civilized" and "no civilization" in the basis of an intensive psychiatric examination. Determine the classification criterion using the discriminate function for the results recorded in the table below.

	Group Mean		
Subset	No civilization	Civilization	
Information	12.75	8.75	12.75 - 8.75 = 3.82
Similarities	9.75	5.33	9.75 - 5.33 = 4.24
Arithmetic	11.49	8.50	2.99
Picture Competition	7.97	4.75	3.22

The within-group covariance matrix and its inverse is given as

$$S = \begin{bmatrix} 11.26 & 9.41 & 7.16 & 3.38 \\ 9.41 & 13.53 & 7.38 & 2.50 \\ 7.16 & 7.38 & 11.58 & 2.62 \\ 3.38 & 2.50 & 2.62 & 5.81 \end{bmatrix} \quad S^{-1} = \begin{bmatrix} 0.260 & -0.137 & -0.059 & -0.066 \\ -0.137 & 0.187 & -0.038 & 0.016 \\ -0.059 & -0.038 & 0.151 & -0.017 \\ -0.066 & 0.016 & -0.017 & 0.211 \end{bmatrix}$$

Solution:

The vector of the mean differences for the two diagnostic groups is $(\bar{x}_1 - \bar{x}_2)' = [3.82 \quad 4.24 \quad 2.99 \quad 3.22]$ and discriminant function is

$$y = (\bar{X}_1 - \bar{X}_2)S_x^{-1} = [3.82 \quad 4.24 \quad 2.99 \quad 3.22] \begin{bmatrix} 0.260 & -0.137 & -0.059 & -0.066 \\ -0.137 & 0.187 & -0.038 & 0.016 \\ -0.059 & -0.038 & 0.151 & -0.017 \\ -0.066 & 0.016 & -0.017 & 0.211 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{bmatrix}$$

$$= 0.0259X_1 + 0.2083X_2 + 0.0086X_3 + 0.4456X_4$$

The mean discriminant scores for non-civilization factor & civilization factor groups are:

$$\bar{Y}_1 = (\bar{X}_1 - \bar{X}_2)S^{-1}\bar{X}_1 = [3.82 \quad 4.24 \quad 2.99 \quad 3.22] \begin{bmatrix} 0.260 & -0.137 & -0.059 & -0.066 \\ -0.137 & 0.187 & -0.038 & 0.016 \\ -0.059 & -0.038 & 0.151 & -0.017 \\ -0.066 & 0.016 & -0.017 & 0.211 \end{bmatrix} \begin{bmatrix} 12.57 \\ 9.57 \\ 11.49 \\ 7.97 \end{bmatrix} = 5.970$$

$$\bar{Y}_2 = (\bar{X}_1 - \bar{X}_2)S^{-1}\bar{X}_2 = [3.82 \quad 4.24 \quad 2.99 \quad 3.22] \begin{bmatrix} 0.260 & -0.137 & -0.059 & -0.066 \\ -0.137 & 0.187 & -0.038 & 0.016 \\ -0.059 & -0.038 & 0.151 & -0.017 \\ -0.066 & 0.016 & -0.017 & 0.211 \end{bmatrix} \begin{bmatrix} 8.57 \\ 5.33 \\ 8.50 \\ 4.75 \end{bmatrix} = 3.527$$

The midpoint of these 2 means on the discriminant function scale is given as

$$\frac{1}{2}(\bar{X}_1 - \bar{X}_2)'S^{-1}(\bar{X}_1 + \bar{X}_2) = \frac{1}{2}(5.970 + 3.527) = \frac{9.497}{2} = 4.75$$

Decision:

Assign i^{th} individual to civilization factor diagnostic group if $y_i \leq 4.75$ and to no-civilization factor group if $y_i > 4.75$

Canonical Correlations.

- In this section we shall consider the use of the quantities of coefficient of ē Linear compounds and discuss the test statistic in its squared correlation sense - Suppose that the $p+q$ variates $[X_1' \; X_2']$ of some multi-dimensional population have been divided so that their covariance matrix Σ has the partitioned form. In this section we shall drop ē requirement that the distribution is multi-normal and merely specify that:
  - The elements of Σ are finite.
  - Σ is of the full rank $p+q$
  - The first $r \leq \min(p+q)$ xtic roots of $\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}'$ are distinct.

The N observation vectors have been drawn randomly from ē pplth and their sample covariance matrix is partitioned as:

$$S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \quad \ldots \quad (4.6)$$

In conformance with, Σ, in an attempt to reduce ē no. of variates and their non-zero correlations to more parsimonious degrees we seek to answer ē following qstns.

What are the linear compounds?

$u_1 = a_1' x_1$, $v_1 = b_1' x_2$ ... $u_s = a_s' x_1$, $v_s = b_s' x_2$ ... (4.7)

- With ē property that ē sample correlation of u_1 & v_1 is the greatest, the sample correlation of u_2 & v_2 greatest among all linear compounds uncorrelated with u_1 & v_1, and so on, for all $s = \min(p+q)$ possible pairs. Clearly, the first pairs have ē coefficients determined implicity in the union-intersection test of set independence, and its squared correlation is ē test statistic. We introduce ē constraints in ē maximization problem

$a_i' S_{11} a_j = 0$, $b_i' S_{22} b_j = 0$, $a_i' S_{12} b_j = 0$, $a_j' S_{12} b_i = 0$, $i \neq j$ ... (4.8)

- It can be shown in ē maximization problem that ē coefficients of ē i^{th} pair are given by the homogeneous linear eqtn

$(S_{12} S_{22}^{-1} S_{12}' - c_i S_{11}) a_i = 0$, $(S_{12}' S_{11}^{-1} S_{12} - c_i S_{22}) b_i = 0$

- Where c_i is ē largest root of ē determinant eqtn $|S_{12} S_{22}^{-1} S_{12}' - \lambda S_{11}| = 0$ or $|S_{12}' S_{11}^{-1} S_{12} - \lambda S_{22}| = 0$

- Rewrite c_i as ē squared product-moment correlation of ē i^{th} linear compounds:

$$c_i = r_{u_i, v_i}^2 = \frac{(a_i' S_{12} b_i)^2}{a_i' S_{11} a_i b_i' S_{22} b_i}$$

- If c_i is a distinct root, ē coefficient vectors a_i & b_j will be unique, and their linear compounds will be uncorrelated with ē other canonical variates. In order to demonstrate this property we assume that c_i & c_j are distinct roots & write ē homogeneous eqtns for their respective pairs of coefficients as:

$(S_{12} S_{22}^{-1} S_{12}' - c_i S_{11}) a_i = 0$, $(S_{12}' S_{11}^{-1} S_{12} - c_i S_{22}) b_i = 0$

$(S_{12} S_{22}^{-1} S_{12}' - c_j S_{11}) a_j = 0$, $(S_{12}' S_{11}^{-1} S_{12} - c_j S_{22}) b_j = 0$

- Pre-multiply these equations by a_j', b_j', a_i' and b_i', respectively. Then,

$(c_i - c_j) a_i' S_{11} a_j = 0$ and $(c_i - c_j) b_i' S_{22} b_j = 0$

- Since $c_i \neq c_j$ it must follow that ē billinear forms equal to ē covariance of u_i, u_j, v_i, v_j are zero. Similarly, pre-multiplication of ē 1st & 4th eqtns by $b_j' S_{12}' S_{11}^{-1}$ and $a_i' S_{12} S_{22}^{-1}$, respectively, and substraction of ē resulting 1st scalar eqtn from the second lead to ē condition $(c_i - c_j) b_j' S_{12}' a_i = 0$ of zero correlation of

u_i & v_i and c_i & c_j are distinct. If, as in the population, the r largest c_i are distinct, r pairs of unique canonical variables are formed.

Summary:

- We begin with a sample whose observation vectors have the covariance matrix $\Sigma = \Sigma_0$. Through the canonical variate transformation we have passed to the new scores $u_1 \ldots u_s$, $v_1 \ldots v_s$ with conditional matrix

$$\begin{bmatrix} 1 & \cdots & 0 & c_1^{1/2} & \cdots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & & 1 & 0 & & c_s^{1/2} \\ c_1^{1/2} & & 0 & 1 & & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & & c_s^{1/2} & 0 & & 1 \end{bmatrix}$$

- All the correlation btwn $\bar{e}$ sets of $\bar{e}$ original variables has been channelled thro' the s canonical correlations. Identical canonical correlation will be obtained for $S_{11}^{-1} S_{12} S_{22}^{-1} S_{12}'$ or $R_{11}^{-1} R_{12} R_{22}^{-1} R_{12}'$. In the first case the elements of a_i and b_i will be in units proportional to those of the respective responses in $\bar{e}$ sets and $\bar{e}$ dimensions of u_i & v_i will be meaningless. Canonical variates based on $\bar{e}$ correlation matrix and dimensionless should be evaluated in terms of the standard score $Z_{ij} = \dfrac{X_{ij} - \bar{X}_j}{s_j}$ of $\bar{e}$ original observations.

Exa 4.4

Determine the canonical correlation and variates for the following correlation matrix which has been partitioned into four matrices of order 2X2.

$$R = \begin{bmatrix} 1.00 & 0.45 & \cdots & -0.19 & 0.43 \\ 0.45 & 1.00 & \cdots & -0.02 & 0.62 \\ \cdots & \cdots & & \cdots & \cdots \\ -0.19 & -0.02 & & 1.00 & -0.29 \\ 0.43 & 0.62 & & -0.29 & 1.00 \end{bmatrix}$$

Exam

Solution

$$R_{11}^{-1} = \frac{1}{1-0.45^2}\begin{bmatrix} 1.00 & -0.45 \\ -0.45 & 1.00 \end{bmatrix} = \begin{bmatrix} 1.254 & -0.564 \\ -0.564 & 1.254 \end{bmatrix}$$

$$R_{22}^{-1} = \frac{1}{1-0.29^2}\begin{bmatrix} 1.00 & 0.29 \\ 0.29 & 1.00 \end{bmatrix} = \begin{bmatrix} 1.092 & 0.317 \\ 0.317 & 1.092 \end{bmatrix}$$

$$R_{11}^{-1}R_{12}R_{22}^{-1}R_{12}' = \begin{bmatrix} 1.254 & -0.564 \\ -0.564 & 1.254 \end{bmatrix}\begin{bmatrix} -0.19 & 0.43 \\ -0.02 & 0.62 \end{bmatrix}\begin{bmatrix} 1.092 & 0.317 \\ 0.317 & 1.092 \end{bmatrix}\begin{bmatrix} -0.19 & -0.02 \\ 0.43 & 0.62 \end{bmatrix}$$

Check multiplication of matrices using calc.

$$= \begin{bmatrix} 0.094 & 0.088 \\ 0.213 & 0.373 \end{bmatrix}$$

$$|A-\lambda I| = \begin{vmatrix} 0.094-\lambda & 0.087 \\ 0.213 & 0.373-\lambda \end{vmatrix} = 0$$

$$(0.094-\lambda)(0.373-\lambda) - (0.213)(0.087) = \lambda^2 - 0.467\lambda + 0.0165 = 0$$

$$\lambda = \frac{0.467 \pm \sqrt{0.152}}{2}$$

$\lambda_1 = 0.428$, $\lambda_2 = 0.038$,

The first xtic , $\lambda_1 = 0.428$ and ē largest canonical correlation is $\sqrt{0.428} = 0.655$

The second xtic, $\lambda_2 = 0.038$ and ē second " " " $\sqrt{0.038} = 0.195$

The 1st coefficients of ē canonical associated with 1st xtic root are given by the equations

$$R_{12}R_{22}^{-1}R_{12}' = \begin{bmatrix} -0.19 & 0.43 \\ -0.02 & 0.62 \end{bmatrix}\begin{bmatrix} 1.092 & 0.317 \\ 0.317 & 1.092 \end{bmatrix}\begin{bmatrix} -0.19 & -0.02 \\ 0.43 & 0.62 \end{bmatrix}$$

$$= \begin{bmatrix} 0.190 & 0.255 \\ 0.255 & 0.412 \end{bmatrix}$$

$$\begin{bmatrix} 0.190 & 0.255 \\ 0.255 & 0.412 \end{bmatrix}\begin{bmatrix} a_{11} \\ a_{12} \end{bmatrix} = 0.428\begin{bmatrix} 1.00 & 0.45 \\ 0.45 & 1.00 \end{bmatrix}\begin{bmatrix} a_{11} \\ a_{12} \end{bmatrix}$$

Then,

$0.190\,a_{11} + 0.255\,a_{12} = 0.428\,a_{11} + 0.193\,a_{12}$ or $0.238\,a_{11} = 0.062\,a_{12} = 0$

$0.255\,a_{11} + 0.413\,a_{12} = 0.193\,a_{11} + 0.428\,a_{12}$ or $-0.062\,a_{11} + 0.016\,a_{12} = 0$

Let $a_{12} = 1$ then $a_{11} = 0.261$ or $a_{11} = 0.252$ and $a_{12} = 0.967$

The 2nd coefficients of canonical associated with the 1st xtic root are given by the eqtns:

$$R_{12}'R_{11}^{-1}R_{12} = \begin{bmatrix} -0.19 & -0.02 \\ 0.43 & 0.62 \end{bmatrix}\begin{bmatrix} 1.254 & -0.564 \\ -0.564 & 1.254 \end{bmatrix}\begin{bmatrix} -0.19 & 0.43 \\ -0.02 & 0.62 \end{bmatrix}$$

$$= \begin{bmatrix} 0.041 & -0.047 \\ -0.047 & 0.413 \end{bmatrix}$$

realtouch · artificial wicker decor

$$\begin{bmatrix} 0.041 & -0.047 \\ -0.047 & 0.413 \end{bmatrix} \begin{bmatrix} b_{11} \\ b_{12} \end{bmatrix} = 0.429 \begin{bmatrix} 1.00 & -0.29 \\ -0.29 & 1.00 \end{bmatrix} \begin{bmatrix} b_{11} \\ b_{12} \end{bmatrix}$$

Then

$0.041 b_{11} - 0.047 b_{12} = 0.429 b_{11} - 0.124 b_{12}$ or $0.388 b_{11} - 0.077 b_{12} = 0$

$-0.047 b_{11} + 0.413 b_{12} = -0.124 b_{11} + 0.428 b_{12}$ or $0.077 b_{11} - 0.015 b_{12} = 0$

Let $b_{12} = 1$ then $b_{11} = 0.198$ or $b_{11} = 0.194$ and $b_{12} = 0.981$

Replacement of 0.428 by the smaller root 0.038 in the preceeding systems of equations to solve for coefficients for the 2nd canonical variates

The 1st coefficients of ē canonical associated with 2nd characteristic root are given by the equations

$$R_{12} R_{22}^{-1} R_{12}' = \begin{bmatrix} -0.19 & 0.43 \\ -0.02 & 0.62 \end{bmatrix} \begin{bmatrix} 1.092 & 0.317 \\ 0.317 & 1.092 \end{bmatrix} \begin{bmatrix} -0.19 & -0.02 \\ 0.43 & 0.62 \end{bmatrix}$$

$$= \begin{bmatrix} 0.190 & 0.255 \\ 0.255 & 0.412 \end{bmatrix}$$

$$\begin{bmatrix} 0.190 & 0.255 \\ 0.253 & 0.412 \end{bmatrix} \begin{bmatrix} q_{21} \\ q_{22} \end{bmatrix} = 0.038 \begin{bmatrix} 1.00 & 0.45 \\ 0.45 & 1.00 \end{bmatrix} \begin{bmatrix} q_{21} \\ q_{22} \end{bmatrix}$$

Then

$0.189 q_{21} + 0.255 q_{22} = 0.038 q_{21} + 0.017 q_{22}$ or $0.151 q_{21} + 0.238 q_{22} = 0$

$0.255 q_{21} + 0.413 q_{22} = 0.017 q_{21} + 0.038 q_{22}$ or $0.238 q_{21} + 0.375 q_{22} = 0$

Let $q_{21} = -1.00$ then $q_{22} = 0.63$ or $q_{21} = -0.844$ then $q_{22} = 0.53$

The second coefficients of canonical associated with the 2nd xtic root are given by the equations

$$R_{12}' R_{11}^{-1} R_{12} = \begin{bmatrix} -0.19 & -0.02 \\ 0.43 & 0.62 \end{bmatrix} \begin{bmatrix} 1.254 & -0.564 \\ -0.564 & 1.254 \end{bmatrix} \begin{bmatrix} -0.19 & 0.43 \\ -0.02 & 0.62 \end{bmatrix}$$

$$= \begin{bmatrix} 0.041 & -0.047 \\ -0.047 & 0.413 \end{bmatrix}$$

$$\begin{bmatrix} 0.041 & -0.047 \\ -0.047 & 0.413 \end{bmatrix} \begin{bmatrix} b_{21} \\ b_{22} \end{bmatrix} = 0.038 \begin{bmatrix} 1.00 & -0.29 \\ -0.29 & 1.00 \end{bmatrix} \begin{bmatrix} b_{21} \\ b_{22} \end{bmatrix}$$

Then

$0.041 b_{21} - 0.047 b_{22} = 0.038 b_{21} - 0.011 b_{22}$ or $0.003 b_{21} - 0.036 b_{22} = 0$

$-0.047 b_{21} - 0.413 b_{22} = -0.011 b_{21} + 0.038 b_{22}$ or $-0.036 b_{21} + 0.375 b_{22} = 0$

Let $b_{21} = -1.00$ then

$b_{22} = -0.083$ or $b_{21} = -0.997$ and $b_{22} = -0.083$

5 FACTOR ANALYSIS.

Introduction

- In many real-life applications, the number of independent variables used in predicting a response variables will be many. The difficulties in having too many independent variables include ē following

(i) Increased computational time to get the solution

(ii) Increased time in data collection.

iii) Too much expenditure in data collection.

iv) Presence of redundant independent variables

v) Difficulty in making inferences.

- In order to avoid the difficulties associated with ē many input variables, factor analysis may be considered for analysis that aims at grouping the original variables into factors which underlie the input variables such that each factor will account for one or more input variables. The total number of input variables, but after performing factor analysis, the total no. of factors in the study can be reduced by dropping ē insignificant factors based on certain criterion.

Exa 5.1

In the problem of studying ē customer feedback about a two-wheeler produced by an industry, the marketing manager designed a questionnaire to study ē customer's' feedback about ē two-wheel product in order to identify ē factors of ē study. The initial 6 variables for ē study identified include ē following:

(i) X_1 - Fuel efficiency

(ii) X_2 - Life span of the 2-wheeler

(iii) X_3 - Handling convenience

(iv) X_4 - Quality of original spares

(v) X_5 - Breakdown rate

(vi) X_6 - Price

- After administering a questionnaire among respondents, the opinion of customers can be obtained on ē variables in which ē range of ē scores for each of ē variables is assumed to be btwn 1 and 10. The score 1 means the lowest rating while 10 means the highest rating. If the application of factor analysis groups. Assume that the application of factor analysis groups these variables as follows:

(i) X_1, X_2, X_4 and X_5 into factor_1

(ii) X_6 into factor_2

(iii) X_3 into factor_3

- If all ē factors are significant, then they are retained for future analysis. A careful examination of these groupings of variables into factors reveals that the factor_1, factor_2 and factor_3 can be named as technical factor, price factor and personal factor respectively. The factor analysis reveals that in future while conducting a detailed study, it is sufficient to get opinion of ē customers

on the three factors which are obtained through factor analysis. If there is n variables, there will be n factors, each factor say K is represented by a linear composite. Let F_K be the linear composite of the factor K given as

$$F_K = W_{1K} X_1 + W_{2K} X_2 + W_{3K} X_3 + \dots + W_{iK} X_i + \dots + W_{nK} X_n = \sum_{i=1}^{n} W_{ik} X_i, \quad k = 1, 2, \dots, n$$

- Where Wik is ē weight of ē original variable X_i in the linear composite of the factor k.
- There are several methods of factor analysis, but they do not necessarily give ē same results and ē therefore factor analysis is not a single unique method but a set of techniques. The factor analysis methods include ē centroid method, principal components method and maximum likelihood method. In this section the centroid method is discussed.

Centroid Method.

- The method maximizes the sum of absolute loadings of each factor. In this method, the coefficients of ē terms in ē linear composite in each factor will be either +1 or -1. This method is less cumbersome when compared to methods. The commonly used terms relating to factor analysis include the following:

(i) Factor - It is an underlying dimension that account for several observed variables that can be one or more factors depending upon the nature of the study and ē no. of variables involved in it.

(ii) Factor - loadings - Are those values which explain how closely the variables are related to each one of the factors discovered.

(iii) Communality (h^2) - It indicates how much of each variable is accounted for by ē underlying factor taken together. A high value of communality means that not much of ē variable is left over after whatever ē factors represent is taken into consideration.

(iv) Eigen value - When ē sum of squared values of factor loadings relating to a factor, then such sum is referred to as Eigen value or latent root. Eigen value indicates ē relative importance of each factor in accounting for ē particular set of variables being analyzed.

(v) Total sum of squares - The resulting value of ē sum of Eigen values of all factors.

(vi) Rotation - In the context of factor analysis is that which reveals different structures in ē data and though diff. rotations give results that appear to be diff., but from a statistical point of view, all results are taken as equal. If ē factors are independent orthogonal rotation is done and if ē factors are correlated, an oblique rotation is made.

(vii) Factor scores - it represents ē degree to which each respondent gets high scores on ē group of items that load high on each factor.

Centroid Method Procedure.

(a) Compute ē matrix of correlation, R, wherein unities are place in ē diagonal spaces. The product moment formula is used for working out ē correlation coefficients.

(b) If ē correlation matrix obtained happens to be positive manifold (that is disregarding ē diagonal elements each variable has a large sum of positive correlations than of negative correlations). ē centroid method requires that ē weights for all ē variables be ±1.0, ē variables are not weighted but simply summed. In case ē correlation matrix is not a positive manifold, ēn reflections must be made before ē 1st centroid factor is obtained.

(c) The 1st centroid factor is determined as:

(i) The sum of coefficients (including diagonal unity) in each column of ē correlation matrix is worked out.

(ii) The sum of these column sums (T) is obtained.

(iii) The sum of each column obtained in (i) above is divided by ē sq. root of T obtained in (ii) above, resulting in what is called centroid loadings. The full set of loadings so obtained constitute ē 1st centroid factor (say A).

d) In order to obtain ē 2nd centroid factor (say B), one must first obtain a matrix of residual coefficients, for this purpose, the loading for ē 2 variables on ē 1st centroid factor are multiplied. This is done for all possible pairs of variables and ē resulting matrix of factor cross products are named as Q_1. The Q_1 is substracted element by element from ē original matrix of ē correlation, R, and ē result is ē 1st matrix of ē residual coefficients R_1. After obtaining R_1, one must reflect some of the variables in it, meaning thereby that some of the variables are given negative signs in ē sum (this is usually done by inspection)

Exa 5.2

The random variables X_1, X_2 and X_3 have ē variance-covariance matrix $\Sigma = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix}$. Determine the new factors if the factors X_1, X_2 and X_3 are to be grouped into ē demographic and economical factors using the centroid factor analysis method.

Solution.

The correlation matrix R for ē 6 variables is as given below:

$$r_{ik} = \frac{\sum_{j=1}^{n} X_{ij} X_{kj} - n\bar{X}_i \bar{X}_k}{\sqrt{\sum_{j=1}^{n} X_{ij}^2 - n\bar{X}_i^2}\ \sqrt{\sum_{j=1}^{n} X_{kj}^2 - n\bar{X}_k^2}}$$

$$R_1 = \begin{bmatrix} 1.000 & -0.894 & 0.000 \\ -0.894 & 1.000 & 0.000 \\ 0.000 & 0.000 & 1.000 \end{bmatrix}$$

$r_{11} = \frac{s_{11}}{\sqrt{s_{11}}\sqrt{s_{11}}}$ r_{12} s_{12}

Since ē matrix R_1 is not positive manifold, the variable X_2 is reflected to have ē following correlation matrix

EXAM

$$R_1 = \begin{bmatrix} 1.000 & 0.894 & 0.000 \\ 0.894 & 1.000 & 0.000 \\ 0.000 & 0.000 & 1.000 \end{bmatrix} \quad R_1$$

Determine the sum of columns.

$$S_i = \sum_{k=1}^{3} r_{ik},$$

$S_1 = 1.000 + 0.894 + 0.000 = 1.894$ ✓

$S_2 = 0.894 + 1.000 + 0.000 = 1.894$ ✓

$S_3 = 0.000 + 0.000 + 1.000 = 1.000$ ✓

Determine ē sum of all columns

$$T = \sum_{k=1}^{3} S_i = 1.894 + 1.894 + 1.000$$

$$= 4.789$$ ✓✓

The 1st centroid factor A_1 is obtained by dividing column total by sq. root of sum of all columns (T), that is,

$$UL_1(j) = S_i/\sqrt{T} = S_i/\sqrt{4.789}$$

$S_i/\sqrt{T}$ $S_1/\sqrt{T}$

S_i	1.894	1.894	1.000
$S_i/\sqrt{T}$	0.866	0.866	0.457

✓✓

In order to obtain ē 2nd centroid factor B, develop ē matrix of factor cross product Q_1

First Centroid Factor	↓→	0.866	-0.866	0.457
	0.866	0.749	-0.749	0.396
	-0.866	-0.749	0.749	-0.396
	0.457	0.396	-0.396	0.209

Q_1 ✓✓

Next obtain ē matrix of residual coefficients (R_2) by subtracting Q_1 from R_1 as recorded below.

$R_2 = R_1 - Q_1$

0.251	-0.145	-0.396
-0.145	0.251	0.396
-0.396	0.396	0.791

✓✓

Reflect the variables X_2 and X_3 and obtain ē reflected matrix of residual coefficients R_2'

0.251	0.145	0.396
0.145	0.251	0.396
0.396	0.396	0.791

Determine the sum of columns

$S_i = \sum_{k=1}^{3} r_{ik}$,

$S_1 = 0.251 + 0.145 + 0.396 = 0.791$

$S_2 = 0.145 + 0.251 + 0.396 = 0.791$

$S_3 = 0.396 + 0.396 + 0.791 = 1.582$

Determine the sum of all columns

$T = \sum_{k=1}^{3} S_i = 0.791 + 0.791 + 1.582 = 3.165$

The 2nd centroid factor (B) is obtained by dviding each of ē column total by sq. root of sum of all columns (T), that is.

$UL_1(i) = S_i/\sqrt{T} = S_i/\sqrt{3.165}$

S_i	0.791	0.791	1.582
$S_i/\sqrt{T}$	0.445	0.445	0.889

The loadings of all the two factors are given below.

Variables	Centroid Factor Loadings: Factor A	Factor B	Communality (h^2)
X_1	0.866^2	0.445^2	$(0.866)^2 + (0.445)^2 = 0.947$
X_2	-0.866^2	-0.445^2	$(0.866)^2 + (0.445)^2 = 0.947$
X_3	0.457^2	-0.889^2	$(0.457)^2 + (0.889)^2 = 1.000$
	1.708	1.187	2.894

Proportions of total and common variance.

Variables	Centroid Factor Loadings: Factor A	FACTOR B	Communality (h^2)
Eigen value (Variance accounted)	1.708	1.187	2.894
Proportion of total variance	1.708/3 = 0.427	1.187/3 = 0.297	
Proportion of common variance	1.708/2.894 = 0.590	1.187/2.894 = 0.410	1.708/2.894

The common variance, i.e. $\sum_{i=1}^{3} h^2 = 2.894$

Total variance is equal to no. of variables = 3.000

Conclusion:

Assign each variable to the factor with which it has the maximum absolute loadings

Factor	Name of factor	Variable and description
1	Demographic	X_1
		X_2
2	Economical	X_3

The proportion of total variance of factor_1 = 0.427 (42.7%)
" " " " " " factor_2 = 0.297 (29.7%)
The " " common variance of factor_1 = 0.590 (59.0%)
" " " " " " factor_2 = 0.410 (41.0%)

Example 5.3

The rating of expenditure, income, education, religion, marital, age and weights for persons in a town were collected in the range 1-10 (1-least rated, while 10 is highest rated) and recorded below.

X_1	6	4	4	1	4	4	3	7	5	5
X_2	8	4	1	2	3	4	3	7	3	4
X_3	9	6	6	6	5	6	0	6	1	2
X_4	9	8	5	3	5	8	9	9	8	83

If the variables are grouped into demographic, economic and social factors, determine the first, second and third centroid factors using the method of factor analysis.

Solution

The correlation matrix R for ē 6 variables is as given below.

$$r_{ik} = \frac{\sum_{i=1}^{n} X_{ij} X_{ik} - n \bar{X}_i \bar{X}_k}{\sqrt{\sum_{i=1}^{n} X_{ij}^2 - n \bar{X}_i^2} \sqrt{\sum_{i=1}^{n} X_{kj}^2 - n \bar{X}_k^2}}$$

$$R_1 = \begin{bmatrix} 1.000 & 0.742 & 0.168 & 0.496 \\ 0.742 & 1.000 & 0.424 & 0.568 \\ 0.168 & 0.424 & 1.000 & 0.050 \\ 0.496 & 0.568 & 0.050 & 1.000 \end{bmatrix}$$

Since ē matrix R_1 is positive manifold, non of ē variables is reflected and correlation matrix given as below

$$R_1 = \begin{bmatrix} 1.000 & 0.742 & 0.168 & 0.496 \\ 0.742 & 1.000 & 0.424 & 0.568 \\ 0.168 & 0.424 & 1.000 & 0.050 \\ 0.496 & 0.568 & 0.050 & 1.000 \end{bmatrix}$$

$$r = \frac{\sum (X_1 - \bar{X}_1)(X_2 - \bar{X}_2)}{\sqrt{\sum (X_1 - \bar{X}_1)} \sqrt{\sum (X_2 - \bar{X}_2)}}$$ ← Indirect relation ← Pearsons

Determine ē sum of columns

$S_i = \sum_{k=1}^{4} r_{ik}$,

$S_1 = 1.000 + 0.742 + 0.168 + 0.496 = 2.406$

$S_2 = 0.742 + 1.000 + 0.424 + 0.568 = 2.734$

$S_3 = 0.168 + 0.424 + 1.000 + 0.050 = 1.642$

$S_4 = 0.496 + 0.568 + 0.050 + 1.000 = 2.114$

Determine ē sum of all columns

$T = \sum_{k=1}^{4} S_i = 2.406 + 2.734 + 1.642 + 2.114 = 8.895$

The first centroid factor A, is obtained by dividing column total by sq. root of sum of all columns (T), that is,

$UL(i) = S_i/\sqrt{T} = S_i/\sqrt{8.895}$

S_i	2.406	2.734	1.642	2.114
$S_i/\sqrt{T}$	0.807	0.917	0.551	0.709

In order to obtain ē 2nd centroid factor B, develop ē matrix of factor cross product Q_1.

	↓→	0.807	0.917	0.551	0.709
First Centroid factors	0.807	0.651	0.739	0.444	0.572
	0.917	0.739	0.840	0.505	0.650
	0.551	0.444	0.505	0.303	0.590
	0.709	0.572	0.650	0.390	0.502

Next obtain the matrix of residual coefficients (R_2) by subtracting Q_1 from R_1 as recorded below.

0.349	0.003	-0.276	-0.076
0.003	0.160	-0.081	-0.082
-0.276	-0.081	0.697	-0.340
-0.076	-0.082	-0.340	0.498

Reflect the variables X_3 and X_4 and obtain the reflected matrix of residual coefficient R_2'

0.349	0.003	0.276	0.076
0.003	0.160	0.081	0.082
0.276	0.081	0.697	0.340
0.076	0.082	0.340	0.498

Determine the sum of columns.

$S_1 = 0.349 + 0.003 + 0.276 + 0.076 = 0.704$

$S_2 = 0.003 + 0.160 + 0.081 + 0.082 = 0.325$

$S_3 = 0.276 + 0.081 + 0.697 + 0.340 = 1.394$

$S_4 = 0.076 + 0.082 + 0.340 + 0.498 = 0.996$

Determine the sum of all columns

$T = 0.704 + 0.325 + 1.394 + 0.996 = 3.419$

The second centroid factor B, is obtained by dividing each of $\bar{e}$ column total by sq. root of sum of all columns (T), that is,

$UL(i) = S_i/\sqrt{T} = S_i/\sqrt{3.419}$

S_i	0.704	0.325	1.394	0.996
$S_i/\sqrt{T}$	0.381	0.176	0.754	0.538

In order to obtain the third centroid factor C, develop the matrix of factor cross product Q_2.

	↓ →	0.381	0.176	-0.754	-0.538
First centroid factors	0.381	0.145	0.067	-0.287	-0.205
	0.176	0.067	0.031	-0.133	-0.095
	-0.754	-0.287	-0.133	0.568	0.406 ~~0.046~~
	-0.538	-0.205	-0.095	0.406	0.290

Next obtain $\bar{e}$ matrix of residual coefficients (R_3) by substracting Q_2 from R_2 ;

0.204	-0.064	0.011	0.129
-0.064	0.129	0.052	0.013
0.011	0.052	0.129	-0.746
0.129	0.013	-0.746	0.208

Reflect the variables X_2, X_3 and X_4 and obtained reflected matrix of residual coeff. R_2'

0.204	0.064	0.011	0.129
0.064	0.129	0.052	0.013
0.011	0.052	0.129	0.746
0.129	0.013	0.746	0.208

Determine the sum of columns

$S_1 = 0.204 + 0.064 + 0.011 + 0.129 = 0.409$

$S_2 = 0.064 + 0.129 + 0.052 + 0.013 = 0.258$

$S_3 = 0.011 + 0.052 + 0.129 + 0.746 = 0.937$

$S_4 = 0.129 + 0.013 + 0.746 + 0.208 = 1.096$

Determine the sum of all columns

$T = 0.409 + 0.258 + 0.937 + 1.096 = 2.699$

Elaine - Youtie [illegible]

The third centroid factor C_i is obtained by dividing column total by sq. root of T

$UL(i) = Si/\sqrt{T} = Si/\sqrt{2.699}$

Si	0.409	0.258	0.937	1.096
$Si/\sqrt{T}$	0.249	0.157	0.570	0.667

The loadings of all ē 3 factors are given in table below

Centroid Factor Loadings

Variables	Factor A	Factor B	Factor C	Communality (h^2)
X_1	0.807	0.381	0.249	$(0.807)^2 + (0.381)^2 + (0.249)^2 = 0.796$
X_2	0.917	0.176	-0.157	$(0.917)^2 + (0.176)^2 + (0.157)^2 = 0.871$
X_3	0.551	-0.754	-0.570	$(0.551)^2 + (0.754)^2 + (0.570)^2 = 0.871$
X_4	0.709	-0.538	-0.667	$(0.709)^2 + (0.538)^2 + (0.667)^2 = 0.792$
Are these totals?	2.296	1.034	0.857	3.330

Why ē negatives?

Proportion of total and common variance.

Centroid Factor Loadings

Variables	Factor A	Factor B	Factor C	Communality (h^2)
Eigen value (Variance accounted)	2.296	1.034	0.857	3.330
Proportion of total variance	0.574	0.259	0.214	
Proportion of common variance	0.689	0.311	0.257	

The common variance, i.e $\sum_{i=1}^{4} h^2 = 3.330$

Total variance is equal to no. of variables = 4.000

Conclusion

Assign each variable to the factor with which it has ē maximum absolute loadings.

Factor	Name of factor	Variable and description
1	Demographic	X_1
		X_2
		X_4.
2	Economical	X_3
3	Social	

The proportion of total variance of factor_1 = 0.574 (57.4%)

" " " " " " " _2 = 0.259 (25.9%)

" " " " " " " _3 = 0.214 (21.4%)

The proportion of common variance of factor_1 = 0.689 (68.9%)

" " " " " " " _2 = 0.311 (31.1%)

" " " " " " " _3 = 0.257 (25.7%)

6. ANALYSIS OF VARIANCE AND COVARIANCE.

Analysis of Variance (ANOVA)

- It is an analysis technique that is used to compare more than 2 population such as in comparing the yield of 4 varieties, the duration of fluorescent tubes from 5 vendors or engine performance from 4 modes. In this section we consider ē one-way analysis of variance.

- In designing the experiment the experimental units are arranged in surpentine manner then treatments allocated randomly to the experimental units. Label the experimental units from 1-16 treatments T_1, T_2, T_3 and T_4 for 4 treatments. In the case of 16 experimental units and 4 treatments, the number of experiments units for each treatment is $N/t = 16/4 = 4$. The arrangement of experimental units is as shown below for 16 experimental units.

01	02	03	04
08	07	06	05
09	10	11	12
16	15	14	13

In order to assign the treatments randomly, one can use the lottery method or random number method. In the lottery method, label the 16 marbles from 1 to 16, shuffle the marbles and select one marble randomly at a time from an urn. The first 4 marbles denote the first 4 experimental units assigned treatment T_1, second 4 marbles denote the 4 experimental units to treatment T_2, ... last 4 marbles denote the experimental units assigned to treatment T_4.

- An alternative method to the lottery method is the random number method. In this method select 16 3-digit numbers (if the ~~rise~~ size of total experimental units is two-digit, select three-digit to avoid repetition), rank the random numbers and assign the first $n=4$ ranks that denote the experimental units to treatment T_1, second $n=4$ ranks to treatment T_2..., last $n=4$ ranks to treatment T_4 resp. The three-digit random numbers selected from table 6.2 below are recorded in ē next table 6.3.

6.2

3	2	8	5	7	1	4	5	8	3	5	2	8	4	2	1	2	9	2	4	4
6	7	6	8	3	8	5	4	3	0	9	8	2	5	8	8	8	0	2	7	8
3	9	4	4	4	1	8	6	5	0	0	0	7	8	7	3	9	7	5	9	1
7	6	5	8	7	8	3	0	4	4	8	2	7	6	2	5	9	2	2	5	2
8	8	3	9	5	0	9	2	4	9	5	4	3	7	2	1	8	6	6	5	6
6	6	4	0	9	1	0	0	7	1	3	8	4	0	9	4	5	2	4	6	0
0	5	5	5	1	9	4	7	1	2	8	2	8	6	6	5	7	1	2	6	7
4	0	2	6	0	4	6	3	5	1	0	7	2	3	6	7	9	9	5	6	0
9	5	1	4	4	6	0	1	5	5	5	7	6	4	7	7	4	8	5	6	9
8	6	2	0	2	9	8	1	8	0	0	9	2	5	9	5	7	7	5	2	9
5	3	3	8	8	1	8	7	8	5	6	3	7	8	1	6	4	0	4	7	2
2	6	6	6	0	7	3	7	9	3	3	9	1	4	3	8	7	0	1	6	0
4	0	0	4	2	3	1	8	2	7	9	2	2	0	9	7	5	3	2	4	5
8	0	7	4	3	1	6	5	4	3	8	0	1	0	9	4	7	4	6	5	4
9	6	6	9	6	1	6	7	9	2	5	5	5	1	0	2	8	2	3	1	0
5	9	5	7	9	8	4	8	6	5	1	2	0	2	5	8	8	7	5	4	1

6.3

Treatment	Random Numbers	Rank	Treatment	Random Numbers	Rank
T_1	5 7 1	08	T_3	9 5 0	15
	8 3 8	12		0 9 1	02
	4 4 1	05		5 1 9	07
	8 7 8	13		6 0 4	09
T_2	4 4 6	06	T_4	4 2 3	03
	0 2 9	01		4 3 1	04
	8 8 1	14		9 6 1	16
	6 0 7	10		7 9 8	11

The ranks 08, 12, 05, 13 that denote the experimental units are assigned treatment T_1

" 06, 01, 14, 10 " " " " " " T_2

" 15, 02, 07, 09 " " " " " " T_3

" 03, 04, 16, 11 " " " " " " T_4

The allocation is as shown below.

01, T_3	02, T_2	03, T_4	04, T_4
08, T_1	07, T_2	06, T_3	05, T_1
09, T_2	10, T_3	11, T_4	12, T_1
16, T_4	15, T_2	14, T_3	13, T_1

- In investigating the differences in groups in one-way analysis of variance we follow the steps discussed below;

(i) Arrangement of the observations

Group	Observations	Total
T_1	y_{11} y_{12} ... y_{1n}	T_1
T_2	y_{21} y_{22} ... y_{2n}	T_2
⋮		⋮
T_t	y_{t1} y_{t2} ... y_{tn}	T_t
		G

(ii) Analysis of Variance Model.

$y_{ij} = \mu + t_i + e_{ij}$ $\begin{cases} i = 1, 2, \ldots t \\ j = 1, 2, \ldots, n \end{cases}$

Where: y_{ij} - The j^{th} observation for the i^{th} group, μ - mean, t_i = i^{th} group effect.

e_{ij} - The random error that is normally and independently distributed.

(iii) Hypothesis

$H_0: = \mu_1 = \mu_2 = \ldots = \mu_t$ against $H_a: = \mu_i \neq \mu_k$ for atleast 1 pair (i,k) at α 1·0·s

(iv) Analysis of variance components

Consider corrected total sum of squares.

$$\sum_{i=1}^{t}\sum_{j=1}^{n}(y_{ij}-\bar{y})^2 = \sum_{i=1}^{t}\sum_{j=1}^{n}(y_{ij}-\bar{y}_i+\bar{y}_i-\bar{y})^2$$

$$= \sum_{i=1}^{t}\sum_{j=1}^{n}(y_{ij}-\bar{y}_i)^2 + 2\sum_{i=1}^{t}\sum_{j=1}^{n}(y_{ij}-\bar{y}_i)(\bar{y}_i-\bar{y}) + \sum_{i=1}^{t}\sum_{j=1}^{n}(y_{ij}-\bar{y})^2$$

$$= \sum_{i=1}^{t}\sum_{j=1}^{n}(y_i-\bar{y})^2 + \sum_{i=1}^{t}\sum_{j=1}^{n}(y_{ij}-y_i)^2 \quad \text{since } 2\sum\sum(y_{ij}-\bar{y}_i)(\bar{y}_i-\bar{y})=0$$

$$= \frac{1}{n}(T_1^2+T_2^2+\ldots+T_t^2) - \frac{G^2}{N} + \sum_{i=1}^{t}\sum_{j=1}^{n}(y_{ij}-\bar{y}_i)^2$$

$$SS_{Total} = SS_{Treatments} + SS_{Error}$$

Where, $SS_{Total} = \sum_{i=1}^{t}\sum_{j=1}^{n} y_{ij}^2 - \frac{G^2}{N}$ $\quad SS_{Error} = \sum_{i=1}^{t}\sum_{j=1}^{n}(y_{ij}-\bar{y}_i)^2$

$$SS_{Treatments} = \frac{1}{n}(T_1^2+T_2^2+\ldots+T_t^2) - \frac{G^2}{N}$$

(v) Analysis of variance table.

Source of variation	df	SS	MS	F
Treatments	t-1	$SS_{Treatments}$	$\frac{SS_{Treatments}}{t-1}$	$\frac{MS_{Treatments}}{MS_{Error}}$
Error	N-t	By subtraction	$\frac{SS_{Error}}{N-t}$	
Total	N-1	SS_{Total}		

(vi) Test statistic

$F = \frac{MS_{Treatments}}{MS_{Error}}$ that has approximately F-distribution with t-1 and N-t degrees of freedom

(vii) Decision.

Reject H_0 if $F > F_\alpha(t-1, N-t)$ and conclude that the means of treatment are significantly different at α 1·0·s.

Turkey's Test

- In many practical situations, we wish to compare pairs of means and determine which means differ by testing the difference btwn all ē pairs of treatment means. Thus we are interested in contrasts of ē form $\Gamma = \mu_i - \mu_j$ for all $i \neq j$. The popular method for making such comparison is the Turkey's procedure proposed in 1953 for testing hypotheses for which the overall significance level is exactly α when the sample sizes are equal and at most α when the sample sizes are unequal. The procedure can also be used to construct intervals on the differences in all pairs of means. The Turkey's procedure makes use of ē distribution of the standard range statistic

$$q = \frac{\bar{y}_{max} - \bar{y}_{min}}{\sqrt{\frac{MS\,Error}{n}}}$$

Where $\bar{y}_{max}$ and $\bar{y}_{min}$ are the largest and smallest sample means, resp, out of a group of p sample means Turkey's test declares two means significantly different if ē absolute value of their sample differences exceeds:

$$T_\alpha = \begin{cases} q_\alpha(t,f)\sqrt{\frac{MS\,Error}{n}} & \text{for equal sample sizes} \\ \frac{q_\alpha(t,f)}{\sqrt{2}}\sqrt{MS\,Error\left(\frac{1}{n_i}+\frac{1}{n_j}\right)} & \text{for unequal sample sizes.} \end{cases}$$

Where, t = The number of treatments

f = The number of d.f associated with the MS Error.

n = The number of observations for each treatment when observations are equal

(n_i, n_j) = The sample sizes for unequal sample size pair (i,j)

Exa 6.1

The yields of rice for different varieties T_1, T_2, T_3 and T_4 were collected and recorded below

Varieties	Yield				Totals.	$\bar{y}$
T_1	39	25	16	20	100	25
T_2	33	45	52	30	160	40
T_3	45	45	40	50	180	45
T_4	41	45	60	54	200	50

Determine whether the varieties have different effects at $\alpha = 0.05$ l.o.s

Solution

(i) Analysis of variance model

$$y_{ij} = \mu + t_i + e_{ij} \quad \begin{cases} i = 1,2,\ldots,t \\ j = 1,2,\ldots,n \end{cases}$$

(ii) Hypothesis

$H_0: \mu_1 = \mu_2 = \ldots = \mu_t$ against $H_a: \mu_i \neq \mu_k$ for atleast one pair (i,k) at α l.o.s

(iii) Analysis of variance components.

$SS_{Total} = \sum_{i=1}^{t}\sum_{j=1}^{n} y_{ij}^2 - \frac{G^2}{N} = 39^2 + \dots + 54^2 - \frac{(640)^2}{16} = 2292.00$

$SS_{Treatments} = \frac{1}{n}(T_1^2 + T_2^2 + \dots + T_t^2) - \frac{G^2}{N} = \frac{1}{4}(100^2 + \dots + 200^2) - \frac{(640)^2}{16} = 1400.00$

$SS_{Error} = SS_{Total} - SS_{Treatments} = 2292 - 1400 = 892.00$

(iv) Analysis of variance table.

Source of variation	df	SS	MSS	F
Treatments	t-1 = 3	1400	1400/3 = 466.67	6.28
Error	N-t = 12	892	892/12 = 74.33	
Total	N-1 = 15	2292		

t = 4, N = 16

(v) Test statistic

$F = \frac{MS\ Treatment}{MS_{Error}} = \frac{466.67}{74.33} = 6.28$

(vi) Decision

Tabulated value $F_{0.05}(3,12) = 3.49$, since $F_{calc} > F_{tab}$ reject H_0 and conclude that the means of varieties are significantly different at $\alpha = 0.05$ i.e. 5%

- Since the null hypothesis has been rejected we wish to determine the pairs of means that are significantly different using Turkey's Test. That is determine the statistic

$T_q = q_\alpha(t,f)\sqrt{\frac{MS_E}{n}} = 4.20\sqrt{\frac{74.33}{4}} = 18.11$

$q_{0.05(4,12)}$

		$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_3$
		25	40	45
$\bar{y}_2$	40	-15	0	5
$\bar{y}_3$	45	-20	-5	0
$\bar{y}_4$	50	-25	-10	-5

The pairs that are significantly different include $(\bar{y}_1, \bar{y}_3)$ and $(\bar{y}_1, y_4)$

Confidence Interval

$\bar{y}_i - \bar{y}_j - q_\alpha(t,f)\sqrt{\frac{MS_E}{n}} \leq \mu_i - \mu_j \leq \bar{y}_i - \bar{y}_j + q_\alpha(t,f)\sqrt{\frac{MS_E}{n}}$

$= -18.11 \leq \mu_i - \mu_j \leq 18.11$

Exa 6.2

The lifespan of fluorescent tubes from 4 vendors V_1, V_2, V_3 and V_4 were collected & recorded below

Vendors	Duration				Totals	
V_1	15	20	25	12	72	$\bar{y}_1 = 18$
V_2	18	16	20	10	64	$\bar{y}_2 = 16$
V_3	45	45	40	50	180	$\bar{y}_3 = 45$
V_4	15	15	15	39	84	$\bar{y}_4 = 21$

Determine whether there are differences of lifespan of fluorescent tubes due to the vendors at $\alpha = 0.05$ l.o.s

Solution

(i) Analysis of variance model

$y_{ij} = \mu + t_i + e_{ij}$ $\begin{cases} i = 1, 2, \ldots, t \\ j = 1, 2, \ldots, n \end{cases}$

(ii) Hypothesis

$H_0: \mu_1 = \mu_2 = \ldots = \mu_t$ against $H_a: = \mu_i \neq \mu_k$ for atleast one pair (i, k) at α l.o.s

(iii) Analysis of variance components

$SS_{Total} = \sum_{i=1}^{t} \sum_{j=1}^{n} y_{ij}^2 - \frac{G^2}{N} = 15^2 + \ldots + 39^2 - \frac{400^2}{16} = 2820$

$SS_{Treatments} = \frac{1}{n}(T_1^2 + T_2^2 + \ldots + T_t^2) - \frac{G^2}{N} = \frac{1}{4}(72^2 + \ldots + 84^2) - \frac{400^2}{16} = 2184$

$SS_{Error} = SS_{Total} - SS_{Treatments} = 2820 - 2184 = 636$

(iv) ANOVA table

Source of variation	df	SS	MS	F
Treatments	$t-1$ 3	2184	728	13.74
Error	$N-t$ 12	636	53	
Total	$N-1$ 15	2820		

(v) Test statistic; $F = 13.74$

(vi) Decision

Tabulated value $F_{0.05}(3, 12) = 3.49$, since $F_{calc} > F_{tab}$ reject H_0 and conclude that the means of vendors are significantly different at $\alpha = 0.05$ l.o.s

- Since H_0 has been rejected we wish to determine the pairs of means that are significantly different using Turkey's test. Determine $\bar{e}$ statistic.

$T_q = q_\alpha(t, f)\sqrt{\frac{MSE}{n}} = 4.20\sqrt{\frac{53}{4}} = 15.29$

		$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_3$
		18	16	45
$\bar{y}_2$	16	2	0	29
$\bar{y}_3$	45	-27	-29	0
$\bar{y}_4$	21	-3	-5	24

The pairs that are significantly different include: $(\bar{y}_1, \bar{y}_3)$, $(\bar{y}_2, \bar{y}_3)$ and $(\bar{y}_3, \bar{y}_4)$

Confidence Interval

$\bar{y}_i - \bar{y}_j - q_\alpha(t, f)\sqrt{\frac{MSE}{n}} \leq \mu_i - \mu_j \leq \bar{y}_i - \bar{y}_j + q_\alpha(t, f)\sqrt{\frac{MSE}{n}}$

$= -15.29 \leq \mu_i - \mu_j \leq 15.29$

Analysis of Covariance.

- In " " " one may be interested in studying one independent variable, X and may wish to control the influence of uncontrolled variable (covariate), Z, which is known to be correlated with the dependent variable Y. Then the technique of analysis of covariance is considered for this type of analysis for a valid evaluation of ē outcome of analysis. The arrangement of data is given below.

Treatments	T_1		T_2			T_t	
	X	Y	X	Y	---	X	Y
	X_{11}	Y_{11}	X_{21}	Y_{21}		X_{t1}	Y_{t1}
Observations	X_{12}	Y_{12}	X_{22}	Y_{22}		X_{t2}	Y_{t2}
	⋮	⋮	⋮	⋮		⋮	⋮
	X_{1n}	Y_{1n}	X_{2n}	Y_{2n}		X_{tn}	Y_{tn}
Total	T_{x1}	T_{y1}	T_{x2}	T_{y2}		T_{xt}	T_{yt}

Analysis of covariance model

$$y_{ij} = \mu + t_i + b(x_{ij} - \bar{X}) + e_{ij}$$

Analysis of covariance components

$$G_{xx} = \sum_{i=1}^{t}\sum_{j=1}^{n} X_{ij}^2 - \frac{G_x^2}{N}, \quad G_{xy} = \sum_{i}^{t}\sum_{j}^{n} X_{ij} Y_{ij} - \frac{G_x G_y}{N}, \quad G_{yy} = \sum_{i}^{t}\sum_{j}^{n} y_{ij}^2 - \frac{G_y^2}{N}$$

$$T_{xx} = \frac{1}{n}\sum_{i=1}^{t}\sum_{j=1}^{n} T_{xi}^2 - \frac{G_x^2}{N}, \quad T_{xy} = \frac{1}{n}\sum_{i}^{t} T_{xi} T_{yi} - \frac{G_x G_y}{N}, \quad T_{yy} = \frac{1}{n}\sum_{i}^{t} T_{y.}^2 - \frac{G_y^2}{N}$$

$$E_{xx} = G_{xx} - T_{xx}, \quad E_{xy} = G_{xy} - T_{xy}, \quad E_{yy} = G_{yy} - T_{yy}$$

Adjusted values of X are given as

~~$G'_{xx} = G_{xx} - \frac{G_{xy}^2}{G_{xx}}$, $E'_{xx} = E_{xx} - \frac{E_{xy}^2}{E}$~~

$$G'_{xx} = G_{xx} - \frac{G_{xy}^2}{G_{yy}}, \quad E'_{xx} = E_{xx} - \frac{E_{xy}^2}{E_{yy}}, \quad T'_{xx} = G'_{xx} - E'_{xx}$$

Adjusted values of Y are given as

$$G'_{yy} = G_{yy} - \frac{G_{xy}^2}{G_{xx}}, \quad E'_{yy} = E_{yy} - \frac{E_{xy}^2}{E_{xx}}, \quad T'_{yy} = G'_{yy} - E'_{yy}.$$

Adjusted degrees of freedom.

The degrees of freedom for the error is adjusted for adjusted values of X and Y by reducing it by one and therefore d.f for adjusted values of X & Y for total are also reduced by one.

- If X is not affected by ē treatments then there should not be significant difference btwn treatments with reference to X. Then the regression coefficient with treatments is computed as $b = \frac{E_{xy}}{E_{yy}}$ and therefore adjust treatment means for Y as:

$$\bar{X}_i = \bar{X}_i - b(\bar{Y}_i - \bar{Y})$$

Conventionally the results are combined and given in a single table as shown below.

Source of Variation	Sum of products				Adjusted values of Y			
	df	XX	XY	YY	df	SS	MS	F
Treatments	t-1	T_{XX}	T_{XX}	T_{XX}				
Error	n-t	E_{XX}	E_{XX}	E_{XX}	n-t-1	E_{XX}	EMS	
Total	n-1	G_{XX}	G_{XX}	G_{XX}	n-2	G_{XX}		
Treatments adjusted for the average regression within treatments.					t-1	T_{XX}	TMS	$\frac{TMS}{EMS}$

The regression coefficient within treatments is computed as $b = \frac{E_{XY}}{E_{XX}}$ if X is affected by treatments and therefore adjust the treatment means of Y as

$$\bar{Y}_i' = \bar{Y}_i - b(\bar{X}_i - \bar{X})$$

Exa. 6.3

The data of the variables X, Y and treatments T_1, T_2 & T_3 were collected & recorded as shown

Treatments	X	y	X	y	X	y	X	Y	$\bar{X}_\cdot$	$\bar{Y}$
T_1	10	2	16	5	14	7	40	14	13.33	4.67
T_2	15	8	14	2	15	5	44	15	14.67	5.0
T_3	20	5	25	6	21	9	66	20	22	6.67

Determine whether the treatments are significant at $\alpha = 0.05$ l.o.s

Solution.

(i) Analysis of covariance model.

$$y_{ij} = \mu + t_i + b(x_{ij} - \bar{x}) + e_{ij}$$

(ii) Hypothesis

$H_0: \mu_1 = \mu_2 = \mu_3$ against $H_a: \mu_i \neq \mu_k$ for atleast one pair (i,k) at $\alpha = 0.05$ l.o.s

(iii) Analysis of variance components for X.

$$G_{XX} = \sum_i^t \sum_j^n X_{ij}^2 - \frac{G_x^2}{N} = 10^2 + \ldots + 21^2 - \frac{150^2}{9} = 164$$

$$G_{XY} = \sum_i^t \sum_j^n X_{ij} Y_{ij} - \frac{G_x G_y}{N} = 10(2) + \ldots + 21(9) - \frac{150(49)}{9} = 43.33$$

$$G_{YY} = \sum_i^t \sum_j^n Y_{ij}^2 - \frac{G_y^2}{N} = 2^2 + \ldots + 9^2 - \frac{49^2}{9} = 46.22$$

$$T_{XX} = \frac{1}{n} \sum_i^t T_{xi}^2 - \frac{G_x^2}{N} = \frac{1}{3}(40^2 + \ldots + 66^2) - \frac{150^2}{9} = 130.67$$

$$T_{XY} = \frac{1}{n} \sum_i^t T_{xi} T_{yi} - \frac{G_x G_y}{N} = \frac{1}{3}(40(14) + \ldots + 66(20)) - \frac{150(49)}{9} = 30$$

$$T_{YY} = \frac{1}{n} \sum_i^t T_{y_i}^2 - \frac{G_y^2}{N} = \frac{1}{3}(14^2 + \ldots + 20^2) - \frac{49^2}{9} = 6.89$$

$E_{xx} = G_{xx} - T_{xx} = 164 - 130.67 = 33.33$

$E_{xy} = G_{xy} - T_{xy} = 43.33 - 30 = 13.33$

$E_{yy} = G_{yy} - T_{yy} = 46.22 - 6.89 = 39.33$

(iv) Adjusted values of X are given as

$G'_{xx} = G_{xx} - \frac{G_{xy}^2}{G_{yy}} = 164 - \frac{43.33^2}{46.22} = 123.38$

$E'_{xx} = E_{xx} - \frac{E_{xy}^2}{E_{yy}} = 33.33 - \frac{13.33^2}{39.33} = 28.81$

$T'_{xx} = G'_{xx} - E'_{xx} = 123.38 - 28.81 = 94.56$

Source of variation		df	SS_x	SS_y	SS_{xy}	Adjusted df	Adjusted SS_x	Adjusted MS_x	Adjusted F_x
Treatments	t-1	2	130.67	6.89	30.00	2	94.56	47.28	8.20
Error	n-t	6	33.33	39.33	13.33	n-t-1 = 5	28.81	5.76	
Total	n-1	8	164.00	46.22	43.33	n-2 = 7	123.38		

v) Test statistic

$F = 8.20$

vi) Decision (reject Ho)

Tabulated value $F_{0.05}(2,5) = 5.79$, since $F_{calc} > F_{tab}$ then ē differences in treatments means are significant at $\alpha = 0.05$ l.o.s and therefore there are variations due to factor X.

vii) Adjusted means of X groups

$\bar{X}'_i = \bar{X}_i - b(\bar{Y}_i - \bar{Y})$

$b = \frac{E_{xy}}{E_{yy}} = \frac{13.33}{39.33} = 0.3389$

$\bar{X}'_1 = \bar{X}_1 - b(\bar{Y}_1 - \bar{Y}) = 13.33 - 0.3389(4.67 - 5.44) = 13.59$

$\bar{X}'_2 = \bar{X}_2 - b(\bar{Y}_2 - \bar{Y}) = 14.67 - 0.3389(5.00 - 5.44) = 14.82$

$\bar{X}'_3 = \bar{X}_3 - b(\bar{Y}_3 - \bar{Y}) = 22.00 - 0.3389(6.67 - 5.44) = 21.58$

viii) Adjusted components of Y are given as

$G'_{yy} = G_{yy} - \frac{G_{xy}^2}{G_{xx}} = 46.22 - \frac{43.33^2}{164} = 34.77$

$E'_{yy} = E_{yy} - \frac{E_{xy}^2}{E_{xx}} = 39.33 - \frac{13.33^2}{33.33} = 34.00$

$T'_{yy} = G'_{yy} - E'_{yy} = 34.77 - 34.00 = 0.77$

Source of variation	df	SS_x	SS_y	SS_{xy}	df	SS_y	MS_y	F_y
Treatments	2	130.7	6.89	30.00	2			
Error	6	33.3	39.33	13.33	n-t-1 = 5	E'_{yy} 34.00	6.80	
Total	8	164.0	46.22	43.33	n-2 = 7	34.77		
Treatments adjusted for the ave. regression within treatments					t-1 = 2	0.77	0.39	0.057 = 0.39/6.80

ix) Decision

Tabulated value $F_{0.05}(2,5) = 5.79$, since $F_{calc} > F_{tab}$ then the differences in treatments means are significant at $\alpha = 0.05$ 1.0.5

* Exercise.

Don't even read this topic!

7. CLUSTER AND CONJOINT ANALYSIS.

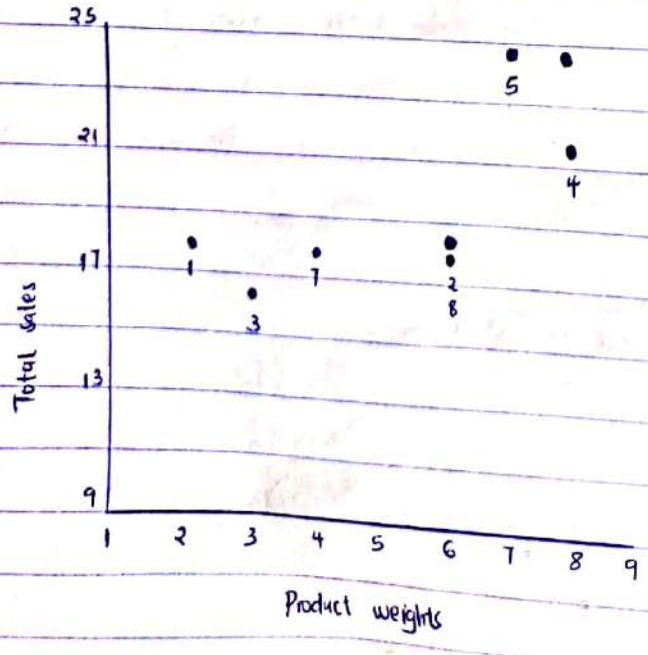
Cluster Analysis.

- The " " technique groups persons, items or objects into unknown number of groups such that members of each group have similar attributes or simply groups that are homogeneous. The different cluster analysis techniques from groups such that the similarity among ē members of the groups is maximised.
- The cluster analysis techniques are similar to discriminant analysis except that whereas in discriminant analysis the number of groups are known, in cluster analysis the groups are unknown. In order to explain the concept of cluster analysis in illustration is conducted using the data of weights of products (X_1) and total sales (X_2) recorded in table below for vendors in a duration of one week.

Sales of product for vendors

Id	1	2	3	4	5	6	7	8
Weight (X_1)	2	6	3	8	7	8	4	6
Sales (X_2)	18	18	18	18	18	18	18	18

Graph of product sales against product weights



The observations in the table are represent in a scatter diagram and the graph reveals the possible 3 groups that may be given as group one with individuals 1, 3, 7, group two with individual 2, 8 and group 3 with individuals 4, 5, 6 resp. The measures of similarity include distance measure, correlation coefficients and association coefficients.

Distance Measure.

- The similarity measure commonly used is the Euclidean distance measure that is given by the formula for instance for k points is given as $d_{ij} = \sqrt{\sum_{k=1}^{n} (X_{ik} - X_{jk})^2}$

Where d_{ij} is the distance btwn ē points i & j

X_{ik} is the coordinate of ē point i along the axis k

X_{jk} is the " " " " j " " " k

n is the total number of variables (axes)

- Considering 2 points (1, 2), the distance btwn the 2 points is given as:

$$d_{12} = \sqrt{(X_{11} - X_{21})^2 + (X_{12} - X_{22})^2}$$

Exa. 7.1

The data of product weights and sales were collected and recorded in table below.

Id	1	2	3	4	5	6	7	8
Weights (X_1)	2	6	3	8	7	8	4	6
Sales (X_2)	18	18	18	18	18	18	18	18

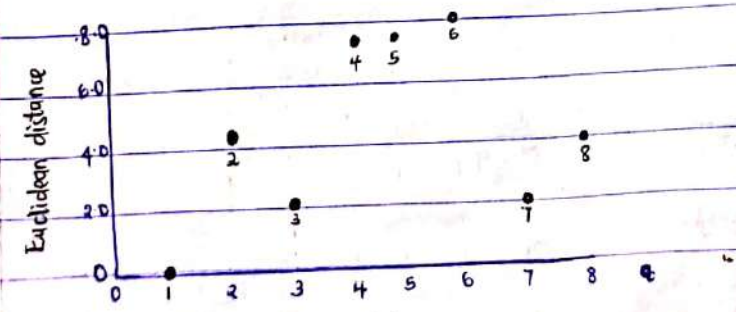
Determine the possible cluster groups for the data using the Euclidean distance measure.

Solution.

$$d_{ij} = \sqrt{\sum (X_{ik} - X_{jk})^2}$$

Id	X_1	X_2	Euclidean, d_{ij}
1	2	18	0.0
2	6	20	4.5
3	3	16	2.2
4	8	22	7.2
5	7	24	7.8
6	8	24	8.5
7	4	18	2.0
8	6	19	4.1

Graph of Euclidean distance against Id



Groups when using Euclidean distance are given as

Group 1: 1, 3, 7 Group 2: 2, 8 Group 3: 4, 5, 6

The representative point for all the objects within a cluster may be the centroid of that cluster. The formula for the coordinates of the centroid of a cluster is given as;

$$Y_k = \frac{1}{a} \sum_{i=1}^{a} X_{ik}$$

Where Y_k is the coordinate of the centroid of the cluster along the axis k, X_{ik}, the coordinate of the point i along the axis k, and a is the total no. of points in the cluster.

If the distance btwn a point / partial cluster and another point / partial cluster is the least among all such values, then they can be grouped together.

Id	X_1	X_2	1	2	3	4	5	6	7	8
1	2	18	0.0	4.5	2.2	7.2	7.8	8.5	2.0	4.1
2	6	20	4.5	0.0	5.0	2.8	4.1	4.5	2.8	1.0
3	3	16	2.2	5.0	0.0	7.8	8.9	9.4	2.2	4.2
4	8	22	7.2	2.8	7.8	0.0	2.2	2.0	5.7	3.6
5	7	24	7.8	4.1	8.9	2.2	0.0	1.0	6.7	5.1
6	8	24	8.5	4.5	9.4	2.0	1.0	0.0	7.2	5.4
7	4	18	2.0	2.8	2.2	5.7	6.7	7.2	0.0	2.2
8	6	19	4.1	1.0	4.2	3.6	5.1	5.4	2.2	0.0

Correlation Coefficient

The correlation coefficient btwn variables can be used as a measure in cluster analysis, however, the measure is not commonly used in many fields. The measure when used in grouping variables considers variables with maximum coefficient as the variables that belong to same group

Association Coefficient.

The measure that considers the presence or absence of attributes in objects such that if a particular attribute is present in an object, then the respective association coefficient is assumed as 1 otherwise it is assumed as 0. The association coefficient is a kind of similarity coefficient and grouping when done using this measure groups objects on the maximum value of such coefficient.

d_{ij}

Id	X_1	X_2	1	2	3	4	5	6	7	8
1	2	18	1.4	0.0	0.0	0.0	0.0	0.0	1.0	0.0
2	6	20	0.0	1.4	0.0	0.0	0.0	0.0	0.0	1.0
3	3	16	0.0	0.0	1.4	0.0	0.0	0.0	0.0	0.0
4	8	22	0.0	0.0	0.0	1.4	0.0	1.0	0.0	0.0
5	7	24	0.0	0.0	0.0	0.0	1.4	1.0	0.0	0.0
6	8	24	0.0	0.0	0.0	1.0	1.0	1.4	0.0	0.0
7	4	18	1.0	0.0	0.0	0.0	0.0	0.0	1.4	0.0
8	6	19	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.4

Conjoint Analysis

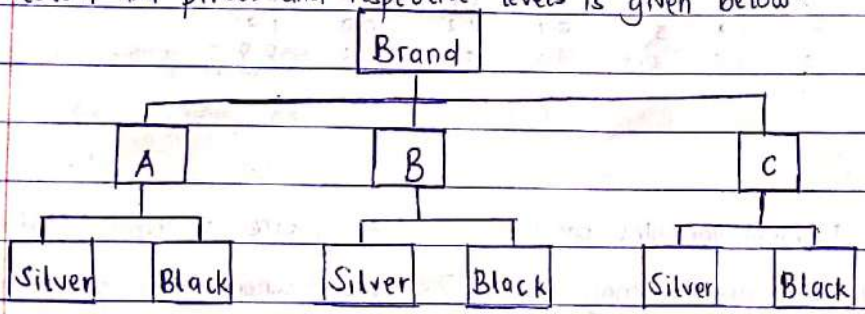
- The " " is a technique normally used to measure the customer preferences on products. Let the number of factors for the product be denoted by m and number of levels for each factor be denoted by $n_1, n_2, \ldots, n_m$, then the total no. of profiles for the product is $n_1 \times n_2 \times \ldots \times n_m$ (Kothari, 2004). A research project manager wish to buy ē computers for a project but before awarding the contract to supplier he/she wish to carry out a study on preference of ē project participants. The preference is affected by the brand, color, and price and ē levels of each of the factors of the product are recorded as below;

Brand : Acer (A), HP (B) and Toshiba (C)

Color : Silver and Black

Price : 50, 120 and 150

The complete factorial combinations of the product profiles in for the factors (brand, color, and price) and respective levels is given below.



	50	q	b	c	d	e	f
Price	120	g	h	i	j	k	l
	150	m	n	o	p	q	r

The description of the profiles are given as:

Profile q : Brand - A, Color - Silver, Price - 50, Profile b : Brand - A, Color - Black, Price - 50

⋮

Profile q : Brand - C, Color - Silver, Price - 150, Profile r : Brand - C, Color - Black, Price - 150

The total no. of product profile is $3 \times 2 \times 3 = 18$, i.e., all ē 18 product profiles are considered for ē study. A card defining ē product profile, for instance ē one shown in Figure 5.4 is represented to ē customer and he/she is required to give preference value using ē interval or ordinal scale. In ē interval scale, the respondent responce is in the scale 0 to 10 where 0 means that ē customer has absolutely no interest on ē product with ē profile shown on ē respective card and 10 means ē customer has extremely high interest to buy that product.

If one choose to use ē ordinal scale, then each respondent is expected to rank all ē product profiles shown to him/her from ē most desirable product profile (rank 1) to ē least desirable product profile (rank 18) such that rank 1 means that ē customer has extremely high interest to buy ē product with ē profile shown on ē respective card and ē rank 18 means that ē customer has absolutely no interest to buy that product.

7.4 Product profile card for the combination a (Brand - A, Color - silver, Price - $ 50)

Card 1	
Brand:	Acer
Color:	Silver
Price:	$ 50
Respondent Rating:	

The traditional conjoint analysis techniques is thought of as a multiple regression problem whose responden't's ratings for ē product concepts are observations on ē dependent variable The xties of ē product or attribute levels are observations on ē independent or predictor variables. The estimated regression coeffiecients associated with ē independent variables are ē part-worth utilities or preference scores for the levels and coefficient of determination R^2 for ē regression xtics ē internal consistency of ē respondent.

- Consider the following coding for ē levels of factors given as:

A-1, B-2, C-3, Silver-1, Black-2

- The factorial experiment design is then given in table below with respective codes.

	Raw				Coding		
Id	Brand	Color	Price	preference	Brand	Color	Price
1	A	Silver	50	5.0	1	1	50
2	A	Silver	120	6.0	1	1	120
3	A	Silver	150	7.0	1	1	150
4	A	Black	50	9.5	1	2	50
5	A	Black	120	8.5	1	2	120
6	A	Black	150	9.0	1	2	150
7	B	Silver	50	4.0	2	1	50
8	B	Silver	120	5.5	2	1	120
9	B	Silver	150	6.5	2	1	150
10	B	Black	50	8.0	2	2	50
11	B	Black	120	7.5	2	2	120
12	B	Black	150	9.5	2	2	150
13	C	Silver	50	3.0	3	1	50
14	C	Silver	120	4.5	3	1	120
15	C	Silver	150	5.0	3	1	150
16	C	Black	50	6.5	3	2	50
17	C	Black	120	7.0	3	2	120
18	C	Black	150	9.0	3	2	150

Now record the responds preference using ē interval scale in which ē respondent rates ē profile in ē scale 0 to 10. Assume the first card is made up of 1st levels of brand, color and price, that is Acer, silver and $ 50 and ē respondent gives preference of 4, then 2nd card has profile brand, color and price of Acer, silver and $ 120, and respondent has preference rating of 5 and so on, ē preferences are recorded resp. for each of ē profiles

- Consider using ē dummy coding procedure for ē independent variables or product xtics. The procedure considers ē coding in which ē dummy coding uses 0 to represent absence of xtic and 1 to represent ē presence of ē xtic.
- The brand factor wil be coded in 3 separate columns (Acer, HP and Toshiba), color would be recorded in 2 column (silver & black) and price in 3 columns (50, 120 & 150)

Dummy coding for the factor levels.

Acer	HP	Toshiba	Silver	Black	50	120	150	Preference
1	0	0	1	0	1	0	0	4
1	0	0	1	0	0	1	0	5
1	0	0	1	0	0	0	1	0
1	0	0	0	1	1	0	0	8
1	0	0	0	1	0	1	0	5
1	0	0	0	1	0	0	1	2
0	1	0	1	0	1	0	0	6
0	1	0	1	0	0	1	0	5
0	1	0	1	0	0	0	1	3
0	1	0	0	1	1	0	0	9
0	1	0	0	1	0	1	0	6
0	1	0	0	1	0	0	1	5
0	0	1	1	0	1	0	0	8
0	0	1	1	0	0	1	0	7
0	0	1	1	0	0	0	1	5
0	0	1	0	1	1	0	0	8
0	0	1	0	1	0	1	0	7
0	0	1	0	1	0	0	1	6
6	6	6	6	6	6	6	6	99

Consider fitting multiple linear regression

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_{p-1} x_{p-1i} + e_i$$

- In multiple regression analysis for ē conjoint analysis there is no independent variable may be perfectly predicted based on ē state of any other independent variable or combination of independent variables and therefore regression procedure cannot separate the effects of ē confounded variables, for instance, we can perfectly predict the state of Brand A based on ē states of Brand B and C, this is called linear dependency.
- In order to solve ē problem of linear dependency, omit one column from each attribute (any of ē attributes, it does not matter which attribute is to be omitted) and therefore omit ē first attribute for each of ē variables (factors) and produce a table as shown in table 7.4

In order to estimate for ē regression parameters ē least squares o4 Maximum likelihood method may be this section ō least squares method is used to solve for ē parameters using ē following steps;

Rewrite the model as

$$Y = X\beta + e$$

Table 7.4 Recorded data for attributes after omitting one attribute on each factor.

HP	Toshiba	Black	120	150	Preference
0	0	0	0	0	4
0	0	0	1	0	5
0	0	0	0	1	0
0	0	1	0	0	8
0	0	1	1	0	5
0	0	1	0	1	2
1	0	0	0	0	6
1	0	0	1	0	5
1	0	0	0	1	3
1	0	1	0	0	9
1	0	1	1	0	6
1	0	1	0	1	5
0	1	0	0	0	8
0	1	0	1	0	7
0	1	0	0	1	5
0	1	1	0	0	8
0	1	1	1	0	7
0	1	1	0	1	6

Multiply both sides by X' such that

$$X'Y = X'X\beta$$

Multiply both sides by inverse of $X'X$ such that

$$\beta = (X'X)^{-1}X'Y$$

$$X'X = \begin{bmatrix} 18 & 6 & 6 & 9 & 6 & 6 \\ 6 & 6 & 0 & 3 & 2 & 2 \\ 6 & 0 & 6 & 3 & 2 & 2 \\ 9 & 3 & 3 & 9 & 3 & 3 \\ 6 & 2 & 2 & 3 & 6 & 0 \\ 6 & 2 & 2 & 3 & 0 & 6 \end{bmatrix} \quad (X'X)^{-1} = \begin{bmatrix} 0.333 & -0.167 & -0.167 & -0.111 & -0.167 & -0.167 \\ -0.167 & 0.333 & 0.167 & 0.000 & 0.000 & 0.000 \\ -0.167 & 0.167 & 0.333 & 0.000 & 0.000 & 0.000 \\ -0.111 & 0.000 & 0.000 & 0.222 & 0.000 & 0.000 \\ -0.167 & 0.000 & 0.000 & 0.000 & 0.333 & 0.167 \\ -0.167 & 0.00 & 0.000 & 0.000 & 0.167 & 0.333 \end{bmatrix}$$

$$X'Y = \begin{bmatrix} 99 \\ 34 \\ 41 \\ 56 \\ 35 \\ 21 \end{bmatrix} \quad \hat{\beta} = \begin{bmatrix} 4.944 \\ 1.667 \\ 2.833 \\ 1.444 \\ -1.333 \\ 3.667 \end{bmatrix}$$

$$\hat{y} = 4.944 + 1.667x_1 + 2.833x_2 + 1.444x_3 - 1.333x_4 - 3.667x_5$$

Hypothesis

$H_0: \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$ against $H_a: \beta_i \neq \beta_k$ for at least one pair (i, k) at $\alpha = 0.05$ l.o.s

From ē table 7.4 above, include

$(\bar{y} - y)^2$: 0.892 | 1.929 | 1.633 | 2.596 | 0.003 | 0.522 | 0.373 | 0.077 | 0.003 | 0.892 | 0.522 | 0.373 | 0.049 | 0.309 | 0.790 | 1.494 | 0.790 | 0.198 | = 13.44

Analysis of Variance table.

Source of variation	df	SS	MS	F
Regression	5	75.06	15.011	13.398
Error	12	13.44	1.120	
Total	17	88.50		

Test statistic

$F = 13.398$

Decision

Tabulated value of $F_{0.05}(5,12) = 3.106$. Since $F_{calc} > F_{tab}$, reject H_0 and conclude $\beta_i \neq \beta_k$ for atleast one pair (i, k) at $\alpha = 0.05$ l.o.s

Performance of model

Coefficient of determination, R^2

$$R^2 = \frac{SS_{Regression}}{SS_{Total}} = \frac{75.06}{88.50} = 0.8481$$ (the regression accounts for 84.81% of variation)

Adjusted coefficient of determination, R_{adj}^2

$$R_{adj}^2 = \left| 1 - \frac{SS_{Error}/df_{Error}}{SS_{Total}/df_{Total}} \right| = \left| 1 - \frac{13.44/12}{88.50/17} \right| = 0.7848$$ (the regression accounts for 74.48% of variation)

In this problem we have considered solving separate regression eqtns for each respondent, therefore to estimate utilities, the respondent must have to evaluate at least as many profiles as parameters to be evaluated.

5/7/2021 Distributions of Linear functions of a Random Vector of Quadratic Forms.

Distribution of Linear functions of a Random Vector.

Properties of Linear Function of a random vector.

a) If a single random variable such as X_1 is multiplied by a constant c, then,

$E(cX_1) = cE(X_1) = c\mu_1$

$V(cX_1) = E(cX_1 - c\mu_1)^2 = c^2 E(X_1 - \mu_1)^2$

$= c^2 V(X_1) = c^2 \delta_{11}$.

b) If X_2 is a second random variable where a & b are constants then using additional properties of expectation we get

$Cov(aX_1, bX_2) = E(aX_1 - a\mu_1)(bX_2 - b\mu_2) = aE(X_1 - \mu_1)(X_2 - \mu_2)b$

$= a\, cov(X_1, X_2)\, b = a\delta_{12} b$

c) If $Z = aX_1 + bX_2$ is a linear combination,

$E(Z) = E(aX_1 + bX_2) = aE(X_1) + bE(X_2)$

$= a\mu_1 + b\mu_2$

$V(Z) = V(aX_1 + bX_2) = E(aX_1 + bX_2 - a\mu_1 - b\mu_2)^2$

$= E(aX_1 - a\mu_1 + bX_2 - b\mu_2)^2$

$= a^2 E(X_1 - \mu_1)^2 + abE(X_1 - \mu_1)(X_2 - \mu_2) + b^2 E(X_2 - \mu_2)^2$

$= a^2\delta_{11} + 2ab\delta_{12} + b^2\delta_{22}$

d) If $c' = [a\ b]$ and variance covariance matrix of $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ is $\Sigma = \begin{bmatrix} \delta_{11} & \delta_{12} \\ \delta_{21} & \delta_{22} \end{bmatrix}$ then we have

$E(c'X) = c'E\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = [a\ b]\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$

$= a\mu_1 + b\mu_2$

$V(c'X) = c'V\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}c = [a\ b]\begin{bmatrix} \delta_{11} & \delta_{12} \\ \delta_{21} & \delta_{22} \end{bmatrix}\begin{bmatrix} a \\ b \end{bmatrix}$

$= c^2\delta_{11} + 2ab\delta_{12} + b^2\delta_{22}$

e. Consider a general linear combination of p random variables $X_1, X_2, \ldots, X_p$ such that

$Z_1 = C_{11}X_1 + C_{12}X_2 + \ldots + C_{1p}X_p$

$Z_2 = C_{21}X_1 + C_{22}X_2 + \ldots + C_{2p}X_p$

$\vdots$

$Z_q = Z_{q1}X_1 + C_{q2}X_2 + \ldots + C_{qp}X_p$

OR

$$Z = \begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_q \end{bmatrix} = \begin{bmatrix} C_{11} & C_{12} & \cdots & C_{1p} \\ C_{21} & C_{22} & \cdots & C_{2p} \\ & \vdots & & \\ C_{q1} & C_{q2} & \cdots & C_{qp} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix} = CX_1$$

$$E(z) = E[C\,X] = CE(X) = CE\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix} = C\mu_X$$

$$V(z) = VCX = C\,V(X)\,C' = CV\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix} C' = C\begin{bmatrix} \delta_{11} & \delta_{12} & \cdots & \delta_{1p} \\ \delta_{21} & \delta_{22} & \cdots & \delta_{2p} \\ & \vdots & & \\ \delta_{p1} & \delta_{p2} & \cdots & \delta_{pp} \end{bmatrix} C'$$

where μ_X and Σ_X are the mean vector and variance covariance matrix of X.

Examples.

1. Let $X' = [X_1\ X_2\ X_3\ X_4]$ be a random vector with mean vector $\mu_X' = [\mu_1\ \mu_2\ \mu_3\ \mu_4]$ and variance covariance matrix.

$$\Sigma_X = \begin{bmatrix} \delta_{11} & \delta_{12} & \delta_{13} & \delta_{14} \\ \delta_{21} & \delta_{22} & \delta_{23} & \delta_{24} \\ \delta_{31} & \delta_{32} & \delta_{33} & \delta_{34} \\ \delta_{41} & \delta_{42} & \delta_{43} & \delta_{44} \end{bmatrix}$$

i) Find the mean vector and variance covariance matrix for the linear combination given as

$Z = X_1 - X_3$

Solution:

$$E(z) = E(CX) = CE(X) = [1\ 0\ -1\ 0]\ E\begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{bmatrix}$$

$$= [1\ 0\ -1\ 0]\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{bmatrix}$$

$$= \mu_1 - \mu_3$$

$$V(z) = V(CX) = CV(X)\,C'$$

$$= [1\ 0\ -1\ 0]\begin{bmatrix} \delta_{11} & \delta_{12} & \delta_{13} & \delta_{14} \\ \delta_{21} & \delta_{22} & \delta_{23} & \delta_{24} \\ \delta_{31} & \delta_{32} & \delta_{33} & \delta_{34} \\ \delta_{41} & \delta_{42} & \delta_{43} & \delta_{44} \end{bmatrix}\begin{bmatrix} 1 \\ 0 \\ -1 \\ 0 \end{bmatrix}$$

$$= [1\ 0\ -1\ 0]\begin{bmatrix} \delta_{11} - \delta_{13} \\ \delta_{21} - \delta_{23} \\ \delta_{31} - \delta_{33} \\ \delta_{41} - \delta_{43} \end{bmatrix}$$

$$= \delta_{11} - \delta_{13} - \delta_{31} + \delta_{33}$$

ii) Find the mean vector and variance covariance matrix for the linear combination given as:

$Z_1 = X_1 - X_3$, $Z_2 = X_2 + X_4$

Exam

$$Z = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix}$$

$$E(z) = CE(x) = \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} E \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{bmatrix} = \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{bmatrix}$$

$$= \begin{bmatrix} \mu_1 - \mu_3 \\ \mu_2 + \mu_4 \end{bmatrix}$$

$$V(z) = VC(x) = CV(x)C'$$

$$= \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \delta_{11} & \delta_{12} & \delta_{13} & \delta_{14} \\ \delta_{21} & \delta_{22} & \delta_{23} & \delta_{24} \\ \delta_{31} & \delta_{32} & \delta_{33} & \delta_{34} \\ \delta_{41} & \delta_{42} & \delta_{43} & \delta_{44} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \delta_{11} - \delta_{13} \\ \delta_{21} - \delta_{23} \\ \delta_{31} - \delta_{33} \\ \delta_{41} - \delta_{43} \end{bmatrix} \begin{bmatrix} \delta_{12} + \delta_{14} \\ \delta_{22} + \delta_{24} \\ \delta_{32} + \delta_{34} \\ \delta_{42} + \delta_{44} \end{bmatrix}$$

$$= \begin{bmatrix} \delta_{11} - \delta_{13} - \delta_{31} + \delta_{33} & \delta_{12} + \delta_{14} - \delta_{32} - \delta_{34} \\ \delta_{21} - \delta_{23} - \delta_{41} - \delta_{43} & \delta_{22} + \delta_{24} + \delta_{42} + \delta_{44} \end{bmatrix}$$

2. Derive expressions of the mean and variance covariance for the following linear combinations in terms of the mean and variance of the random variables X_1, X_2, X_3, X_4

a) $X_1 - X_2 + X_3$

b) $X_1 + X_2 - X_3 + X_4$

If X_1, X_3 and X_4 are independent.

Solution

a) $$E(z) = E(cx) = \begin{bmatrix} 1 & -1 & 1 & 0 \end{bmatrix} E \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{bmatrix} = \begin{bmatrix} 1 & -1 & 1 & 0 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{bmatrix}$$

$$= \mu_1 - \mu_2 + \mu_3$$

$$V(z) = VC(x) = CV(x)C' = \begin{bmatrix} 1 & -1 & 1 & 0 \end{bmatrix} V(x) \begin{bmatrix} 1 \\ -1 \\ 1 \\ 0 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & -1 & 1 & 0 \end{bmatrix} \begin{bmatrix} \delta_{11} & \delta_{12} & \delta_{13} & \delta_{14} \\ \delta_{21} & \delta_{22} & \delta_{23} & \delta_{24} \\ \delta_{31} & \delta_{32} & \delta_{33} & \delta_{34} \\ \delta_{41} & \delta_{42} & \delta_{43} & \delta_{44} \end{bmatrix} \begin{bmatrix} 1 \\ -1 \\ 1 \\ 0 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & -1 & 1 & 0 \end{bmatrix} \begin{bmatrix} \delta_{11} - \delta_{12} + \delta_{13} \\ \delta_{21} - \delta_{22} + \delta_{23} \\ \delta_{31} - \delta_{32} + \delta_{33} \\ \delta_{41} - \delta_{42} + \delta_{43} \end{bmatrix}$$

$$= \delta_{11} - \delta_{12} + \delta_{13} - \delta_{21} + \delta_{22} - \delta_{23} + \delta_{31} - \delta_{32} + \delta_{33}$$

$$= \delta_{11} - 2\delta_{12} + 2\delta_{13} + \delta_{22} - 2\delta_{23} + \delta_{33}$$

b) $E(z) = E(cx) = cE(x) = cE\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = c\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{bmatrix}$

$= [1 \quad 1 \quad -1 \quad 1]\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{bmatrix}$

$= \mu_1 + \mu_2 - \mu_3 + \mu_4$

or

$V(z) = V(cx) = cv(x)c' = [1 \quad 1 \quad -1 \quad 1]\begin{bmatrix} \delta_{11} & \delta_{12} & \delta_{13} & \delta_{14} \\ \delta_{21} & \delta_{22} & \delta_{23} & \delta_{24} \\ \delta_{31} & \delta_{32} & \delta_{33} & \delta_{34} \\ \delta_{41} & \delta_{42} & \delta_{43} & \delta_{44} \end{bmatrix}\begin{bmatrix} 1 \\ 1 \\ -1 \\ 1 \end{bmatrix}$

$= [1 \quad 1 \quad -1 \quad 1]\begin{bmatrix} \delta_{11} + \delta_{12} - \delta_{13} + \delta_{14} \\ \delta_{21} + \delta_{22} - \delta_{23} + \delta_{24} \\ \delta_{31} + \delta_{32} - \delta_{33} + \delta_{34} \\ \delta_{41} + \delta_{42} - \delta_{43} + \delta_{44} \end{bmatrix}$

$= \delta_{11} + \delta_{12} - \delta_{13} + \delta_{14} + \delta_{21} + \delta_{22} - \delta_{23} + \delta_{24} + \delta_{31} + \delta_{32} - \delta_{33} + \delta_{34} + \delta_{41} + \delta_{42} - \delta_{43} + \delta_{44}$

$= \delta_{11} + 2\delta_{12} - 2\delta_{13} + 2\delta_{14} + \delta_{22} - 2\delta_{23} + \delta_{33} - 2\delta_{34} + 2\delta_{24} + \delta_{44}$

$= \delta_{11} + 2\delta_{12} + \delta_{22} - 2\delta_{23} + \delta_{33} + 2\delta_{24} + \delta_{44}$

Exercise:

Derive the expression for the mean & variance-covariance for the following linear combinations in terms of the mean and variance of the random variables X_1, X_2, X_3 and X_4.

$X_1 - X_2 + X_3 + X_4$. If X_2, X_3 and X_4 are independent.

The equation of the best line of fit is an approximation of the simple linear regression where β_0 is the y-intercept and A is the slope/gradient of the best line representation.

Example

The age of the household head (x) and expenditure (y) for households were collected and recorded below.

X	25	28	30	25	30	50	60	72
y	10	12	15	10	20	15	20	18

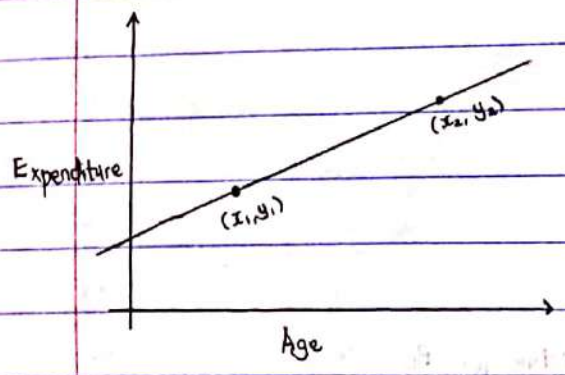
Determine the simple linear regression for estimating the expenditure of the household.

Solution

The simple linear regression model takes the general form

$y_i = \beta_0 + \beta_1 x_1 + \varepsilon_i$

Construct a scatter diagram and draw the best line of fit. Estimate the unknown parameters.



$\hat{\beta}_0$ = y- intercept = 9.27

$\hat{\beta}_1 = \frac{\Delta y}{\Delta x} = \frac{y_2 - y_1}{x_2 - x_1} = \frac{20.70 - 9.27}{30 - 0} = 0.143$

Multivariate Normal Linear Regression and Correlation Analysis.

Simple Linear Regression.

- Simple linear regression model is one that has single response & single predictor variable that takes the general form

$y_i = \beta_0 + \beta_1 x_1 + \theta_i \quad , i = 1, 2, \ldots, n$

[Response] = [mean depending on x_i] + [error]

- In order to predict ē response variable, one has to estimate first the unknown parameters β_0 & β_1
- The methods commonly used to estimate ē unknown parameters include ; graphical, least squares and maximum likelihood methods as described below.

a) Graphical Method.

Construct scatter diagram for ē response & predictor variables. The best line of fit is drawn such that no. of points above ē best line and no. of points below ē $\hat{}$ best line are nearly equal. The equation of the best line of fit is an approximation of the simple linear regression where the β_0 is the y- intercept & β_1 is the slope / gradient of the best line representation

Example:

The The example & solution is above.

b) Least Square Method.

Consider the simple linear regression

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

Make e_i the subject, square both sides:

$$e_i^2 = (y_i - \beta_0 - \beta_1 x_1)^2$$

Sum over all possible sample observations

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_1)^2$$

Let $\phi = \sum_{i=1}^{n} e_i^2$

$$\phi = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_1)^2$$

Differentiate ϕ w.r.t β_0 and evaluate to zero to solve for β_0.

$$\frac{\partial \phi}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\hat{\beta_0} = \sum_{i=1}^{n} (y_i - \beta_1 x_i) = \bar{y} - \beta_1 \bar{x}$$

Differentiate ϕ w.r.t. β_1 and evaluate to 0 to solve for β_1.

$$\frac{\partial \phi}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = 0$$

But, $\beta_0 = \bar{y} - \beta_1 \bar{x}$ so that,

$$(y_i - \bar{y} + \beta \bar{x} - \beta_1 x_i) x_i = 0$$

$$\beta_1 (x_i - \bar{x}) x_i = \sum_{i=1}^{n} (y_i - \bar{y}) x_i$$

$$= \frac{\sum_{i=1}^{n} (y_i - \bar{y}) x_i}{\sum_{i=1}^{n} (x_i - \bar{x}) x_i}$$

$$= \frac{\Sigma (y_i - \bar{y})(x_i - \bar{x})}{\Sigma (x_i - \bar{x})^2}$$

Example:

The age and weight of persons were collected and recorded below.

x	25	28	30	25	30	50	60	72
y	10	12	15	10	20	15	20	18

Repere the simple linear regression to estimate weight.

Solution:

x	y	xy	x^2
25			
28			
30			
25			
30			
50			
60			
72			
320			
$\bar{x}$ = 120		5132	15118

c) Maximum Likelihood Method.

Consider the probability function:

$$f(x_i, \beta, \delta) = \frac{1}{\sqrt{2\pi}\sigma} \exp -1/2\sigma^2 (x_i - \beta_0 - \beta_1 x_i)^2$$

Consider the joint probability function for $x_1, x_2, \ldots, x_n$

$$f(\underline{x}, \beta, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp -\frac{1}{2\sigma^2} (y_1 - \beta_0 - \beta_1 x_1)^2$$

$$\frac{1}{\sqrt{2\pi}\sigma} \exp -\frac{1}{2\sigma^2} (y_2 - \beta_0 - \beta_1 x_2)^2$$

$$\vdots$$

$$\frac{1}{\sqrt{2\pi}\sigma} \exp -\frac{1}{2\sigma^2} (y_n - \beta_0 - \beta_1 x_n)^2$$

$$L = (2\sigma)^{-n/2} \sigma^{2-n/y} \exp -\frac{1}{2\sigma^2} \sum (y - \beta_0 - \beta_1 x_i)^2$$

Find natural way of both sides

$$\ln L = \frac{-n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum (y_i - \beta_0 - \beta_1 x_i)^2$$

Differentiate w.r.t. β_0 and equate to zero.

$$\frac{\partial}{\beta_0} \ln L = -2 \sum (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

Differentiate w.r.t. β_1 and equate to zero.

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = 0$$

$$\hat{\beta_1} = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$$

Differentiate w.r.t. to σ^2

$$n/2\, \sigma^2 = 1/2\sigma \sum (y_i - \beta_0 - \beta_1 x_i)^2$$

$$n\sigma^2 = \sum (y_i - \beta_0 - \beta_1 x_i)^2 \quad \hat{\sigma}^2 = 1/n \sum (y_i - \beta_0 - \beta_1 x_i)^2$$

Under Least Squares.

Example:

x	
y	

We refer the simple linear regression to estimate weight.

Solution:

x	y	xy	x^2
25	10	250	625
28	12	336	784
30	15	450	900
25	10	250	625
30	20	600	900
50	15	750	2500
60	20	1200	3600
72	18	1296	5184
		5132	15118

$$\beta_1 = \frac{\sum xy - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{5132 - 8\left(\frac{320}{8}\right)\left(\frac{120}{8}\right)}{15118 - 8\left(\frac{320}{8}\right)^2} = 0.1432$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x} = 9.271$$

Test significance of simple Linear Regression Models

Hypothesis: $H_0: \beta_1 = 0$ against $H_a: \beta_1 \neq 0$ at α level of significance.

Analysis of Variance Components.

Consider the corrected total sum of squares.

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 = \sum (y_i - \hat{y})^2 + 2\sum (y_i - \hat{y})(\hat{y}_i - \bar{y}) + \sum (\hat{y}_i - \bar{y})^2$$

$$= \sum (y_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

since $2\sum (y_i - \hat{y})(\hat{y}_i - \bar{y}) = 0$

$SS_{Total} = SS_{Error} + SS_{Reg}$

$= SS_{Reg} + SS_{Error}$.

Degrees of freedom.

The total sum of squares has $n-1$, regression has $p-1$ and error has $n-p$. Where p is the value of regression parameters, n is the sample size.

Source of variation	df	SS	MS	F
Regression	$p-1$	By substraction	$\frac{SS_{Reg}}{p-1}$	$\frac{MS_{Reg}}{MSE}$
Error	$n-p$	$\sum (y_i - \hat{y}_i)^2$	$\frac{SSE}{n-p}$	
Total	$n-1$	$\sum (y_i - \bar{y})^2$		

Test statistic

$F = \frac{MS_{Reg}}{MS_{Error}}$ that has approximately F distribution with $p-1, n-p$ d.f.

Decision.

Reject H_0 if $F > F_\alpha (p-1, n-p)$ and conclude that $\beta_1 \neq 0$ at α level of significance.

Coefficient of determination (R^2)

$$R^2 = 1 - \frac{SS_{Error}}{SS_{total}} = \frac{SS_{Reg}}{SS_{Total}}$$

Adjusted Coefficient of determination.

$$R^2_{adj} = 1 - \frac{SS_{error}/n-p}{SS_{Total}/n-1}$$

Test significance of rejection parameter using t - test.

Suppose we wish to test the hypothesis. Hypothesis: $H_0: \beta_1 = \beta_{10}$ against $H_a: \beta_1 \neq \beta_{10}$ at α level of significance.

Test statistic:

$t_1 = \dfrac{\beta_1 - \beta_{10}}{\sqrt{v(\beta_1)}}$ but $v(\hat{\beta}_1) = \dfrac{\sigma^2}{S_{xx}}$ where $S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$

t is approximately t distribution with $n-2$ degrees of freedom.

Decision:

Reject H_0 if $t_1 > t_{\alpha/2, n-p}$ at α level of significance.

Similarly, we wish to test the hypothesis,

Hypothesis; $H_0: \beta_0 = \beta_{00}$ against $H_a: \beta_0 \neq \beta_{00}$ at α 1.0.6

Test statistic

$t_0 = \dfrac{\beta_0 - \beta_{00}}{\sqrt{v(\beta_0)}}$ but $v(\hat{\beta}_0) = \sigma^2 \left(1/n + \dfrac{\bar{x}^2}{S_{xx}}\right)$ where $S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$

t_0 is approximately t distribution with $n-\beta$ degrees of freedom.

Decision:

Reject H_0 if $t_0 > t_{\alpha/2, n-p}$ at α-level of significance.

Confidence Intervals

The $100(1-\alpha)\%$ C.I for $\bar{e}$ parameters are given as

$\hat{\beta}_0 \pm t_{\alpha/2, n-p} \sqrt{\sigma^2 (1/n + \frac{\bar{x}^2}{S_{xx}}}$ so that,

$\hat{\beta}_0 - t_{\alpha/2, n-p} \sqrt{\sigma^2 \left(1/n + \dfrac{\bar{x}^2}{S_{xx}}\right)} < \beta_0 < \hat{\beta}_0 + t_{\alpha/2, n-p} \sqrt{\sigma^2 \left(1/n + \dfrac{\bar{x}^2}{S_{xx}}\right)}$

$\hat{\beta}_1 \pm t_{\alpha/2, n-p} \sqrt{\dfrac{\sigma^2}{S_{xx}}}$ such that

$\hat{\beta}_1 - t_{\alpha/2, n-p} \sqrt{\dfrac{\sigma^2}{S_{xx}}} < \beta_1 < \hat{\beta}_1 + t_{\alpha/2, n-p} \sqrt{\sigma^2/S_{xx}}$

Example

The amount of sand and floor area in urban area construction were collected and recorded below.

Sand (x)	25	28	30	25	30	50	60	72
Area (y)	10	12	15	10	20	15	20	18

Determine the following for determining the floor area:

a) The simple linear regression model using least squares

b) Whether the SLR model is significant.

c) Whether the SLR parameters are significant.

d) 95% C-I of the regression parameters.

Solution.

Consider the simple linear regression model.

$y_i = \beta_0 + \beta_1 x_i + e_i$

x	y	xy	x^2	y^2	$\hat{y}$
25	10	250	625	100	
28	12	336	784	144	
30	15	450	900	225	
25	10	250	625	100	
30	20	600	900	400	
50	15	750	2500	225	
60	20	1200	3600	400	
72	18	1296	5184	324	
320	120	5132	15118	1918	

a) $\hat{y}_i = \beta_0 + \beta_1 x_i$ where,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} xy - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{5132 - 8\left(\frac{320}{8}\right)\left(\frac{120}{8}\right)}{15118 - 8\left(\frac{320}{8}\right)^2} = 0.1432.$$

$$\beta_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$= \frac{120}{8} - 0.1432\left(\frac{320}{8}\right)$$

$$= 9.271$$

Simple linear regression model is given as:

$\hat{y}_i = 9.271 + 0.1432x$

b) Hypothesis ; $H_0: \beta_1 = 0$ against $H_a: \beta_1 \neq 0$.

Analysis of variance component:

Corrected total sum of squares.

$$SS_{Total} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - n\bar{y}^2$$

$$= 10^2 + 12^2 + \ldots + 18^2 - 8\left(\frac{120}{8}\right)^2 = 1918 - 8\left(\frac{120}{8}\right)^2$$

$SS_{error} = \sum (y_i - \hat{y}_i)^2$

$= 2.85^2 + 1.28^2 + \ldots + 0.68^2 = 70.45$

$SS_{reg} = SS_{total} - SS_{error}$.

$= 118 - 70.45 = 47.55$

Anova table.

Source of variation	D.F	SS	MS	F
Regression	1	47.55	47.55	4.05
Error	6	70.45	11.74	
Total	7	118		

T.S $= F = \frac{MS_R}{MSE} = \frac{47.55}{11.74} = 4.05$

Decision

Tabulated value $f_{0.05}(1,6) = 5.99$, since $f < f_\alpha(1,6)$ fail to reject H_0 & conclude that $\beta_1 = 0$ at $\alpha = 0.05$ l.o.s

Significance of the Simple Regression Parameters

Hypothesis : $H_0: \beta_0 = \beta_{00}$ vs $H_a: \beta_0 \neq \beta_{00}$ at $\alpha = 0.05$ l.o.s

$$t_0 = \frac{\beta_0 - \beta_{00}}{\sqrt{v(\beta_0)}} = \frac{9.271}{\sqrt{\sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} = \frac{9.271}{\sqrt{11.74\left(\frac{1}{8} + \frac{(320/8)^2}{15118 - 8(320/8)^2}\right)}}$$

$= 2.997$

Decision

Tabulated $t_{\alpha/2}(n-p) = t_{0.025}(6) = 2.45$, since $t_0 > t_{\alpha/2}(6)$ reject H_0 and conclude that $\beta_0 \neq 0$ at $\alpha = 0.05$ l.o.s.

Hypothesis : $H_0: \beta_1 = \beta_{10}$ vs $H_a: \beta_1 \neq \beta_{10}$ at $\alpha = 0.05$ l.o.s

$$t_1 = \frac{\beta_1 - \beta_{10}}{\sqrt{v(\beta_1)}} = \frac{0.1432}{\sqrt{\frac{11.74}{15118 - 8(\frac{320}{8})^2}}} = 2.012$$

Decision: $t_{\alpha/2, n-p} = t_{0.025}(6) = 2.45$

Since $t_1 < t_{0.025}(6)$ fail to reject H_0 and conclude that $\beta_1 = 0$ at $\alpha = 0.05$ l.o.s.

d) Confidence Intervals:

- The $100(1-\alpha)\%$ confidence intervals are given as:

$\hat{\beta}_0 = 9.271, \quad 9.271 \pm t_{0.025}(6)\sqrt{\sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}$

$= 9.271 \pm 2.45\sqrt{11.74\left(\frac{1}{8} + \left(\frac{320}{8}\right)^2\right)\left(\frac{1}{15118 - 8\left(\frac{320}{8}\right)^2}\right)}$

$= 9.271 \pm 7.580$

$= (1.7000, 10.841)$

$\beta_1 = 0.1432, \quad 0.1432 \pm t_{0.025}(6)\sqrt{\sigma^2/S_{xx}}$

$= 0.1432 \pm 2.45 \sqrt{\frac{11.74}{15118 - 8\left(\frac{320}{8}\right)^2}}$

$= 0.1432 \pm 0.174$

$= (0.031, 0.13172)$

Multiple EXAM

Multivariate Normal Linear Regression and Correlation Analysis.

Multivariate Normal Linear Regression Model.

- The " " " " analysis is statistical methodology for predicting values of one or more response variables from the collection of predictor (independent) variables.

- The multivariate linear regression model with a single response takes the form:

$y_i = \underbrace{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_{p-1} x_{p-1\,i}}_{\text{Mean}} + \underbrace{e_i}_{\text{error}}$

(Response)

- The term linear refers to the fact that the mean is a linear function of the unknown parameters $\beta_0, \beta_1, \dots, \beta_{p-1}$

- Consider, $p = 2$

$y_i = \beta_0 + \beta_1 x_{1i} + e_i$

Consider $p = 3$

$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + e_i$

When n independent observations on y and the associated values of x are collected then complete model is given as:

$y_1 = \beta_0 + \beta_1 x_{11} + \dots + \beta_{p-1} x_{p-11} + e_1$

$y_2 = \beta_0 + \beta_1 x_{12} + \dots + \beta_{p-1} x_{p-12} + e_2$

$\vdots$

$y_n = \beta_0 + \beta_1 x_{1n} + \dots + \beta_{p-1} x_{p-1n} + e_n$

Where $\beta_0, \beta_1, \ldots, \beta_{p-1}$ are the unknown parameters and the error term e_i has the following properties: $E(e_i) = 0$

$$Cov(e_i, e_j) = 0$$

$$V(e_i) = \sigma^2$$

a) Least Squares Method.

- Consider multivariate linear regression.

$$y_i = \beta_0 + \beta_1 x_1 + \ldots + \beta_{p-1} x_{p-1i} + e_i$$

- Make e_i the subject, square both sides over the observations

$$\sum^{n} e_i^2 = \sum^{n} (y_i - \beta_0 - \beta_1 x_{1i} - \ldots - \beta_{p-1} x_{p-1i})^2$$

Let $\phi = \sum_{i=1}^{n} e_i^2$

- Differentiate ϕ w.r.t. β_0 and equate to zero to solve for β_0.

$$\sum^{n} (y_i - \beta_0 - \beta_1 x_{1i} - \ldots - \beta_{p-1} x_{p-1i}) = 0$$

$$\sum^{n} (n\beta_0 + \beta_1 x_{1i} + \ldots + \beta_{p-1} x_{p-1i}) = \sum^{n} y_i$$

Differentiate ϕ w.r.t β_1 and equate to zero to solve for β_1.

$$\sum^{n} (y_i - \beta_0 - \beta_1 x_{1i} - \ldots - \beta_{p-1} x_{p-1i}) x_{1i} = 0$$

$$\sum^{n} (\beta_0 + \beta_1 x_{1i} + \ldots + \beta_{p-1} x_{p-1i}) x_{1i} = \sum y_i \cdot x_{1i}$$

Differentiate ϕ w.r.t. β_{p-1} and equate to zero to solve for β_{p-1}.

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i} - \ldots - \beta_{p-1} x_{p-1i}) x_{p-1i} = 0$$

$$\sum_{i=1}^{n} (\beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{p-1i}) x_{p-1i} = \sum y_i x_{p-1i}$$

$$\begin{bmatrix} n & \sum^{n} x_{1i} & \cdots & \sum^{n} x_{p-1i} \\ \sum^{n} x_{1i} & \sum^{n} x_{1i}^2 & \cdots & \sum^{n} x_{1i} x_{p-1i} \\ \vdots & & & \\ \sum^{n} x_{p-1i} & \sum^{n} x_{1i} x_{p-1i} & \cdots & \sum^{n} x_{p-1i}^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum y_i x_{1i} \\ - \\ \sum y_i x_{p-1i} \end{bmatrix}$$

Such that:

$$X\beta = Y$$

Multiply both sides by X'

$$X'X\beta = X'Y$$

Multiply both sides by $(X'X)^{-1}$

$$\hat{\beta} = (X'X)^{-1} X'Y$$

Where $\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_{p-1} \end{bmatrix}$ $X'X = \begin{bmatrix} n & \hat{\Sigma} x_{1i} & \cdots & \hat{\Sigma} x_{p-1i} \\ \hat{\Sigma} x_{1i} & \hat{\Sigma} x_{1i}^2 & \cdots & \hat{\Sigma} x_{1i} x_{p-1i} \\ \vdots & & & \\ \hat{\Sigma} x_{p-1i} & \Sigma x_{p-1i} x_i & \cdots & \hat{\Sigma} x_{p-1i}^2 \end{bmatrix}$

$X'Y = \begin{bmatrix} \Sigma y_i \\ \Sigma x_{1i} y_i \\ \vdots \\ \Sigma x_{p-1i} y_i \end{bmatrix}$

Analysis of Variance.

Hypothesis: $H_0: \beta_1 = \beta_2 \ldots = \beta_{p-1} = 0$ against $H_a: \beta_i \neq \beta_i$ for atleast one pair.

Analysis of variance component.

Corrected sum of squares.

$$\hat{\sum_{i=1}^{n}} (y_i - \bar{y})^2 = \hat{\sum_{i=1}^{n}} (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2$$

$$= \hat{\sum_{i=1}^{n}} (y_i - \hat{y}_i)^2 + 2\hat{\sum} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + \hat{\sum_{i=1}^{n}} (\hat{y}_i - \bar{y})^2$$

$$= \hat{\sum_{i=1}^{n}} (y_i - \hat{y}_i)^2 + \hat{\sum_{i=1}^{n}} (\hat{y}_i - \bar{y})^2$$

$SS_{Total} = SS_{Reg} + SS_{Error}$

ANOVA table.

Source of variation	df	SS	MS	F
Regression	p-1	by substraction	$SS_{Reg}/p-1$	MS_{Reg}/MS_{Error}
Error	n-p	$\hat{\sum_{i=1}^{n}} (y_i - \hat{y}_i)^2$	$SS_{Error}/n-p$	
Total	n-1	$\hat{\sum_{i=1}^{n}} (y_i - \bar{y})^2$		

Test significance of the Regression Model.

Test statistic: $F = MS_{Reg}/MS_{Error}$ approximately with p-1, n-p d.f.

Decision:

Reject H_0 if $F > F_\alpha (p-1, n-p)$ at α level of significance.

Confidence Interval of Parameters.

The $100(1-\alpha)\%$ confidence interval of the parameters are given as:

$\hat{\beta}_1 \pm t_{\alpha/2,\, n-p} \sqrt{s^2 c_{ii}}$

where: $s^2 = \frac{1}{n-p} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

c_{ii} = i^{th} value of $(X'X)^{-1}$

Coefficient of Determination.

$R^2 = 1 - \frac{SS_{Error}}{SS_{Total}} = \frac{SS_{Reg}}{SS_{Total}}$

Adjusted coefficient of Determination.

$R^2_{adj} = 1 - \frac{SS_{error}/df}{SS_{Total}/df} = 1 - \frac{SS_{Error}/n-p}{SS_{Total}/n-1}$

Example:

EXAM

The data of the dependent variable (Y) and independent variables X_1 & X_2 were collected and recorded below.

X_1	15	15	16	12	15	17	14	15	15	14
X_2	5	6	7	6	6	8	6	7	6	6
Y	25	24	23	20	25	28	22	24	25	24

Determine the following:

a) The multiple linear regression model for estimating Y using the sample data.

b) Test the significance of the multiple linear regression, $\alpha = 0.05$ 1.05

c) The 95% C.I of the regressor parameters.

Solution.

Consider multiple linear regression model:

$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i$

Rewrite the model in matrix form as:

$Y = X\beta$

X_1	X_2	Y	X_1^2	X_1X_2	X_2^2	X_1y	X_2y	y^2	$\hat{y}$	$(y-\hat{y})^2$
15	5	25	225	75	25	375	125	625		
15	6	24	225	90	36	360	144	576		
16	7	23	256	112	49	363	161	529		
12	6	20	144	72	36	240	120	400		
15	6	25	225	90	36	375	150	625		
17	8	26	289	136	64	476	224	784		
14	6	22	196	84	36	308	132	484		
15	7	24	225	105	49	360	168	576		
15	6	25	225	90	36	375	150	625		
14	6	24	196	84	36	336	144	576		
			2206	938	403	3573	1518	5800		