MULTIVARIATE ANALYSES

INTRODUCTION

- Multivariate analysis is used to describe analyses of data where there are multiple variables or observations for each unit or individual.

- Often times these data are interrelated and statistical methods are needed to fully answer the objectives of our research.

Examples Where Multivariate Analyses May Be Appropriate

- The study to determine the possible causes of a medical condition, such as heart disease. An initial survey of non-disease males is conducted and data are collected on age, body weight, height, serum cholesterol, phospholipids, blood glucose, diet, and many other putative factors. The history of these males is followed and it is determined if and when they may be diagnosed with heart disease.

- Determining the value of an apartment. Factors possibly related to the value are size of the apartment, age of the building, number of bedrooms, number of bathrooms, and location (e.g. floor, view, etc.).

- A medical study is conducted to determine the effects of air pollution on lung function. Because you can't assign people randomly to treatment groups (i.e. a rural environment with no air quality concerns vs. a large city with air quality issues), a research chose four cohorts that live in areas with very different air quality and each location is close to an air-monitoring device. The researchers took measures of lung function on each individual at two different time periods by recording one breath. Data collected on this breath included length of time for the inhale and exhale, speed and force of the exhale, and amount of air exhaled after one second and at the mid-point of the exhale. The air quality at these two times was also recorded.

- Political surveys to determine which qualities in a candidate are most important in garnering popularity.

Types of Multivariate Analyses To Be Taught

- *Multiple linear regression*: A linear regression method where the dependent variable Y is described by a set of X independent variables. An example would be to determine the factors that predict the selling price or value of an apartment.

- *Multiple linear correlation*: Allows for the determination of the strength of the strength of the linear relationship between Y and a set of X variables.

- *Multivariate nonlinear regression*: A form of regression analysis in which the dependent variable Y is described by a nonlinear combination of the independent variables X.

- *Response Surface Regression*: A form of multivariate non-linear regression where the influences of several independent or “response” variables on a dependent variable are determined. The goal of response surface regression is to optimize a response.

- *Discriminant analysis:* In an original survey of males for possible factors that can be used to predict heart disease, the researcher wishes to determine a linear function of the many putative causal factors that would be useful in predicting those individuals that would be likely to have a heart attack within a 10-year period.

- *Principal component analysis (PCA)*: Is used to simplify the description of a set of interrelated variables. PCA considers all variables equally; they are not divided into dependent and independent variables. In PCA, the interrelated variables are in essence transformed into new, uncorrelated values. Using the data from the lung function example, the data for each individual are highly interrelated since they were all recorded on one breath. Because the data are interrelated, you need to use a method that develops a new set of measurements that are uncorrelated with each other. PCA allows development of new uncorrelated measurements called principal components. It is hoped that the first 2-3 of the principal components can be used to explain the original variation in lung function. Use of PCA may allow you to use fewer principal components than the number of variables in the original data set and help to simply the interpretation and explanation of the results.

- *Factor analysis*: Is similar to PCA in that it allows one to determine the interrelationships among a set of variables. Like PCA, factor analysis does not have a dependent variable that is described by a set of independent variables. Using our political survey example, factor analysis will allow you to group each of the questions into subgroups that are uncorrelated with each other.

- *Cluster analysis*: Is a method for grouping individuals or objects into unknown groups. This method differs from discriminant analysis in that the number and the characteristics of the groups are unknown prior to the analysis.

CHARACTERIZING DATA

Types of Variables

1. Qualitative variable:
    - One in which numerical measurement is not possible.
    - An observation is made when an individual is assigned to one of several mutually exclusive categories (i.e. cannot be assigned to more than one category).
    - Non-numerical data.
    - Observations can be neither meaningfully ordered nor measured (e.g. hair color, resistance vs. susceptibility to a pathogen, etc.).
2. Quantitative variable:
    - One in which observations can be measured.
    - Observations have a natural order of ranking.
    - Observations have a numerical value (e.g. yield, height, enzyme activity, etc.)

- Quantitative variables can be subdivided into two classes:
    1. *Continuous*: One in which all values in a range are possible (e.g. yield, height, weight, etc.).
    2. *Discrete*: One in which all values in a range are not possible, often counting data (number of insects, lesions, etc.).

Steven's Classification of Variables

- Stevens (1966)[1] developed a commonly accepted method of classifying variables.

    1. *Nominal variable*:
        - Each observation belongs to one of several distinct categories.
        - The categories don't have to be numerical.
        - Examples are sex, hair color, race, etc.
    2. *Ordinal variable*:
        - Observations can be placed into categories that can be ranked.

[1] Stevens, S.S. 1966. Mathematics, measurement and psychophysics. pp. 1-49. *In* S.S. Stevens (ed.) Handbook of experimental psychology. Wiley, New York.

- An example would be rating for disease resistance using a 1-10 scale, where 1=very resistant and 10=very susceptible.

- The interval between each value in the scale is not certain.

3. *Interval variables*:
   - Differences between successive values are always the same.

   - Examples would be temperature and date.

4. *Ratio variables*:
   - A type of interval variable where there is a natural zero point or origin of measurement.

   - Examples would be height and weight.

   - The difference between two interval variables is a ratio variable.

Descriptive measures depending on Steven's scale†

Classification	Graphical measures	Measures of central tendency	Measures of dispersion
Nominal	Bar graphs Pie charts	Mode	Binomial or multinomial variance
Ordinal	Bar graphs Histogram	Median	Range
Interval	Histogram areas are measurable	Mean	Standard deviation
Ratio	Histogram areas are measurable	Geometric mean Harmonic mean	Coefficient of variation

†Table adapted from Afifi, A., S. May, and V.A. Clark. 2012. Practical multivariate analysis 5th edition. CRC Press, Taylor and Francis Group, Boca Raton, FL.

Presenting Variables

1. Y_i notation
   a) In this course, we are going to use the letter Y to signify a variable using the Y_i notation.

   b) Y_i is the $i^{\underline{th}}$ observation of the data set Y. (Y_1, Y_2, Y_3 . . . Y_n).

   c) If Y=1, 3, 5, 9, then Y_1=___ and Y_3=___.

2. Vector notation
   - The modern approach to presenting data uses vectors.

   - Specifically, a vector is an ordered set of n elements enclosed by a pair of brackets.

$$\text{Y}=\begin{vmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \ldots \\ Y_n \end{vmatrix}$$

Using numbers from the previous example,

$$\text{Y}=\begin{vmatrix} 1 \\ 3 \\ 5 \\ 9 \end{vmatrix}$$

   - Y' is called the transpose of Y.

   - The transpose of a column vector is a row vector.

   - Using the previous example, $Y' = \begin{vmatrix} 1 & 3 & 5 & 9 \end{vmatrix}$

3. Matrices
   - Have numbers or values arranged in rows and columns.

   - The size of matrices is described using the nomenclature of *a x b*, where the first number is the number of rows and the second number is the number of columns.

   - The nomenclature used in naming elements in vectors is X_{ij} where *i*=row and *j*=column.

   - A square matrix has the same number of columns and rows.

   - For example, a 4 x 4 square matrix would take on the form:

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

Vector and Matrix Math

1. Multiplying two vectors
   - A row and column vector can be multiplied if each vector has the same number of elements.
   - The product of vector multiplication is the sum of the cross products of the corresponding entries.
   - Multiplication between two column vectors requires taking the transpose of one of the vectors.
   - For example, if

$$X = \begin{bmatrix}1\\3\\4\end{bmatrix} \text{ and } Y = \begin{bmatrix}2\\4\\5\end{bmatrix} \qquad \text{then } X'Y = \begin{bmatrix}1 & 3 & 4\end{bmatrix} \text{ x } \begin{bmatrix}2\\4\\5\end{bmatrix}$$

$X'Y$=(1*2) + (3*4) + (4*5) = 34

2. Multiplying a scalar (λ) and a matrix
   - The multiplication between a scalar (a number) and matrix gives a new matrix where each element is multiplied by the scalar.
   - For example, if

$$\lambda\text{=4 and X=}\begin{bmatrix}1 & 4 & 7\\2 & 5 & 8\\3 & 6 & 9\end{bmatrix}$$

$$\text{Then 4X=}\begin{bmatrix}4*1 & 4*4 & 4*7\\4*2 & 4*5 & 4*8\\4*3 & 4*6 & 4*9\end{bmatrix} = \begin{bmatrix}4 & 16 & 28\\8 & 20 & 32\\12 & 24 & 36\end{bmatrix}$$

3. Multiplying a matrix and a vector
   - The number of columns in the first matrix must be equal to the number of rows in the vector.
   - For example, a matrix of size *n x m* can be multiplied with a vector with *m* rows (*m x 1)*.
   - In multiplication, the elements of the rows in the first matrix are multiplied with the corresponding elements in the vector. The sum of these products becomes the element in the product matrix.
   - The product matrix is one with a size of *n x 1*.

- For example,

  Maxtrix $X = \begin{vmatrix} 1 & 3 \\ 2 & 4 \end{vmatrix}$ and Vector $Y = \begin{vmatrix} a \\ b \end{vmatrix}$, then

  XY will be a 2 x 1 matrix equal to:

  $\begin{vmatrix} (1a + 3b) \\ (2a + 4b) \end{vmatrix}$

4. Multiplying two matrices
   - The number of columns in the first matrix must be equal to the number of rows in the second matrix.
   - For example, a matrix of size *n x m* can be multiplied with a matrix of size *m x p*.
   - In multiplication, the elements of the rows in the first matrix are multiplied with the corresponding columns in the second matrix. The sum of these products becomes the element in the product matrix.
   - The product matrix is one with a size of *n x p*.
   - For example, multiplying a 2 x 3 by a 3 x 2 matrix would, would give you a product matrix of size 2 x 2

$$X = \begin{bmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{bmatrix} \text{ and } \begin{bmatrix} a & d \\ b & e \\ c & f \end{bmatrix} = \begin{bmatrix} (1a + 3b + 5c) & (1d + 3e + 5f) \\ (2a + 4b + 6c) & (2d + 4e + 6f) \end{bmatrix}$$

5. Identity Matrix
   - A matrix of size n x n with values of one on the diagonals and zero elsewhere.
   - The identity matrix is denoted with the nomenclature I_n
   - For example

$$I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \text{ and } I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

6. Eigenvectors and Eigenvalues
   - Eigenvectors and eigenvalues are dependent on the concepts of vectors and linear transformations.
   - Vectors can be thought of as arrows of fixed length and direction.

- Vectors can be described by a set of Cartesian coordinates that are numbers.

- A linear transformation can be described by a square matrix.

- If the multiplication of the vector by the square matrix results in a change in length of the vector but does not result in the change of the direction or changes the vector to the opposite direction, the vector is called an eigenvector of that matrix.

- Example
  - Given that A is an n x n matrix, X is a non-zero vector, and λ is a scalar such that $AX= \lambda X$, then we can call X an eigenvector of A and λ an eigenvalue of A.

  - Let $A = \begin{bmatrix} 6 & 16 \\ -1 & -4 \end{bmatrix}$ and $X = \begin{bmatrix} -8 \\ 1 \end{bmatrix}$.

    $AX = \begin{bmatrix} 6 & 16 \\ -1 & -4 \end{bmatrix} \quad \begin{bmatrix} -8 \\ 1 \end{bmatrix} = \begin{bmatrix} -32 \\ 4 \end{bmatrix} = 4\begin{bmatrix} -8 \\ 1 \end{bmatrix}$

    $\begin{bmatrix} -8 \\ 1 \end{bmatrix}$ is an eigenvector with an eigenvalue of 4

  - Let $A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$ and $X = \begin{bmatrix} 3 \\ -3 \end{bmatrix}$.

    $AX = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \quad \begin{bmatrix} 3 \\ -3 \end{bmatrix} = \begin{bmatrix} 3 \\ -3 \end{bmatrix} = 1\begin{bmatrix} 3 \\ -3 \end{bmatrix}$

    $\begin{bmatrix} 3 \\ -3 \end{bmatrix}$ is an eigenvector with an eigenvalue of 1

- **The determination of an eigenvector and eigenvalue cannot be done for all vectors.**

Description of Variables in Data Analysis

1. *Dependent or outcome variable:*

- Most of the data you collect in an experiment.

- Can be thought of as the outcome of the experiment.

- Often referred to in the data using the letter *Y*.

- The term dependent does not necessarily mean that there necessarily is a causal relationship between dependent and independent variables.

Standardizing Variables in Multivariate Analyses

- Often times in multivariate analyses the independent variables have different scales (e.g. monetary value, area, temperature, concentration, etc.).

- Given that some of these variables may have very large values while others are very small, the variables may not contribute equally to the analysis due to differences in scale.

- For example, variable *X1* may have a scale of 1-100 and variable *X2* has a scale of 1-10. Using these variables without standardization could result in variable *X1* a larger influence on the results.

- To minimize this scaling effect, the original data can be transformed so the ranges of the new variables are comparable and data variability is reduced.

- Common standardization methods include:

    - **0-1 scaling**: each variable in the dataset is recalculated as $(X_n - min\ X_n)/range\ X_n$.
        - Variables have different means and standard deviations, but equal ranges. Also at least one variable has a value of *0* and another has a value of *1*.

    - Dividing each value by the range (*X/range X)*.
        - Variables have different means and standard deviations, but similar ranges.

    - **Z-score scaling**: each variable in the dataset is recalculated as $(X_n\text{-}mean\ X_n)/s$. All variables in the dataset have a mean equal to *0*, variance equal to *1*, but different ranges.

- Dividing each variable by the standard deviation. This results in variables with equal to *1*, but different means and ranges.