Multivariate Analysis 2019-2020

Prof Dr NAJLAA FAWZI

Multivariate statistics is a form of statistics encompassing the simultaneous observation and analysis of more than one statistical variable.

The application of multivariate statistics is multivariate analysis

Multivariate statistics concerns understanding the different aims and background of each of the different forms of multivariate analysis, and how they relate to each other.

The practical implementation of multivariate statistics to a particular problem may involve several types of univariate and multivariate analysis in order to understand the relationships between variables and their relevance to the actual problem being studied.

Multivariate analysis simultaneously analyze the effect of a number of variables while analyzing data to measure association.

Multivariate Analysis

There may be multiple factors affecting the outcome. Then we have to quantify each of them in the order of importance. For example, in carcinoma breast a number of factors affect the outcome like estrogen receptors, nodal status, tumor size, menopausal status, etc.

Multivariate analysis is more realistic in a variety of medical conditions.

It can be used for more complex data where univariate analysis is not possible.

If univariate analysis is done, it ignores many variables, and outcome prediction is somewhat less accurate.
On the conside, multivariate analysis is more complex and requires a larger number of observations to accurately predict the outcome.

The aim of multivariate analysis is to predict the outcome based on some existing information.

For example, if TNM status and grade of the tumor of a particular cancer are known, 5-year survival can be predicted. The second aim is to explain. For example, we can explain which variable out of four variables, is the most important variable (factor analysis is a type of multivariate analysis:

A factor can have many variables. So number of variables are taken together to form a factor to analyze), which is the most important variable that can be explained. Different methods (like MANOVA, logistic regression, multiple regression, ANCOVA, etc.) are applied for different types of data.

Examples where multivariate analysis is applicable

Example-1

To study the factors affecting weight reduction, researcher collects data on dietary habits, calorie intake per day, vegetarian or non vegetarian, the type of work (sedentary or manual), and exercise habits.

How do these data affect outcome (weight reduction) after an intervention?

Example-2

The risk of stroke and cardiac event is dependent upon many factors like hypercholesterolemia, diabetic or not, whether hypertensive or not, and smoker or not.

A study as to how much these factors contribute individually to myocardial infarction

had to consider all these variables in the analysis.

Multivariate analysis is not a single test.

Different types of analysis are required for different types of data.

Multivariate analysis can quantify the importance of multiple factors in the order of importance.

Functions of multivariate analysis

- **Control for confounders**
- **Test for interactions between predictors (effect modification)**
- **Improve predictions**

Two broad types of multivariate analysis are

1- linear regression model [when the dependent variable (outcome) is continuous]

2-logistic regression model [when the dependent variable (outcome) is discrete].

In addition, when the time of follow up was not even between cases; then we use proportional hazard model.

Other types of multivariate regression

- Multiple linear regression is for normally distributed outcomes
- Logistic regression is for binary outcomes
- Cox proportional hazards regression is used when time-to-event is the outcome

Uses:

1-Explanatory uses: explains the relationship between the outcome (disease), and predictors (exposure and other factors).

2-Predictive uses: the estimated coefficient can be used to calculate the probability that an individual (who is having a specific set of predictors' values) will experience the outcome of interest.

Multiple regression is an extension of simple linear regression.
It is used when we want to predict the value of a variable based on the value of two or more other variables.
The variable we want to predict is called the dependent variable (or sometimes, the outcome, target or criterion variable).

Multiple regression generally explains the relationship between multiple independent or predictor variables and one dependent or criterion variable.
A dependent variable is modeled as a function of several independent variables with corresponding coefficients, along with the constant term.
Multiple regression requires two or more predictor variables, and therefore it is called multiple regression.

Multiple regression

Is the procedure for quantifying the relationship of one variable with two or more variables.

Utilities of multiple regression

1- It helps to assess or study the causative factors in the variable of interest.

Regression is a good devise for causal relationship.

2- The scientific laws can be described with regression procedures.

Height or weight or nutritional intake and oxygen consumption varies by age, sex, and grades of nutritional status or climate.

Patterns and laws can be quantified with careful studies.

3- Regression helps to have a substitution of some specific variables for others which are costly or time consuming to measure.

Variables like lean body mass is difficult to measure , but it is related to age, height , body mass index.

Regression procedure helps to estimate lean body mass from other variables.

4- Statistical control of variation in various biochemical parameters and comparison of various dependent variables can be studied with use of multiple or univariate regression analysis.

Multiple Regression: To Predict an Outcome

Multiple regression is like linear regression. Here the outcome (dependent variable or target variable) is predicted depending upon two more input variables.

Interpretation of the results of multiple regression is complex and difficult as multiple variables are involved and different variables may have different degrees of influence on the outcome.

For example, predicting the 5-year survival of a cancer patient depending upon the *TNM* status of the patient. Here, *T* status, *N* status, and *M* status are the three independent variables. Predicting the 5-year survival rate is the outcome.

Multiple Linear Regression

The concept of multiple regression used for analyzing the association among several variables are the extension of the linear regression .

$$y = a + b\,x$$

$$y = a + b1x1 + b2x2 + b3x3.....+ b\,n\,x\,n$$

The concept of simple linear regression where a single predictor variable X was used to model the response variable Y . In many applications, there is more than one factor that influences the response. Multiple regression models thus describe how a single response variable Y depends linearly on a number of predictor variables.

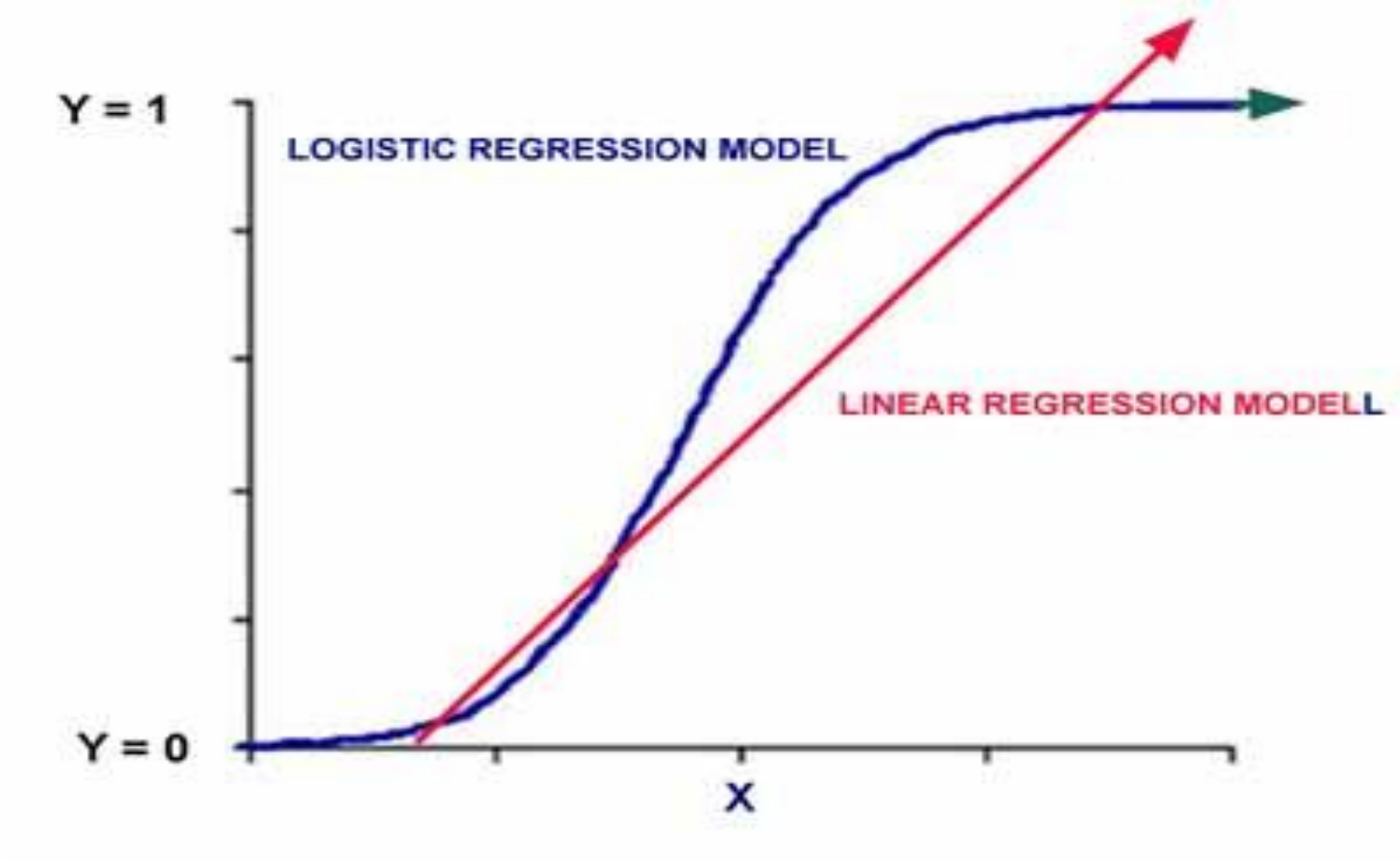
- **More than one predictor…**

E(y)= α + β_1*X + β_2 *W + β_3 *Z…

Each regression coefficient is the amount of change in the outcome variable that would be expected per one-unit change of the predictor, if all other variables in the model were held constant.

The height of a child can depend on the height of the mother, the height of the father, nutrition, and environmental factors.

The coefficient of each independent (b) can be interpreted as the magnitude of change in the value of the mean of the dependent variable for every unit change in that predictor variable, taking into account the effect of all other variables in the model.


Y = 1
LOGISTIC REGRESSION MODEL
LINEAR REGRESSION MODELL
Y = 0
X

Logistic Regression

It is a statistical method of analyzing variables (one or more) to predict whether an outcome falls into a category or not.

Logistic regression analysis studies the association between a categorical dependent variable and a set of independent (explanatory) variables.

The name logistic regression is used when the dependent variable has only two values, such as 0 and 1 or Yes and No.

In other words, it predicts *dichotomous outcome* like survival or death, male or female, recurrence or no recurrence, wound gets infected or no infection, etc. Dependent variable is categorical. For example, let us assume there is a condition where mortality is related to age.

Logistic Regression Example: Tumor Prediction

A Logistic Regression classifier may be used to identify whether a tumor is malignant or if it is benign. Several medical imaging techniques are used to extract various features of tumors. For instance, the size of the tumor, the affected body area, etc. These features are then fed to a Logistic Regression classifier to identify if the tumor is malignant or if it is benign.

Age as a predictor of survival (in a condition): Logistic regression applicable

Age group (in years)	Average no. of survivors in a group of 100 patients	Probability of surviving
11–20	21	21
21–30	26	26
31–40	8	8
41–50	2	2
51–60	2	2
61+	1	1

This data may be plotted as scatter graph as in the above example. From the graph, the probability of survival or death can be predicted for a patient when the age is known.

2,677 adults referred to a sleep clinic with suspected sleep apnoea. They developed an apnoea severity index, and related this to the presence or absence of hypertension. They wished to answer two questions:
i) Is the apnoea index predictive of hypertension, allowing for age, sex and body mass index?
ii) Is sex a predictor of hypertension, allowing for the other covariates?
The results are given in table 1 below.

Risk factors for hypertension

Risk factor	Estimate (log odds)	(95% CI)	Odds ratio
Age (10 years)	0.805	(0.718 to 0.892)	2.24
Sex (male) 0.161	(-0.061 to 0.383)	1.17	
BMI (5 kg/m2)	(0.256 to 0.409)	0.332	1.39
Apnoea index (10 units)	0.116	(0.075 to 0.156)	1.12

The Purpose Of Logistic Regression

The crucial limitation of linear regression is that it cannot deal with dependent variable's that are dichotomous and categorical.

Many interesting variables are dichotomous: for example, consumers make a decision to buy or not buy, a product may pass or fail quality control, there are good or poor credit risks, an employee may be promoted or not.

Like ordinary regression, logistic regression provides a coefficient ‘b’, which measures each independent variable’s partial contribution to variations in the dependent variable.

Logistic regression

In statistics, logistic regression (sometimes called the logistic model or logit model) is used for prediction of the probability of occurrence of an event by fitting data to a logistic function.

It is a generalized linear model used for binomial regression.

Like other forms of regression analysis, it makes use of one or more predictor variables that may be either numerical or categorical.

For example, the probability that a person has a stroke within a specified time period might be predicted from knowledge of the person's age, sex and body mass index.

Logistic regression is used extensively in the medical and social sciences fields, as well as marketing applications such as prediction of a customer's propensity to purchase a product or cease a subscription.

Logistic regression takes S shape distribution.

The probability of having the outcome (expressed between 0 to 1).

The logistic regression model is widely used in health science research , the model is frequently used by epidemiologist as a model for the probability (interpreted as the risk), that an individual will acquire a disease during some specified time period during which he or she is exposed to a condition (called risk factor) known to be or suspected of being associated with the disease.

One very practical advantage of logistic regression over multiple linear regression in epidemiologic research is that these coefficients can be directly converted to an odds ratio that provides an estimate of the relative risk that is adjusted for confounding.

Logistic regression is a powerful statistical tool for estimating the magnitude of the association after adjusting simultaneously for number of potential confounding factors.

A. y = patients survive (y = 0) or die (y =1)
x1 = therapy (x1 = A, B; nominal)
x2 = age (in years; continuous)
x3, . . . = laboratory parameters.

case–control–study (epidemiology)
y = case (y = 1) or control (y = 0)
x1 = exposed (x1 = 1) or not (x1 = 0)

x2, . . . = confounder.

Example The application of a logistic regression may be illustrated using a fictitious example of death from heart disease. This simplified model uses only three risk factors (age, sex, and blood cholesterol level) to predict the 10-year risk of death from heart disease. These are the parameters that the data fit:

$\beta_0 = -5.0$ (the intercept)
$\beta_1 = +2.0$
$\beta_2 = -1.0$
$\beta_3 = +1.2$
x_1 = age in years, above 50
x_2 = sex, where 0 is male and 1 is female
x_3 = cholesterol level, in m mol/L above 5.0

In this model, increasing age is associated with an increasing risk of death from heart disease (z goes up by 2.0 for every year over the age of 50), female sex is associated with a decreased risk of death from heart disease (*z* goes down by 1.0 if the patient is female), and increasing cholesterol is associated with an increasing risk of death (z goes up by 1.2 for each 1 mmol /L increase in cholesterol above 5 mmol /L).

We wish to use this model to predict subject's risk of death from heart disease: he is 55 years old and his cholesterol level is 7.0 mmol/L. The subject's risk of death is:

Step wise Regression:

The procedure consists of a series of steps.

At each step of procedure each variable then in the model is evaluated to see if , according to specified criteria it should remain in the model.

Example: A nursing director would like to use nurse personal characteristics to develop a regression model for predicting the job performance (dependent) , while the following variables are variables from which to choose the independent variables to include in the model:

confidence, interest, ambition , communication skills , problem solving skills, creativity.

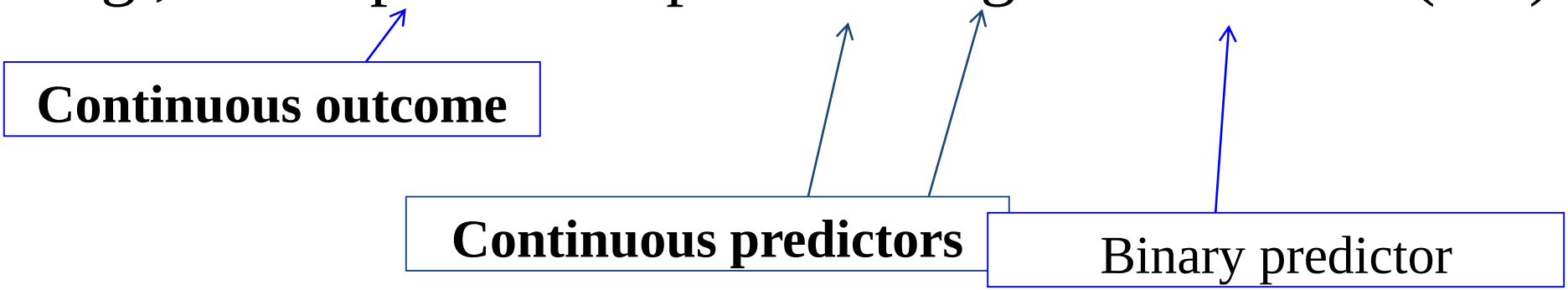
Common multivariate regression models

Outcome (dependent variable)	Example outcome variable	Appropriate multivariate regression model	Example equation	What do the coefficients give you?
Continuous	Blood pressure	Linear regression	blood pressure (mmHg) = α + βsalt*salt consumption (tsp/day) + βage*age (years) + βsmoker*ever smoker (yes=1/no=0)	slopes—tells you how much the outcome variable increases for every 1-unit increase in each predictor.
Binary	High blood pressure (yes/no)	Logistic regression	ln (odds of high blood pressure) = α + βsalt*salt consumption (tsp/day) + βage*age (years) + βsmoker*ever smoker (yes=1/no=0)	odds ratios—tells you how much the odds of the outcome increase for every 1-unit increase in each predictor.
Time-to-event	Time-to-death	Cox regression	ln (rate of death) = α + βsalt*salt consumption (tsp/day) + βage*age (years) + βsmoker*ever smoker (yes=1/no=0)	hazard ratios—tells you how much the rate of the outcome increases for every 1-unit increase in each predictor.

Review of statistical tests

- **The following table gives the appropriate choice of a statistical test or measure of association for various types of data (outcome variables and predictor variables) by study design.**

e.g., blood pressure= pounds + age + treatment (1/0)



Types of variables to be analyzed

Cross-sectional/case-control studies

Predictor variable/s	Outcome variable	Statistical procedure or measure of association
Binary (two groups)	Continuous	T-test
Binary	Ranks/ordinal	Wilcoxon rank-sum test
Categorical (>2 groups)	Continuous	ANOVA
Continuous	Continuous	Simple linear regression
Multivariate (categorical and continuous)	Continuous	Multiple linear regression

Categorical	**Categorical**	**Chi-square test (or Fisher's exact)**
Binary	**Binary**	**Odds ratio, risk ratio**
Multivariate	**Binary**	**Logistic regression**

Cohort Studies/Clinical Trials

Binary	Binary	Risk ratio

Alternative summary: statistics for various types of outcome data

Outcome Variable	Are the observations independent or correlated?		Assumptions
	independent	correlated	
Continuous (e.g. pain scale, cognitive function)	Ttest ANOVA Linear correlation Linear regression	Paired ttest Repeated-measures ANOVA Mixed models/GEE modeling	Outcome is normally distributed (important for small samples). Outcome and predictor have a linear relationship.
Binary or categorical (e.g. fracture yes/no)	Difference in proportions Relative risks Chi-square test Logistic regression	McNemar's test Conditional logistic regression GEE modeling	Chi-square test assumes sufficient numbers in each cell (>=5)
Time-to-event (e.g. time to fracture)	Kaplan-Meier statistics Cox regression	n/a	Cox regression assumes proportional hazards between groups

Continuous outcome (means)

Outcome Variable	Are the observations independent or correlated?		Alternatives if the normality assumption is violated (and small sample size):
	independent	correlated	
Continuous (e.g. pain scale, cognitive function)	**Ttest:** compares means between two independent groups **ANOVA:** compares means between more than two independent groups **Pearson's correlation coefficient** (linear correlation): shows linear correlation between two continuous variables **Linear regression:** multivariate regression technique used when the outcome is continuous; gives slopes	**Paired ttest:** compares means between two related groups (e.g., the same subjects before and after) **Repeated-measures ANOVA:** compares changes over time in the means of two or more groups (repeated measurements) **Mixed models/GEE modeling**: multivariate regression techniques to compare changes over time between two or more groups; gives rate of change over time	Non-parametric statistics **Wilcoxon sign-rank test**: non-parametric alternative to the paired ttest **Wilcoxon sum-rank test** (=Mann-Whitney U test): non-parametric alternative to the ttest **Kruskal-Wallis test:** non-parametric alternative to ANOVA **Spearman rank correlation coefficient:** non-parametric alternative to Pearson's correlation coefficient

Binary or categorical outcomes (proportions)

Outcome Variable	Are the observations correlated?		Alternative to the chi-square test if sparse cells:
	independent	correlated	
Binary or categorical (e.g. fracture, yes/no)	**Chi-square test:** compares proportions between two or more groups **Relative risks:** odds ratios or risk ratios **Logistic regression:** multivariate technique used when outcome is binary; gives multivariate-adjusted odds ratios	**McNemar's chi-square test:** compares binary outcome between correlated groups (e.g., before and after) **Conditional logistic regression:** multivariate regression technique for a binary outcome when groups are correlated (e.g., matched data) **GEE modeling:** multivariate regression technique for a binary outcome when groups are correlated (e.g., repeated measures)	**Fisher's exact test:** compares proportions between independent groups when there are sparse data (some cells <5). **McNemar's exact test:** compares proportions between correlated groups when there are sparse data (some cells <5).