What is “multivariate”?

- Multivariate data analysis is a set of statistical models that examine patterns in multidimensional data by considering, at once, several data variables. It is an expansion of bivariate data analysis, which considers only two variables in its models. As multivariate models consider more variables, they can examine more complex phenomena and find data patterns that more accurately represent the real world.

- Consider as an example the regression model — a method to analyze correlations in data. The non-multivariate case of regression is the analysis between two variables, and it is called a bivariate regression. It could be used, for instance, to see how the h*eight* of a swimmer correlates to its s*peed***.** By doing a bivariate regression, the analyst could find that taller swimmers tend to swim faster. Although it is right, we know that the *height* is not the only thing influencing *speed,* so the bivariate model hardly explains the complete phenomena of swimming.

Multivariate analysis is typically used for:

- Quality control and quality assurance
- Process optimisation and process control
- Research and development
- Consumer and market research

The Objective of multivariate analysis

- (1) **Data reduction or structural simplification**: This helps data to get simplified as possible without sacrificing valuable information. This will make interpretation easier.
- (2) **Sorting and grouping**: When we have multiple variables, Groups of “similar” objects or variables are created, based upon measured characteristics.
- (3) **Investigation of dependence among variables**: The nature of the relationships among variables is of interest. Are all the variables mutually independent or are one or more variables dependent on the others?

- (4) **Prediction Relationships between variables**: must be determined for the purpose of predicting the values of one or more variables based on observations on the other variables.
- (5**)** **Hypothesis construction and testing**. Specific statistical hypotheses, formulated in terms of the parameters of multivariate populations, are tested. This may be done to validate assumptions or to reinforce prior convictions.

- Sometimes the output in a data set is a vector of variables rather than a single variable. Where the vector of variables consists of the same quantity, blood pressure say, at several different times, recorded on the same individual. The output variables need not be the same quantity; they could be something like heights and weights for a set of children. Given this sort of data, we might be able to analyse it using a multivariate linear model, which is

$$\underset{(c\times 1)}{\boldsymbol{y}_j} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_{1j} + \boldsymbol{\beta}_2 x_{2j} + \cdots + \boldsymbol{\beta}_n x_{nj} + \boldsymbol{\epsilon}_j \quad j = 1, 2, \ldots, m$$

Sometimes the output in a data set is a vector of variables rather than a single variable. We will see a special case of this in Section 3.3.2, where the vector of variables consists of the same quantity, blood pressure say, at several different times, recorded on the same individual. The output variables need not be the same quantity; they could be something like heights and weights for a set of children.

Given this sort of data, we might be able to analyse it using a multivariate linear model, which is

$$\underset{(c\times 1)}{\boldsymbol{y}_j} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_{1j} + \boldsymbol{\beta}_2 x_{2j} + \cdots + \boldsymbol{\beta}_n x_{nj} + \boldsymbol{\epsilon}_j \quad j = 1, 2, \ldots, m \tag{3.26}$$

where the $\boldsymbol{\varepsilon}_j$'s are independently and identically distributed as[4] $\mathcal{N}^c(\mathbf{0}, \Sigma)$ and m is the number of data points. The '$(c \times 1)$' under '$\boldsymbol{y}_j$' indicates the dimensions of the vector, in this case c rows and 1 column; the $\boldsymbol{\beta}$'s are also $(c \times 1)$ vectors.

This model can be fitted in exactly the same way as a linear model (by least squares estimation). One way to do this fitting would be to fit a linear model to each of the c dimensions of the output, one-at-a-time.

Having fitted the model, we can obtain fitted values

$$\hat{\boldsymbol{y}}_j = \hat{\boldsymbol{\beta}}_0 + \sum_{i=1}^{n} \hat{\boldsymbol{\beta}}_i x_{ij} \quad j = 1, 2, \ldots, m$$

and hence residuals

$$\boldsymbol{y}_j - \hat{\boldsymbol{y}}_j \quad j = 1, 2, \ldots, m.$$

The analogue of the residual sum of squares from the (univariate) linear model is the matrix of residual sums of squares and products for the multivariate linear

Two types of variables

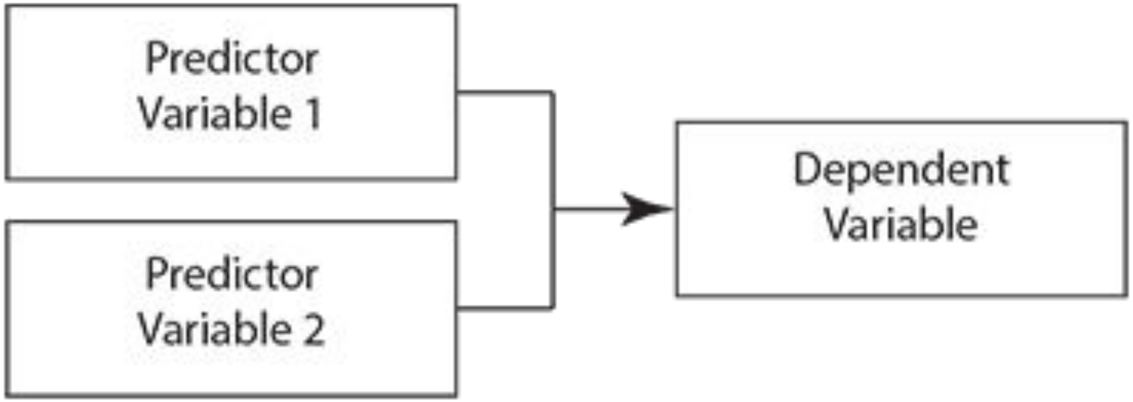
- **In multivariate analysis, the first thing to decide is the role of the variables.**

There are two possibilities:

- The variable **causes** an effect: *predictor variable*
- The variable is **affected**: *dependent variable*

This is a function of your model, not of the variables themselves, and the same variable may be either in different studies.

- The relationships between variables are usually represented by a picture with arrows:



- The variables can be observe directly, or infer them from what is happening. These are known as **latent variables**.

Example: Success at School

- It is hard to measure ‘*success at school*’: it is a **latent variable**.
- You might decide that '*success at school*' consists of academic success, together with some measure of social success (perhaps average duration of friendships, or size of ‘friendship group’) plus one of effort put in (which you could measure as perceptions of either students or teachers). These are your observed variables.
- The *measurement model* examines the relationship between the observed and latent variables.

Causal Models

- The strength or weakness of any causal model is the selection of the variables. If you miss out a major causal factor, then your conclusions will be either limited or incorrect. It is therefore worth taking time on defining your model as carefully as possible.
- There is a balance to be struck between simplicity and including more variables to obtain a better fit. Obviously you do not want to miss out a major causal variable, and including more variables will always give a better fit. But you need to consider whether the additional complexity is worth it for the gain in quality of the model.