Multivariate statistics

Wan Nor Arifin

Unit of Biostatistics and Research Methodology, Universiti Sains Malaysia.

email: wnarifin@usm.my

December 1, 2018

1. Multivariate
2. Screening of data for accuracy
3. Normality, linearity and homoscedasticity
4. Data transformation

Multivariate

Introduction

Multivariate?

Strictly speaking:

- variate = outcome/dependent variable (DV)
- *uni*variate = one DV
- *bi*variate = two DVs
- *multi*variate = > two DVs

$\rightarrow$ regardless of the number of independent variables (IVs)/predictors

In general, analysis involving > 2 variables = multivariate analysis.

Introduction

Why bother?

- most studies and research involve many variables.
- consider many predictors and many outcomes at the same time.
- computer!
  - availability of software
  - processing power

Screening of data for accuracy

Methods

- proofreading – compare data collection form with dataset.
- exploratory data analysis:
  - descriptive statistics.
  - graphical exploration.

Descriptive statistics

- numerical variables
  - mean, median
  - SD, IQR, MAD
  - minimum, maximum
- categorical variables
  - n, %

Practical

multivariate.R

Graphical exploration

- numerical variables
  - histogram, box-and-whisker plot, Q-Q plot.
  - more details in normality.
- categorical variables
  - bar charts, pie charts etc.
  - descriptive statistics are more informative.

Practical

multivariate.R

Normality, linearity and homoscedasticity

Normality, linearity and homoscedasticity

- All are concerned with numerical variables

Normality

Normal distribution of data of DV and IV.

- graphical
  - Univariate: histogram, box-and-whisker plot
  - Bivariate: scatter plot
  - Multivariate:
    - ★ Q-Q plot (multivariate) – Mahalanobis distance[1] vs expected normal distribution values.
    - ★ χ^2 vs Mahalanobis distance plot (Arifin, 2015).

[1] The distance of a case from the centroid, where centroid is the intersection of the means of all variables (Tabachnick & Fidell, 2007).

Practical

multivariate.R

Normality

- statistical
  - skewness – symmetry
    - ★ $<$ 2-3 times of its SE:

$$SE = \sqrt{\frac{6}{N}}$$

  - kurtosis – peakness/flatness
    - ★ $<$ 2-3 times of its SE:

$$SE = \sqrt{\frac{24}{N}}$$

  - statistical tests – Shapiro-Wilk test.
  - Multivariate – Mardia's skewness and kurtosis.

Practical

multivariate.R

Linearity

Linear relationship between two variables.

- graphical
  - Bivariate: scatter plot.
- statistical
  - Linear regression, correlations.

Practical

multivariate.R

Homoscedasticity

Equality/homogeneity of variances:

- across groups (categorical IV).
- for each levels of IV (numerical IV).
- graphical
  - Univariate per group: Compare histograms and box-and-whisker plots.
  - Bivariate: scatter plot.
- statistical
  - Tests of equality of variance.

Practical

multivariate.R

Data transformation

Data transformation

- whenever numerical data are not normally distributed.
- to turn these data into normally distributed data.

Common data transformation

- square root – $\sqrt{X}$
- natural log – lnX
- log 10 – $log_{10}X$
- reciprocal – $\frac{1}{X}$
- power of k – X^k, e.g. X^2, X^3

Suitable data transformation

Depending on the tail of the skewness, we may try suitable transformations[2]:

Table 1: Skewness tail and suitable transformations.

Tail	Transformation (R format)	Purpose
Right	sqrt(x), log(x), log10(x), 1/x	Make larger values smaller
Left	x^k	Make smaller values larger

[2] More details can be referred to Kutner, Nachtsheim, Neter, & Li (2005), Hair, Black, Babin, & Anderson (2010) and Tabachnick & Fidell (2007), i.e. transformation of Y/X/both to handle normality, heteroscedasticity and normality.

Practical

multivariate.R

Other issues[3] for self-study:

- Missing data
- Multivariate outliers
- Multicollinearity and singularity

[3] Not covered in your syllabus.

References

Arifin, W. N. (2015). The graphical assessment of multivariate normality using SPSS. *Education in Medicine Journal*, *7*(2), e71–e75.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis*. New Jersey: Prentice Hall.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical model (5th ed.)*. Singapore: McGraw-Hill Education (Asia).

Tabachnick, B., & Fidell, L. (2007). *Using multivariate statistics (5th ed.)*. USA: Pearson.