Deirdre Lawless, School of Computer Science, TU Dublin City Campus

NUMBERS TO REMEMBER

Missing Data

- If missing data represent less than 5% of the total and is missing in a random pattern from a large data set, almost any procedure for handling missing values yields similar results.

Outliers

- If the sample size is small (80 or fewer cases), a case is an outlier if its standard score is ±2.5 or beyond.
- If the sample size is larger than 80 cases, a case is an outlier if its standard score is ±3.29 or beyond

Assessing Normality

- Skewness and Kurtosis
- Standardised scores for skewness and kurtosis between -2 and +2 are considered acceptable in order to prove normal univariate distribution.
- Standardised scores
  - Check impact of outliers
  - At 0.05 level if 95% of your data is within outlier bounds when converted to standardised scores – it is likely your data is safe to treat as normal.

Effect Sizes

- Cohen's d 0.1 Weak, 03. Moderate, 0.5 and above Strong
- Eta 0.01 Small, 0.06 Moderate, 0.14 Large

Dummy/Indicator Variables

- 0 Reference, 1 Category of interest
- To transform a nominal variable with n levels requires n-1 dummy/indicator variables.

Collinearity

- Correlation coefficient above 0.8
- VIF above 2.5
- Tolerance lower than 0.4

Cooks Distance

- Investigate cases with values above 1

Factor Analysis/PCA

- Correlations above 0.3
  - Concern if 0.8 or above
- Determinant >0.00001
- KMO above 0.6
- Bartlett Significant
- Cronbach's alpha above 0.6

Strong Factors/Components

- Extract factors/components with Eigenvalues above 1
- Need loading of 3 or more at level 0.3 or above for strong factors/components
- 0 or few cross loadings
- Communalities above 0.8 v strong - likely to be between 0.4-0.7