Society of Actuaries in Ireland

Multivariate Analysis & PCA

ISL Presentation

John Nolan, FSAI

Agenda

- Multivariate Analysis:
  - Intro
  - Simpson’s Paradox
  - Techniques Used - PCA
  - Simple Example

Multivariate Analysis - Intro

- Three types of analysis:

1. **Univariate analysis**
   - The examination of the distribution of cases on only one variable at a time (e.g. college graduation)

2. **Bivariate analysis**
   - The examination of two variables simultaneously (e.g. the relationship between gender and college graduation)

3. **Multivariate analysis**
   - The examination of more than two variables simultaneously (e.g., the relationship between gender, race, and college graduation)
   - Multivariate Analysis allow the separate and combined effects of the independent variable to be examined

Multivariate Analysis – Simpson’s Paradox

- Using Simpson’s Paradox to show why Multivariate analysis is necessary
- Simpson's paradox occurs when groups of data show one particular trend, but this trend is reversed when the groups are combined together.
- Example:
  - ❑ 44% of male applicants are admitted by a university, but only 33% of female applicants
  - ❑ Men more likely to get admitted? Difference too large to be down to chance
  - ❑ Does this mean there is unfair discrimination?

	Male	Female
Accepted	35	20
Refused Entry	45	40
Total	80	60
% Accepted	**44%**	**33%**

- ❑ University decided to investigate by further breaking down by degree

Multivariate Analysis – Simpsons Paradox

- Results by degree:

Engineering	Male	Female
Accepted	30	10
Refused Entry	30	10
Total	60	20
% Accepted	**50%**	**50%**

English	Male	Female
Accepted	5	10
Refused Entry	15	30
Total	20	40
% Accepted	**25%**	**25%**

- → No relationship between sex and acceptance for either programme, i.e. no discrimination

- Why?
  - More females apply for English programme, but it is hard to get in to (25% success)
  - More males apply for engineering, but it is easier to get in to (50% success)
  - Degree is the confounding variable
  - Demonstrates why we shouldn’t just scratch the surface.

Multivariate Analysis – What's its all about?

- Definition:
  - ❑ "The simultaneous analysis of several variables"

- MVA uses ALL available data to capture the most information possible. Never a simple uni/bivariate analysis.The basic principle is **to boil down hundreds of variables to a mere handful**.

- Making sense of large mases of data -> Data-rich but knowledge poor.

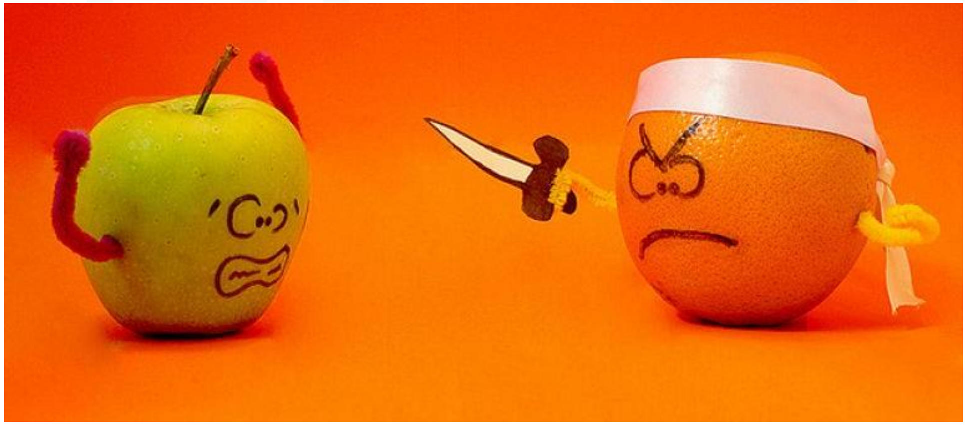
- Multivariate analysis can help summarise the data and avoid spurious results as seen in previous examples.

- MVA is based on "**Ockham's Razor**":
  - ❑ "Everything should be kept as simple as possible, but no simpler."

- Simpson's paradox shows how we need to consider more variables, Ockham's Razor tells us to consider less… Need to find a balance.

Multivariate Analysis – Example

- Apples Versus Oranges



- Could come up with 100’s of different factors to compare them:
  - ❑ Colour, shape, texture, firmness, …
  - ❑ Skin: smoothness, thickness,…
  - ❑ Juice: PH, taste, composition
  - ❑ Seeds, etc.
- Ultimately, there will never be more than one difference: is it an apple or an orange?

Multivariate Analysis – Techniques

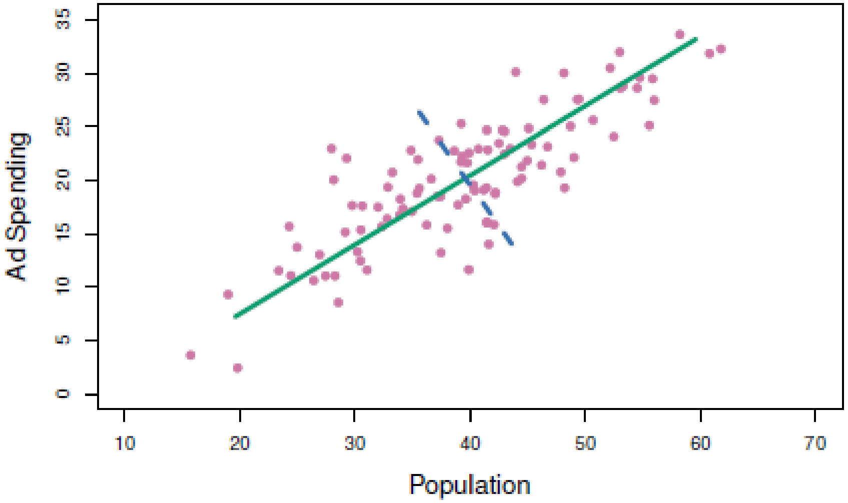
- Many different techniques used to perform a multivariate analysis:
    - Principal Component Analysis (PCA)
    - Singular Value Decomposition (SVD)
    - Multiple regression
    - Logistic regression
    - Discriminant Analysis
    - Multivariate Analysis of Variance (MANOVA)

- Most of these are pretty complex, with heavy maths behind them

Multivariate Analysis – Principal Component Analysis (PCA)

- Used to identify the underlying dimensions or “Principle Components” for **sources of variation**.

- An unsupervised learning algorithm, it finds patterns by itself. In particular, PCA finds (mutually orthogonal) directions of greatest variance.



- The green solid line indicates the first principal component direction, and the blue dashed line indicates the second principal component direction.
- Essentially finding new variables that are linear functions of those in the original dataset, that successively maximize variance and that are uncorrelated with each other

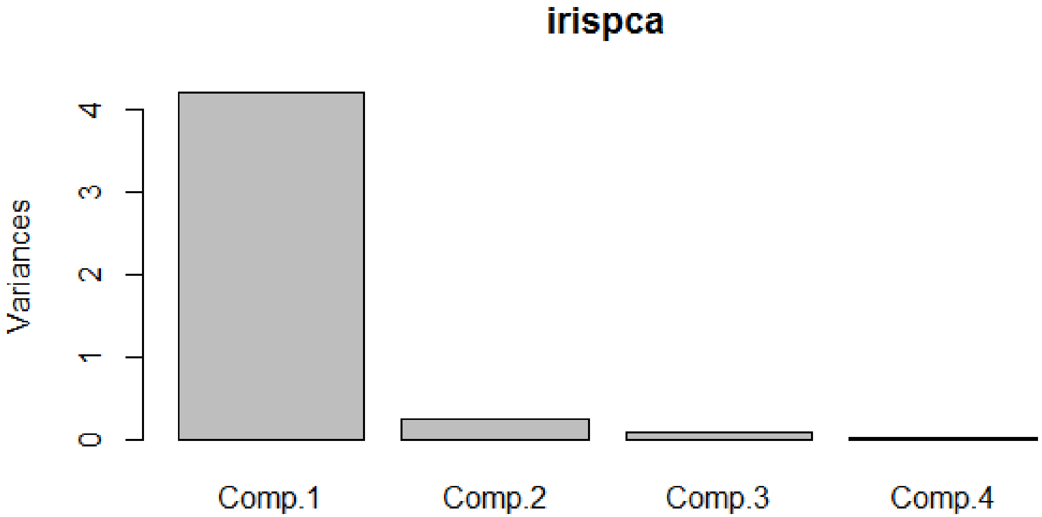
Multivariate Analysis – Principal Component Analysis (PCA)

- Form of data compression without much loss of information
- First principle component accounts for as much of the **variability** as possible, and each succeeding component accounts for as much of the remaining variability as possible.
- …. Reduces effect of **multi-collinearity**.
  - Refers to predictors that are correlated with other predictors.
  - Occurs when model includes multiple factors that are correlated to both response and other variables
  - Results when you have factors that are a bit redundant.

Multivariate Analysis – Principal Component Analysis (PCA)

- Can be performed in R using prcomp(dataset) or princomp(dataset)

- Create a scree plot:
  - *“A scree plot displays the proportion of the total variation in a dataset that is explained (**PVE = Proportion of Variance Explained**) by each of the components in a principle component analysis. It helps you to identify how many of the components are needed to summarise the data.”*



- See R code for example “10.4 Lab 1: Principal Components Analysis”