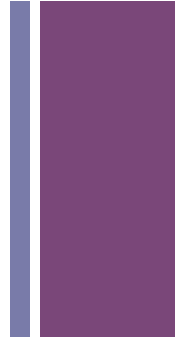
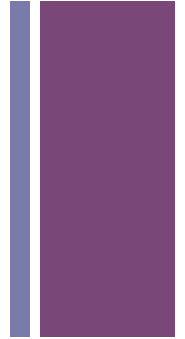


# + Multivariate analysis



- Concerned with datasets that have more than one response variables for each observational unit
- N rows (cases) and P columns (variables)
  - Relationships among cases
  - Relationships among variables
- First, visualize
  - Pairs plot – plot scatter plot matrix
  - pairs plots can easily miss interesting structure
  - multivariate methods explore the data in a less coordinate-dependent way

# + The number of datasets to analyze: one or two



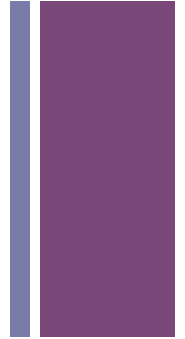
## ■ One dataset

- Find relationships among variables or cases
  - Principal component analysis (PCA) – continuous variables
  - Correspondence analysis (CA) – categorical variables
  - Multidimensional scaling (MDS) and cluster analysis - proximity

## ■ Two datasets

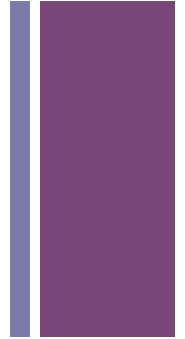
- Find relationships between independent variable set and dependent variable set, or between two dependent variable sets
  - Multiple linear regression – not really multivariate, just one y
  - Discriminant analysis (DA) and logistic regression – categorical DV
  - Multivariate analysis of variance (MANOVA) – math same as DA
  - Confirmatory factor analysis – create models (SEM)
  - Exploratory factor analysis – no constraints on models
  - Canonical correlation analysis (CC) – extension of multiple linear regression
  - Multiple factor analysis – perform PCA on each data table

# + Principal component analysis (PCA)



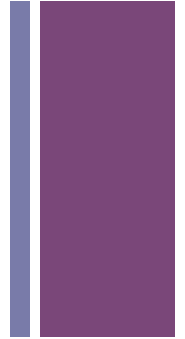
- Interval or ratio level of measurement
  - Nominal scale: No ordering of values (e.g., Male or female)
  - Ordinal scale: Can infer ordering of values (e.g., low, medium, and high self-esteem)
  - Interval scale: Can infer ordering of values, the values are evenly spaced (e.g., Celcius scale, intelligence tests)
  - Ratio scale: Same as interval scale but zero is special, can make ratio (e.g., Weight, Kelvin scale)
- The goal of PCA is to decompose a data table with correlated measurements into a new set of uncorrelated (i.e., orthogonal) variables.
  - Step 1: Subtract the mean
  - Step 2: Calculate the covariance matrix (or correlation)
  - Step 3: Calculate the eigenvectors and eigenvalues
  - Step 4: Choose components and form a feature vector
  - Step 5: Derive the new data
- The importance of each component is expressed by the variance (i.e., eigenvalue) of its projections or by the proportion of the variance explained.
- PCA is useful when you want to develop an index to summarizes complex data - rank students by examination scores, rank cities by cost of living
- PCA is useful for visualization when there are too many explanatory variables relative to the number of observations and when the explanatory variables are highly correlated.

# + Correspondence analysis (CA)



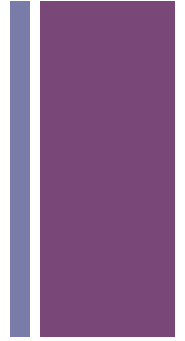
- Nominal or ordinal level of measurement
- Used to display the associations among a set of categorical variables
- Is a generalization of PCA to contingency tables
- The factors of CA give an orthogonal decomposition of the Chi-square associated to the table.

# + Multidimensional scaling (MDS), additive tree, cluster analysis



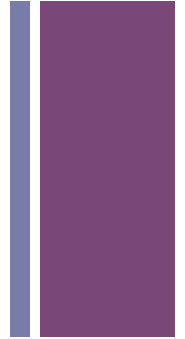
- Similarity or distance
- Applied when the rows and the columns of the data table represent the same units and when the measure is a distance or a similarity.
- The goal of the analysis is to represent an observed proximity matrix geometrically.
- MDS is used to represent the units as points on a map such that their Euclidean distances on the map approximate the original similarities
  - Classic MDS, which is equivalent to PCA, is used for distances
  - Non-metric MDS for similarities
- Additive tree analysis and cluster analysis are used to represent the units as “leaves” of a tree with the distance “on the tree” approximating the original distance or similarity.
  - Hierarchical clustering: a series of partitioning steps from a single cluster containing everyone to n clusters each containing a single individual.
  - K-means clustering: find initial partitioning and start moving individuals.

# + Multiple linear regression analysis



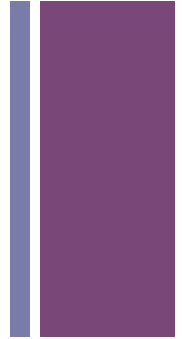
- Several IVs are used to predict one DV.
  - $y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 \dots + e$
  - Generalization of simple linear regression
  - When the IVs are orthogonal, the problem reduces to a set of univariate regressions.
  - When the IVs are correlated, their importance is estimated from the partial coefficient of correlation.

# + Discriminant analysis (DA)



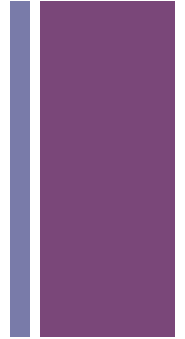
- Predicting a nominal variable
  - Used to classify a case into one of two or more populations.
  - You need to know which population the individual belongs to for the initial sample.
  - You classify future individuals whose membership is unknown (prediction).
  - You identify which variables contribute to making the classification (description).
- Mathematically equivalent to MANOVA
- Used when a set of IVs are used to predict the group to which a given unit belongs (a nominal DV).
  - Deciding whether to approve a loan - age, income, marital status, outstanding debt, home ownership, etc..
  - Deciding whether an individual is more or less likely to be depressed - age, income, education, etc..

# + Logistic regression



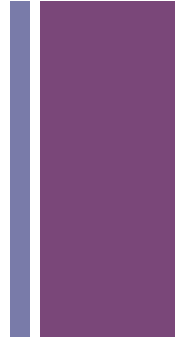
- Binomial: for binary or dichotomous response variable, 0 and 1, such as presence and absence, success and failure, etc..
- Multinomial: for more than two categories.
- As in linear regression, we are looking for a relationship between our response variable and a set of independent variables, which are often called covariates.
- We can transform the output of a linear regression to be suitable for probabilities
  - $\text{logit}(p) = \log(o) = \log(p/(1-p)) = b_0 + b_1x_1 + b_2x_2 + \dots$
  - The odds:  $p = .8$ , then the odds =  $.8 / .2 = 4$  to 1  $\rightarrow$  4 times as likely
  - $\text{odds} = p / (1 - p)$ ,  $p = \text{odds} / (1 + \text{odds})$
  - The inverse of the logit function is the logistic function.
    - if  $\text{logit}(p) = z$ , then  $p = \exp(z) / (1 + \exp(z))$
- Some examples
  - Depression level (0 or 1)
  - People's occupational choices
  - Brand choices based on gender and age.

# + Multivariate analysis of variance



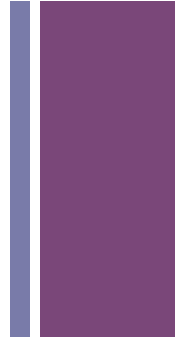
- IVs have the same structure as in a standard ANOVA
- Used to predict a set of DVs.
- MANOVA computes a series of ordered orthogonal linear combinations of the DVs (i.e., factors) with the constraint that the first factor generates the largest F if used in an ANOVA.

# + Factor analysis



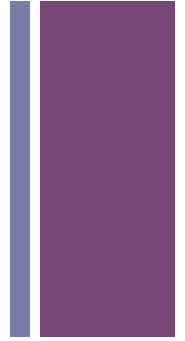
- Examines the interrelationships among a large number of variables to determine underlying dimensions (factors)
  - Assumption: there are some underlying common factors (e.g., intelligence, social class, social/soft drug)
  - These factors cannot be directly observed – latent variables – concepts that cannot be measured directly
  - Measurable variables – manifest variables – are expected to be related to the latent variables
  - Looks at the relationships between assumed latent variables and the manifest variables.
- Confirmatory
  - Fitting a model
  - Generates one or a few models of an underlying explanatory structure, which is often expressed as a graph (how IVs related to DVs).
  - Then the correlations among the DV's are fit to this structure.
  - Models are evaluated by comparing how well they fit the data.
  - Structural equation modeling (SEM)
- Exploratory
  - No constraints on which of the manifest variables load on the common factors

# + Canonical correlation analysis



- Extension of multiple regression – more than one  $y$ 
  - $x = [x_1, x_2, \dots, x_{q_1}]$ ,  $y = [y_1, y_2, \dots, y_{q_2}]$
  - $R_{11}$  = cor matrix of variables in  $x$
  - $R_{22}$  = cor matrix of variables in  $y$
  - $R_{12}$  = cor between  $x$  and  $y$  ( $q_1$  by  $q_2$  matrix)
  - $E_1 = R_{11}^{-1} R_{12} R_{22}^{-1} R_{21}$
- CC combines the DVs to find pairs of new variables (called canonical variables, one for each data table) which have the highest correlation.
- CV's, even when highly correlated, do not necessarily explain a large portion of the variance of the original tables.
  - This make the interpretation of the CV sometimes difficult, but CC is nonetheless an important theoretical tool

# + Multiple factor analysis



- MFA combines several data tables into one single analysis.
- Perform a PCA of each table.
- Each data table is normalized by dividing all the entries of the table by the first eigenvalue of its PCA.
  - Akin to the univariate Z-score
  - Equalizes the weight of each table