

Canonical Correlation Analysis

Univariate Correlation

- If x & y are two variables upon which n observations are obtained, then the Pearson Product Moment Correlation Coefficient is given by:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

Pearson Product Moment Correlation Coefficient

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

where s_{xy} = covariance between x & y

s_x^2 = variance of x

s_y^2 = variance of y

Multiple Correlation Coefficient

- If \mathbf{x} is a vector of q variables, and y is another variable, then the multiple correlation coefficient between \mathbf{x} & y is:

$$R_{xy} = \frac{\mathbf{b}'\mathbf{s}_{xy}}{\sqrt{\mathbf{b}'\mathbf{S}_{xx}\mathbf{b}s_y^2}}$$

Multiple Correlation Coefficient

$$R_{xy} = \frac{\mathbf{b}'\mathbf{s}_{xy}}{\sqrt{\mathbf{b}'\mathbf{S}_{xx}\mathbf{b}s_y^2}}$$

where $\mathbf{s}_{xy} = p \times 1$ covariance matrix between \mathbf{x} & y

\mathbf{S}_{xx} = variance-covariance matrix of \mathbf{x}

s_y^2 = variance of y

$$\mathbf{b} = \mathbf{S}_x^{-1} \mathbf{s}_{xy}$$

Multiple Correlation Coefficient

- $\mathbf{b}'\mathbf{x}$ represents that linear combination of the q variables in \mathbf{x} which is maximally correlated with y . The square of R_{xy} is the R^2 statistic of multiple regression.

$$R_{xy}^2 = \mathbf{s}'_{xy} \mathbf{S}_{xx}^{-1} \mathbf{s}_{xy} (s_y^2)^{-1}$$

Correlation between 2 random vectors

- Consider the case where we have n observations on vectors $\mathbf{x}_{q \times 1}$ and $\mathbf{y}_{p \times 1}$.
- We still wish to assess the relationship between the set of q variables in \mathbf{x} and the set of p variables in \mathbf{y} .

$$\text{Let } u = \mathbf{b}'\mathbf{x}$$

$$v = \mathbf{c}'\mathbf{y}$$

Correlation between 2 random vectors

- Then the correlation between u & v is:

$$r_{uv} = \frac{\mathbf{b}'\mathbf{S}_{xy}\mathbf{c}}{\sqrt{(\mathbf{b}'\mathbf{S}_{xx}\mathbf{b})(\mathbf{c}'\mathbf{S}_{yy}\mathbf{c})}}$$

where \mathbf{S}_{xy} = covariance matrix between \mathbf{x} & \mathbf{y}

\mathbf{S}_{xx} = variance-covariance matrix of \mathbf{x}

\mathbf{S}_{yy} = variance-covariance matrix of \mathbf{y}

Canonical Correlation

- A logical extension of the multiple correlation case would be to determine the vectors \mathbf{b} & \mathbf{c} such that r_{uv}^2 is maximized.

Canonical Correlation

The squared first canonical correlation is defined by:

$$\hat{\theta}_1 = \max_{\substack{\mathbf{b} \neq \mathbf{0} \\ \mathbf{c} \neq \mathbf{0}}} \left[r_{(\mathbf{b}'\mathbf{x}, \mathbf{c}'\mathbf{y})}^2 \right]$$

with

$$\mathbf{b}'\mathbf{S}_{xx}\mathbf{b} = \mathbf{c}'\mathbf{S}_{yy}\mathbf{c} = 1$$

Canonical Correlation

- In general the squared canonical correlations are the solutions to the determinantal equation:

$$\left| \mathbf{S}_{yx} \mathbf{S}_{xx}^{-1} \mathbf{S}_{xy} \mathbf{S}_{yy}^{-1} - \hat{\theta} \mathbf{I} \right| = 0$$

Relationship with MANOVA

- The multivariate general linear model is:

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E}$$

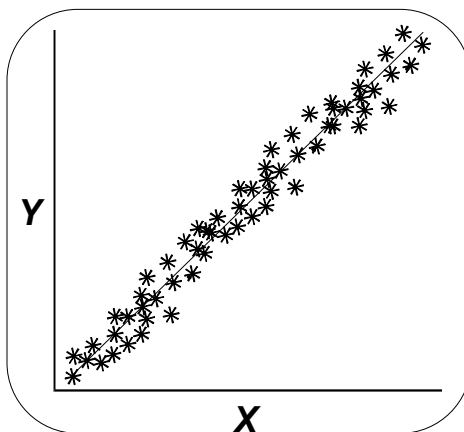
- Whether this model represents a multivariate regression model, or a MANOVA, we seek to assess the relationship between \mathbf{X} (ignoring the first column corresponding to the intercept) and \mathbf{Y} .

Relationship with MANOVA

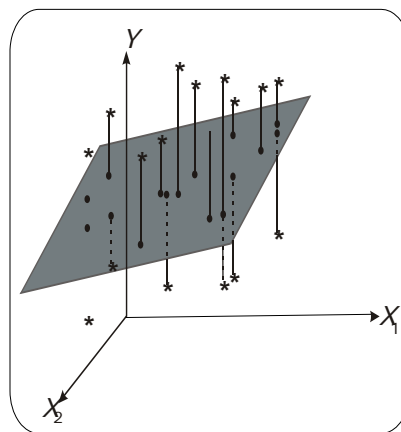
- The multivariate test statistics are functions of the eigenvalues of \mathbf{HE}^{-1} or $\mathbf{H}(\mathbf{H}+\mathbf{E})^{-1} = \mathbf{HT}^{-1}$. These eigenvalues are the squares of the canonical correlations between \mathbf{X} and \mathbf{Y} .

Regression Analysis

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon_i$$



$$Y_i = \beta_0 + \beta_1 X + \varepsilon_i$$



Multivariate Multiple Regression

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}$$

where

\mathbf{Y} $n \times p$ matrix of p response variables for n subjects

\mathbf{X} $n \times (k + 1)$ matrix of k independent variables (plus the intercept) for n subjects

$\boldsymbol{\beta}$ $(k + 1) \times p$ matrix of coefficients for the k independent variables used to predict the p dependent variables.

Another General Linear Model

- Like MANOVA, multivariate regression is a type of general linear model.
 - The hypotheses are similar.
 - The estimates are obtained in a similar way.
 - The assumptions are similar.
 - The multivariate test statistics are the same.

Multivariate Regression Assumptions

- Multivariate normality of the responses
- Common covariance structure across observations
- Independent observations
- Randomly sampled observations from the population of interest
- Linear association between the predictors and the responses

The REG Procedure for Multivariate Regression

- General form of the REG procedure:

```
PROC REG <options>;  
  MODEL  $Y_s = <X_s>$  </options>;  
  MTEST <equation,...,equation> </ options>;  
RUN;
```

Canonical Correlation Using PROC CANCORR

- In addition to the canonical correlations, you see the following:
 - raw and standardized canonical coefficients for each canonical variate
 - simple correlations for all variable and variate combinations
 - step-down likelihood-ratio tests for canonical correlations
 - multivariate tests for H_0 : all canonical correlations = 0
 - adjusted canonical correlations and standard errors of the canonical correlations.

Likelihood-ratio Tests of Canonical Variates

- The likelihood-ratio test of \mathbf{HE}^{-1} is equal to Wilks' Lambda and can be converted to an approximate F test.
 - This is a step-down method.
 - The null hypothesis is that the canonical correlation R_i and all those smaller than $R_i = 0$.
 - This test is performed sequentially for all canonical variate pairs.

Interpreting the Canonical Correlation

- Canonical coefficients are just like regression coefficients:
 - They depend on the other variables in the model.
 - They depend on the scale of measurement.
 - Standardized coefficients address the scaling issue, but they do not address the problem of dependencies among variables.
 - Coefficients are useful for prediction but not for interpretation.

Interpreting the Canonical Correlation

- A more useful way to interpret the canonical correlation in terms of the input variables is to look at the simple correlation statistics. For each pair of variates, look at
 - the correlation between each variable and its canonical variate
 - the correlation between each variable and the canonical variate for the other set of variables.

The CANCELL Procedure

- General form of the CANCELL procedure:

```
PROC CANCELL <options>;  
    <VAR v-variables>;  
    WITH w-variables;  
RUN;
```