STAT 419 Study Guide

This is mising the conversion from Wilks to F stats

Introduction

Matrix Algebra

Length of a vector:

$$\sqrt{\vec{x}'\vec{x}} \tag{1}$$

Linear Independence

$c_1a_1 + ... = 0$

If c's can be found then the values ... are linearly dependent.

If x1 and x2 are linearly dependent the correlation betwee n the two is 1.

Rank

The rank of a matrix is the numboer of linearly independent rows or columns.

Inverse of a Matrix

If a matrix A is square, 2x2, and of full rank than the inverse is found by:

$$A^{-1} = \frac{1}{ad - bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \tag{2}$$

If A has an inverse than A is nonsingular otherwise singular.

Determinant

$$|A| = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{21}a_{12} \tag{3}$$

Properties

If D is a diagonal matrix:

$|D| = \Pi$ of diagonals

If A is singular:

$$|A| = 0 \tag{4}$$

If A and B are square and of the same size:

|AB| = |A| |B|

(5)

Trace

tr(A) = sum along the diagonal

Orthogonal Vectors

Two n * 1 vectors a and b are orthogonal if:

$$a'b = 0 \tag{6}$$

Orthogonal vectors are perpendicular.

Normalized Vector

If $a'a = 1$ then a is normalized (length equals 1)

Can normalize any vector **a** dividing by each element by the length of **a**.

Eigenvalues and Eigenvectors

For every square matrix **A** and a scaler λ a nonzero vector x can be found such that :

$$Ax = \lambda x \tag{7}$$

Can be found by solving the *characteristic equation*:

$$|A - \lambda I| = 0 \tag{8}$$

In a n x n matrix it will produce n roots.

Eigenvalue and Eigenvector Properties

$$tr(A) = \sum_{i=1}^{n} \lambda_i \tag{9}$$

$$|A| = \prod_{i=1}^{n} \lambda_i \tag{10}$$

Characterizing and Displaying Mulltivariate Data

Population covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{yx} & \sigma_y^2 \end{bmatrix} \tag{11}$$

$$\sigma_i^2 = variances \tag{12}$$

Population correlation coeffcient:

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{13}$$

Sample correlation matrix:

$$r_{jk} = \frac{s_{jk}}{\sqrt{S_{jj} S_{kk}}} \tag{14}$$

Linear Combinations

$$z = \vec{a}' \vec{y} \tag{15}$$

The sample mean of the linear combinations is:

$$\bar{z} = \vec{\bar{a}}' \vec{\bar{y}} \tag{16}$$

The sample variances of the combinatins is:

$$s_z^2 = a' S a \tag{17}$$

Vector of Linear Combinatoions

A contains a series of weighting schemes.

The covariance matrix of the vector of linear combinations

$$cov(z) = A\Sigma A' \tag{18}$$

Measures of Overall Variablity

Measuring how scattered the observation vectors y ... are around $\bar{y}$

Generalized sample variance: |S|

$$|S| = \prod_{j=1}^{p} \lambda_j \tag{19}$$

Total Sample Variance: tr(S)

Large values imply scattering about $\bar{y}$ and low values imply close concentation However, if |S| is equal or near zero it could indicate multicollinearity

Properties of Linear Combinations

The mean of the linear combinations $E(z)$:

$$E(z) = E(A\vec{y}) = A\mu \tag{20}$$

The covariance marix of the vector of linear combinations $cov(z)$:

$$cov(z) = A\Sigma A' \tag{21}$$

Mahalanobis Distance

$$d^2 = (y_1 - y_2)' S^{-1} (y_1 - y_2) \tag{22}$$

The Multivariate Normal Distribution

Normal quantile-quantile (qq) Plots

Line shows theoretical quantiles of the normal distribution. If off the line then data is non-normal.

Bivariate scatterplots

If any curves then the data is non-normal

Chi-Square Probablity Plot

Line is the chi-squared quantile plot. If off the line then data is non-normal.

Tests on One or Mean Vectors (Σ Unknown)

Hotelling's T^2 Test

One Sample

What are you testing:

Testing to see if a mean vector jointly equals a certain known value

Null/Alternative Hypothesis:

$$H_0 : \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} 5.1 \\ 5.2 \end{bmatrix} \tag{23}$$

$$H_A : \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \neq \begin{bmatrix} 5.1 \\ 5.2 \end{bmatrix} \tag{24}$$

Test Statistic:

$$T^2 = n(\vec{\bar{y}} - \mu)' S^{-1} (\vec{\bar{y}} - \mu) \tag{25}$$

Conclusions:

Can conclude that the true mean var1 and var2 are not jointly equal to 5.1 and 5.2.

Assumptions:

We assume that the multivariate random sample y1 ... is drawn from normal distribution.

Two Sample

What are you testing:

Testing to see if the mean variables are identical in two groups.

Null/Alternative Hypothesis:

$$H_0 : \mu_1 = \mu_2 \tag{26}$$

$$H_A : \mu_1 \neq \mu_2 \tag{27}$$

Test Statistic:

$$T^2 = \frac{n_1 n_2}{n1 + n2}(\vec{y_1} - \vec{y_2})' S_p^{-1} (\vec{y_1} - \vec{y_2}) \tag{28}$$

$$S_p = \frac{1}{n1 + n2 - 2}[(n_1 - 1)S_1 + (n_2 - 1)S_2] \tag{29}$$

Conclusions:

Can conclude that the mean variables for group 1 and 2 are not jointly equal.

Assumptions:

Assume that the two samples are multivariate normal and that the population covariances are equal.

After Rejection:

Can perform univariate two-sample t-tests on each variable to see which variables have different means for the two groups

Paired Two Sample

What are you testing:

Testing to see if two dependent groups have jointly equal characteristics.

Null/Alternative Hypothesis:

$$H_0 : \mu_d = 0 \tag{30}$$

$$H_A : \mu_d \neq 0 \tag{31}$$

Where an example of μ_d is:

$$\mu_d = \mu_y - \mu_x = \begin{bmatrix} \mu_{y1} - \mu_{x1} \\ \mu_{y2} - \mu_{x2} \end{bmatrix} \tag{32}$$

Test Statistic:

$$T^2 = n\vec{d}' S_d \vec{d} \tag{33}$$

$$\bar{d} = \frac{1}{n} \sum_{i=1}^{n} d_i \tag{34}$$

$$S_d = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})(d_i - \bar{d})' \tag{35}$$

Conclusions:

Can conclude that the true mean var1 and var2 are not jointly equal between the two groups.

Assumptions:

We assume that the differences are multivariate normal.

Profile Analysis

One Sample

What are you testing:

Testing to see if the population means are flat, all equal.

Null/Alternative Hypothesis:

$$H_0 : C\mu = 0 \tag{36}$$

$$H_A : C\mu \neq 0 \tag{37}$$

Where C is (p-1) x p contrast matrix

Test Statistic:

$$T^2 = n(C\vec{\bar{y}})'(CSC')^{-1}(C\vec{\bar{y}})' \tag{38}$$

Conclusions:

Can conclude that the true mean of var at the different measures are not all equal.

Two Sample

Test of Parallelism

What are you testing:

Testing to see if two groups have equal mean value for different variables.

Null/Alternative Hypothesis:

$$H_0 : C\mu_1 = C\mu_2 \tag{39}$$

$$H_A : C\mu_1 \neq C\mu_2 \tag{40}$$

Conclusions:

Can conclude that the profiles are not parallel. Interaction between group and type of test. Association between type of test and score performace depends on group.

After Fail to Rejection:

Perform test of coincidence.

Test of Coincidence

What are you testing:

Testing to see if the two profiles are equal?

Null/Alternative Hypothesis:

$$H_0 : \bar{j}'\mu_1 = \bar{j}'\mu_2 \tag{41}$$

$$H_A : \bar{j}'\mu_1 \neq \bar{j}'\mu_2 \tag{42}$$

Where j = (p x 1) of all 1's

Conclusions:

Can conclude that the profiles are not identical. There is a significant association between group and score performance.

Test of Flatness

What are you testing:

Testing to see if the profiles are both flat.

Null/Alternative Hypothesis:

$$H_0 : \frac{1}{2}C(\mu_{11} + \mu_{21}) = 0 \tag{43}$$

$$H_A : \frac{1}{2}C(\mu_{11} + \mu_{21}) \neq 0 \tag{44}$$

Conclusions:

Can conclude that there is an association between type of test and score performance.

Multivariate Analysis of Variance

Manova

One Way

What are you testing:

Compare the mean vectors from k multivariate normal populations. Determine if there is a treatment effect on several variables.

Null/Alternative Hypothesis:

$$H_0 : \mu_1 = \mu_2 = \mu_3 \tag{45}$$

H_A : at least two μ_i's differ

Test Statistic:

k = number of groups p = number of variables dim(H) = dim(E) = pxp rank(H) = min(p, (k - 1))

If observations are balanced:

$$s = rank(E) = min(p, k(n-1)) \tag{46}$$

$$\nu_E = k(n-1) \tag{47}$$

If observations are unbalanced:

$$rank(E) = min(p, \sum_{i=1}^{k} n_i - k) \tag{48}$$

$$\nu_H = k - 1 \tag{49}$$

$$\nu_E = \sum_{i=1}^{k} n_i - k \tag{50}$$

Wilks

$$\Lambda = \frac{|E|}{|E+H|} \tag{51}$$

$$\Lambda = \prod_{i=1}^{s} \frac{i}{1+\lambda_i} \tag{52}$$

$$0 \leq \Lambda \leq 1 \tag{53}$$

Reject H_0 if $\Lambda < \Lambda_{\alpha,p,\nu_H,\nu_E}$

Roy

$$\Theta = \frac{\lambda_1}{1+\lambda_1} \tag{54}$$

Pillai

$$V^(s) = \sum_{i=1}^{s} \frac{\lambda_i}{1+\lambda_i} \tag{55}$$

Lawley-Hotelling

$$U^{(s)} = \sum_{i=1}^{s} \lambda_i \tag{56}$$

test stat $= \frac{\nu_E}{\nu_H} U^{(s)}$

Follow up F-tests

Perform individual F-tests after rejecting manova null hypothesis.

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k \tag{57}$$

Each of these tests can be done at the 0.05 level.

Two Way

Wilks

What are you testing:

Investigate the effects of two factors on several dependent variables. Interaction: The effect of one factor depends on the level(s) of the other(s).

Null/Alternative Hypothesis:

Different for each of the tests

$$H_0 : \vec{\mu_1} = \vec{\mu_2} = \vec{\mu_a} \tag{58}$$

Test Statistic:

$$\Lambda_{AB} = \frac{|E|}{|E+H_AB|} \ \Lambda_{p,(a-1)(b-1),ab(a-1)} \tag{59}$$

$$\Lambda_A = \frac{|E|}{|E+H_A|} \ \Lambda_{p,a-1,ab(a-1)} \tag{60}$$

$$\Lambda_B = \frac{|E|}{|E + H_B|} \Lambda_{p,b-1,ab(a-1)} \tag{61}$$

Reject H_0 if the test statistic is less than the critical value.

Conclusions:

$$\Lambda_{AB} \tag{62}$$

Can conclude that there is an interaction effect between factor A and factor B. The effect of factor A on the result depends on factor B.

$$\Lambda_A \tag{63}$$

Can conclude that results depend on factor A.

$$\Lambda_B \tag{64}$$

Can conclude that results depend on factor B

k-sample Profile

Same as the two-sample profile tests with more variables in the null hypothesis.

Tests on Covariance Matrices

Testing Specific Structure for Σ

What are you testing:

Testing to see if covariance matrix equals a specific value.

Null/Alternative Hypothesis:

$$H_0 : \Sigma = \Sigma_0 \tag{65}$$

$$H_A : \Sigma \neq \Sigma_0 \tag{66}$$

Conclusions:

Can conclude that covariance does not equal a particular value.

Box's M test

What are you testing:

Testing to see if two covariance matrices are equal.

Null/Alternative Hypothesis:

$$H_0 : \Sigma_1 = \Sigma_2 = ... = \Sigma_k \tag{67}$$

H_A : At least two $\Sigma_j' s$ differ

Test Statistic:

$$M = [(\frac{|S_1|}{|S_p|})^{\nu_1}(\frac{|S_2|}{|S_p|})^{\nu_2}...(\frac{|S_k|}{|S_p|})^{\nu_k}]^{0.5} \tag{68}$$

M small to reject. If all of the Sigmas are equal than each of the terms in the M equation will be equal to one. If the sample covariance matrices are different than each determinant will be smaller than the pooled covariance and result in a small value for M.

$$S_p = \frac{\sum_{i=1}^{k} \nu_i S_i}{\sum_{i=1}^{k} \nu_i} = \frac{E}{\nu_E} \tag{69}$$

$$\nu_i = \nu_i - 1 \tag{70}$$

Conclusions:

Can conclude that at least one population covariance pair differes significantly.

Discriminant Analysis

Technique used to investigate how the response variables combine to optimally separate the groups. Main goal is to find a vector **a** such that the linear combination of it maximized the squared standarized distance.

Total number of discriminant functions:

$$s = min(p, k-1) \tag{71}$$

Finding the discriminant function:

$$a = S_p^{-1}(\bar{y_1} - \bar{y_2}) \tag{72}$$

Relative Importance of Discriminant Functions

$$\frac{\lambda_i}{\sum_{j=1}^{s} \lambda_j} \tag{73}$$

Standardized Discriminant Functions

$$a^* = [diag(S_p)]^{\frac{1}{2}} a \tag{74}$$

Test for at Least One Significant Discriminant Function

Let $\alpha_1, \alpha_2...$ be population discrim functions

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 \tag{75}$$

H_A : at least two disciminant functions are not equal

$$\Lambda_1 = \prod_{i=1}^{s} \frac{1}{1+\lambda_i} \tag{76}$$

$$\Lambda_2 = \prod_{i=2}^{s} \frac{1}{1+\lambda_i} \tag{77}$$

If we reject we more onto partial F test.

Conclusion: At least the first discriminant function is needed for group separation.

The pvalue for multiple eigenvalue tests requires bonferroni correction:

$$\frac{\alpha}{s-1} \tag{78}$$

Partial F Test

See how each variable contributes to overall group separation.

$$\Lambda = \frac{\Lambda_p}{\Lambda_{p-1}} \tag{79}$$

Λ_p = all variables Λ_{p-1} = all variables except the one we are testing

Conclusions: Can conclude that missing variable doesn't significantly contribute to group separation in the presence of remaining variables.

Requires Bonferroni correction

Classifican Analysis

Linear Classification Rule

$$\frac{\bar{z_1} + \bar{z_2}}{2} \tag{80}$$

Linear Classification Function

$$Li(y) = \bar{y_i}' S_p^{-1} y - \frac{1}{2} \bar{y_i}' S_p^{-1} \bar{y_i} \tag{81}$$

Select classification function for which Li is largest.

Error Rates

AER = off diagonals / total observations ACCR = 1 - AER

Principal Component Analysis

Linear combinations of variables that best preserve the primary sources of variablity. Can be thought of as translating the coordinate system to the sample mean vector and then rotating the coordinate axes until they maximize variance.

The principal components are orthogonal.

The number of principal components is equal to the number of variables.

The principal components are all uncorrelated and also orthognal to one another. Perpendicular. Loadings are the coefficients of the eigenvactors.

Variablity Accounted for by a PC

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{j=1}^{p} \lambda_j} \tag{82}$$

What is going on with page 408.

Understanding the components

The first pc is an overall measure of size. If the second component contains both positive and negative values it can be thought of as a contrast between the sets of variables. An overall measure of laundry.

Number of Principal Components to keep

Components account for a specified percentage of the total variance

Retain components with eigen values greater than the average of eigenvalues *The standard deviation is the square root of the eigen value.*

Use a scree plot and look for an obvious drop

Test the significance of the components

What are you testing:

Testing to see if any of the principal components are significant.

Null/Alternative Hypothesis:

$$H_0 : \gamma_{p-k+1} = \gamma_{p-k+2} = ... \tag{83}$$

H_A : At least one γ is not equal

Test Statistic:

Start from the last test and work up until the first test that rejects the test. Use Bonferonni correction. $\frac{\alpha_e}{p-1}$

Cluster Analysis

Main equation to find distance between variables is euclidean distance:

$$d(x, y) = \sqrt{\sum_{j=1}^{p}(x_j - y_j)^2} \tag{84}$$

Agglomerative Approach

Start with n clusters. Work down until there is only 1 cluster while keeping track of the merge distances.

Single Linkage

Find the minimum distance between two clusters. Then merge the minimum of all of the distances.

Complete Linkage

Find the maximum distance between two clusters. Then merge the minimum of all of the distances.

Record each of the merge distances.