Stockholm University

Autumn 2020
20201015

Department of Statistics

Exam: Multivariate Analysis, Advanced level, 7.5 ECTS credits

The exam consists of 4 exercises giving a total of 50 points. In order to get full score for an exercise provide detailed and well motivated solutions. In order to pass the exam at least 25 points are needed.

Exercise 1. (10p)

Let $\boldsymbol{X} \sim N_5(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{X} = (\boldsymbol{X}_1^{\mathrm{T}}, \boldsymbol{X}_2^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{X}_1 : 2 \times 1$ and $\boldsymbol{X}_2 : 3 \times 1$, and $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ be partitioned correspondingly

$$\boldsymbol{\mu} = \left(\begin{array}{r} 3 \\ -2 \\ \hline 4 \\ -3 \\ 5 \end{array}\right), \boldsymbol{\Sigma} = \left(\begin{array}{rr|rrr} 8 & -2 & 1 & 0 & 3 \\ -2 & 4 & 1 & 6 & -2 \\ \hline 1 & 1 & 10 & -1 & -1 \\ 0 & 6 & -1 & 4 & 0 \\ 3 & -2 & -1 & 0 & 1 \end{array}\right).$$

(a) Find $E(\boldsymbol{X}_1|\boldsymbol{X}_2)$ and $Cov(\boldsymbol{X}_1|\boldsymbol{X}_2)$.

(b) Using the data on bone mineral content in the file **Exercise1.txt** (*Source*: Data courtesy of Everett Smith, Johnson & Wichern, 2013), construct and sketch a 95% confidence ellipse for the pair $(\mu_1, \mu_2)^{\mathrm{T}}$, where $\mu_1 = E(X_1)$ and $\mu_2 = E(X_2)$, variables X_1 and X_2 represent "Dominant Radius" and "Radius", respectively. Construct the 95% Bonferroni intervals for the individual means μ_1 and μ_2. Also, find the 95% simultaneous T^2-intervals. Compare the two sets of intervals. What advantage, if any, do the T^2-intervals have over the Bonferroni intervals?

Exercise 2. (10p)

Use the data in the file **Exercise2.txt** on the pulp fiber characteristics, Z_1 = arithmetic fiber length, Z_2 = long fiber fraction, Z_3 = fine fiber fraction, Z_4 = zero span tensile, and the paper properties, and Y_1 = breaking length, Y_2 = elastic modulus, Y_3 = stress at failure, Y_4 = burst strength.

(a) Perform a multivariate multiple regression analysis using all four response variables, Y_1,Y_2, Y_3 and Y_4, and the four independent variables, Z_1, Z_2, Z_3 and Z_4.
(i) Suggest and fit an appropriate linear regression model. Specify the matrix of estimated coefficients $\hat{\boldsymbol{B}}$ and estimated error covariance matrix $\boldsymbol{\Sigma}$.
(ii) Analyze the residuals. Check for outliers.

Exercise 3. (15p)

File **Exercise3.txt** contains data on snap beans (Keuls, Martakis and Magid, 1984). There are four response variables: y_1 = yield earliness, y_2 = specific leaf area (SLA) earliness, y_3 =total yield, and y_4 = average SLA, and two factors: sowing date (S) and variety (V).

(a) Conduct a two-way MANOVA: test for main effects and interaction. What are the appropriate hypotheses for your analysis? Use $\alpha = 0.05$.
(b) Provide appropriate plots to illustrate the obtained results.
(c) What conclusions can you draw from your data analysis?
(d) Check the assumptions of MANOVA using graphical tools and formal statistical tests.
(e) In previous experiments, the second variety gave higher yields. Compare variety 2 with varieties 1 and 3 by means of a test on a contrast.
(f) If any of the tests in part (a) rejects H_0, carry out ANOVA F-tests on the four variables.

Exercise 4. (15p)

For 30 brands of Japanese Seishu wine, Siotani et al. (1963) studied the relationship between y_1 = taste, y_2 = odor, and

X_1 = pH,	X_5 = direct reducing sugar,
X_2 = acidity 1,	X_6 = total sugar,
X_3 = acidity 2,	X_7 = alcohol,
X_4 = sake meter,	X_8 = formyl-nitrogen.

Using the data in **Exercise4.txt**, examine the relationship between (y_1, y_2) and $(x_1, \ldots, x_8)$ using the canonical correlation analysis.
(a) Test for the significance of the canonical relations at the $\alpha = 0.05$ level.
(b) Determine the sample canonical variates and the sample correlations of the canonical variates with their component variables. Interpret the canonical variates.
(c) Are the canonical variates good summary measures of their respective sets of variables? Explain.