Ma 439 — Linear Models — Fall 2010

Problem Set #3 — Due November 16, 2010

Prof. Sawyer — Washington University

Arrange your homework in three parts, the following order:

Part I. Your answers to all questions, either written by hand or using a word processor.

Part II. The SAS programs (*.sas files) that you used for all problems in which you used SAS.

Part III. The output from the SAS programs in Part II.

For all problems in which you use SAS, either copy or transcribe answers from the SAS output to Part I or else refer in Part I to specific pages in Part III by saying (for example) "The scatterplot or matrix for Problem 3 is on page 17 of the SAS output (Part III). See page 19 in the SAS output for more details."

Make sure that you have consecutive page numbers on the SAS output in Part III by adding your own page numbers to the SAS output if necessary, so that (for example) you don't have several different page 1s in Part III. If you like, you can number pages as (for example) "Page 3-2" for the second page of output for Problem 3.

Five (5) problems on 4 pages. Use SAS (or an equivalent statistical package) for problems 3–5.

1. Let $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$ and $B = \begin{pmatrix} 3 & 5 \\ 7 & 11 \end{pmatrix}$. Let $C = A \otimes B$ be the 4×4 Kronecker or tensor product of A and B

(i) Show carefully from the definition what are the four numbers in the first row of the matrix C.

(ii) Derive all entries for $C = A \otimes B$, and show that it is consistent with the partitioned form

$$C \;=\; A \otimes B \;=\; \begin{pmatrix} a_{11}B & a_{12}B \\ a_{21}B & a_{22}B \end{pmatrix}$$

(*Hint:* See the definition of $A \otimes B$ in Section 3 of the Multivariate Linear Models notes on the Math 439 Web site.)

2. Let e be a $n \times d$ matrix whose rows are independent and distributed as $N(0, \Sigma)$ where Σ is a $d \times d$ positive definite matrix. (See e.g. Sections 1–3 (equations (1.3), (1.4), or (3.5)) of the Multivariate Linear Models notes.)

Let R_1 be an arbitrary $n \times n$ orthogonal matrix, and let R_2 be a $d \times d$ orthogonal matrix such that $\Sigma = R_2' D R_2$ where $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ is a diagonal matrix with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d > 0$. Define a $n \times d$ matrix $Z = \{ Z_{ij} \}$ by

$$Z_{ij} \;=\; \sum_{k=1}^{n} \sum_{b=1}^{d} (R_1)_{ik} (R_2)_{jb} e_{kb}$$

(i) Find the mean and covariance matrix of Z.

(ii) Show that $\{ Z_{ij} : 1 \le i \le n, 1 \le j \le d \}$ are nd independent, real-valued normally-distributed random variables with mean zero and variances depending on (i, j). Find a formula for $\mathrm{Var}(Z_{ij})$.
(*Hint:* See Section 3 and Lemma 3.3 of the Multivariate Linear Models notes, version November 8, 2010, or later.)

3. The file `RatWeights.dat` on the Math439 Web site contains weekly gains over four weeks for three groups of rats. The first group (Group=1) was the control and was given a normal rat diet with no extra chemicals. Group 2 was given thyroxin in their drinking water, and Group 3 was given thiouracil.

(i) Is there a significant difference in weekly gains among the groups of rats, viewing the four weekly gains as vector-valued data? Carry out a MANOVA procedure to find out (using SAS or an equivalent statistical package). Why does the output give 4 different P-values for the vector-valued procedure? Are they all significant or all nonsignificant? What are the P-values?

(ii) What are the two nonzero eigenvalues of the test matrix involved? What functions of the two eigenvalues are the four tests in part (i) based upon? (*Hint:* See Section 6.1 in the text or Section 11 in the Multivariate Linear Model notes.)

(iii) Which of the four individual coordinates (weekly gains) are significant, using the corresponding univariate ANOVAs with three treatment groups? What are the P-values for the coordinates that are significant?

(iv) Find the means of the four weekly weight gains for each group. Does any one group stand out as being different in weekly gains, particularly in the last few weeks? (*Hint:* Use `Proc Means`.)

(v) Carry out the MANOVA test in part (i) for Groups 1 and 2 only. What are the P-values of the four tests in part (i) in this case? Why are they now the same?

(*Hints:* See the program `MAEgyptSkulls.sas` on the Math439 Web site and the discussion of one-way multivariate analysis of variance in the text (Chapter 6). For part (v), re-open the data set and exclude Group=3.)

4. A naturalist makes 4 measurements (Height, Width, Tail_Length, Length) on 50 lizards of a particular species along with the Altitude at which the lizard was collected. The data is in the file `AltLizards.dat` on the Math439 Web site.

(i) Carry out a multivariate regression of the four lizard measurements on Altitude using SAS or a comparable statistical package. What is the P-value for the multivariate regression? Why does the output list *four* different

P-values for the multivariate regression on Altitude, and why are all of them the same?

(ii) What are the degrees of freedom of the F-distribution that is behind the four tests in part (i), both numerator and denominator? How were these calculated in this case? Do the numbers of degrees of freedom in the output agree with your predictions?

(*Hint:* See the discussion of the Hotelling T^2 test in the handout on Multivariate Linear Models.)

(iii) Which of the four simple univariate regressions for the four physical measurements on Altitude are significant? For the significant univariate regressions, what are their P-values? what are the estimated slopes as a function of Altitude?

(iv) Construct a two-dimensional scatterplot of the lizard measurements with Width on the Y-axis and Length on the X-axis, with a plotting symbol `L` for Altitude less than or equal to 10 and `H` for Altitude greater than 10. Can you see a trend from the upper left to the lower right in the scatterplot as the altitude increases or decreases? If so, in which way?

(*Hint:* You can define a plotting symbol in a data step by an if-then-else statement like, "`if Altitude<10 then ASym='L'; else Asym='H';`".)

5. Consider the lizards represented in the data set `AltLizards.dat` whose Altitude is either less than 8.0 (call these Type=1) or else greater than 12.0 (call these Type=2). (Discard the remaining lizards.) For simplicity, call the 4 lizard measurements Y_1 Y_2 Y_3 Y_4 instead of Height Width Tail_Length Length. The naturalist is interested in finding a rule that depends only on Y_1–Y_4 and that will classify most Type=1 lizards as Type=1 (i.e., low altitude) and most Type=2 lizards as Type=2 (i.e., high altitude).

Use Fisher's Linear Discriminant method for these lizards to find a linear discriminant function

$$L(\text{data}) = c_0 + c_1 Y_1 + c_2 Y_2 + c_3 Y_3 + c_4 Y_4$$

with the property that $L(\text{data}) > 0$ predicts Type=1 and $L(\text{data}) < 0$ predicts Type=2, finding c_0–c_4 to at least two digits beyond the decimal place. Assume that SAS's default assumptions for `proc discrim` holds for these lizards. (That is, the measurements Y_1–Y_4 are joint normal with the same 4×4 covariance matrix within each Type.)

(i) How many lizards are you using to find $L(\text{data})$? That is, how many lizards are of Types either 1 or 2?

(ii) Find coefficients for the linear discriminant function $L(\text{data})$. (*Hint:* See the comments in `DFisherFerns.sas`. In particular, the coefficients of

the function L(data) can be computed as the difference between two columns in a table in SAS's `proc discrim` output.)

(iii) How many mistakes (that is, misclassifications) does the discriminant function make when applied to the lizards that were used to derive it? (This is called a "simple resubstitution" analysis.)

(*Hints:* (i) Try re-opening the dataset and adding a statement like "`if Altitude<8 then Type=1; else if Altitude>12 then Type=2; else delete;`
Check the use of the key word `delete` by a `proc print` statement to display the new dataset.

(ii) Note that you are not using cross-validation nor reading in an additional "moredat" dataset to apply the rule to. Thus the call to "proc discrim" can be much simpler than in `DFisherFerns.sas` on the Math439 Web site.)