

Assignment Sheet No. 6

Reading

Week of October 31 - November 4: J&W, Chapter 8

Week of November 7 - 11: J&W Sec 9.1 - 9.6

Week of November 14 - 18; J&W, Chapter 10

Week of November 21 - 23; J&W, Sec. 11.1 - 11.4

Week of November 28 - December 2; remainder of Chapter 11, Sec. 12.1-12.3

Week of December 5 - 9 J&W, Chapter 12

Written Assignment (due in class Friday, November 18)

Files `cbspots.txt` and `cbbones.txt` are available on the class web page.

1. Here some lines excerpted from matrix `spots` on file `cbspots.txt`

```
spots          50      20 format labels ended
) Density measurements on 19 identifiable spots on each of 50 auto-
) radiographs, each spot corresponding to a particular (probably
) unknown) protein.
)
) The data in each row was derived from the blood of a rat subjected to
) a treatment expected to affect its thyroid hormones.
) There were 10 treatments in all, including a control (treatment 2).
)
) Col. 1: trt = treatment number (1-10)
) Col. 2-20: density measurements on spots 1-19.
)
) A 50 by 10 matrix of indicator dummy variables can be computed in
) MacAnova by
) Cmd> dummys <- 1*(spots[,1] == run(10)')
) A 50 by 9 matrix of contrast dummy variables (values 1, 0 or -1)
) comparing each group with group 10 can be computed from dummys by
) Cmd> design <- dummys[,-10] - dummys[,10]
) This contrast matrix is also available in this file as matrix 'design'
) You can read it by
) Cmd> design <- matread("cbspots.txt","design")
1  0  0  0  0  448 29 20 65 12  0 892 491 122 399 58 0 140 195 396
1  0  6  6  0 683 43  3 49 12 16 928 434 126 324 63 0 171 214 472
1  0  5  0  4 387 50 11 22 36 22 875 231 128 378 28 7 109 160 343
1  0  8  7  0 558 63  7 46 56 36 925 207 245 304 75 0 62 175 408
2 214 96 44 262 323 178 181 7 106 2482 515 179 55 164 37 0 41 137 125
2 259 198 48 202 596 160 182 11 194 2589 498 204 124 225 49 0 118 175 174
2 247 182 53 224 663 140 192 19 131 2678 610 112 111 176 50 0 61 208 172
2 274 148 50 217 704 208 134 8 70 2669 635 45 54 251 57 0 76 240 137
2 270 149 33 230 694 164 154 30 70 1866 558 201 65 155 46 0 76 227 218
2 242 188 36 220 724 202 138 29 72 2110 482 106 53 287 36 0 53 219 137
2 293 97 43 213 687 188 137 37 71 2138 446 127 47 311 46 0 58 239 151
2 189 208 30 203 735 203 162 25 167 2053 674 123 50 217 50 0 92 201 191
2 263 211 58 212 714 217 142 25 164 1916 535 118 108 249 73 0 136 254 199
2 219 200 51 215 683 169 133 12 171 2390 554 59 143 256 63 3 117 166 180
2 208 203 64 196 691 206 145 12 86 1930 511 93 58 152 56 0 97 186 163
3 180 69 22 188 659 130 138 21 2 632 1110 443 381 339 51 0 67 97 253
. . . . .
. . . . .
9 349 204 88 244 186 328 130 38 79 1147 256 54 36 387 75 0 52 202 161
10 215 468 124 134 101 477 202 9 688 1151 363 141 28 124 55 0 100 298 191
```

STAT 5401 Assignment Sheet No. 6

```
10 317 437 150 147 119 659 131 7 527 1745 393 134 37 107 48 0 70 219 155
10 311 285 158 122 110 365 180 13 639 966 306 127 51 109 45 0 94 167 113
10 283 368 106 150 144 682 159 4 703 1202 274 117 42 92 77 0 74 211 183
```

The data arose in a endocrinology study of the effects of 9 treatments on the blood proteins of rats. Each of the $p = 19$ variables is proportional to the density of a spot on an autoradiogram. Treatment 2 corresponds to a **control group** which received no treatment. The interest was in seeing how well the variables could distinguish among the treatments and to determine which were the more important proteins (spots) from this point of view.

The data should be analyzed in terms of the scale $\log(y+1)$ so as to stabilize variances.

(a) Test the hypothesis that the 10 treatment groups have the same spot densities using sequential F tests in the order the responses appear in the file. You can use macro `seqF()` in the revised `Mulvar.mac` macro file to check your results, but you should compute the first 3 sequential F's using `anova()` as illustrated in Lecture 21 (10/25/04) even if fewer would allow you to determine whether the hypothesis could be rejected.

(b) By an analysis of covariance, test the hypothesis that spots 10 through 19 do not differentiate among the treatment groups after adjusting for spots 1 – 9.

(c) Compute the first 3 MANOVA canonical variables z_1 , z_2 , and z_3 , for these data and make scatter plots of z_2 vs z_1 , z_3 vs z_1 , and z_3 vs z_2 .

2. In file `cbbones.txt` are all of the data described in J&W Example 9.14 on p. 558. They consist of 6 bone measurements on 276 White Leghorn chickens. Here is a listing of the header on the matrix and the first three lines of data:

```
bonedata      276      6 format labels
) Bone measurements on n = 276 outbred female chickens, all in mm.
) Col. 1:  skull length
) Col. 2:  skull breadth
) Col. 3:  femur length (leg bone)
) Col. 4:  tibia length (leg bone)
) Col. 5:  humerus length (wing bone)
) Col. 6:  ulna length (wing bone)
)"3x%lf %lf %lf %lf %lf %lf"
(3x,3f5.1,f6.1,2f5.1)
 3 40.3 31.0 80.3 116.5 78.6 73.8
219 41.0 31.0 77.4 119.0 74.7 70.8
147 40.3 30.3 84.5 125.5 79.3 73.6
. . . . .
```

The data have been reordered in random order. The first item on each data line (which is omitted by `read()`) is the original case number.

An error in the original data has been corrected.

(a) Use appropriate techniques to identify any remaining outliers. This can involve bivariate plots, rankit plots, and chi-squared Q-Q plots. So that subsequent analyses are consistent, do *not* eliminate the outliers.

(b) Use the singular value decomposition of the data matrix with means subtracted to find the best rank 2 approximation to the data matrix. Print out only the first 10 and last 10 rows of the approximation. Also print out the first and last 10 rows of t_1L_1 and t_2L_2 (t_j the singular values and L_j the left singular vectors). Make plots against case number of the residuals from the approximation for each variable similar to those in Lecture 24, and a plot against case number of the sum of squared residuals for each case.

(c) Determine two sets of principal components, one set based on the sample variance matrix and the other on the sample correlation matrix. Rescale the coefficients of the latter so that they may be directly applied to the original measurements. That is, find vectors $\mathbf{v}_1 = [v_{11}, \dots, v_{61}]'$, $\mathbf{v}_2 = [v_{12}, \dots, v_{62}]'$, ... of coefficients v_{ij} such that the j^{th} principal component is $\mathbf{v}_j' \mathbf{y} = \sum_{1 \leq i \leq 6} v_{ij} y_i$. Print out only the first 10 and last 10 rows of each set of principal components.

Hint: The eigenvectors of the sample correlation matrix \mathbf{R} are the coefficients of the normalized variables $y_i/\sqrt{s_{ii}}$, *not* of y_i .

(d) For both sets of principal components determined in (c), make scatter plots of the first two principal components against each other.

(e) Find that best rank 2 approximations to the sample variance matrix \mathbf{S} and sample correlation matrix \mathbf{R} . Compare the rank 2 approximation to \mathbf{S} to the sample variance matrix of the best rank 2 approximation to the data matrix found in (b).

MacAnova note

Here's one way to compute a lower rank approximation to a matrix X along with residuals using the SVD of the matrix after subtracting the mean vector. X contained artificial data having nothing to do with the bone data.

```
Cmd> list(X) # n = 20, p = 5
X                REAL    20    5

Cmd> xbar <- sum(X)/nrows(X) # mean vector as a row vector
Cmd> svdX <- svd(X - xbar,all:T) # compute singular value decomposition
Cmd> left <- svdX$leftvectors # columns are left singular vectors
Cmd> vals <- dmat(svdX$values) # diagonal matrix of singular values
Cmd> right <- svdX$rightvectors # columns are right singular vectors
Cmd> m <- 2 # Number of singular vectors to use
Cmd> J <- run(m) # selector subscript
Cmd> X2 <- xbar + left[,J] %**% vals[J,J] %**% right[,J]' # Approximation
Cmd> resids <- X - X2 # residuals

Cmd> list(X2,resids) # both have same dimensions as X
resids                REAL    20    5
X2                    REAL    20    5

Cmd> sum(vector(resids^2)) # sum of squared residuals
(1)                306.16

Cmd> sum(diag(vals)[-J]^2) #sum of squares of left out sing. values
(1)                306.16

Cmd> rowss <- vector(sum(resids'^2)) # row sums of squared residuals
Cmd> list(rowss) # an element for every row
rowss                REAL    20
```