MAT4375 HW#2, 1999

Due to html limitations, ``X- bar'' will be noted $\underline{X}$.
From the third edition of Johnson and Wichern

- Problem 3.4
- Problem 3.6
- Problem 4.4
- Problem 4.14 (#4.19 in the fourth edition)
- Problem 4.20 (#4.27 in the fourth edition)
- Problems 4.21 and 4.22 (#4.28 and 4.29 in the fourth edition)

Problem 3.4

For visual simplicity we work in millions of dollars rather than dollars.

- (a) Here $p_1 = (x_1'\cdot 1/\|1\|)\, 1/\|1\|$, which equals $\underline{x}_1\cdot 1 \cong 2.10121$.
- (b) $e_1 = y_1 - \underline{x}_i 1 \cong (1.3967, .3843, -.3183, -.3757, -.4556, -.6314)'$. Note that $(s_{11})^{1/2} = \|y_1 - \underline{x}_1\|/6^{1/2}$.
- (c) By construction p_1 and e_1 are orthogonal. We have that $\|e_1\| \cong 1.717$ and $\|p_1\| \cong 5.147$. We can sketch the relationship among y_1, e_1 and p_1 as a right-angle triangle with hypotenuse y_1, short side e_1 and long side p_1.
- (d) We obtain $p_2 = .5141$, $e_2 \cong (.109, .079, -.002, -.014, -.0051, -.119)'$, $\|p_2\| \cong 1.259$ and $\|e_2\| \cong 0.187$.
- (e) The angle between e_1 and e_2 is given by $e_1'e_2 = \|e_1\|\cdot\|e_2\|\cos(\theta)$. Here $\cos\theta = .8936$ and so θ is about 27 degrees.

Problem 3.6

- (a) Here

$$X - \underline{X}1' = \begin{bmatrix} -3 & 0 & 3 \\ 0 & 1 & -1 \\ -3 & 1 & 2 \end{bmatrix}$$

  Since the determinant is 0, it is not of full rank. In fact, the third row is the sum of the two others.
- (b) Using $S = (n-1)^{-1}\sum_{j=1}^{3}(X_i - \underline{X})(X_i - \underline{X})'$ we obtain

$$S = \begin{bmatrix} 9 & -3/2 & 15/2 \\ -3/2 & 1 & -1/2 \\ 15/2 & -1/2 & 7 \end{bmatrix}$$

  and the determinant of S is 0. Thus the generalized sample variance is 0. This follows from the fact that the rows of the matrix in (a) lie in a plane within R^3 - i.e. define a region of 0 volume.
- (c) The total sample variance is the trace of S, i.e. 17.

Problem 4.4

Here $X \sim N(\mu, \Sigma)$ with $\mu = (2, -3, 1)'$ and

$$S = = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 3 & 2 \\ 1 & 2 & 2 \end{bmatrix}.$$

- (a) $l = (3, -2, 1)'$. Thus $l'X$ has a $N(l'\mu, l'\Sigma l) = N(13,9)$ distribution.
- (b) X_2 and $Y = X_2 - a'(X_1, X_3)'$ will be independent if they have 0 covariance. That is, we need

$$0 = Cov(X_2,Y) = Cov(X_2,X_2) - Cov(X_2,a'(X_1, X_3)').$$

This is equivalent to

$$0 = Var(X_2) - a_1 Cov(X_2,X_1) - a_2 Cov(X_2,X_3)$$

. Take $a_1 = a_2 = 1$ to fulfill this condition.

Problem 4.14 (4.19 in fourth edition)

Here $X_1, X_2, ..., X_{20}$ are iid $N_6(\mu, \Sigma)$ random vectors.

- (a) $\chi^2(6)$ distribution.
- (b) $\underline{X} \sim N_6(\mu, \Sigma/20)$ and $(\underline{X}-\mu)/n^{1/2} \sim N_6(0,\Sigma)$.
- (c) $(n-1)S \sim W_{19}(\Sigma)$.

Problem 4.20 (4.27 in the 4th edition)

Program Output

The p-values of the univariate Shapiro-Wilks tests, r_Q value for the χ^2 QQ plot and comments on the χ^2 QQ plot followbelow. It seems that the original data is strongly non-normal, and that the log-transformed data is not yet sufficiently close to normal. If $\gamma = .25$, the transform seems adequate - .046 is not too bad. The Mardia statistics on Skewness and Kurtosis have similar behaviour. One should note that the QQ plot line seems to be strongly influenced by the relatively few data points far removed from the center. Perhaps some of these should be deleted and the analysis re-done.

Case	univariate p-values		r_Q	comments on plot
untransformed	< .0001	< .0001	.939	strongly S-shaped
$\gamma = .25$	.184	.046	.9883	fairly linear
$\gamma = 0$	.02	.002	.9693	bad for 5 ``large" points

Problems 4.21 and 4.22 (4.28 and 4.29 in the 4th edition)

Program Output

- 4.21 (4.28) The r_Q value is 0.969. For a sample of 42 this has p-value between .01 and .05 (Table 4.2), indicating non-normality of solar radiation. The normal QQ plot seems to show a heavy lower tail to the distribution.

- 4.22 (4.29) (a,b) The generalized distances are shown in the output of the sas program. These appear in a sorted table together with and indicator (``dd'') of those squared distances (``mahdist'') in the 50% envelope (``1'') and in the 95% envelope (``2''). The bivariate plot of NO_2 against O_3 with these labels is also shown. Notice the cluster of observations for O_3 around 25. These might be outliers.
- 4.22 (4.29) (c) The 6 furthest removed observations from the center seem to part of a separate group from the remaining observations. This reinforces the need to consider splitting the data

---