1. (15 pts) Ex. 3.3 Find the principal components of the following correlation matrix given by MacDonnell (1902) from measurements of seven physical characteristics in each of 3 000 convicted criminals:

$$R = \begin{array}{l} \text{Head length} \\ \text{Head breadth} \\ \text{Face breadth} \\ \text{Left finger length} \\ \text{Left forearm length} \\ \text{Left foot length} \\ \text{Height} \end{array} \begin{pmatrix} 1.000 & & & & & & \\ 0.402 & 1.000 & & & & & \\ 0.396 & 0.618 & 1.000 & & & & \\ 0.301 & 0.150 & 0.321 & 1.000 & & & \\ 0.305 & 0.135 & 0.289 & 0.846 & 1.000 & & \\ 0.339 & 0.206 & 0.363 & 0.759 & 0.797 & 1.000 & \\ 0.340 & 0.183 & 0.345 & 0.661 & 0.800 & 0.736 & 1.000 \end{pmatrix}$$

How would you interpret the derived components?

2. (10 pts) Ex. 3.4 Not all canonical correlations may be statistically significant. An approximate test proposed by Bartlett (1947) can be used to determine how many significant relationships exist. The test statistic for testing that at least one canonical correlation is significant is

$$\emptyset_0^2 = -\left\{n - \frac{1}{2}(q_1 + q_2 + 1)\right\} \sum_{i=1}^{s} \log(1 - \lambda_i)$$

where the λ_i are the eigenvalues of E1 and E2. Under the null hypothesis that all correlations are zero, $\emptyset_0^2$ has a chi-square distribution with $q1 \times q2$ degrees of freedom. Write R code to apply this test to the headsize data (Table 3.1) and the depression data (Table 3.3).

3. Ex. 3.5 (15 pts) Repeat the regression analysis for the air pollution data described in the text after removing whatever cities you think should be regarded as outliers. For the results given in the text and the results from the outliers-removed data, produce scatterplots of Sulphur dioxide concentration against each of the principal component scores. Interpret your results.

4. (30 pts) Construct PCA for heptathlon data (in textbook section 3.10.2) as the examples did in class. Interpret all the results you obtained from PCA package in R.

5. (20 pts) Measurements of properties of pulp fibers and the paper made from them are contained in propertydata. There are $n = 62$ observations of the pulp fiber characteristics.

   Let the paper characteristics be
   $x_1 = \text{breaking length}$, $x_2 = \text{elastic modulus}$, $x_3 = \text{stress at failue}$, $x_4 = \text{burst strength}$.

   Let the pulp fiber characteristics be
   $y_1 = \text{arithmetic fiber length}$, $y_2 = \text{long fiber fraction}$, $y_3 = \text{fine fiber fraction}$,
   $y_4 = \text{zero span tensile}$.

   a) Determine the number of significant canonical variate pairs;
   b) Compute the canonical variates from the data;
   c) Interpret each member of a canonical variate pair using its correlations with the member variables;
   d) Use the results of canonical correlation analysis to describe the relationships between two sets of variables.