# 8
# Multiple Regression and Canonical Correlation

## 8.1 Introduction

In this chapter we discuss two related but separate techniques, *multiple regression* and *canonical correlation*. The first of these is not strictly a multivariate procedure; the reasons for including it in this book are that it provides some useful basic material both for the discussion of canonical correlation in this chapter and modelling longitudinal data in Chapter 9.

## 8.2 Multiple Regression

Multiple linear regression represents a generalization, to more than a single explanatory variable, of the simple linear regression model met in all introductory statistics courses. The method is used to investigate the relationship between a dependent variable, $y$, and a number of explanatory variables $x_1, x_2, \ldots, x_q$. Details of the model, including the estimation of its parameters by least squares and the calculation of standard errors are given in Display 8.1. Note in particular that the explanatory variables are, strictly, not regarded as random variables at all so that multiple regression is essentially a *univariate* technique with the only random variable involved being the response, $y$. Often the technique is referred to as being *multivariable* to properly distinguish it from genuinely multivariate procedures.

As an example of the application of multiple regression we can apply it to the air pollution data introduced in Chapter 3 (see Table 3.1), with $SO_2$ level as the dependent variable and the remaining variables being explanatory. The model can be applied in R and S-PLUS® and the results summarized using

```
attach(usair.dat)
usair.fit<-lm(SO2~Neg.Temp + Manuf + Pop + Wind +
   Precip + Days)
summary(usair.fit)
```

**Display 8.1**
**Multiple Regression Model**

- The multiple linear regression model for a response variable $y$ with observed values $y_1, y_2, \ldots, y_n$ and $q$ explanatory variables, $x_1, x_2, \ldots, x_q$, with observed values $x_{i1}, x_{i2}, \ldots, x_{iq}$ for $i = 1, 2, \ldots, n$, is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_q x_{iq} + \varepsilon_i.$$

- The regression coefficients $\beta_1, \beta_2, \ldots, \beta_q$ give the amount of change in the response variable associated with a unit change in the corresponding explanatory variable, *conditional* on the other explanatory variables in the model remaining unchanged.
- The explanatory variables are strictly assumed to be fixed; that is, they are not random variables. In practice, where this is rarely the case, the results from a multiple regression analysis are interpreted as being *conditional* on the observed values of the explanatory variables.
- The residual terms in the model, $\varepsilon_i, i = 1, \ldots, n$, are assumed to have a normal distribution with mean zero and variance $\sigma^2$. This implies that, for given values of the explanatory variables, the response variable is normally distributed with a mean that is a linear function of the explanatory variables and a variance that is not dependent on these variables. Consequently an equivalent way of writing the multiple regression model is as $y \sim N(\mu, \sigma^2)$ where $\mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q$.
- The "linear" in multiple linear regression refers to the parameters rather than the explanatory variables, so the model remains linear if, for example, a quadratic term for one of these variables is included. (An example of a *non-linear model* is $y = \beta_1 e^{\beta_2 x_{i1}} + \beta_3 e^{\beta_4 x_{i2}} + \varepsilon_i$.)
- The aim of multiple regression is to arrive at a set of values for the regression coefficients that makes the values of the response variable predicted from the model as close as possible to the observed values.
- The least-squares procedure is used to estimate the parameters in the multiple regression model.
- The resulting estimators are most conveniently written with the help of some matrices and vectors. By introducing a vector $\mathbf{y}' = [y_1, y_2, \ldots, y_n]$ and an $n \times (q + 1)$ matrix $\mathbf{X}$ given by

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1q} \\ 1 & x_{21} & x_{22} & \ldots & x_{2q} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nq} \end{pmatrix},$$

we can write the multiple regression model for the $n$ observations concisely as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon}' = [\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n]$ and $\boldsymbol{\beta}' = [\beta_0, \beta_1, \ldots, \beta_q]$.

- The least-squares estimators of the parameters in the multiple regression model are given by the set of equations

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

- More details of the least-squares estimation process are given in Rawlings et al. (1998).
- The variation in the response variable can be partitioned into a part due to regression on the explanatory variables and a residual as for simple linear regression. The can be arranged in an analysis of variance table as follows:

| Source | DF | SS | MS | $F$ |
|---|---|---|---|---|
| Regression | $q$ | RGSS | RGSS/$q$ | RGMS/RSMS |
| Residual | $n - q - 1$ | RSS | RSS/$n - q - 1$ | |

- The residual mean square $s^2$ is an estimator of $\sigma^2$.
- The covariance matrix of the parameter estimates in the multiple regression model is estimated from

$$\mathbf{S}_{\hat{\beta}} = s^2(\mathbf{X}'\mathbf{X})^{-1}.$$

  The diagonal elements of this matrix give the variances of the estimated regression coefficients and the off-diagonal elements their covariances.
- A measure of the fit of the model is provided by the *multiple correlation coefficient*, $R$, defined as the correlation between the observed values of the response variable, $y_1, K, y_n$, and the values predicted by the fitted model, that is,

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_q x_{iq}$$

- The value of $R^2$ gives the proportion of variability in the response variable accounted for by the explanatory variables.

The results are shown in Table 8.1. The $F$ statistic for testing the hypothesis that all six regression coefficients in the model are zero is 11.48 with 6 and 34 degrees of freedom. The associated $p$-value is very small and the hypothesis should clearly be rejected. The $t$-statistics suggest that *Manuf* and *Pop* are the most important predictors of sulphur dioxide level. The square of the multiple correlation coefficient is 0.67 showing that 67% of the variation in $SO_2$ level is accounted for by the six explanatory variables.

When applying multiple regression in practice, of course, analysis would continue to try to identify a more parsimonious model, followed by examination of residuals from the final model to check assumptions. We shall not do this since

**Table 8.1** Results of Multiple Regression Applied to Air Pollution Data

| Covariate | Estimated regression coefficient | Standard error | $t$-value | $p$ |
|---|---|---|---|---|
| (Intercept) | 111.7285 | 47.3181 | 2.3612 | 0.0241 |
| Neg. temp | 1.2679 | 0.6212 | 2.0412 | 0.0491 |
| Manuf | 0.0649 | 0.0159 | 4.1222 | 0.0002 |
| Pop | −0.0393 | 0.0151 | −2.5955 | 0.0138 |
| Wind | −3.1814 | 1.8150 | −1.7528 | 0.0887 |
| Precip | 0.5124 | 0.3628 | 1.4124 | 0.1669 |
| Days | −0.0521 | 0.1620 | −0.3213 | 0.7500 |

Residual standard error: 14.64 on 34 degrees of freedom.
Multiple $R$-squared: 0.6695.
$F$-statistic: 11.48 on 6 and 34 degrees of freedom, the $p$-value is 5.419e − 007.

here we are largely interested in the univariate multiple regression model merely as a convenient stepping stone to discuss a number of multivariate procedures beginning with canonical correlation analysis.

## 8.3 Canonical Correlations

Multiple regression is concerned with the relationship between a single variable $y$ and a set of variables $x_1, x_2, \ldots, x_q$. Canonical correlation analysis extends this idea to investigating the relationship between two sets of variables, *each* containing more than a single member. For example, in psychology an investigator may measure a set of aptitude variables and a set of achievement variables on a sample of students. In marketing, a similar example might involve a set of price indices and a set of prediction indices. The objective of canonical correlation analysis is to find linear functions of one set of variables that maximally correlate with linear functions of the other set of variables. In many circumstances one set will contain multiple dependent variables and the other multiple independent or explanatory variables and then canonical correlation analysis might be seen as a way of predicting multiple dependent variables from multiple independent variables. Extraction of the coefficients that define the required linear functions has similarities to the process of finding principal components as described in Chapter 3. Some of the steps are described in Display 8.2.

To begin we shall illustrate the application of canonical correlation analysis on a data set reported over 80 years ago by Frets (1921). The data are given in Table 8.2 and give head measurements (in millimeters) for each of the first two adult sons in 25 families. Here the family is the "individual" in our data set and the four head measurements are the variables. The question that was of interest to Frets was whether there is a relationship between the head measurements for pairs of sons? We shall address this question by using canonical correlation analysis.

Here we shall develop the canonical correlation analysis from first principles as detailed in Display 8.2. Assuming the head measurements data are contained in the

**Display 8.2**
**Canonical Correlation Analysis (CCA)**

- The purpose of canonical correlation analysis is to characterize the independent statistical relationships that exist between two sets of variables, $\mathbf{x}' = [x_1, x_2, \ldots, x_{q_1}]$ and $\mathbf{y}' = [y_1, y_2, \ldots, y_{q_2}]$.
- The overall $(q_1 + q_2) \times (q_1 + q_2)$ correlation matrix contains all the information on associations between pairs of variables in the two sets, but attempting to extract from this matrix some idea of the association between the two sets of variables is not straightforward. This is because the correlations between the two sets may not have a consistent pattern; and these between-set correlations need to be adjusted in some way for the within-set correlations.
- The question is "How do we quantify the association between the two sets of variables **x** and **y**?"
- The approach adopted in CCA is to take the association between **x** and **y** to be the largest correlation between two single variables $u_1$ and $v_1$ derived from **x** and **y**, with $u_1$ being a linear combination of $x_1, x_2, \ldots, x_{q_1}$ and $v_1$ being a linear combination of $y_1, y_2, \ldots, y_{q_2}$.
- But often a single pair of variables, $(u_1, v_1)$ is not sufficient to quantify the association between the $x$ and $y$ variables, and we may need to consider some or all of $s$ pairs $(u_1, v_1), (u_2, v_2), \ldots, (u_s, v_s)$ to do this, where $s = \min(q_1, q_2)$.
- Each $u_i$ is a linear combination of the variables in **x**, $u_i = \mathbf{a}_i'\mathbf{x}$, and each $v_i$ is a linear combination of the variables **y**, $v_i = \mathbf{b}_i\mathbf{y}$, with the coefficients $(\mathbf{a}_i, \mathbf{b}_i)(i = 1, \ldots, s)$ being chosen so that the $u_i$ and $v_i$ satisfy the following:

  (1) The $u_i$ are mutually uncorrelated, i.e., $\text{cov}(u_i, u_j) = 0$ for $i \neq j$.
  (2) The $v_i$ are mutually uncorrelated, i.e., $\text{cov}(v_i, v_j) = 0$ for $i \neq j$.
  (3) The correlation between $u_i$ and $v_i$ is $R_i$ for $i = 1, \ldots, s$, where $R_1 > R_2 > \cdots > R_s$.
  (4) The $u_i$ are uncorrelated with all $v_j$ except $v_i$, i.e., cov $(u_i, v_j) = 0$ for $i \neq j$.

- The linear combinations $u_i$ and $v_i$ are often referred to as *canonical variates*, a name used previously in Chapter 7 in the context of multiple discriminant function analysis. In fact there is a link between the two techniques. If we perform a canonical correlation analysis with the data **X** defining one set of variables and a matrix of group indicators, **G**, as the other we obtain the linear discriminant functions. Details are given in Mardia et al. (1979).
- The vectors $\mathbf{a}_i$ and $\mathbf{b}_i$ $i = 1, \ldots, s$, which define the required linear combinations of the $x$ and $y$ variables, are found as the eigenvectors of matrices $\mathbf{E}_1(q_1 \times q_1)$ (the $\mathbf{a}_i$) and $\mathbf{E}_2(q_2 \times q_2)$ (the $\mathbf{b}_i$) defined as

$$\mathbf{E}_1 = \mathbf{R}_{11}^{-1}\mathbf{R}_{12}\mathbf{R}_{22}^{-1}\mathbf{R}_{21}, \qquad \mathbf{E}_2 = \mathbf{R}_{22}^{-1}\mathbf{R}_{21}\mathbf{R}_{11}^{-1}\mathbf{R}_{12},$$

where $\mathbf{R}_{11}$ is the correlation matrix of the variables in $\mathbf{x}$, $\mathbf{R}_{22}$ is the correlation matrix of the variables in $\mathbf{y}$, and $\mathbf{R}_{12}(=\mathbf{R}_{21})$ the $q_1 \times q_2$ matrix of correlations across the two sets of variables.

- The canonical correlations $R_1, R_2, \ldots, R_s$ are obtained as the square roots of the nonzero eigenvalues of either $\mathbf{E}_1$ or $\mathbf{E}_2$.
- The $s$ canonical correlations $R_1, R_2, \ldots, R_s$ express the association between the $\mathbf{x}$ and $\mathbf{y}$ variables after removal of the within-set correlation.
- More details of the calculations involved and the theory behind canonical correlation analysis are given in Krzanowski (1988).
- Inspection of the coefficients of each original variable in each canonical variate can provide an interpretation of the canonical variate in much the same way as interpreting principal components (see Chapter 3). Such interpretation of the canonical variates may help to describe just how the two sets of original variables are related (see Krzanowski 2004).
- In practice, interpretation of canonical variates can be difficult because of the possibly very different variances and covariances among the original variables in the two sets, which affects the sizes of the coefficients in the canonical variates. Unfortunately there is no convenient normalization to place all coefficients on an equal footing (see Krzanowski, 2004).
- In part this problem can be dealt with by restricting interpretation to the standardized coefficients, that is, the coefficients that are appropriate when the original variables have been standardized.

data frame, `headsize`, the necessary R and S-PLUS code is:

```
headsize.std<-sweep(headsize,2,
  sqrt(apply(headsize,2,var)),FUN="/")
#standardize head measurements by
#dividing by the appropriate standard deviation
#
#
headsize1<-headsize.std[,1:2]
headsize2<-headsize.std[,3:4]
#
#find all the matrices necessary for calculating the
#canonical variates and canonical correlations
#
R11<-cor(headsize1)
R22<-cor(headsize2)
R12<-c(cor(headsize1[,1],headsize2[,1]),cor(headsize1[,1],
  headsize2[,2]),
cor(headsize1[,2],headsize2[,1]),cor(headsize1[,2],
  headsize2[,2]))
```

**Table 8.2** Head Sizes in Pairs of Sons (mm)

| $x_1$ | $x_2$ | $x_3$ | $x_4$ |
|---|---|---|---|
| 191 | 155 | 179 | 145 |
| 195 | 149 | 201 | 152 |
| 181 | 148 | 185 | 149 |
| 183 | 153 | 188 | 149 |
| 176 | 144 | 171 | 142 |
| 208 | 157 | 192 | 152 |
| 189 | 150 | 190 | 149 |
| 197 | 159 | 189 | 152 |
| 188 | 152 | 197 | 159 |
| 192 | 150 | 187 | 151 |
| 179 | 158 | 186 | 148 |
| 183 | 147 | 174 | 147 |
| 174 | 150 | 185 | 152 |
| 190 | 159 | 195 | 157 |
| 188 | 151 | 187 | 158 |
| 163 | 137 | 161 | 130 |
| 195 | 155 | 183 | 158 |
| 186 | 153 | 173 | 148 |
| 181 | 145 | 182 | 146 |
| 175 | 140 | 165 | 137 |
| 192 | 154 | 185 | 152 |
| 174 | 143 | 178 | 147 |
| 176 | 139 | 176 | 143 |
| 197 | 167 | 200 | 158 |
| 190 | 163 | 187 | 150 |

$x_1$ = head length of first son; $x_2$ = head breadth of first son; $x_3$ = head length of second son; $x_4$ = head breadth of second son.

```
#
R12<-matrix(R12,ncol=2,byrow=T)
R21<-t(R12)
#
#see display 8.2 for relevant equations
E1<-solve(R11)%*%R12%*%solve(R22)%*%R21
E2<-solve(R22)%*%R21%*%solve(R11)%*%R12
#
E1
E2
#
eigen(E1)
eigen(E2)
```

The results are shown in Table 8.3. Here the four linear functions are found to be

$$u_1 = 0.69x_1 + 0.72x_2, \quad v_1 = 0.74x_1 + 0.67x_2,$$
$$u_2 = 0.71x_1 - 0.71x_2, \quad v_2 = 0.70x_1 - 0.71x_2.$$

**Table 8.3** Canonical Correlation Analysis Results on Headsize Data

$$\mathbf{E}_1 = \begin{bmatrix} 0.306 & 0.305 \\ 0.314 & 0.319 \end{bmatrix}$$

$$\mathbf{E}_2 = \begin{bmatrix} 0.330 & 0.324 \\ 0.295 & 0.295 \end{bmatrix}$$

Eigenvalues of $\mathbf{E}_1$ and $\mathbf{E}_2$ are 0.62 and 0.0029, giving the canonical correlations as $\sqrt{0.6215} = 0.7885$ and $\sqrt{0.0029} = 0.0537$ The respective eigenvectors are;

$$\begin{aligned} \mathbf{a}_1' &= [0.695, 0.719], \\ \mathbf{a}_2' &= [0.709, -0.705], \\ \mathbf{b}_1' &= [0.742, 0.670], \\ \mathbf{b}_2' &= [0.705, 0.711]. \end{aligned}$$

The first canonical variate for both first and second sons is simply a weighted sum of the two head measurements and might be labelled "girth"; these two variates have a correlation of 0.79. Each second canonical variate is a weighted difference of the two head measurements and can be interpreted roughly as head "shape"; here the correlation is 0.05. (Girth and shape are defined to be uncorrelated within first and second sons, and also between first and second sons.)

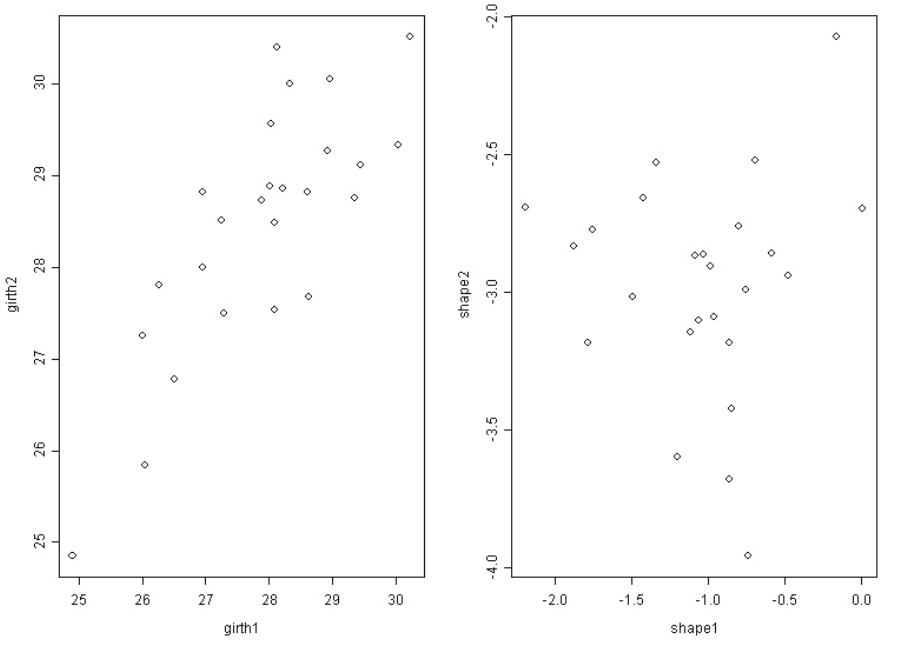


**Figure 8.1** Scatterplots of girth and shape for first and second sons.

In this example it is clear that the association between the two head measurements of first and second sons is almost entirely expressed through the "girth" variables with the two "shape" variables being almost uncorrelated. The association between the two sets of measurements is essentially one-dimensional. A scatterplot of girth for first and second sons and a similar plot for shape reinforce this conclusion. The plots are both shown in Figure 8.1 which is obtained as follows;

```
girth1<-0.69*headsize.std[,1]+0.72*headsize.std[,2]
girth2<-0.74*headsize.std[,3]+0.67*headsize.std[,4]
shape1<-0.71*headsize.std[,1]-0.71*headsize.std[,2]
shape2<-0.70*headsize.std[,3]-0.71*headsize.std[,4]
#
cor(girth1,girth2)
cor(shape1,shape2)
#
par(mfrow=c(1,2))
plot(girth1,girth2)
plot(shape1,shape2)
```

The correlations between girth for first and second sons and similarly for shape calculated by this code are included to show that they give the same values (apart from rounding differences) as the canonical correlation analysis.

We can now move on to a more substantial example taken from Afifi et al. (2004), and also discussed by Krazanowski (2004). The data for this example arise from a study of depression amongst 294 respondents in Los Angeles. The two sets of variables of interest were "health variables," namely the CESD (the sum of 20 separate numerical scales measuring different aspects of depression) and a measure of general health and "personal" variables, of which there were four, gender, age, income and educational level (numerically, coded from the lowest "less than high school," to the highest, "finished doctorate"). The sample correlation matrix between these variables is given in Table 8.4. Here the maximum number of canonical variate pairs is 2, and they can be found using the following R and S-PLUS code:

```
r22<-matrix(c(1.0,0.044,-0.106,-0.180,0.044,1.0,-0.208,
   -0.192,-0.106,-0.208,1.0,0.492,-0.180,-0.192,0.492,1.0),
   ncol=4,byrow=T)
r11<-matrix(c(1.0,0.212,0.212,1.0),ncol=2,byrow=2)
```

**Table 8.4** Sample Correlation Matrix for the Six Variables in the Los Angeles Depression Study

| | CESD | Health | Gender | Age | Education | Income |
|---|---|---|---|---|---|---|
| CESD | 1.0 | 0.121 | 0.124 | −0.164 | −0.101 | −0.158 |
| Health | 0.212 | 1.0 | 0.098 | 0.308 | −0.270 | −0.183 |
| Gender | 0.124 | 0.098 | 1.0 | 0.044 | −0.106 | −0.180 |
| Age | −0.164 | 0.308 | 0.044 | 1.0 | −0.208 | −0.192 |
| Education | −0.101 | −0.270 | −0.106 | −0.208 | 1.0 | 0.492 |
| Income | −0.158 | −0.183 | −0.180 | −0.192 | 0.492 | 1.0 |

```
r12<-matrix(c(0.124,-0.164,-0.101,-0.158,0.098,0.308,
   -0.270,-0.183),ncol=4,byrow=T)
r21<-t(r12)
#
E1<-solve(r11)%*%r12%*%solve(r22)%*%r21
E2<-solve(r22)%*%r21%*%solve(r11)%*%r12
#
E1
E2
#
eigen(E1)
eigen(E2)
```

The results are shown in Table 8.5. The first canonical correlation is 0.409 which if tested as outlined in Exercise 8.3 and has an associated $p$-value that is very small. There is strong evidence that the first canonical correlation is significant. The corresponding variates, in terms of standardized original variables, are

$$u_1 = 0.461\ \text{CESD} - 0.900\ \text{Health},$$
$$v_1 = 0.024\ \text{Gender} + 0.885\ \text{Age} - 0.402\ \text{Education} + 0.126\ \text{Income}.$$

High coefficients correspond to CESD (positively) and health (negatively) for the perceived health variables, and to age (positively) and education (negatively)

**Table 8.5** Canonical Variates and Correlation for Los Angeles Depression Study Variables

```
eigen(R1)
$values:
[1] 0.16763669 0.06806171

$vectors:
numeric matrix: 2 rows, 2 columns.
           [,1]       [,2]
[1,]  0.4610975 -0.9476307
[2,] -0.8998655 -0.3193681

eigen(R2)
$values:
[1] 1.676367e-001 6.806171e-002 -1.734723e-018 0.000000e+000

$vectors:
numeric matrix: 4 rows, 4 columns.
            [,1]       [,2]        [,3]       [,4]
[1,]  0.02424121  0.6197600 -0.03291919 -0.9378101
[2,]  0.88498865 -0.6301703 -0.16889507 -0.1840554
[3,] -0.40155454 -0.6503368 -0.53979845 -0.3193533
[4,]  0.12576714 -0.8208262  0.49453453 -0.3408145

sqrt(eigen(R1)$values)
[1] 0.4094346 0.2608864
```

for the personal variables. It appears that relatively older and medicated people tend to have a lower depression score, but perceive their health as relatively poor, while relatively younger but educated people have the opposite health perception. (I am grateful to Krzanowski, 2004, for this interpretation.)

The second canonical correlation is 0.261 which is again significant (see Exercise 8.3 and 8.4). The corresponding canonical variates are

$$u_2 = 0.95 \text{ CESD} - 0.32 \text{ Health},$$
$$v_2 = 0.620 \text{ Gender} - 0.630 \text{ Age} - 0.650 \text{ Education} - 0.821 \text{ Income}.$$

Since the higher value of the gender variable is for females, the interpretation here is that relatively young, poor, and uneducated females are associated with higher depression scores and, to a lesser extent, with poor perceived health (again this interpretation is due to Krzanowski, 2004).

## 8.4 Summary

Canonical correlation analysis has the reputation of being the most difficult multivariate technique to interpret. In many respects it is a well earned reputation! Certainly one has to know the variables involved very well to have any hope of extracting a convincing explanation. But in some circumstances (the heads measurement data is an example), CCA does provide a useful description of the association between two sets of variables.

## *Exercises*

8.1 If $\mathbf{x}$ is a $q_1$-dimensional vector and $\mathbf{y}$ a $q_2$-dimensional vector, show that they linear combinations $\mathbf{a}'\mathbf{x}$ and $\mathbf{b}'\mathbf{y}$ have correlation

$$p(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a}'\boldsymbol{\Sigma}_{12}\mathbf{b}}{(\mathbf{a}'\boldsymbol{\Sigma}_{11}\mathbf{a}\mathbf{b}'\boldsymbol{\Sigma}_{22}\mathbf{b})^{1/2}},$$

where $\boldsymbol{\Sigma}_{11}$ is the covariance matrix of the $\mathbf{x}$ variables, $\boldsymbol{\Sigma}_{22}$ the corresponding matrix for the $\mathbf{y}$ variables, and $\boldsymbol{\Sigma}_{12}$ the covariances across the two sets of variables.

8.2 Table 8.6 contains data from O'Sullivan and Mahon (1966) (data also given in Rencher, 1995), giving measurements on blood glucose for 52 women. The $y$'s represent fasting glucose measurements on three occasions and the $x$'s are glucose measurements one hour after sugar intake. Investigate the relationship between the two sets of variables using canonical correlation analysis.

8.3 Not all canonical correlations may be statistically significant. An approximate test proposed by Bartlett (1947) can be used to determine how many significant

**Table 8.6** Blood Glucose Measurements on Three Occasions. From *Methods of Multivariate Analysis*, Rencher, A.C. Copyright © 1995. Reprinted with permission of John Wiley & Sons, Inc.

| Fasting | | | One hour after sugar intake | | |
|---|---|---|---|---|---|
| $y_1$ | $y_2$ | $y_3$ | $x_1$ | $x_2$ | $x_3$ |
| 60 | 69 | 62 | 97 | 69 | 98 |
| 56 | 53 | 84 | 103 | 78 | 107 |
| 80 | 69 | 76 | 66 | 99 | 130 |
| 55 | 80 | 90 | 80 | 85 | 114 |
| 62 | 75 | 68 | 116 | 130 | 91 |
| 74 | 64 | 70 | 109 | 101 | 103 |
| 64 | 71 | 66 | 77 | 102 | 130 |
| 73 | 70 | 64 | 115 | 110 | 109 |
| 68 | 67 | 75 | 76 | 85 | 119 |
| 69 | 82 | 74 | 72 | 133 | 127 |
| 60 | 67 | 61 | 130 | 134 | 121 |
| 70 | 74 | 78 | 150 | 158 | 100 |
| 66 | 74 | 78 | 150 | 131 | 142 |
| 83 | 70 | 74 | 99 | 98 | 105 |
| 68 | 66 | 90 | 119 | 85 | 109 |
| 78 | 63 | 75 | 164 | 98 | 138 |
| 103 | 77 | 77 | 160 | 117 | 121 |
| 77 | 68 | 74 | 144 | 71 | 153 |
| 66 | 77 | 68 | 77 | 82 | 89 |
| 70 | 70 | 72 | 114 | 93 | 122 |
| 75 | 65 | 71 | 77 | 70 | 109 |
| 91 | 74 | 93 | 118 | 115 | 150 |
| 66 | 75 | 73 | 170 | 147 | 121 |
| 75 | 82 | 76 | 153 | 132 | 115 |
| 74 | 71 | 66 | 413 | 105 | 100 |
| 76 | 70 | 64 | 114 | 113 | 129 |
| 74 | 90 | 86 | 73 | 106 | 116 |
| 74 | 77 | 80 | 116 | 81 | 77 |
| 67 | 71 | 69 | 63 | 87 | 70 |
| 78 | 75 | 80 | 105 | 132 | 80 |
| 64 | 66 | 71 | 86 | 94 | 133 |
| 67 | 71 | 69 | 63 | 87 | 70 |

(*Continued*)

**Table 8.6** (*Continued*)

| Fasting | | | One hour after sugar intake | | |
|---|---|---|---|---|---|
| 78 | 75 | 80 | 105 | 132 | 80 |
| 64 | 66 | 71 | 83 | 94 | 133 |
| 67 | 71 | 69 | 63 | 87 | 70 |
| 78 | 75 | 80 | 105 | 132 | 80 |
| 64 | 66 | 71 | 83 | 94 | 133 |
| 71 | 80 | 76 | 81 | 87 | 86 |
| 63 | 75 | 73 | 120 | 89 | 59 |
| 90 | 103 | 74 | 107 | 109 | 101 |
| 60 | 76 | 61 | 99 | 111 | 98 |
| 48 | 77 | 75 | 113 | 124 | 97 |
| 66 | 93 | 97 | 136 | 112 | 122 |
| 74 | 70 | 76 | 109 | 88 | 105 |
| 60 | 74 | 71 | 72 | 90 | 71 |
| 63 | 75 | 66 | 130 | 101 | 90 |
| 66 | 80 | 86 | 130 | 117 | 144 |
| 77 | 67 | 74 | 83 | 92 | 107 |
| 70 | 67 | 100 | 150 | 142 | 146 |
| 73 | 76 | 81 | 119 | 120 | 119 |
| 78 | 90 | 77 | 122 | 155 | 149 |
| 73 | 68 | 90 | 102 | 90 | 122 |
| 72 | 83 | 68 | 104 | 69 | 96 |
| 65 | 60 | 70 | 119 | 94 | 89 |
| 52 | 70 | 76 | 92 | 94 | 100 |

NOTE: Measurements are in mg/100 ml.

relationships exist. The test statistic for testing that at least one canonical correlation is significant is

$$\Phi_0^2 = -\left\{n - \frac{1}{2}(q_1 + q_2 + 1)\right\} \sum_{i=1}^{s} \log(1 - \lambda_i)$$

where the $\lambda_i$ are the eigenvalues of $\mathbf{E}_1$ and $\mathbf{E}_2$. Under the null hypothesis that all correlations are zero $\Phi_0^2$ has a chi-square distribution with $q_1 \times q_2$ degrees of freedom. Write R and S-PLUS code to apply this test to the headsize data and to the depression data.

8.4 If the test in the previous exercise is significant, then the largest canonical correlation is removed and the residual is tested for significance using

$$\phi_1^2 = -\left\{n - \frac{1}{2}(q_1 + q_2 +)\right\} \sum_{i=2}^{s} \log(1 - \lambda_i).$$

Under the hypothesis that all but the largest canonical correlation is zero $\phi_1^2$ has a chi-square distribution with $(q_1 - 1)(q_2 - 1)$ degrees of freedom. Amend the function written for Exercise 8.3 to include this further test and then extend it to test for the significance of all the canonical correlations in both the headsize and depression data sets.