Chapter 8: Canonical Correlation Analysis and Multivariate Regression

- We now will look at methods of investigating the association between sets of variables.
- When exactly two variables are measured on each individual, we might study the association between the two variables via *correlation analysis* or *simple linear regression analysis*.
- When one *response* (or *dependent*) variable and several *explanatory* variables (a.k.a. *independent variables* or *predictors*) are observed for each individual, then the method of *multiple linear regression analysis* could be used to study the relationship between the response and the predictors.

Canonical Correlation Analysis and Multivariate Regression

- In this chapter, we consider having two *sets* of variables, say, one set $X_1, \ldots, X_{q_1}$ and another set $Y_1, \ldots, Y_{q_2}$.

- When one set is considered "response variables" and the other set is considered "predictor variables", then we could use *multivariate regression*.

- When there is not a clear response-predictor relationship, we could use *canonical correlation analysis* (CCA) to analyze the associations.

Canonical Correlation Analysis (CCA)

- In CCA, we wish to characterize distinct statistical relationships between a set of q_1 variables and another set of q_2 variables.
- For example, we may have a set of “aptitude variables” and a set of “achievement variables” for a sample of individuals.
- Another example: We may have a set of “job duty variables” and a set of “job satisfaction variables” for a sample of employees.
- Another example: We may have a set of “head measurements” and a set of “body measurements” for a sample of individuals or animals.
- How are the sets associated?

The CCA Approach

- While the $(q_1 + q_2) \times (q_1 + q_2)$ correlation matrix contains the sample correlations between *all pairs* of variables, it does not directly tell us about within-set associations and between-set associations.

- Let the first set of variables be denoted as $\mathbf{x} = x_1, \ldots, x_{q_1}$ and the second set be denoted as $\mathbf{y} = y_1, \ldots, y_{q_2}$.

- We will seek the linear combination of the x variables and the linear combination of the y variables that are most highly correlated.

- After that, we will seek other linear combinations of the x's and y's that have high correlations.

- We want each pair of combinations to tell us something distinct, so we require that the combinations be mutually uncorrelated with the rest *except for their "partner" combination!*

Mathematics Behind CCA

- *Step 1:* Choose $u_1 = \mathbf{a}_1^{'}\mathbf{x} = a_{11}x_1 + a_{21}x_2 + \cdots + a_{q_1 1}x_{q_1}$ and $v_1 = \mathbf{b}_1^{'}\mathbf{y} = b_{11}y_1 + b_{21}y_2 + \cdots + b_{q_2 1}y_{q_2}$ such that $R_1 = corr(u_1, v_1)$ is greater than the correlation between any other linear combinations of the x's and y's.

- *Step 2:* Choose $u_2 = \mathbf{a}_2^{'}\mathbf{x} = a_{12}x_1 + a_{22}x_2 + \cdots + a_{q_1 2}x_{q_1}$ and $v_2 = \mathbf{b}_2^{'}\mathbf{y} = b_{12}y_1 + b_{22}y_2 + \cdots + b_{q_2 2}y_{q_2}$ such that $R_2 = corr(u_2, v_2)$ is as large as possible, subject to the restrictions on the next slide.

- We can continue doing this for s steps, getting s pairs of linear combinations, where $s = \min(q_1, q_2)$.

- In practice, we may focus on a smaller number of pairs of linear combinations than s.

Restrictions on the Linear Combinations

- We place the following restrictions on the possible linear combinations:
  1. $cov(u_i, u_j) = 0$ for all $i \neq j$ (the u_i's are all uncorrelated)
  2. $cov(v_i, v_j) = 0$ for all $i \neq j$ (the v_i's are all uncorrelated)
  3. $cov(u_i, v_j) = 0$ for all $i \neq j$ (the u_i is uncorrelated with all v_j *except* v_i)
  4. $R_1 > R_2 > \cdots > R_s$ (the earlier pairs of linear combinations have the higher correlations)
- The linear combinations $(u_1, v_1), \ldots, (u_s, v_s)$ are called the *canonical variates*.
- The correlations $R_1, R_2, \ldots, R_s$ between the canonical variates are called the *canonical correlations*.

Decomposition of the Full Sample Correlation Matrix

- If we arrange all $q_1 + q_2$ variables into one combined data set in the order $x_1, \ldots, x_{q_1}, y_1, \ldots, y_{q_2}$, then we could write the full sample correlation matrix as

$$\mathbf{R} = \left(\begin{array}{c|c} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \hline \mathbf{R}_{21} & \mathbf{R}_{22} \end{array} \right)$$

- Here, $\mathbf{R}_{11}$ is the $q_1 \times q_1$ sample correlation matrix of the first set of variables (the x's) alone.
- $\mathbf{R}_{22}$ is the $q_2 \times q_2$ sample correlation matrix of the second set of variables (the y's) alone.
- $\mathbf{R}_{12}$ is the $q_1 \times q_2$ matrix of correlations between the x's and the y's.
- Note that $\mathbf{R}_{21} = \mathbf{R}_{12}'$, i.e., the transpose of $\mathbf{R}_{12}$.

Coefficients of the Linear Combinations

- The vectors $\mathbf{a}_i$ and $\mathbf{b}_i$ ($i = 1, \ldots, s$) that contain the coefficients of the s pairs of linear combinations can be derived from $\mathbf{R}_{11}, \mathbf{R}_{12}, \mathbf{R}_{22}$.
- The vectors $\mathbf{a}_1, \ldots, \mathbf{a}_s$ are the eigenvectors of the $q_1 \times q_1$ matrix $\mathbf{E}_1 = \mathbf{R}_{11}^{-1}\mathbf{R}_{12}\mathbf{R}_{22}^{-1}\mathbf{R}_{21}$.
- The vectors $\mathbf{b}_1, \ldots, \mathbf{b}_s$ are the eigenvectors of the $q_2 \times q_2$ matrix $\mathbf{E}_2 = \mathbf{R}_{22}^{-1}\mathbf{R}_{21}\mathbf{R}_{11}^{-1}\mathbf{R}_{12}$.
- The canonical correlations $R_1, R_2, \ldots, R_s$ are the square roots of the (nonzero) eigenvalues of either $\mathbf{E}_1$ or $\mathbf{E}_2$.

Interpreting the Canonical Variables and Correlations

- The canonical correlations $R_1, R_2, \ldots, R_s$ represent the associations between the set of x's and the set of y's after the within-set correlations have been removed.
- Canonical variables are typically somewhat artificial, being combinations of possibly disparate variables.
- Thus they do not typically have meaningful units of measurement.
- It is common to *standardize* all the variables before performing the CCA.
- We may interpret the coefficients of the canonical variables similarly to how we interpret the coefficients of principal components.
- Understanding which variables "load heavily" on the various u_i's and v_i's can help us describe the associations between the sets of variables.

Other Facts About CCA

- There is a relationship between multiple discriminant function analysis and CCA.
- Suppose $\mathbf{X}$ is a data matrix with several variables and $\mathbf{G}$ is a matrix of indicators assigning each individual to one of several groups.
- Then if we perform a CCA to investigate the association between $\mathbf{X}$ and $\mathbf{G}$, we obtain the linear discriminant functions as the result (Mardia et al., 1979).
- The i-th **squared** canonical correlation is the proportion of the variance of u_i explained by $y_1, \ldots, y_{q_2}$.
- It is also the proportion of the variance of v_i explained by $x_1, \ldots, x_{q_1}$.
- The largest **squared** canonical correlation, R_1^2, is sometimes used to measure "set overlap."

Inference in CCA

- It may be of interest to formally test whether the canonical correlations are significantly different from zero.
- Problems 8.3 and 8.4 of the textbook outline (likelihood-ratio-based) χ^2 tests proposed by Bartlett.
- The first of these tests H_0 : All (population) canonical correlations are zero vs. H_a : At least one canonical correlation significantly differs from zero.
- If H_0 is rejected, then Bartlett proposes a sequence of procedures that test whether the second-largest canonical correlation significantly differs from zero, then the third-largest, etc.

Inference in CCA (Continued)

- In R and SAS we can implement a nearly equivalent series of (likelihood-ratio-based) F-tests (due to Rao) that test the null hypothesis that the current (population) canonical correlation and all smaller ones are zero.
- We judge each canonical correlation (taken from largest to smallest) to be significant if its accompanying P-value is small enough.
- Once a nonsignificant P-value is obtained, that canonical correlation (and all smaller ones) are judged not significantly different from zero.
- Note that the overall family significance level of this series of sequential tests cannot easily be determined, so we should use the procedure as a rough guideline.
- This procedure is appropriate for large samples from an approximately multivariate normal population.

Multivariate Regression

- In *multivariate regression* we wish to predict or explain a set of r response (or dependent) variables $Y_1, \ldots, Y_r$ via a set of p predictor (or independent) variables $X_1, \ldots, X_p$.
- For example, the military may have several outcome variables that can be measured for enlistees.
- These outcome variables may be related to predictor variables (such as scores on physical tests and/or intelligence tests) through a multivariate regression model.
- The multivariate regression model extends the multiple regression model to the situation in which there are several different response variables.

The Multivariate Regression Model

- The ordinary *multiple linear regression* model equation can be written in matrix-vector form as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\mathbf{Y}$ and $\boldsymbol{\epsilon}$ are $n \times 1$ vectors, $\mathbf{X}$ is a matrix containing the observed values of the predictor variables (plus a column of 1's), and $\boldsymbol{\beta}$ is a vector containing the regression coefficients.

- The *multivariate linear regression* model equation can be written similarly:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- Here, $\mathbf{Y}$ and $\boldsymbol{\epsilon}$ are $n \times r$ matrices, $\mathbf{X}$ is still an $n \times (p+1)$ matrix containing the observed values of the predictor variables (plus a column of 1's), and $\boldsymbol{\beta}$ is now a $(p+1) \times r$ matrix containing the regression coefficients.

Further Explanation of the Multivariate Regression Model

- The n rows of $\mathbf{Y}$ correspond to the n different individuals.
- The r columns of $\mathbf{Y}$ correspond to the r different response variables.
- Note that the first row of $\boldsymbol{\beta}$ is a row of intercept terms corresponding to the r response variables.
- Then the $(i+1, j)$ entry of $\boldsymbol{\beta}$ measures the marginal effect of the i-th predictor variable on the j-th response variable.

The Multivariate Regression Model Assumptions

- We assume that all of the nr elements of $\boldsymbol{\epsilon}$ have mean 0.
- Any single row of $\boldsymbol{\epsilon}$ has covariance matrix $\boldsymbol{\Sigma}$ (generally non-diagonal).
- This implies that the response variables *within an individual* multivariate observation may be correlated.
- However, we also assume that response values from different individuals are uncorrelated.
- For doing inference about the multivariate regression model, we further assume that each column of $\boldsymbol{\epsilon}$ has a multivariate normal distribution.

Fitting the Multivariate Regression Model

- We can fit the multivariate regression model using least squares, analogously to multiple linear regression.
- The matrix of estimated regression coefficients $\hat{\boldsymbol{\beta}}_{LS}$ is found by:

$$\hat{\boldsymbol{\beta}}_{LS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

- This choice of estimated coefficients $\hat{\boldsymbol{\beta}}_{LS}$ is the value of $\hat{\boldsymbol{\beta}}$ that minimizes $tr[(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})]$.
- From now on, we will typically drop the LS subscript and simply refer to the least-squares estimate of $\boldsymbol{\beta}$ as $\hat{\boldsymbol{\beta}}$.

More on the Fitted Multivariate Regression Model

- Computationally, $\hat{\boldsymbol{\beta}}$ may be found by computing separate least-squares multiple regression equations for each of the r response variables.
- We then combine the resulting vectors of regression estimates into a matrix $\hat{\boldsymbol{\beta}}$.
- The matrix of fitted response values (containing the "predicted" response vectors for the observed individuals) is $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$.
- The matrix of residual values is simply $\mathbf{Y} - \hat{\mathbf{Y}}$.
- The multivariate regression model can be estimated in R with the `lm` function and in SAS with `PROC REG` (or `PROC GLM`).

Inference in the Multivariate Regression Model

- If the error vectors have a multivariate normal distribution, then $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimator of $\boldsymbol{\beta}$ and each column of $\hat{\boldsymbol{\beta}}$ has a multivariate normal sampling distribution.
- We can use these facts to make various inferences about the regression model.
- For example, we may wish to test whether one (or some) of the predictor variables are not related to the set of response variables.
- To test whether the i-th predictor is related to the set of response variables, we test whether the i-th row of $\boldsymbol{\beta}$ equals the zero vector.
- This can be done with a likelihood-ratio test (either a χ^2 test or an F-test).

More Inference in the Multivariate Regression Model

- Furthermore, we may we may wish to test whether a set of several predictor variables is not related to the set of response variables.

- For example, label the predictors as $X_1, X_2, \ldots, X_p$. We can test whether, say, only the first p_1 of the predictors are related to the set of response variables, and the last $p - p_1$ are useless in predicting the set of response variables.

- For this type of test, we can decompose the $\boldsymbol{\beta}$ matrix into 2 pieces:

$$\boldsymbol{\beta} = \left(\frac{\boldsymbol{\beta}_{(1)}}{\boldsymbol{\beta}_{(2)}} \right)$$

  where $\boldsymbol{\beta}_{(1)}$ contains the first $p_1 + 1$ rows of $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_{(2)}$ contains the last $p - p_1$ rows of $\boldsymbol{\beta}$.

- We test $H_0 : \boldsymbol{\beta}_{(2)} = \mathbf{0}$, where $\mathbf{0}$ here is a matrix (the same size as $\boldsymbol{\beta}_{(2)}$) of zeroes.

- Of course, if the predictors we want to test about are not the last few, we simply pick out the appropriate rows of $\boldsymbol{\beta}$ and test whether those rows all equal the zero vector.

Test Statistic for Testing Hypotheses Involving $\boldsymbol{\beta}$

- Whether we want to test that *one* predictor, *some* predictors, or *all* predictors is/are not related to the set of responses, we can use a likelihood ratio approach.

- The test statistic is based on the discrepancy between $\mathbf{E}_{full}$ and $\mathbf{E}_{reduced}$.

- $\mathbf{E}_{full}$ is the matrix of sums of squares and cross products of residuals for the full model (containing all the predictor variables) and $\mathbf{E}_{reduced}$ is that matrix for the reduced model (without the predictor(s) are are testing about).

- Under H_0, for large samples, the test statistic

$$-[n - p - 1 - 0.5(r - p + p_1 + 1)] \ln\left(\frac{|\mathbf{E}_{full}|}{|\mathbf{E}_{reduced}|}\right)$$

has an approximate χ^2 distribution with $r(p - p_1)$ degrees of freedom, so a χ^2 test can be done.

- A similar test statistic has an approximate F-distribution, so an essentially equivalent F-test will test these hypotheses.

Prediction Ellipsoids in Multivariate Regression

- Suppose we have a new individual whose values of the predictor variables are known but whose values for the response variables are not available (yet).
- A point prediction of $[Y_1, Y_2, \ldots, Y_r]$ for this individual is simply $\mathbf{x}_0^{'}\hat{\boldsymbol{\beta}}$, where $\mathbf{x}_0^{'} = [1, x_{10}, \ldots, x_{p0}]$ contains the known values of the predictor variables for that individual.
- An r-dimensional $100(1-\alpha)\%$ *prediction ellipsoid* can be constructed based on the F-distribution (see Johnson and Wichern, 2002, pp. 395-396 for details).
- These ellipsoids are 2-D ellipses when there are $r = 2$ response variables, and they can be plotted fairly easily in R.

Confidence Ellipsoids in Multivariate Regression

- Also, we may wish to estimate the mean response vector $[E(Y_1), E(Y_2), \ldots, E(Y_r)]$ corresponding to the values $\mathbf{x}_0' = [1, x_{10}, \ldots, x_{p0}]$ of the predictor variables.
- The point estimate of $[E(Y_1), E(Y_2), \ldots, E(Y_r)]$ is again $\mathbf{x}_0'\hat{\boldsymbol{\beta}}$.
- An r-dimensional $100(1-\alpha)\%$ *confidence ellipsoid* for the mean response vector can be constructed based on the F-distribution.
- For a given $\mathbf{x}_0$, the confidence ellipsoid for the mean response vector will always be tighter than the corresponding prediction ellipsoid for the response vector of a new individual.

Checking Model Assumptions in Multivariate Regression

- The model assumptions should be checked in multivariate regression using techniques similar to those used in simple linear regression or multiple linear regression.

- To check the normality of the error terms, a normal Q-Q plot of the residual vectors $\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_r$ for each response variable can be examined.

- For each response variable, the residual vector can be plotted against the vector of fitted values to look for outliers or unusual patterns.

- Transformations of one or more response variables may be tried if violations of the model assumptions are apparent.