A MULTIVARIATE ANALYSIS OF FAMILY DATA

ALLAN DONNER' AND JOHN J. KOVAL


The authors describe the application of multivariate analysis to the problem of estimating intra-family correlations and testing them for statistical significance. This is illustrated by a re-analysis of survey data collected by Miall and Oldham (Clin Sci 1958;17:459-87) on the familial aggregation of blood pressure. The multivariate analysis provides collective tests of significance for parent-child and child-child correlations based on likelihood ratio theory, and maximum likelihood estimates for individual intra-family correlations.

The analysis of data arising from family studies of continuous attributes, such as height, life span and blood pressure, has often been based on the computation of simple correlation and regression coefficients (1-3). Although this approach is satisfactory when a single familial relationship is being investigated, it is inefficient when applied to the investigation of several intra-family relationships simultaneously. Nonetheless, this is the aim of most studies in which information is collected on each of several family members, for example, the mother, father and one or more children. The purpose of this paper is to show how multivariate analysis can be used to efficiently estimate intra-family correlations and test them for statistical significance. This is done by applying the appropriate techniques to a well known data set published by Miall and Oldham (4) that arose from their classic study of the familial aggregation of blood pressure. The methodology itself is based on the underlying assumption that measurements within a family follow a multivariate normal distribution and provides the user with maximum likelihood estimates and likelihood ratio tests for sex-specific parent-child and child-child correlations. The computations are done by a portable Fortran program. (The latter is available from the authors at a nominal cost.)

MATERIALS AND METHODS

The purpose of the Miall-Oldham study (4) was to investigate the familial resemblance of blood pressure in a group of subjects representative of the general population from which they were drawn. This was done by selecting at random 250 individuals from the population of the Rhonnda-Fach mining valley in South Wales as "propositi" for the study. Attempts were made to contact all first degree relatives of these individuals living...
within a certain distance from the valley, and blood pressures were then recorded on each member of the so-constructed families. The dependence of the blood pressure score of a propositus upon the scores of his relatives was investigated using regression analysis.

Our intention here is to perform a multivariate re-analysis of the Miall-Oldham data, mainly for illustrative purposes. These data cannot be strictly regarded as information collected on a "random sample" of families in the Rhonnda-Fach valley. For example, an individual propositus found to be deceased would be replaced, while deceased relatives, of course, could not be replaced. As Miall and Oldham point out, this could selectively lead to lower blood pressures among relatives than propositi. Furthermore, in an ordinary random sample, spouses of propositi would be included. However, the Miall-Oldham survey was carefully executed, and the data are well known and published. We have therefore decided to use their readings on diastolic blood pressure (rounded by Miall and Oldham to the nearest 5 or 10 mm mark below the actual reading) to illustrate methodology that is relevant to the investigation of family resemblance in general. For this re-analysis, we have grouped together members of a "family," making no distinctions between "propositus" and "relative." Instead, we have identified each family member as either a mother, father, son or daughter, omitting data on the youngest generation in those families having three generations represented, and omitting data on single-generation families entirely. This leads to information on 216 families, of which 122 contain information on only one parent.

Underlying model

Let

$$X_i = (X_{im}, X_{if}, X_{id_1}, \ldots, X_{id_t}, X_{is_1}, \ldots, X_{is_t})$$

denote measurements on the ith family, $i = 1, 2, \ldots, N$, where $X_{im}$ is the mother's score, $X_{if}$ the father's score, $X_{id_1}, \ldots, X_{id_t}$ are the scores on the $D_t$ daughters, and $X_{is_1}, \ldots, X_{is_t}$ are the scores on the $S_t$ sons, respectively. We assume that the measurements within a family follow a multivariate normal distribution with the following parametric structure:

- $\mu_M =$ mean mother score
- $\mu_F =$ mean father score
- $\mu_D =$ mean daughter score
- $\mu_S =$ mean son score
- $\sigma_M^2 =$ variance of the mother score
- $\sigma_F^2 =$ variance of the father score
- $\sigma_D^2 =$ variance of the daughter score
- $\sigma_S^2 =$ variance of the son score
- $\rho_{MF} =$ mother-father correlation
- $\rho_{MD} =$ mother-daughter correlation
- $\rho_{MS} =$ mother-son correlation
- $\rho_{FD} =$ father-daughter correlation
- $\rho_{FS} =$ father-son correlation
- $\rho_{DS} =$ daughter-son correlation
- $\rho_{DS} =$ daughter-daughter correlation
- $\rho_{SS} =$ son-son correlation

The model assumes that these parameters are constant among the families in the population under study and in particular that they are independent of family size. The primary aim of our re-analysis is to make inferences about the eight correlation parameters, although in some studies interest may focus on estimation of the means and variances as well.
MULTIVARIATE ANALYSIS OF FAMILY DATA

Procedure

The model may be written formally as:

\[ X_i \sim N(\mu_i, \Sigma_i), \ i = 1, 2, \ldots, N, \]

where \( N \) denotes the multivariate normal density function,

\[ \mu_i = (\mu_M, \mu_F, \mu_D, \mu_P, \mu_{SD}, \mu_{SP}, \mu_{DP}, \mu_{FP}), \]

and \( \Sigma \) denotes the variance-covariance matrix of \( X_i \), with structure specified as above.

Let \( (X_1, X_2, \ldots, X_N) \) denote a sample from the model above. The likelihood for the \( i \)th family with two parents, \( S \) sons and \( D \) daughters, \( i = 1, 2, \ldots, N \), is given by

\[
2 \prod_{i=1}^{N} l(\mu_i, \Sigma_i | X_i) = (2\pi)^{-1/2} (\Sigma_i + X_i^T \Sigma_i X_i)^{1/2} \exp\{-1/2(X_i - \mu_i)^T \Sigma_i^{-1}(X_i - \mu_i)\}. \tag{2}
\]

The likelihood of the entire sample is then given by the product

\[ L = \prod_{i=1}^{N} l(\mu_i, \Sigma_i | X_i). \]

Note that the likelihood component for a particular family depends on the structure of the family, i.e., the number of daughters and the number of sons.

In the case of a variable number of sons and daughters per family, iterative methods are required to compute the maximum likelihood estimates (MLEs) of the eight correlation parameters, i.e., the values of the correlations that minimize \(-2 \log L\). One method is to directly minimize \(-2 \log L\) with respect to all parameters in the model, with the exception of \( \mu_M \) and \( \mu_F \), whose MLEs are given respectively, by

\[ \bar{X}_M = \frac{\sum_{i=1}^{N} X_{im}}{N} \]

and

\[ \bar{X}_F = \frac{\sum_{i=1}^{N} X_{if}}{N}. \]

For this purpose, we used the ZXMIN subroutine from the IMSL library (5) on a Cyber 73 computer. The method of Atwood and Foster (6) was used to handle the complication that all 14 variables are bounded. Our program prints out the MLEs of all 16 parameters in the model, and for each of the eight correlation parameters, provides a likelihood ratio \( \chi^2 \) statistic for testing whether the estimate of the parameter differs significantly from zero. This test statistic is computed by a) setting the value of the parameter in question equal to 0 in \( L \), b) minimizing the resulting expression for \(-2 \log L\) with respect to all remaining variables and c) subtracting this minimum from the minimum value of \(-2 \log L\) as computed over all variables in the model. The resulting test statistic has an approximate chi-square distribution with one degree of freedom under the null hypothesis that the true value of the parameter is equal to zero (7).

Large-sample 95 per cent confidence limits may also be obtained, based on the matrix of values of second derivatives of the logarithm of the likelihood function evaluated at the MLEs. Since the diagonal elements of the inverse of this matrix are asymptotic variances of the MLEs (7), the assumption of normality of the distribution of the estimators together with the asymptotic variances provide large sample confidence intervals.

Results

Miall and Oldham used polynomial regression methods to adjust their blood pressure data for age and sex. In our reanalysis,
adjustment for age only is required, and we have used the standard-score method, now widely used in epidemiologic research to control for categorical confounding factors. For each individual in the age groups 5–14, 15–24, 25–34, 35–44, 45–54, 55–64, 65–74 and 75+ years, the score \( Z = (X - M)/S \) was computed, where \( M \) and \( S \) are the mean and standard deviation, respectively, of all scores in the given age group. This transformation ensures that differences in means and variances are removed among the eight sub-classes.

A logical first step in our re-analysis was to group the genetic relationships as either inter-generational or intra-generational, and to perform overall significance tests in each group. Thus, the following null hypothesis were tested:

\[
H_{01}: \rho_{MD} = \rho_{MS} = \rho_{FD} = \rho_{FS} = 0 \\
H_{02}: \rho_{DD} = \rho_{SS} = \rho_{DS} = 0
\]

Our program can be used to provide likelihood ratio tests for these hypotheses, i.e., multiple degree of freedom tests of significance based on the chi-square distribution that have optimal properties in large samples (7). The likelihood ratio chi-square statistics for \( H_{01} \) and \( H_{02} \) are 17.4 and 31.9, respectively, with four and three degrees of freedom. Thus, it can be reasonably concluded at any conventional level of significance that there exists a familial aggregation of diastolic blood pressure a) between parents and their children and b) among siblings. These results are consistent with those obtained by Miall and Oldham.

It may also be of interest to conduct separate tests of significance for mother and offspring and for father and offspring. This suggests the following null hypotheses:

\[
H_{03}: \rho_{MD} = \rho_{MS} = 0 \\
H_{04}: \rho_{FD} = \rho_{FS} = 0
\]

The likelihood ratio chi-squares for \( H_{03} \) and \( H_{04} \) each have two degrees of freedom and are given by 14.32 and 0.94, respectively. Thus, the mother-child aggregation of blood pressure is statistically significant, while the father-child aggregation is not.

To assess the strength of individual intra-family correlations, point and interval estimates are required. Table 1 supplies maximum likelihood estimates and corresponding 95 per cent confidence intervals for each of these correlations, based on the theory described above. These intervals are only approximate, but should be satisfactory in studies of reasonably large size (>100 families).

For the sake of comparison, we also present in table 1 the "usual" estimates of the eight correlation parameters, i.e., estimates which ignore information on family members not directly involved in the correlation. For \( \rho_{MF} \), we have simply calculated the mother-father product-moment correlation. For each of \( \rho_{MD}, \rho_{MS}, \rho_{FD} \) and \( \rho_{FS} \), we have calculated the "pair-wise" estimator, obtained by repeating each child score with the corresponding parent score, and computing the product-moment correlation over all such pairs. Rosner et al. (8) have shown that this estimator, which requires multiple counting of a given parent score, is the maximum likelihood estimator in the case of families with a single parent and a fixed number of children. For the sib-sib correlations \( \rho_{SS} \) and \( \rho_{MM} \), we have computed standard intra-class correlations as the usual estimates, based on a one-way analysis of variance of blood pressure Z-score grouped by family (9). Estimation of the remaining parameter, \( \rho_{DS} \), has been given little attention in the literature, but a reasonable procedure would seem to be the computation of the product-moment correlation over all possible brother-sister pairs (with each pair of observations being counted twice in order to avoid the necessity of designating one member of a pair as X and the other as Y). This estimator is the maximum likelihood estimator of \( \rho_{DS} \) in the case of a fixed number of brothers and sisters per family (12).
MULTIVARIATE ANALYSIS OF FAMILY DATA

Table 1

Maximum likelihood estimates, "usual" estimates, and approximate 95 per cent confidence intervals for eight intra-family correlations in the Miall-Oldham study (4)

<table>
<thead>
<tr>
<th>Correlation parameter</th>
<th>Maximum likelihood estimate*</th>
<th>Usual estimate†</th>
<th>95% confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho_{MM}$</td>
<td>0.254</td>
<td>0.208</td>
<td>(0.17, 0.34)</td>
</tr>
<tr>
<td>$\rho_{MF}$</td>
<td>0.142</td>
<td>0.116</td>
<td>(0.06, 0.22)</td>
</tr>
<tr>
<td>$\rho_{FF}$</td>
<td>0.004</td>
<td>-0.042</td>
<td>(-0.08, 0.08)</td>
</tr>
<tr>
<td>$\rho_{SS}$</td>
<td>0.072</td>
<td>0.135</td>
<td>(-0.03, 0.17)</td>
</tr>
<tr>
<td>$\rho_{MS}$</td>
<td>0.231</td>
<td>0.226</td>
<td>(0.12, 0.33)</td>
</tr>
<tr>
<td>$\rho_{FS}$</td>
<td>0.125</td>
<td>0.162</td>
<td>(0.02, 0.23)</td>
</tr>
<tr>
<td>$\rho_{OS}$</td>
<td>-0.102</td>
<td>-0.008</td>
<td>(-0.24, 0.04)</td>
</tr>
</tbody>
</table>

* From multivariate analysis using information on all family members.
† See text.

The results in table 1 show that the two mother-child correlations and the three sib-sib correlations are statistically significant, as judged by the 95 per cent confidence interval, while the father-child correlations and the mother-father correlation are not statistically significant (single degree of freedom likelihood ratio tests may also be used for this purpose but are not shown here. See Cox and Hinkley (7) for a discussion of the relative merits of these two approaches to hypothesis-testing).

It is interesting to note that in three out of eight cases, the usual estimate and the maximum likelihood estimate differ by 0.05 or more, although both estimates are always contained in the 95 per cent confidence interval. The lower the degree of inter-correlation among the family members, the smaller one would expect these differences to be.

DISCUSSION

We have used the Miall-Oldham data (4) to illustrate a general multivariate approach to the analysis of family data. One advantage of this approach is that it is statistically more efficient than the usual approach of dealing separately with each intra-family relationship. This is because information on all available family members is used to estimate each correlation parameter. Moreover, collective significance-testing is preferable to repeated pairwise testing because it preserves the proper $\alpha$-level for the overall procedure, i.e., it provides a strategy for dealing with the "multiple comparisons" problem. A further advantage of the maximum likelihood approach is that it simplifies estimation and significance-testing problems that often arise with other procedures as a result of the fact that the number of siblings per family tends to be variable. For example, the testing of a parent-child correlation for statistical significance in the variable family size case is not routine (11) and no established method exists as of yet to test a brother-sister correlation for significance. A final advantage concerns the treatment of missing data, e.g., when observations on a subset of fathers are missing. The likelihood function in this case is evaluated as a function of the non-missing scores, and the remainder of the likelihood function is disregarded. Thus the resulting estimates are true maximum likelihood estimates, using all available information. This approach to missing data is particularly useful when the proportion of missing values on one or more family members is large, as is the case here, where the re-constructed families frequently have only one parent. Thus for example, in table 1 the 95 per cent confidence interval for $\rho_{MF}$ obtained
using all available information is \((-0.24, 0.04)\) compared to the much wider interval \((-0.21, 0.20)\) which would be obtained if standard methodology (9) were applied to the scores of the 94 families having complete information on both parents.

The model we have described assumes that the values of the eight correlation parameters are the same in different families. This might not be the case, for example, if twins were present in some families but not in others, or if full sibs reared together were present in some families while half sibs reared apart were present in others. Care must be taken in using the above model that genetic relationships among family members are the same over all families.

A further assumption of the model is that observations within a family follow a multivariate normal distribution. This will often be an appropriate assumption if the attribute under study is continuous. However, it would be of interest to investigate the effect of departures from normality on the efficiency of the estimation procedures presented.

The methodology discussed here necessarily requires the use of a computer. The required software, however, is available from the authors at nominal charge (not including the IMSL subroutines). Our program is oriented toward data having scores for each family on one or both parents and any number of sons or daughters, but can be easily adapted to handle the case of data collected on siblings only. It can also handle inferences conducted on one of several "reduced models." Suppose, for example, the investigator has prior knowledge that the three sib-sib correlations are all of the same magnitude, i.e. that \(\rho_{SS} = \rho_{DD} = \rho_{DS} = \rho_0\). Then all estimation and hypothesis-testing procedures concerning \(\rho_0\) and the five remaining correlation parameters may be conducted under this restriction. Similarly, up to four other reduced models may be specified by the user, or, if desired, any combination of these models.

Finally, the program may be used, through repeated calculation of the log likelihood under different models, to test whether the values of the correlation parameters defined in model 1 are the same in two distinct populations. Such a test might be of interest, for example, in comparing blood pressure correlations in a sample of families having full siblings reared apart and a sample of families having full siblings reared together.

Extension of this methodology based, for example, on theory developed by Lange et al. (11), is needed to deal with more complex family relationships, such as those involving grandparents and/or cousins. Research is also needed on the evaluation of the small-sample properties of the procedures advocated here. It is hoped that these procedures will ultimately lead to richer and more efficient analyses of family data.

References