

Donald F. Morrison, University of Pennsylvania

Dr. Nestel has touched on the problem of the loss of subjects from the original study cohort or panel. The subjects with incomplete data vectors can still be used to estimate population means and covariance matrices: maximum likelihood estimates based on the multivariate normal model have been given and discussed by a number of investigators (see, in particular, Anderson [1]). When all lost subjects are never interviewed again, the data matrix with its blocks of rows ordered from those subjects with complete data to those who dropped from the study at the earliest time is called monotonic, and the estimates of the parameters can be obtained explicitly. However, when the number of complete subjects is very small and the correlation among the responses is low some recent results [8] have indicated that the multivariate maximum likelihood estimates are less efficient than those obtained from the complete data alone.

Under the same assumptions of the multivariate normal model and a monotone data matrix Bhargava [2] developed generalized likelihood ratio tests for hypotheses on the mean vector, covariance matrix, and the multivariate general linear model. Percentage points of the mean vector test statistics have been computed by Mr. Dinesh Bhoj for a forthcoming doctoral dissertation [3].

I would be curious to know what use Dr. Nestel plans to make of information on such exogenous factors as changes in the economy, labor supply, employment practices, the inauguration of training programs, and other environmental conditions that will influence his subjects' experiences in the labor market. Presumably some of those quantities would have to be incorporated as concomitant variables in the analysis of changes and trends in his data.

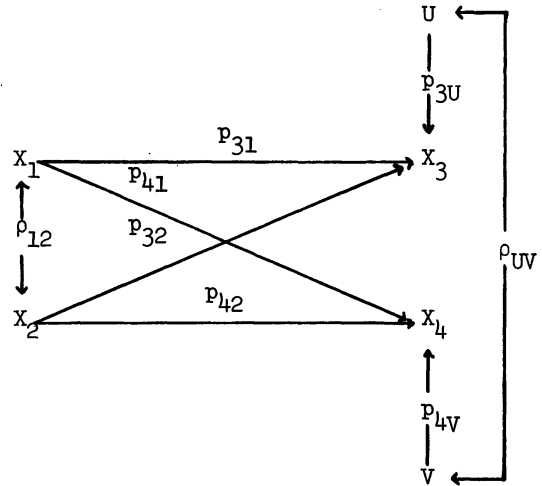
Now let us turn to the correlation models of Drs. Pelz, Faith, and Land. Their investigations suggest to me the following questions:

(1) Are their analyses of longitudinal studies solely concerned with dependence structure? As in the cases of growth functions and repeated measurements experiments, would not the more important conclusions be derived from the mean, or expectation vector, structure?

(2) What are the motivations and consequences of their models for correlation structure? Are their purposes the more parsimonious explanation of the dependence structure, the estimation of components of variance attributable to true scores and measurement error, or the development of hypothesis tests on the means or covariance structure which take advantage of the special properties of patterned covariance matrices?

It is my opinion that more results on sampling properties and hypothesis tests must be developed for the path coefficient models

before they can be applied with confidence in the analysis of data. One of the simplest path coefficient models is amenable to statistical inferences about its coefficients: it is the four-variable model discussed by Heise [5] with the following path diagram:



If we assume that the X_i have a covariance matrix with general element σ_{ij} , the path coefficients are equal to

$$\tilde{B} = \begin{bmatrix} p_{31} & p_{41} \\ p_{32} & p_{42} \end{bmatrix}$$

$$= \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{12}^2} \begin{bmatrix} \sigma_{22}\sigma_{13} - \sigma_{12}\sigma_{23} & \sigma_{22}\sigma_{14} - \sigma_{12}\sigma_{24} \\ \sigma_{11}\sigma_{23} - \sigma_{12}\sigma_{13} & \sigma_{11}\sigma_{24} - \sigma_{12}\sigma_{14} \end{bmatrix},$$

or the matrix of regression coefficients of X_3 and X_4 on the pair of variates X_1 and X_2 . When the X_i have the quadrivariate normal distribution we can test all hypotheses of the kind

$$H_0: \underline{c}' \tilde{B} \underline{a} = 0,$$

for all non-null two-component vectors \underline{a} , \underline{c} , by a variant of the Scheffe'-Roy simultaneous confidence intervals for all bilinear forms of the regression coefficients [7, Sec. 3.6]. Such hypotheses would include

$$H_0: p_{31} = p_{42}, \quad H_0: p_{41} = 0, \quad H_0: p_{32} = 0,$$

and other hypotheses of equality or significance

of the coefficients. This approach errs, of course, on the conservative side, but conversely it provides an exact control over the Type I error probability for tests of all hypotheses generated by the bilinear compounds.

Most path models do not possess the simple one-to-one connection with regression coefficients of our illustration, and this appears to be the case for the Pelz-Faith bivariate model. However, estimation and testing for the general model might be handled by an approach due to Jöreskog [6].

Because the Pelz and Faith investigation has been principally concerned with the case of high autocorrelation, some simplification could have been achieved by beginning with the autoregressive process with unit coefficient. Then the t th realization is equal to the sum of t independent variates with common variance σ^2 . The covariance matrix has the Wiener pattern

$$\sigma^2 \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 2 & \cdots & 2 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 2 & \cdots & p \end{bmatrix}$$

The y_t process can be developed by the same additive model used by the authors, and its cross correlations can be computed in a less involved manner. The cross-correlation function is equal to zero for certain values of the x and y variates' indices. We note, of course, that the processes are no longer stationary.

Let us begin our discussion of Dr. Land's paper with the hypothesis

$$H_0: p_{32} = p_{21}$$

of constant path coefficients from x_1 to x_2 and x_2 to x_3 . We shall restrict our attention to the case of no measurement error ($p_{xe} = 0$) and unit random shock coefficients

($p_{xu_2} = p_{xu_3} = 1$). An alternative test of H_0 can be constructed by the generalized likelihood ratio principle. Following the author's model, we assume that the random shocks u_{t1} , u_{t2} are independently and normally distributed with zero means and common variance σ^2 for $t = 1, \dots, N$. The likelihoods under H_0 and the general alternative of unequal path coefficients can be obtained as products of the univariate normal likelihoods of the u_{tj} variates. The likelihood ratio statistic λ is merely a power of the estimates of the common variance under the null and alternative hypothesis, or

$$\lambda^{2/N} = \frac{\hat{\sigma}_\Omega^2}{\hat{\sigma}_w^2}$$

$$= 1 + \frac{\frac{(\sum x_{t1}x_{t2} + \sum x_{t2}x_{t3})^2}{\sum x_{t1}^2 + \sum x_{t2}^2} - \frac{(\sum x_{t1}x_{t2})^2}{\sum x_{t1}^2} - \frac{(\sum x_{t2}x_{t3})^2}{\sum x_{t2}^2}}{\sum x_{t2}^2 + \sum x_{t3}^2 - \frac{(\sum x_{t1}x_{t2} + \sum x_{t2}x_{t3})^2}{\sum x_{t1}^2 + \sum x_{t2}^2}}$$

Under H_0 - $N \ln (\hat{\sigma}_\Omega^2 / \hat{\sigma}_w^2)$ is asymptotically distributed as a central chi-squared variate with one degree of freedom. Modification of the statistic for unknown, non-zero means of the x variates is evident. The assumption of general random shock coefficients leads to unequal variances in the univariate likelihoods and a cubic equation for the maximum likelihood estimate of the path coefficient under H_0 .

It is unclear to me how the equations (4.3) could be linearized by a logarithmic transformation. An alternative approach would consist of writing the covariance matrix of the measurement variates as

$$\Sigma = \alpha^2 \tilde{P}_1 + \beta^2 \tilde{P}_2$$

where \tilde{P}_1 , \tilde{P}_2 are the respective Markov covariance matrices of the true scores and measurement errors. When \tilde{P}_2 is the identity matrix an estimation technique due to Jöreskog [6] could be employed. More general forms of \tilde{P}_2 would introduce identification problems. If these can be resolved by appropriate constraints on the parameters the estimates might be obtained through an iterative solution of the almost certainly nonlinear maximum likelihood equations. Finally, we note that the equal-correlation matrix might be chosen as an alternative to \tilde{P}_2 : such a matrix might be justified from a variance-components model for the observation variates.

In connection with Dr. Land's models and results we note that a family of tests for the degree of dependence of ordered multinormal variates has been given by Gabriel [4].

References

- [1] Anderson, T. W., "Maximum Likelihood Estimates for a Multivariate Normal Distribution when some Observations are Missing," Journal of the American Statistical Association, 52 (June 1957), 200-203.

- [2] Bhargava, R., "Multivariate Tests of Hypotheses with Incomplete Data," Technical Report No. 3, Applied Mathematics and Statistics Laboratories, Stanford University, 1962 (mimeographed).
- [3] Bhoj, D., Personal communication.
- [4] Gabriel, K. R., "Ante-Dependence Analysis of an Ordered Set of Variables," Annals of Mathematical Statistics, Vol. 33 (March 1962), 201-212.
- [5] Heise, D. R., "Causal Inference from Panel Data," in E. F. Borgatta and G. W. Bohrnstedt, eds., Sociological Methodology: 1970, San Francisco: Jossey-Bass, 1970.
- [6] Jöreskog, K. G., "A General Method for Analysis of Covariance Matrices," Biometrika, 57 (August 1970), 239-251.
- [7] Morrison, D. F., Multivariate Statistical Methods, New York: McGraw Hill Book Company, 1967.
- [8] Morrison, D. F., "Expectations and Variances of Maximum Likelihood Estimates of the Multivariate Normal Distribution Parameters with Missing Data." To be published in the Journal of the American Statistical Association, September 1971.