faculteit psychologie

# Biometric decomposition of phenotypic means in human samples

Conor V. Dolan

UNIVERSITEIT VAN AMSTERDAM

**Summary.**

Quantitative genetic studies of polygenetic human traits focus on genetic and environmental contributions to phenotypic differences between individuals in one and the same well defined population, i.e. on contribution to within group variation. The genetic contribution is attributable to differences between individuals in the genetic constitution or genotype. The environmental contribution is due to differences in the exposure to relevant environmental influences. By means of a number of well established biometrical designs, the contribution of these sources of within group variation can be estimated as variance components contributing to the total phenotypic variance observed with respect to a given phenotype. The identification of these variance components is based on the known percentages of shared alleles between genetically related individuals (e.g. siblings, parents and their offspring, etc.) and on the known history of co-habitation (e.g. sibling growing up together, offspring separated from their biological parents at birth, etc.).

Between group variation arises from differences between phenotypic means observed between well defined groups and is attributable to differences in genetic and environmental influences characterizing the groups in question. Unlike within group variance, between group variance observed in human samples cannot readily be decomposed into genetic and environmental components. The problem is that generally the relevant information to identify genetic and environmental sources of between group variation is absent. Such information pertains to the relevant allele frequencies in each group and the nature of environmental influences at play. In practice, however, very little is known about the precise nature of the environmental influences and genetic influences contributing to continuously varying, polygenetic human traits, such as IQ or weight.

The decomposition of differences in phenotypic means can be undertaken by making certain assumptions concerning the relationship between the sources of within group and between group variation. If one is willing to assume that the within group sources of variation and the between group sources of variation are identical, the between group components of variance are identifiable and may be estimated. The objective of the present dissertation is to investigate various models that can be used to achieve a between group decomposition of variation on the basis of this assumption.

The point of departure for these models is the notion that the groups in questions arise through a process of selection. For instance, groups may be a parent population and a sub-population derived by a selection process from the parent population, or two or more derived sub-populations. Two types of selection can be distinguished: selection based on the observed variable (i.e. the phenotype) and selection based on the latent variables (genetic and environmental variables). The effects of these forms of selection may be investigated using the Pearson-Lawley selection rules and lead to explicit and testable multi-group models that account for differences within the groups and differences between the group by a common set of parameters. The assumption of common causation (i.e. of between and within group variance) thus finds a very precise and testable expression in these models.

As explained in chapter 1 of this dissertation, the work of Meredith (1964), Jöreskog (1971) and Sörbom (1974) concerning the effects of latent selection in the common factor model has resulted in multi-group covariance structure models. A considerable advantage of this approach is that the selection process remains implicit: precise knowledge concerning the latent selection process, apart from the fact that it is indeed latent, is not required. In chapter 3 this approach was explicated and applied to twin data to model individual and gender related mean differences in blood pressure.

In chapter 4, the latent selection was applied to repeatedly measures of weight in a sample of female monozygotic and dizygotic twins. Taking into account the serial dependency of the observations, the developmental process underlying the means trend and the stability of the individual differences is modelled in terms of the changing environmental and genetic influences. From the perspective of selection, the observations at each occasion may be viewed as arising from a process of latent selection on the genetic and environmental variables.

Phenotypic selection again leads to a multi-group covariance model with a common set of parameters for the within and between group variation. As explained in chapter 1 this type of selection was considered by Thomson (1945) and more recently by Muthén (1989a), both in the context of the common factor model (e.g. Lawley and Maxwell, 1971). In chapter 5, this form of selection was applied to multivariate twin data relating to school subjects (English, mathematics etc.). Whereas latent selection leads to a model in which the selection process remains implicit, the model based on phenotypic selection requires exact knowledge concerning selection. In chapter 5, this knowledge was available because the selection was carried out by selecting explicitly individuals for the sample on the basis of a minimum scores on the average of the phenotypic variable.

The general approach taken in this dissertation leads to testable models concerning between and within group sources of variation. As mentioned these models are based on the assumption of common causation of differences between individual and between groups. In the case of latent selection, the plausibility of this assumption finds support in the goodness-of-fit statistics of the models. That is, if the models fit, the assumption is supported. The notion of selection provides a useful and, judging by the confusion apparent in the literature, sorely needed framework for interpreting statistics relating to individual differences and mean differences.

Because the models presented herein are instances of multi-group covariance structure models, a computer program was written as part of this research project to fit covariance structure models. This program has the dual advantage over commercial programs of allowing the user to place a wide variety of constraints on estimated parameters and to fit problem that are too large for such programs. The development of this program has resulted in two chapters that have been included in this dissertation as appendices. Chapter A1 contains a description of the program and a number of illustrations (chapter 4 contains an additional application). Chapter A2 concerns the calculation of standard errors of parameters when estimation is carried out by minimization of the log-likelihood ratio function. Generally such standard errors are obtained by calculating the Fisher's information matrix. In our program, a finite difference approximation of Hessian was used. Both methods are found to be satisfactory in a comparison involving simulated and real data.

In a final chapter, which is also included as an appendix, two recursive methods of calculating factor scores in longitudinal models (a first order autoregression featured as a test case) developed in the field of state-space modelling were compared to the regression method developed in the field of factor analysis. The recursive methods are the Kalman Filter (KF) and the Fixed Interval Smoother (FIS) (Brown, 1983). As it turns out the FIS and the regression method are identical. The KF is based on the same least squares loss function, but uses less information. The factor scores calculated using the KF are therefore characterized by a greater mean square error that those obtained by either of the other two methods. The relevance of this work for quantitative genetics lies in the recent interest in the estimation of individual genetic and environmental scores (Boomsma, Molenaar and Dolan, 1991). These scores are estimated using standard techniques for the estimation of factor scores.