This research was supported by National Institutes of Health, Institute of General Medical Sciences Grants GM-70004-01 and GM-12868-08.

THE GROWTH CURVE MODEL APPROACH TO THE STATISTICAL ANALYSIS OF LARGE DATA FILES

by

Gary G. Koch and Bernard G. Greenberg

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 786

November 1971

The Growth Curve Model Approach to the Statistical Analysis of Large Data Files

by

Gary G. Koch and Bernard G. Greenberg

University of North Carolina, Chapel Hill, N. C.

The development of large-scale, high-speed electronic computers has provided mankind with a powerful tool which can be used to manipulate large volumes of data. As a result, more different types of information are being obtained from more individuals by survey groups and governmental agencies today than ever before.

Unfortunately, the analysis of such large data files is not always clear-cut. One reason for this is that many of the standard statistical procedures which are available have been developed either for small data files or for situations in which the underlying variables tend not to interact. However, when there are complex inter-relationships among variables and the underlying research design involves disproportionate sub-class numbers, a number of difficulties arise. Moreover, such questions are further complicated by the extent to which the measurement scales for the underlying variables tend to be qualitative rather than quantitative. Finally, the large size of such files means that computer manipulations of the data will be expensive unless carefully managed. All of these reasons support a need for a different type of approach to statistical methodology for large data sets rather than small ones in which cost can be minimized, computational singularities can be avoided, and the main effects and interactions can be isolated

and interpreted. One possible strategy for accomplishing these objectives is based upon the "growth curve model" for multivariate data.

In growth curve experiments, one has a moderate number of experimental units or subjects (eg., between 50 and 300) who have been assigned to the different treatment combinations of some factorial design (eg., a factorial with diet effects, crowding effects, temperature effects, etc.) There are several dependent variables of interest like total height and weight, height and weight gain; certain aspects of metabolism, some characteristics of behavior, and finally, several related uncontrolled independent variables (covariables) like initial weight, parental height and weight, and certain environmental effects. Data for all dependent variables and certain independent variables are recorded longitudinally in time for moderate number of distinct time points (eg., between 2 and 100). As a result, for such variables, one has a large number of measurements in the sense that there may be 300 subjects and 100 time points which generates 30,000 observations.

A straightforward way of efficiently analyzing this type of data is to use the "growth curve model" approach. This procedure has two phases. First of all, each subject is considered separately and either multiple regression or some other appropriate method is used to summarize the pattern in variation over time for each longitudinal variable; for example, estimates are obtained for the intercept, the linear trend, the quadratic curvature, the cubic or higher order contortions. These estimates and their corresponding estimated variance-covariance matrix represent the output of

the first phase and the input to the second phase. In terms of computer time, it involves exactly one run through the 30,000 case file. The second phase involves applying multivariate analysis of variance to the vectors of estimated parameters from the first phase for the 300 subjects. Hence, the variation in values of these vectors is to be explained in terms of differences among the subjects with respect to factorial design effects and covariables. Actually, a number of computer runs can be practically carried out at this stage, because one is working with a much smaller data file; (eg., in situations where intercept, linear, and quadratic terms suffice, 30,000 observations have been replaced by 3 estimates for each of 300 cells or 900).

Finally, in certain situations where the time points for different subjects are not the same, there is occasionally interest in using a generalized weighted multivariate analysis of variance. However, this is not always required since unbiased estimates are derived from the unweighted approach. In summary, this two phase methodology is straightforward and direct, is reasonably inexpensive, and leads to estimated parameters and tests of significance which can be clearly interpreted.

The previous discussion has indicated that growth curve experiments can be analyzed in an efficient and systematic manner in spite of the fact that they involve a large number of observations. The question of interest now is whether the same approach can be suitably adapted to other types of large data files. In other words, can such data be

approached in a two phase fashion in which

1. The data file is partitioned into a moderate number of distinct subsets analogous to the subjects of the growth curve experiment; and separate analyses are performed within each subset which lead to a set of estimated parameters. In this context, the data within each subset are transformed into a set of effects and their estimated variances and covariances.
2. The set of estimates for the respective subsets are then synthesized together by suitable methods of multivariate analysis.

With these remarks in mind, it is apparent that the "growth curve approach" can be applied in spirit to almost any large scale research project. The principal issue, in essence, is the extent to which the data can be partitioned into meaningful sub-groups and the extent to which data within the respective sub-groups can be summarized in terms of a simplified set of estimated parameters.

The definition of a sub-group partition for some given data set can usually be based on a classification of variables into types which are descriptive of their roles in the statistical analysis. The first type of variable is the dependent variable. These are the specific entities in terms of which some conclusive statement is to be made. Here, we shall assume that there is a single dependent variable of interest, although the methodology is readily extended to the case of multivariate dependent variables.

The second type of variable are blocking variables which are descriptive of the variety of sub-groups to which a subject belongs. These variables play a role analogous to the differential experimental conditions which are assigned to the experimental units in the standard growth curve experiment. Also, they define the totality of subsets of data within which the first phase of analysis is to be undertaken. Usually, one will interpret the blocking variables to define sub-groups in a completely cross-classified sense; i.e., for every combination of distinct levels of each blocking variable, there will correspond a sub-group which will be analyzed in the first phase. However, in many actual situations, this may not be realistically feasible because of disproportionate frequencies of subjects in the cross-classified scheme and, in particular, missing or empty sub-groups. Here, one may use a system of hierarchial sub-groups in which certain blocking variables are used to define sub-groups only within specific combinations of other blocking variables. The mixed hierarchial and cross-classified scheme which arises in this manner can be analyzed by methods similar to those which are used for analogous kinds of classical experimental designs.

Finally, most of this discussion has implied that blocking variables will be used to define sub-group structures. However, in some cases, one may want to have certain blocking variables play the same role as covariables in the second phase of analysis. In other words, the totality of blocking variables may define such a large number of very small sub-groups (even with hierarchial modifications) that little or no confidence can be placed

in the results of the first phase even after synthesis in the second phase. Hence, one uses some of the blocking variables to define sub-groups; and then for each sub-group obtains the mean (or some other appropriate measure.) for the remaining blocking variables. One then defines what will be called modified blocking variables as the set of values for these means across the totality of sub-groups. These modified blocking variables are functions whose domain of definition is the totality of sub-groups (hence, they may be still labeled as blocking variables) and whose range of values are the corresponding sets of within sub-group means. They are distinguished from strict blocking variables in the sense that in the second phase of analysis, they will be treated as covariables which are descriptive of sub-groups whereas the strict blocking variables define the sub-groups. In this context, it is apparent that such modified blocking variables can play a role which may partially explain any interaction among the strict blocking variables.

The third type of variable is the independent variable in terms of which variations in the dependent variable are to be explained. To some extent, the designations of blocking variable and independent variable are interchangeable although the concepts are different as are the roles they play in the growth curve model approach. The principal distinction is a functional one in that the independent variable is used as a basis of a comparison or other relationship involving the dependent variable while the blocking variables define the totality of sub-groups. More specifically, the independent variables are used in the first phase of analysis to explain the variation of the dependent variable while the blocking variable is used in the second phase to interpret and summarize

the patterns of relationship. Thus, at the first phase, the variation of a dependent variable as a function of several independent variables within a specific sub-group is analogous to the changing growth characteristics of a subject in a growth curve experiment over a longitudinal period of time. As explained previously, such data can be transformed into a set of estimated parameters which reflect the effects of the corresponding independent variables by means of multiple regression or some other suitable procedure. Then, such estimates can be further studied with respect to the blocking variables in the second phase by multivariate analysis of variance (or some appropriate analogue). The primary question which remains for a specific data file is the decision as to which variables are to be blocking variables and which ones are to be independent variables.

The principal criterion for classifying variables which will be recommended here is the extent to which a variable interacts with the dependent variable. However, the way in which this criterion is implemented may vary from one application to another. If there exists a set of variables which can be used to define natural sub-groups, and if all of these variables tend not to interact with the other independent variables according to their effects on the dependent variable, but may interact with each other, then these variables can be used as the blocking variables provided that there are only a few (eg., ≤ 8) independent variables which remain and which may interact with the dependent variable. The resulting analyses from the first phase are complex in this situation; but under such assumptions about the sub-groups, all have the same structure. Moreover, in the second phase of analysis, one should find

that only the "constant" term of the regression is affected by the effects of the blocking variable. Hence, the data can be explained in rather simplified terms after the second phase; namely, in terms of the multivariate analysis of variance of the "constant" term showing the effects of the blocking variables and the "average regression equation" across the sub-groups showing the effects of the independent variables. On the other hand, if it turns out that coefficients other than the "constant" term are influenced by the blocking variables, then interpretation becomes difficult because this implies the existence of a possibly complex interaction between the independent variables and the blocking variables. Hence, an alternative approach is of interest.

The basis of the alternative strategy is the principle that "no interaction" is a well-defined concept whereas "existence of interaction" can mean almost anything. Hence, one chooses the independent variables to be those variables which do not interact with one another with respect to prediction of the dependent variable within each of the specific sub-groups defined by the remaining blocking variables. In other words, the association between the dependent variable and the independent variables can be expressed in simple additive terms once the effects of the blocking variables have been removed. For many applications, the number of variables which have this property will be small (eg., less than 10). Hence, most of the remaining variables will be blocking variables. If all of them were used to form sub-groups, there would be too many sub-groups which were too small. Hence, the ones which are believed to be most important (in the sense of

being the source of greatest variation with respect to the second phase) will be used as the strict blocking variables and the remaining ones will be used as modified blocking variables as defined earlier.

Hence, at the first stage, we obtain a set of sub-groups across which there is interaction but within which there is no interaction. One fits the no interaction multiple regression models to each of the sub-groups defined by the blocking variables. Each of these will have a meaningful and simplified interpretation in its own right because of being based on a no interaction model. At the second stage, multivariate analysis of variance will again be applied to the vector of estimated parameters. Here, not only the variation of the constant term will be of interest, but also the variation of each of the main effects. The analysis of the constant term will provide one with a description of the main effects and interactions of the blocking variables with one another with respect to their roles in explaining the dependent variable. The analyses of the main effect coefficients for the respective independent variables will be indicative of their interactions with the blocking variables. In either context, a clear summary of the relationship of the dependent variable to the independent and blocking variables will be forthcoming. This, in principle, completes the discussion of the methodological strategy. However, some statement should be made about deciding which variables are interacting or not and which ones are important sources of variation. For the present, this question will be deferred by noting that there do exist standard stepwise regression programs which provide certain aspects of this type of information. Otherwise, the remaining components of the "growth curve approach" are rather straightforward to apply. Some illustrations of theory and specific application will be given at a later date.

REFERENCES

[1] Allen, D.M. and Grizzle, J.E. [1969]. Analysis of growth and dose response curves. Biometrics 25, 357-82.

[2] Potthoff, R.F. and Roy, S.N. [1964]. A generalized multivariate analysis of variance model useful especially for growth curve problems. Biometrika 51, 122-27.

[3] Roy, S.N., Gnanadesikan, R., Srivastava, J.N. [1971]. Analysis of certain quantitative multiresponse experiments. Pergamon Press, New York.