MULTIPLE LINEAR REGRESSION VIEWPOINTS
November 1980, Volume 10, Number 4

RELATIONSHIPS AMONG PREDICTORS IN LONGITUDINAL DATA: TEMPORAL-SEQUENTIAL ANALYSIS BY REGRESSION - TSAR

Thomas F. Jordan
University of Missouri - St. Louis

In analyses of longitudinal data attention is appropriately drawn to criterion measures of growth. However, it seems appropriate to consider predictor variables, and to do so in a fashion which helps understand their interrelationships. A method of arranging predictors is described which draws on regression analysis, and so uses powerful inferential tests of statistical significance. Examples are given of patterns of predictors educed by representative analyses. Data are drawn from two data sets with attention to several measures in the first 5-6 years of life of a particular cohort, and to demographic data on childhood from several countries of the world, in the second instance.

INTRODUCTION

In analyses employing data across a span of interest in the cycle of development there is a not unreasonable concentration on the criterion measure. This is because study of the criterion is the step which leads to assessment of a hypothesis about a stage of growth. To a lesser extent we concentrate on the factors which explain the criterion data, with a view to understanding the quantitative influences on a particular level of attainment.

Study of predictor variables tends to focus on them as sources of variance, and combines them in the form of statistical interactions as a particular mode of analysis will permit, and as reasoning can subsequently decipher into non-mathematical, operational propositions. We tend to neglect the total array of predictor variables in a set, except as we engage in path analyses. In such studies the goal, however, tends to be pursuit of the criterion and the linkage via regression weights to the dependent measure. In addition, use of path analysis presumes stability of regression weights; also, it has tended to slight inferential tests of significance and ignores statistical interactions.

It is interesting to consider how predictors in longitudinal data sets may relate to each other. We can give each the status of a criterion whose variance may be understood in the constellation of all predictors, with inferential tests of statistical significance. To this end it is helpful to draw on regression analysis, and on multiple linear regression (Bottenberg and Ward, 1963; McNeil, Kelley, and McNeil, 1975) in particular. The latter technique has been particularly useful in analyses of longitudinal data (Jordan, 1980). The origin of this interest emerges from analyzing the data set of the St. Louis Baby Study (Jordan, 1981a), an inquiry into the developmental span beginning at birth and continuing into adolescence. In these analyses predictor variables are typically arrayed from a variety of domains in order to test hypotheses of relative influence on a developmental criterion (Jordan, 1978). Theoretically, predictors are independent of each other, but there may be dependencies due to the conceptual or practical relevance of a given predictor in the presence of another variable for which the case is equally strong (e.g. from another theoretical domain of influences). This equivalence of a claim to significance may be due to the nature of the model we are examining, despite the statistical shortcomings that situation may precipitate. However, we mention that in passing; the greater point is that predictors as a data set may need as much attention as the larger aggregate which includes the criterion.

PROBLEM

The particular topic we address here is the matter of an empirical structure of relationships among predictors based on inferential statistical significance (F-tests) and employing regression analysis because of its utility for longitudinal data. We start by noting that measures gathered by prospective case studies are time specific, and occur in a sequence as a consequence. Our problem is the search for an empirically derived schema which reflects the sequence of predictor variables in their temporal order. We do so without prior commitment to structures which will stress the dependent variable, as in path analysis to which our approach has a superficial similarity in use of diagrams with arrows.

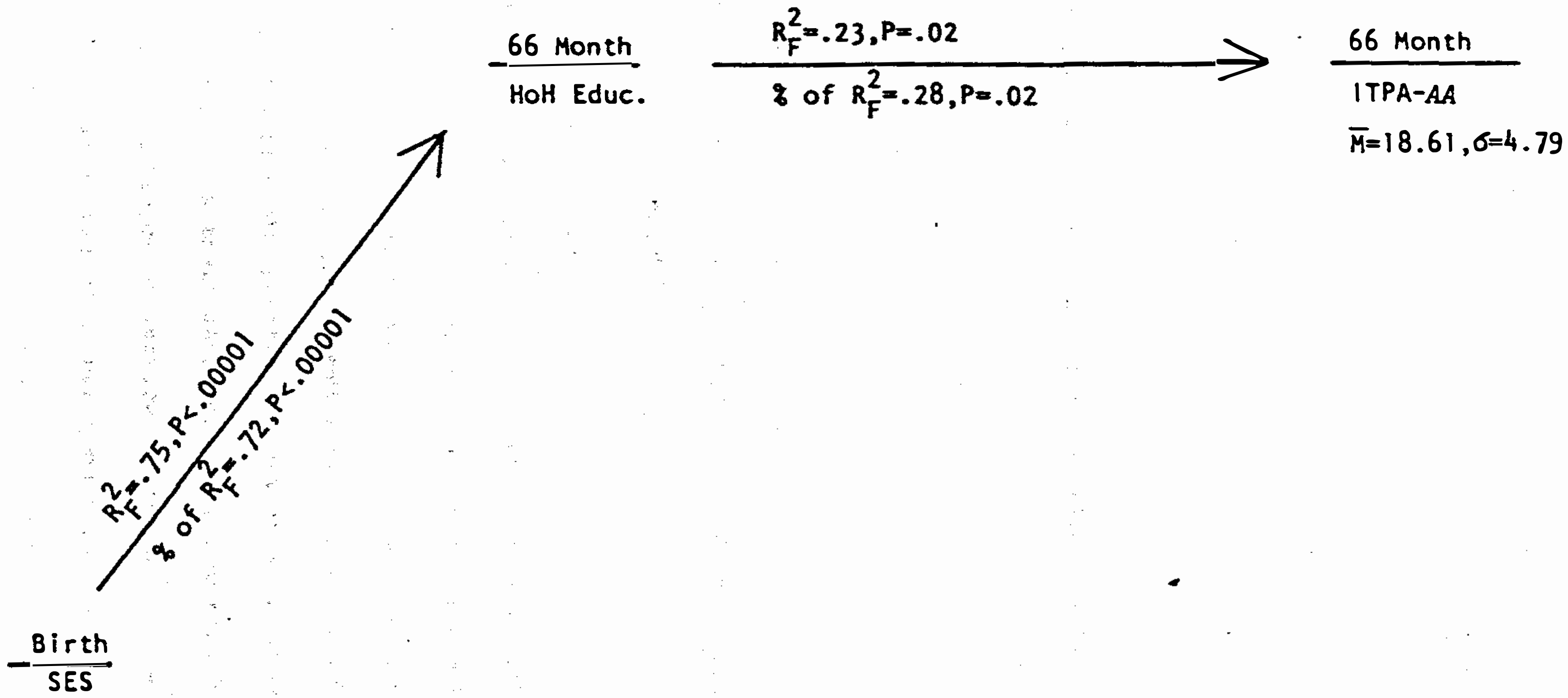
METHOD

In multiple linear regression (McNeil, Kelley, and McNeil, 1975), a regression equation is developed in order to predict a criterion. A critical element is deleted or collapsed, the resulting equation is designated as an alternate mode, and an F value is computed for the loss of predictive efficiency traceable to the altered vector. The basic model may be illustrated as $Y = a_0u + a_1x_1 + a_2x_2 \ldots a_nx_n + e$, where Y = a criterion of continuous or discrete data, u = a unit vector which when multiplied by the weight a_0 yields the regression constant, $a_1a_2 \ldots a_n$ = partial regression weights arrived at by multuple linear regression techniques and calculated to minimize the error sums of squares of prediction (Σe^2), $x_1x_2 \ldots x_n$ = variables in continuous or discrete form, and e = error in predicting a criterion.

When applied to a temporal-sequential data set Y becomes each variable linked to its temporal antecedents and successors, in a regression model which, ideally, incorporates all relevant predictors. Our goal, however, is not to report all significant outcomes, but to schematize those relationships which

FIGURE 1

TSAR ANALYSIS OF PREDICTORS AT BIRTH, 42, 54, AND 66 MONTHS, AND 66 MONTH ITPA *Auditory Association* SCORES-HIGH GROUP (N=54)



are statistically significant in the temporal sequence. By this we mean that (e.g.) SES level at conception and perinatal complications may well be statistically significant as predictors when the other is a criterion. However, the logic of examining perinatal complications as a statistically significant source of SES variance nine months before makes no sense when compared with the opposite proposition. The putative contribution of SES to variance associated with complications at birth nine months after conception would be a rational statement of hypothetical influence and its temporal direction.

FINDINGS

We wish to illustrate some insights gained by using multiple linear regression to derive a temporal-sequential analysis by regression (*TSAR*) with data from the St. Louis Baby Study. We begin by reporting an analysis of data at birth, 42, 54, and 66 months employing some predictors (a TSAR schema) of mental test performance. The predictors shown in Figure I were derived from previous research into salient influences in several independent domains on cognitive attainment.

In Figure I we see data from the developmental histories of 54 bright children. This TSAR schema was derived from a set of five predictors and shows the simplest of all linkages, one in which a predictor in a temporal sequence, perinatal SES (McGuire and White, 1955), is linked statistically to another predictor, the level of education of the head of the household at age 66 months, and then to the criterion. In this analysis three predictors in the full set did not play a statistically significant role, and are not evident in the schema.

In Figure 2 is a schema of four predictors, none of which is the statistically insignificant Apgar score (Apgar and James, 1962) at birth, and a criterion score at child age sixty six months. In this TSAR schema we see

TSAR SCHEMA OF PREDICTORS AT BIRTH, 54 AND 66 MONTHS AND

66 MONTH WPPSI *Vocabulary* CRITERION SCORES (N=348)

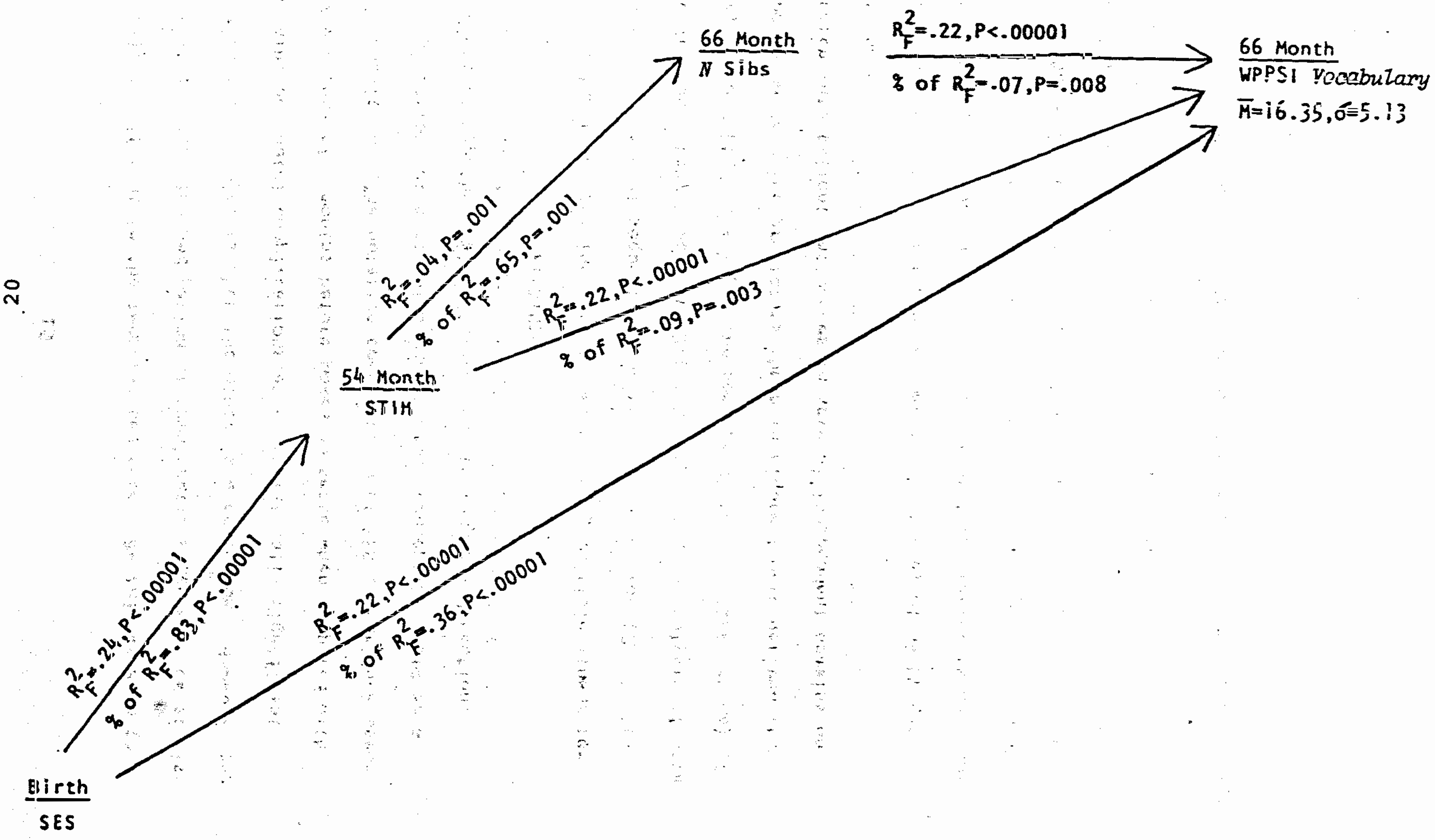
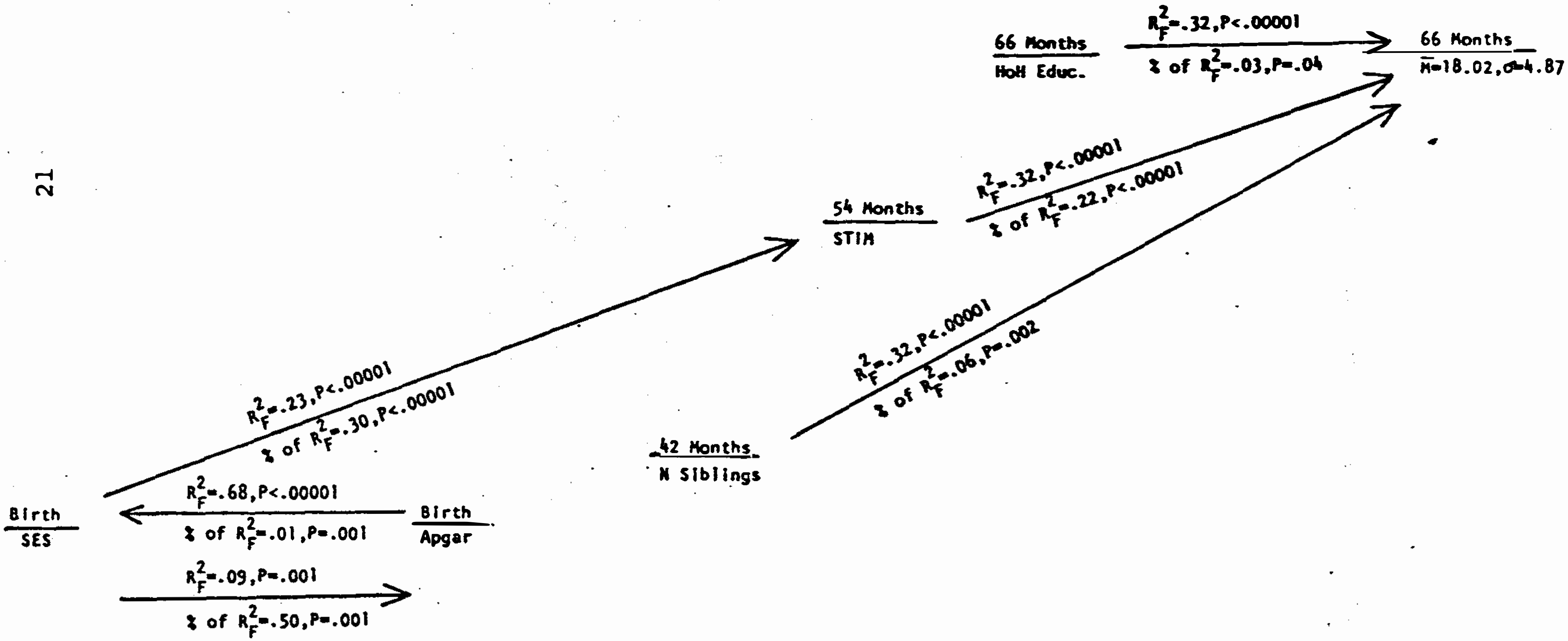


FIGURE 3

TSAR SCHEMA OF PREDICTORS AT BIRTH, 42, 54, AND 66 MONTHS,

AND 66 MONTH ITPA *Auditory Association* SCORES (N=256)



several things. There is a sequential link from perinatal SES score through 54 month STIM (Caldwell, 1970) and the number of sibs at 66 months to the criterion. This schema also shows both direct and indirect links as SES also connects directly to the criterion.

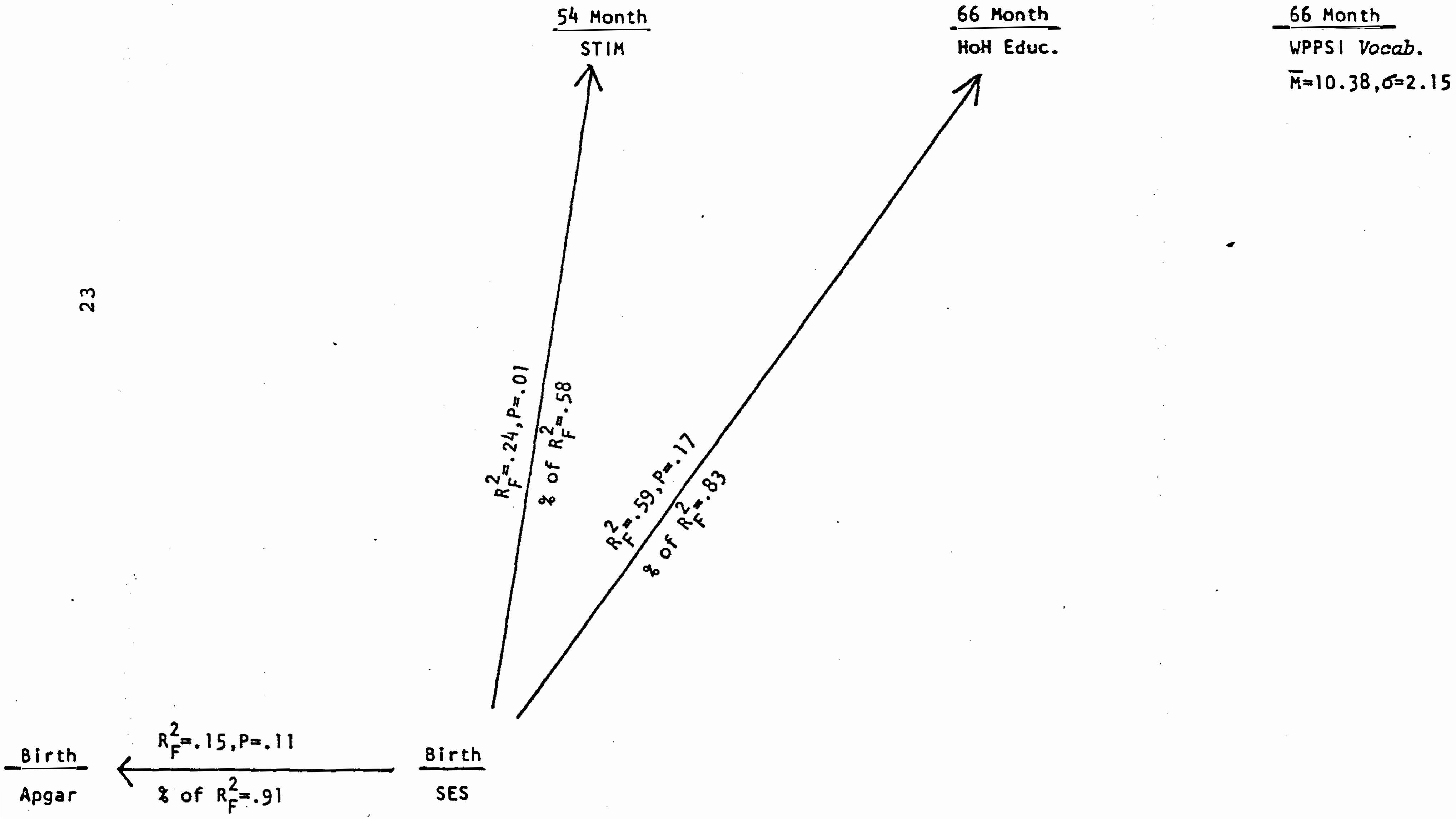
In Figure 3 we see two anomalies illustrated. In the perinatal data we see a two-way statistically significant link between Apgar and SES scores. Of these two mathematically correct findings only one is conceptually valid, since physiological conditions of an infant in the first few minutes of life, the Apgar score, cannot influence an SES score based on parental characteristics. The second anomaly is that we can trace a link backwards from the criterion to a 66 month variable, the educational level of the head of the household, but we can not go back any further since the linkage breaks off. At the same time there are two other antecedents, the 54 and 42 month variables, which go back in the developmental sequence to birth SES score.

In Figure 4's TSAR schema of five predictors we see that four of the predictors can be linked to each other; the only predictor variable not arrayed in Figure 4 is a 42 month variable, number of siblings. More importantly, we see that the schema links predictors, but the set does not connect to the criterion. In this case we see linkages which extend from birth to age 66 months. However, there is no link from any of the predictors in the schema to the criterion. The constellation of four predictors from perinatal Apgar score to education of the head of the household at sixty six months hinges on the perinatal McGuire and White (1955) SES score. However, there is no linkage from the perinatal SES score to the WPPSI *Vocabulary* scores of children selected because of their low scores on Raven's (1950) Colored Progressive Matrices (1947). The practical import of this is that the developmental influences we have seen

FIGURE 4

TSAR SCHEMA OF PREDICTORS AT BIRTH, 42, 54, and 66 MONTHS, AND

66 MONTH WPPSI *Vocabulary* SCORES-LOW GROUP (N=49)



in the previous three schemas continue to cluster among themselves, but have no functional relationship severally or collectively, to the criterion scores of the children.

The examples given so far use clinical data from child development. A parallel example is given in Figure 5; it uses data on social conditions surrounding childhood in fifty six countries and employs 1975 gross national products in United States dollars as the criterion. We provide this analysis of data from a work in progress (Jordan, 1981b) in order to show that economic and social data which have a temporal-sequential flavor can also be explicated in a fashion which is schematic, and which makes use of inferential tests of statistical significance.

DISCUSSION

The TSAR schemas are based in regression analysis and it is important to note that only statistically significant variables and their contribution in full regression models is reported. R^2 values of the full models (R^2_F) vary a great deal in size; this is in contrast to uniformly low R^2 values in models whose criteria are measures of early development (Jordan, 1980).

The TSAR arrangement of data can be applied to any number of predictor variables. The computational and model-building aspect of the regression analysis can be handled by any regression package. In the examples given here the data were analyzed in models in which all predictors were treated as criteria in the presence of the other predictors as co-variables, which is the usual arrangement in multiple linear regression analysis. It is helpful to recall that a number of such combinations could violate the canon of logical order in temporal sequence, with the predictor occurring later in the developmental sequence. At that point the theory guiding selection of variables and their hypothetical relations limits analysis to regression models which are logical and

FIGURE 5

TSAR SCHEMA OF ENVIRONMENTAL INFLUENCES ON $ GNP

IN 56 POOR COUNTRIES