ON THE RELATIONSHIP AMONG VARIABLES IN A LONGITUDINAL STUDY OF PEOPLE CHANGING JOBS

Sidney Cobb, David McFarland, Stanislav V. Kasl, George W. Brooks

Institute for Social Research
University of Michigan
Ann Arbor, Michigan, U.S.A.

This report is concerned with a methodologic problem in the analysis of data from a current longitudinal study of the health of people changing jobs. The objectives of this study are twofold, to describe the effects of the sudden termination of employment in middle life on physical health, mental health and illness behavior, and to study the interrelationship of psychological and physiological variables as people move through a crisis in their lives.

The data available for preliminary analysis involve 66 men who have been observed on five occasions from before the plant closing to one year after the closing. In this analysis we are concerned only with continuous variables collected in identical fashion, once at each time period. As might be expected, we have to deal with a certain amount of missing data because some of the men were unavailable or refused at certain of the time periods. Forgetting about the missing data, we can visualize the material as containing five observations on 66 men or 5 x 66 = 330 data points. Because of missing data, we are in actuality dealing with about 250 data points.

In analyzing the relationship between any two variables we might simply calculate the correlation coefficient for those two variables over all those data points. However, it is of considerably more interest to know if the relationship between the two variables is due to the fact that they are associated

characteristics of individuals at any one time or things that move together through time within individuals or both.

Let us look at the first line of the table. Here we see the relationship between our 7-item self report measure of depression and our 5-item self report measure of anxiety. When we use all the data points in our correlation, we get what we call a raw correlation. This is shown in the first column and for this pair of variables it is 0.66. In order to see if this correlation would hold up in a simple cross sectional survey, we took the mean of the five observations on each man so that we had a mean depression score and a mean anxiety score for each man. When these were correlated we got 0.72 which appears in the second column headed Means.

Next we took the mean depression score and subtracted it from each of the five observations on that man so that we had a depression difference score. Similarly, we obtained an anxiety difference score. When these two were correlated, we obtained a correlation of 0.60 which appears in the last column under the heading Difference Scores. By subtracting out that part of each value which is characteristic of the man, we have left only the changes over time within men. This correlation indicates the degree to which changes over time are correlated.

In the first row of the table we have seen a pair of variables that have high correlations both in the Means and in the Difference Scores. The next pair of variables, the self report of anxiety with the nurses evaluation of the man's anxiety shows a strong correlation in the Means and no correlation in the Difference Scores. A possible interpretation of this finding is that the nurse is well able to distinguish the generally anxious from the normal man but is not able to discriminate the small changes in his anxiety which take place from time to time.

On the third line of the table are the same three correlations between serum uric acid and serum creatinine. Here only the Difference Scores are substantially correlated. This means that uric acid and creatinine have a tendency to go up and down together within men over time but have no association across men. This pattern is rare in our material.

We have pointed out three patterns of association. The first is illustrated by Depression and Anxiety and involves strong correlation in both the Means and the Difference Scores. These are then two variables that are well related across men and across time. The next pair, self report and nurse report of Anxiety, are correlated in the means, i.e. across men but not in the difference scores. The third pair, Uric Acid and Creatinine, are not correlated across men but are correlated across time within men.

There are a variety of other ways this problem might be tackled. In one model each time period and each person, without interaction, was allowed to have an additive effect on the slope and the intercept in a regression equation relating two variables. It turned out that we didn't have enough with values with adequate spread to get reasonable estimates of all the slopes and that the matrix inversions required would be very expensive. Another procedure, formally equivalent to the regression approach, uses what the statistician calls "indicator variables" and the econometrician calls "dummy variables" for time and for person. This would give some computational advantages but still regression coefficients are not as easily interpreted as correlation coefficients, unless one has an explicit causal hypothesis about which variable depends on the other. A somewhat weaker model is two-way analysis of covariance using time and person simultaneously as classification variables. We also might think of the problem as an analysis of variance with repeated measures in which we would partition not only the variation in the several dependent variables but also the covariation between pairs of variables. To date we have shied away from these approaches because of the missing data problem.

Finally, we might have tried to be symmetrical in our correlation approach. For the means this would have been ridiculous for correlating the means across time periods where the number of periods is only five would have no meaning. For the difference scores, it would be possible to be symmetrical working first with the means for each time period subtracted out and then with the means for each man removed. This has the disadvantage of being much more work and of being less intuitively related to the across men and across times concepts with which we started. With these thoughts in mind, we plan to pursue our analysis always studying the interrelationship between variables both in the Means and in the Difference Scores.

Before closing we would like to remind ourselves that other variables may obscure or suppress a relationship or an important relationship may be visible only in a particular subset of the population. Line four in the table illustrates a case in point. Here Serum Uric Acid and Sadness as evaluated by the nurse appear unrelated until we divide the men into those who are flexible and those who are rigid. Among the flexibles there are strong negative relationships between the variables but among the rigids there is a positive relationship in the Difference Scores and a trivial positive relationship among the Means.

We must conclude with the oft forgotten platitude that patience and thoroughness in the analysis of complex data will pay off in the long run.

(Prepared May 28, 1968, for presentation to the Fifth Scientific Meeting of the International Epidemiological Association, August 1968.)

The correlation coefficients for the raw scores, the means and the difference scores for the specified pairs of variables from a longitudinal study of people changing jobs.

				Difference
		Raw	Means	Scores
1	Depression Self Report vs. Anxiety Self Report	.66	.72	.60
2	Anxiety Self Report vs. Anxiety Nurse Report	.43	.60	,07
3	Serum Uric Acid vs. Serum Creatinine	.12	.08	.33
4	Serum Uric Acid vs. Sadness Nurse Report	-,07	-.13	.09
	Same for flexible men	-	-.58	-.25
	Same for rigid men	-	.19	.32

5 P

4/204