DOCUMENT RESUME

ED 106 370 TM 004 488

AUTHOR McNeil, Judy T.
TITLE Regression Analysis for Repeated Measures Designs: Dealing with Missing Data and the Use of Covariates as an Alternative to Person Vectors.
PUB DATE [Apr 75]
NOTE 13p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, D.C., March 30-April 3, 1975)

EDRS PRICE MF-$0.76 HC-$1.58 PLUS POSTAGE
DESCRIPTORS *Analysis of Covariance; Criterion Referenced Tests; Data Processing; *Hypothesis Testing; Individual Characteristics; Individual Differences; Measurement Techniques; *Multiple Regression Analysis; Performance; *Scores; Scoring Formulas; *Statistical Analysis; Testing; Time

ABSTRACT

The method of using person vectors in regression analysis to test repeated measures hypotheses or questions is discussed. These hypotheses involve designs with pre and post scores with one group and with multiple groups. Based on these analyses with person vectors, there are two major focuses of the paper: a proposed solution to the problem of missing data in repeated measures designs, and the use of selected covariates as an alternative to person vectors in controlling for differences between individuals. (Author)

REGRESSION ANALYSIS FOR REPEATED MEASURES DESIGNS: Dealing with Missing Data and the Use of Covariates as an Alternative to Person Vectors

Judy T. McNeil
High/Scope Educational Research Foundation
Ypsilanti, Michigan

Funds for the writing and typing of this manuscript were provided by Educational Monitoring Systems

A paper presented at the annual meeting of the American Educational Research Association's Multiple Linear Regression Special Interest Group

Washington, D. C.
April 1975

REGRESSION ANALYSIS FOR REPEATED MEASURES DESIGNS: Dealing with Missing Data and the Use of Covariates as an Alternative to Person Vectors

Judy T. McNeil
High/Scope Educational Research Foundation

When subjects are measured on a particular criterion variable at more than one point in time, a "repeated measures" analysis is generally used to test hypotheses about the data. One of these hypotheses may be that the group's mean score increased (or decreased) over time. Since it is generally expected that there is a correlation between early and later scores for the same subjects, it is desirable to take this correlation into account, which is in effect controlling for each individual's mean score on the criterion variable. In fact, some writers argue that failure to extract variance attributable to subject differences results in a violation of the assumption of independence of errors (Glass, Peckham, and Sanders, 1973). (Some authors including this one [see also Dixon and Massey, 1969, and Downie and Heath, 1970] would argue instead that the crucial aspect in deciding to use a repeated measures analysis is "accounting for variance," not "meeting assumptions." If a source of variance is known, it is usually wise--from an heuristic point of view as well as for the power of the test--to include that source of variance in the analysis. In those cases where there is no expected correlation between pre and post scores--as would be the case with some criterion-referenced tests with little or no variance on pre or post--then a repeated measures analysis would not be beneficial.)

In regression analysis, controlling for each subject's mean score on the criterion variable (repeated measures analysis) is accomplished by using person vectors, each of which identifies a subject and contains a "1" if the criterion score belongs to that subject and a "0" if not. Using this regression approach, one can accomplish analyses identical to the "correlated" t test or the repeated measures ANOVA; in addition, this approach allows for greater flexibility in the analysis application.

The application of the general linear model to repeated measures problems was discussed in detail in a presentation to the Special Interest Group in Multiple Linear Regression at the AERA annual meeting in 1974 (Pohlmann and McShane, 1974). Therefore, this paper will describe only briefly the method of using person vectors in regression analysis to test research hypotheses regarding repeated measures, for a one group--two trial situation.

Based on these analyses with person vectors there are two particular focuses of this paper: (1) a proposed solution to the problem of missing data, and (2) the use of covariates as an alternative to person vectors in controlling for differences between individuals.

A One Group--Two Trial Research Hypothesis

Although there are many research hypotheses that a researcher may wish to test in this situation, the most typical one would be the following: "The mean score of the group increases from timepoint A to timepoint B, over and above differences between individual mean scores." For example, a researcher may have measured four infants' ability to focus on and visually follow a moving human face at two ages, say 10 days and 30 days. According to developmental theory, the infants' scores should increase from the first timepoint to the second timepoint (if the measure is well constructed). Since she expects a correlation between pre and post scores on this measure at these ages, the researcher wishes to covary these individual differences. The researcher therefore wishes to test the hypothesis that "the mean score on visual following will increase from 10 days to 30 days, over and above differences between individual mean scores." The resulting full model for testing the hypothesis is thus:

Model 1:

$$X_1 = a_0U + a_1T_1 + a_2T_2 + a_3P_1 + a_4P_2 + a_5P_3 + a_6P_4 + E_1$$

where: X_1 = criterion vector of both 10-day and 30-day scores on visual following;

U = the unit vector;

T_1 = 1 if score on X_1 is a 10-day score, 0 otherwise;

T_2 = 1 if score on X_1 is a 30-day score, 0 otherwise;

P_i (i=1 to 4) = a vector containing a 1 if the score is from person i, 0 otherwise;

$a_0, a_1 \ldots a_6$ = a set of least squares weights derived to minimize the sum of the squared elements in E_1; and

E_1 = the error vector.

Sample data that will be used to demonstrate this analysis is shown in Table 1. The data would be organized into vectors as in Figure 1.

Person	10-day Score on Visual Following	30-day Score on Visual Following
1	4	8
2	3	5
3	6	6
4	2	7

Table 1. Sample Data for a One Group--Two Trial Hypothesis

$$X_1 = \begin{bmatrix} 4 \\ 3 \\ 6 \\ 2 \\ 8 \\ 5 \\ 6 \\ 7 \end{bmatrix} \quad U = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \quad T_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad T_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \quad P_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad P_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \quad P_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \quad P_4 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

Figure 1. Vectors for Model 1.

The weights a_1 and a_2 in Model 1 will take on values which will reflect the difference between the 10-day and 30-day means. The restriction placed on Model 1 to test the hypothesis would be $a_1=a_2$, which results in the restricted model, Model 2, which would be compared to the full model using an F test. The full model contains 5 linearly independent vectors, and the restricted model contains 4. The degrees of freedom for the F test are therefore 1 and 3.

Model 2:

$$X_1 = a_0U + a_3P_1 + a_4P_2 + a_5P_3 + a_6P_4 + E_2$$

The type of analysis just described is the basis for the following discussions of missing data and the use of covariates.

Missing Data

When collecting longitudinal data (repeated measures), it is unfortunately all too common to find that some subjects are not available for at least one measurement timepoint. Several possible approaches to handling the problem of missing data, not restricted

to repeated measures, are presented by McNeil, Kelly, and McNeil (1975), and three of them are discussed below. A fourth approach is proposed here as a suggested alternative when dealing with repeated measures.

(1) Insert in place of the missing score the mean value of the variable.

(2) Insert in place of the missing score a random value which is within the range of the variable.

(3) Eliminate subjects for whom any data is missing.

(4) Eliminate a subject for only that timepoint on which data is missing; utilize other data for that subject.

(1) Inserting the mean. Using this approach, the researcher makes the assumption that the person with missing data is like the average subject with data. If, in the example on infant visual following used above, subject number 1 was missing the 10-day score, the researcher would assign the mean of the other three subjects' 10-day scores (3.67) to subject number 1. Sometimes, however, the researcher may not be willing to make the assumption that persons with missing data are like persons with data--it may be that they are absent from testing because they are less healthy or less willing to cooperate or different on some other variable which may be relevant to the construct under investigation. In addition, the insertion of mean values reduces the variance of the predictor variable, resulting generally in a variable with lowered predictive value.

(2) Inserting a random value. One way of getting around the problem of reduced variance due to inserting a mean value is to insert a random score (from the range of observed scores) into the missing data locations. Thus, in the example above, if subject number 1 was missing the 10-day score, the researcher would assign a random value (say 6) to subject 1. However, this procedure tends to decrease the relationship between that predictor variable and any other variable because it adds random variance. It should be pointed out that in a repeated measures analysis, inserting either random scores or mean scores tends to reduce the correlation between pre and post scores--a correlation which is expected to be high and on which the justification of a "repeated measures" or person-vector analysis is based.

(3) Eliminating subjects. Using this approach, all scores for a subject would be eliminated if the subject was missing any score. In repeated measures studies, this method can result in the discarding of much data. If, in the example given above, subject number 1 was missing the 10-day score, the 30-day score for subject number 1 would be eliminated as well. This would eliminate the P_1

vector from Models 1 and 2, and the elements of the remaining vector would be as shown in Figure 2.

X_1	U	T_1	T_2	P_2	P_3	P_4
3	1	1	0	1	0	0
6	1	1	0	0	1	0
2	1	1	0	0	0	1
5	1	0	1	1	0	0
6	1	0	1	0	1	0
7	1	0	1	0	0	1

Figure 2. Elements of vectors when subject 1 is eliminated.

A drawback to this approach is that eliminating subjects because of missing data most likely redefines the population from which one has sampled and hence to which one can generalize. Infants who could not be tested at one timepoint may be the least healthy ones or may come from less organized families who could not arrange to keep the testing appointment. Therefore, if one continually uses only complete data, the population to which one can generalize may be restricted to healthy infants from organized families. This restriction may be unavoidable in some studies. But when one has some data on these differing subjects, it is unfortunate and possibly unnecessary to ignore that data.

(4) Eliminating only missing timepoints. I would propose that in most longitudinal or repeated measures studies the researcher would neither want to reduce the relationship between pre and post scores by inserting mean or random values nor want to give up the use of a subject's data on all timepoints just because the subject is missing a score at one timepoint. I would therefore recommend that a subject be included in the analysis if scores were obtained for him at the earlier timepoint, or at the later timepoint, or at both timepoints. This would mean that all obtained data would be utilized and that for some subjects only pre or only post data would be in the analysis. For example, if infant subject number 1 was missing the 10-day score on visual following, that score would be missing from the analysis but his score at 30 days would be included. The elements of the vectors for the repeated measures analysis would be as shown in Figure 3.

When there are only two timepoints in the analysis, there is no gain in degrees of freedom in using this approach rather than approach number 3 above (eliminating all scores for the subject with missing data). Note that in Figure 2 there are six observations; the full

X_1	U	T_1	T_2	P_1	P_2	P_3	P_4
3	1	1	0	0	1	0	0
6	1	1	0	0	0	1	0
2	1	1	0	0	0	0	1
8	1	0	1	1	0	0	0
5	1	0	1	0	1	0	0
6	1	0	1	0	0	1	0
7	1	0	1	0	0	0	1

Figure 3. Elements of vectors when the 10-day score for subject 1 is eliminated, but the 30-day score for the same subject is retained.

model would contain four linearly independent vectors and the restricted model would contain three. The degrees of freedom when all scores for subject number 1 were eliminated would thus be 1 and 2. In Figure 3 there are seven observations; the full model would contain five linearly independent vectors (because P_1 appears) and the restricted model would contain four. The degrees of freedom when only the 10-day score is eliminated would thus be 1 and 2, the same as for Figure 2. The fact that the degrees of freedom are the same in both instances makes sense when you consider that, by including the P_1 vector along with the 30-day score for subject 1, the 30-day score for this subject is completely accounted for. Note that when there are more than two timepoints in the full model, the elimination of only one timepoint for a subject--rather than all timepoints for that subject--will result in a gain in denominator degrees of freedom. This makes sense because now the person vector for that subject does not completely account for that subject's variance.

The advantage in using this approach when there are only two timepoints is therefore not to be found in degrees of freedom. When there are more than two timepoints, it will yield a gain in df. But an advantage in both cases is that the increased number of observations yields a more stable estimate of the population mean on the criterion score, and this estimate is more representative of the population of subjects that the researcher set out to measure--not just subjects with complete data.

Covariates as an Alternative to Person Vectors

The above discussion has assumed that the use of person vectors is the path a researcher would wish to take when there is an expected correlation between individuals' scores at two timepoints.

But there is another approach that a researcher may wish to consider when dealing with measures that are repeated over time on the same subjects, and that approach is the use of a particular kind of covariate in place of the person vectors.

Using person vectors is essentially covarying for each person's uniqueness. It is acknowledging the expected correlation between pre and post scores by assuming that the reason a particular individual tends to score high (or low) at both timepoints in relation to other subjects is simply because he is that unique individual. But the researcher may know (or have reason to suspect) that there are measurable dimensions along which her subjects vary and which relate to the criterion behavior in explaining why individuals tend to score high (or low) at both timepoints. For example, our researcher may have evidence from prior studies that an infant's ability to focus on and follow a moving human face at both 10 and 30 days is related to the sex ' the birth weight of the infant. Using scores on these two variables to predict the criterion score will probably yield a lower R^2 (greater errors of prediction) than predicting on the basis of which scores belong to which individuals (i.e , using person vectors), but it will be more valuable in a theory-building attempt as well as more generalizable to another group of subjects. It will also be a more parsimonious model. While using person vectors is a powerful technique for testing the hypothesis of differences between timepoints, it does not tell us anything about how individuals' differences on other variables causes them to be different from one another and yet consistent with themselves on the criterion measure. And since person vectors represent each person in the sample individually, they cannot be used to generalize beyond the sample.

Covariates which do not change across timepoints. It is not possible to use both person vectors and one or more covariates which do not differ for a subject across timepoints. An infant's sex and birth weight are this type of potential covariate. Whether one is considering a 10-day or 30-day criterion score, the sex of the infant would be the same, as would the birth weight of that infant. If a variable of this type were used in a regression model in conjunction with person vectors, a linear dependency would be generated. This is illustrated in Figure 4, where it is shown that a set of weights can be found such that the birth weight variable is a linear combination of the person vectors. A choice must be made, then, between the covariate and the person vectors. Another way of looking at this choice is to consider the hypotheses which represent the two choices. For the hypothesis, "the mean score on visual following will increase from 10 days to 30 days, over and above the effect of birth weight," the researcher would use the birth weight variable as a covariate. If the researcher chose instead to state the hypothesis, "the mean score on visual following will increase from 10 days to 30 days, over and above differences between individual mean scores on the criterion," person vectors would be used. This author would argue that, if the researcher has sufficient evidence to expect that a certain covariate

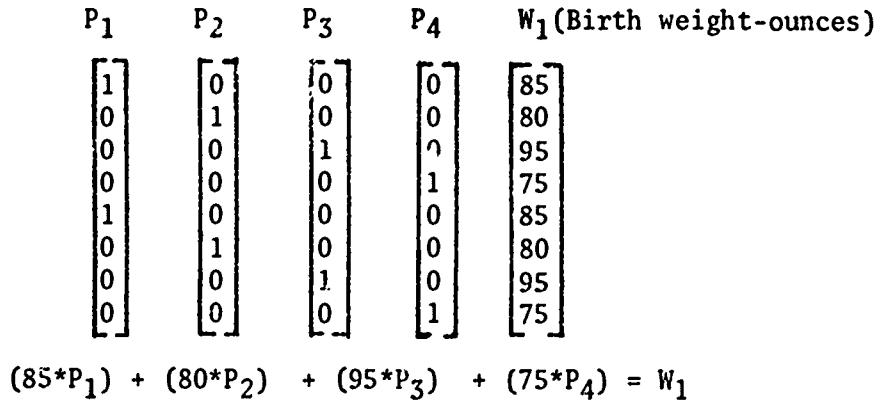

Figure 4. Linear dependency generated by the use of person vectors and a covariate which does not differ across timepoints.

or set of covariates is related to the criterion in a repeated measures design, it would be beneficial to state the hypothesis in terms of the covariate. The benefits are: (1) the covariate model is more parsimonious than the person vector model, (2) the findings are more generalizable to a new sample when the covariate is used than when person vectors are used, and (3) an advance can be made in theory building because the analysis containing the covariate gives a more refined estimate (than person vectors) of what specifically enters into individual differences on the criterion. In short, the researcher may wish to state the hypothesis in terms of a covariate rather than person vectors and thus will give up some of the power of the statistical test of differences over time in return for gaining generalizability of findings that will also aid in theory building. (Even if the researcher chooses to use person vectors, I would recommend that some thought and possibly additional analyses be devoted to accumulating evidence regarding variables which do relate to individual differences on the criterion--for possible use in the future.)

Covariates which change across timepoints. The above discussion has focused on covariates which do not change across timepoints. A short comment should be made about the use in a repeated measures design of covariates which do change across timepoints. It is possible to use in the same analysis both person vectors and covariates which differ across timepoints. If, for example, the covariate of interest were the weight of the infant at the time of testing, this would probably be different at 10 days and 30 days. The researcher might then state the following hypothesis: "the mean score on visual following will increase from 10 days to 30 days, over and above the effects of weight at the time of testing and differences between individual mean scores on the criterion." Examples of

vectors that would be used to test this hypothesis are shown in Figure 5. (A changing-over-time covariate might be useful in a case like this, in which the researcher may wish to know if the infants' scores increase over time, beyond the increase that could be predicted by weight gain alone. It should be pointed out, however, that in a repeated measures design which includes treatment as a factor, one would not wish to include a covariate on which change over time is a possible function of that treatment because the covariate may account for some of the variance in the criterion which is due to treatment.)

X_1	U	T_1	T_2	P_1	P_2	P_3	P_4	W_2 (Ounces at testing)
4	1	1	0	1	0	0	0	85
3	1	1	0	0	1	0	0	84
6	1	1	0	0	0	1	0	97
2	1	1	0	0	0	0	1	75
8	1	0	1	1	0	0	0	90
5	1	0	1	0	1	0	0	95
6	1	0	1	0	0	1	0	99
7	1	0	1	0	0	0	1	85

Figure 5. Example vectors for a situation in which the researcher wishes to use both person vectors and a covariate which changes over time.

Discussion

Two issues regarding the analysis of longitudinal or repeated measures data have been presented--the handling of missing data and the use of covariates. These concepts need not be applied separately in practice; it would be perfectly reasonable to use a covariate instead of person vectors with a set of data containing missing scores. Some additional comments on missing data, covariates, and the basic choice of a repeated measures analysis are presented in this section.

In handling the problem of missing data, the researcher needs to consider how many subjects are missing data on how many timepoints--this is a data-based question. It has been pointed out above that, when there are only two timepoints in an analysis, the use of subjects who have data for one timepoint in addition to those with complete data results in no additional degrees of freedom. The mean obtained is a better estimate of the population mean, but are the total results more generalizable to the population? This writer is not at all certain--my guess is that they are not. But when there are more than two timepoints, it would appear to me

that the use of subjects who have data for one or more timepoints would result in better estimates of both the population mean and the population variance and would yield results that are more generalizable to the population one initially intended to sample.

If one can refer to a question of the "power" of the test when discussing the use of a "repeated measures" analysis versus a "regular" analysis (person vectors versus no person vectors), then it seems appropriate to refer to "power" when the decision is between person vectors and a covariate. This writer would say, then, that the decision regarding the use of a covariate rather than person vectors in a repeated measures analysis is one of weighing theory and statistical power. But if the covariate or set of covariates account for nearly as much variance in the criterion as do the person vectors, and if there are far fewer covariates than person vectors, the covariate analysis may actually be more powerful (rather than less) because it is more parsimonious and thus generates greater degrees of freedom.

It seems appropriate at this point to develop one final issue: when should one use a "repeated measures" analysis. The argument was made early in this paper that, even though data is obtained on the same subjects at two timepoints, one may not necessarily wish to covary for differences in individuals' mean scores (i.e., use a repeated measures analysis). Just as it is not beneficial to covary on any variable which is not related to the criterion, it is not beneficial to control for a correlation between pre and post scores if no such correlation exists. The example was given earlier of a criterion-referenced test--one may expect all students to achieve criterion on the post test. In this case there would be no correlation between pre and post scores. However, the researcher in this case would probably not be interested in a hypothesis regarding an increase in scores from pre to post so the question of repeated measures would not apply. Consider, though, the situation in which a researcher is developing a measure of, say, mathematics concepts. In validating the measure, she wants to show that scores increase from pre to post and also that prescores are correlated with postscores. She is interested in both questions--increasing scores and correlations between scores. The procedure she follows might be this: First she tests a number of subjects at two timepoints. Then she inspects the correlation between pre and post scores. This answers her question regarding the correlation. If there is a substantial correlation, she proceeds with a repeated measures analysis on the difference (increase) between pre and post scores. If on the other hand there is no correlation, she must decide if she is still interested in the question of increasing scores. In the validation of the instrument, this may still be a question of interest. She would therefore test the difference between pre and post scores--without a repeated measures design (it would be an "uncorrelated" t test).

This paper has presented the repeated measures type of analysis as it is formulated in regression models. Issues regarding its use

with missing data and covariates have been presented. Before making decisions on these issues and applying those decisions to data analysis, the researcher must first decide (1) what hypothesis is being asked, and (2) how a lack of correlation between time-points will affect that hypothesis.

REFERENCES

Dixon, W.J. and Massey, F.J. Introduction to Statistical Analysis, Third Edition. New York: McGraw-Hill, 1969, (p. 120).

Downie, N.M. and Heath, R.W. Basic Statistical Methods, Third Edition. New York: Harper and Row, 1970, (p. 177).

Glass, G., Peckham, P., and Sanders, J. Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance. Review of Educational Research, 1972, 42(3), 237-288.

McNeil, K., Kelly, F.J., and McNeil, J. Testing Research Hypotheses Using Multiple Linear Regression. Carbondale, Illinois: Southern Illinois University Press, 1975.

Pohlmann, J.T. and McShane, M.G. Applying the general linear model to repeated measures problems. Paper presented to the Annual Meeting of the American Educational Research Association, Chicago, 1974.