

**Analyzing Repeated Measures Designs:
What You Don't Know CAN Hurt You**

**Richard W. Cole
Computation Center**

The University of Texas at Austin

ABSTRACT

The analysis of data having repeated measures for each observation is typically analyzed using an ordinary least squares model known as Repeated Measures ANOVA where, among other things, equal change across time points is assumed for the observations within a group. When this assumption known as additivity is not met, change across time is likely to be occurring in different amounts for different observations within a group. This "differential change" model is, in reality, very common when the effect of some intervention is being examined. For example, persons with low scores initially, tend to benefit more from an intervention than those having already high scores. This paper examines ordinary least squares regression models for fitting and testing data having repeated measures having two time points. Examples applicable to various behavioral sciences are presented and discussed.

Repeated Measures and Statistical Control

The concept of statistical control for factors that may influence or conceal relationships among variables of interest is of primary importance in data analysis. Many of these factors are due to individual differences among observations in a study. In between-subjects designs, such differences are considered random though, being unexplained, they do increase the error term in the analysis thereby reducing the power of the statistical test (the likelihood that the null hypothesis will be correctly rejected when it is indeed false in the population). One method of controlling for the unique attributes of observations is to assume they will maintain a stable position relative to the mean across all measurement occasions. These consistent differences may then be evaluated and statistically controlled for in any subsequent hypothesis testing. Such within-subjects designs are often referred to as repeated measures designs where subjects are measured repeatedly on some (dependent) variable of interest with their stable individual differences accounted for to reduce the statistical error.

While repeated measures designs may result in more powerful statistical tests than their between subjects counterparts, these advantages come at a price many researchers are not aware of. These assumptions are examined below with suggestions for more appropriate model testing suggested when they are not met.

Assumptions of Repeated Measures Designs

The accuracy of the F statistic for a repeated measures design involves a number of additional assumptions beyond those of normality, equality (across groups) and independence of errors. Such assumptions, when satisfied result in a more powerful statistical test, though violations can produce spurious results.

Compound Symmetry. A common assumption generally discussed in the repeated measures literature is that of Compound Symmetry which states essentially that each dependent variable measurement correlates equally with each other dependent variable measurement. A single correlation coefficient then, is representative of the relationship across all dependent measures. Additionally, the variance of each dependent measure is equal. More technically, this assumption reflects the homogeneity of the variance-covariance matrix of dependent measures, and when satisfied, it reflects the stability of the individual difference discussed above. Failure to meet this assumption necessitates use of a less powerful multivariate approach to analyzing repeated measures. NOTE: there are other approaches to satisfying this assumption such as sphericity which are typically provided in most statistical packages including SAS. Additionally, correction factors are obtainable for moderate violations of this assumption.

Additivity. Less common in the instructional and research literature is an examination of the degree to which the observations within a group have an equal amount of change between two time points. That is, does each observation within the group change an equal amount between two time points. If there is differential change within the group across time, the condition is known as nonadditivity which reflects an observations by time interaction. In this case, where differential change occurs across time, we are able to observe that some observations are perhaps changing while others are not.

Additivity as an assumption typifies the power-assumptions tradeoff. When met, the assumption allows for a more powerful statistical test since fewer parameters are estimated. When violated the researcher is faced with including parameters in the model to estimate the interaction. Inclusion of these parameters may result in a weaker statistical test (though more accurate) unless additional

observations are included. A failure to include these parameters by opting for the more parsimonious additive model may produce biased predicted values and spurious test results. Typically additivity is tested using Tukey's Test for additivity (1946). A rejection of this test is an indication that differential change is occurring within the group across time.

Case Study

As an example consider the case where one is interested in examining whether typing performance changes as a function of instructional courses. Two groups of 50 subjects are randomly selected from some population to which inferences will be drawn. Upon placement into a group, subjects are immediately pretested on their typing performance and then, as a group, either given a traditional (group 1), or an innovative (group 2) typing course to increase typing speed. Upon completion of the course, we again assess each subject on typing performance and wish to examine the following traditional hypothesis appropriately tested with repeated measures ANOVA:

1. Averaging across both groups, there is no change in typing performance from time 1 to time 2.
2. Averaging across time points, the two groups are equal in their typing performance.
3. Both groups have an equal change in average typing speed from time 1 to time 2.

These are the 3 null hypotheses corresponding to a test of the time main effect, group main effect, and time by group interaction. There is no interest in differential change within a group. For our example, the researcher may hope to find in particular that hypothesis 3 is rejected and that the pattern of means for each group and time point is such that group 2 (the innovative group) has a mean equal to group 1 at time point 1 though a mean higher than group 1 at time point 2. This would indicate that, while the groups were equal at time 1, the intervention has had an impact beyond traditional typing instruction as indicated by the difference at time 2. Table 1 contains a summary of this hypothetical data and the results from the three hypothesis tested.

TABLE 1

GLM Default Output

General Linear Models Procedure
Class Level Information

Class	Levels	Values
G	2	1 2

Number of observations in data set = 224

Dependent Variable	T1	T2
Level of TIME	1	2

General Linear Models Procedure
Repeated Measures Analysis of Variance
Tests of Hypotheses for Between Subjects Effects

Source	DF	Type III SS	Mean Square	F Value	Pr > F
G	1	9368.463851	9368.463851	19.34	0.0001
Error	222	107550.908444	484.463552		

Repeated Measures Analysis of Variance
Univariate Tests of Hypotheses for Within Subject Effects

Source: TIME

DF	Type III SS	Mean Square	F Value	Pr > F	Adj G - G	Pr > F H - F
1	13335.18055362	13335.18055362	56.38	0.0001	.	.

Source: TIME*G

DF	Type III SS	Mean Square	F Value	Pr > F	Adj G - G	Pr > F H - F
1	9368.46385075	9368.46385075	39.61	0.0001	.	.

Source: Error(TIME)

DF	Type III SS	Mean Square
222	52508.82339291	236.52623150

Level of		-----T1-----		-----T2-----	
G	N	Mean	SD	Mean	SD
1	112	32.0000000	16.2281038	33.7657770	23.4013269
2	112	32.0000000	16.2281038	52.0575236	19.1743256

Conclusions From Traditional Repeated Measures Study

Based on the results presented above one might conclude that the traditional intervention does not have any substantial impact, while group two's innovative approach shows an increase of about 20 words per minute. Because the modeling approach was with repeated measures ANOVA, this change of 20 wpm is assumed for all observations in the group.

Regression Based Modeling Approach

What follows is a reinterpretation of the data using an alternative model testing procedure (carried out through PROC REG for reasons to be made clear below). Here the researcher has an interest in modeling for differential change within a group. Acceptance of additivity (based on Tukey's test using PROC IML) between subjects and time allows one safely avoid these procedures while nonadditivity suggests their use. At this point the researcher should recognize that they are dealing with a more complex set of relationships among subjects, groups and time. From the rejection of Tukey's test the researcher has found that the data are not be so parsimonious. Not everybody within a particular group is responding similarly to the experimental conditions, though as yet these relationships are still undetermined.

Graphing the Results

Because this is a two time point analysis, time 1 values can be plotted against time 2 as "lines of best fit" output from PROC REG for each of two groups. While the above ANOVA analysis used PROC GLM, we now wish to write an equivalent model using PROC REG. The commands below generate 4 new variables that when used in PROC REG will generate one of many possible models for examining a design having two groups and two time points. The particular model created below contains both of the newly computed intercepts and slopes. The resulting statistics and graph are presented in Table 2 and Figure 1.

```
data new; set rm;
    *compute intercepts for each group;
    g1=0; g2=0 ;
    if g=1 then g1=1 ; else g2=1 ;
    *compute slopes for each group;
    gs1=g1*t1 ;
    gs2=g2*t1 ;
proc glm data=new ;
    class g; model t1 t2 = g /noui ;
    repeated time 2 / nom;
    means g;
```

```

*continuation of SAS program;
proc reg data=new;
  model y=g1 g2 gs1 gs2 / noint ;
  output out=rout p=hats ;
  testlow1: test g1 + 5*gs1=5 ;
  testlow2: test g2 + 5*gs2=5 ;
  testmed1: test g1 + 25*gs1=25;
  testmed2: test g2 + 25*gs2=25;
  testhi1: test g1 + 50*gs1=50;
  testhi2: test g2 + 50*gs2=50;

```

Regression Results

In group 1 it is evident from a visual examination of the data that this intervention resulted in essentially no change across the range of observations. This is not the case with group 2. It is evident from the graph that individuals scoring low at time 1 have had a marked improvement at time 2. For individuals who initially typed 5 words per minute, the predicted time 2 value is over 50 words per minute representing a substantial change. It is important to note however, that such a change in performance is not consistent across the range of time 1 values for group 2. For individuals who initially typed 50 words per minute a modest improvement is seen. It is likely then that Group 2 has differential change based on pretest values calling into question the additivity assumption.

The analysis of this data through PROC REG provides estimates of various parameters generating the lines of best fit displayed in Figure 2. In this output acceptance of the additivity assumption can be roughly checked by examining parameter estimates for the slopes. When these values are close to 1 there is evidence that no differential change exists with in groups. (In fact one could use a test statement of the form: testslps: test gs1=gs1=1 ; for a test similar to Tukey's test of additivity.) It should be mentioned that linearity was assumed within each group between time 1 and time 2 scores for each group.

The equations for the 2 lines are:

$$\text{pred}(t2,g1)=g1+t1*gs1$$

$$\text{pred}(t2,g2)=g2+t1*gs2.$$

From Table 2, it is apparent that group 1 appears to have a less likely violation of additivity than group 2 which has a very low slope value. For group 1 then, we may be willing to assume a constant level of change across group membership while group 2 members have varying levels of change depending on initial values.

Figure 1
TWO-GROUP PLOT OF
BEST
FIT LINES

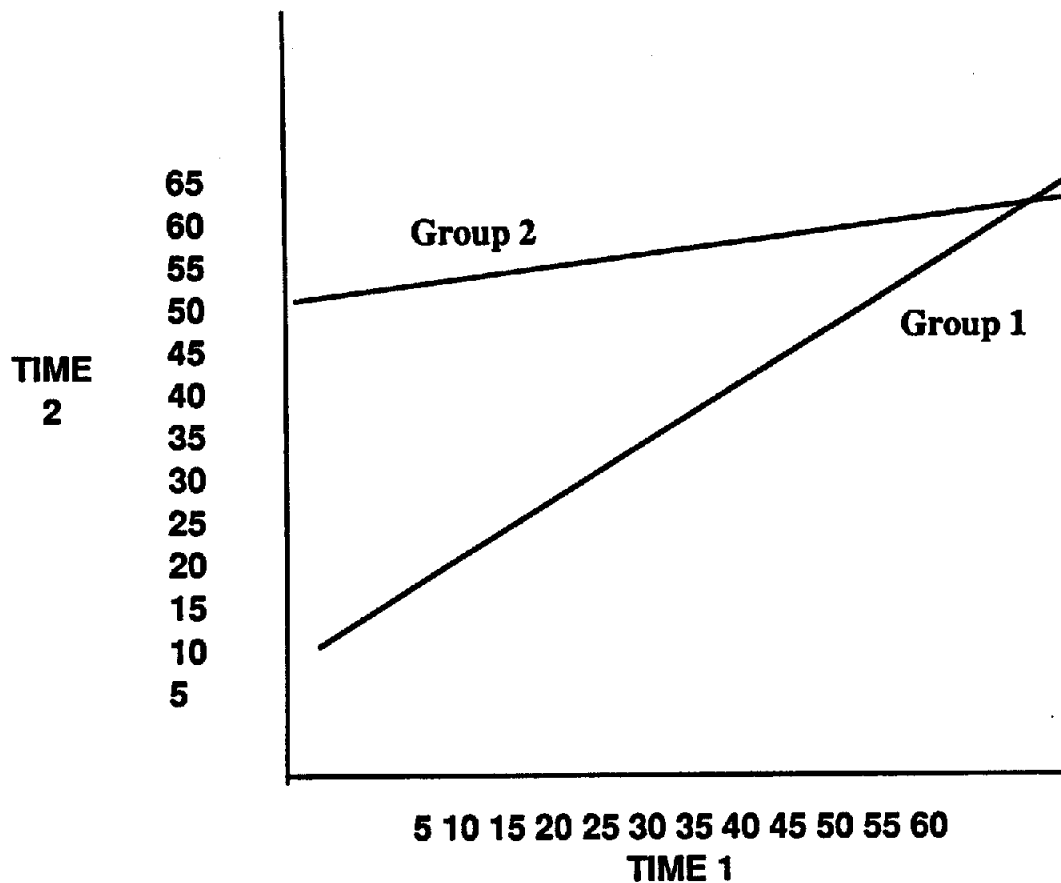


Table 2

DEP VARIABLE: Y
ANALYSIS OF VARIANCE

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB>F
MODEL	4	452709.13	113177.28	310.852	0.0001
ERROR	220	80099.31081	364.08778		
U TOTAL	224	532808.44			
ROOT MSE		19.08108	R-SQUARE	0.8497	
DEP MEAN		42.91165	ADJ R-SQ	0.8469	
C.V.		44.46598			

NOTE: NO INTERCEPT TERM IS USED. R-SQUARE IS REDEFINED.

PARAMETER ESTIMATES

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0: PARAMETER=0	PROB > T
G1	1	6.47109499	4.00060200	1.618	0.1072
G2	1	49.22517492	4.00060200	12.304	0.0001
GS1	1	0.85295881	0.11160246	7.643	0.0001
GS2	1	0.08851090	0.11160246	0.793	0.4286
TEST: TESTLOW1		NUMERATOR: 971.46	DF: 1	F VALUE: 2.6682	
		DENOMINATOR: 364.088	DF: 220	PROB >F : 0.1038	
TEST: TESTLOW2		NUMERATOR: 58913	DF: 1	F VALUE: 161.8099	
		DENOMINATOR: 364.088	DF: 220	PROB >F : 0.0001	
TEST: TESTMED1		NUMERATOR: 736.683	DF: 1	F VALUE: 2.0234	
		DENOMINATOR: 364.088	DF: 220	PROB >F : 0.1563	
TEST: TESTMED2		NUMERATOR: 65910.2	DF: 1	F VALUE: 181.0282	
		DENOMINATOR: 364.088	DF: 220	PROB >F : 0.0001	
TEST: TESTHI1		NUMERATOR: 38.781	DF: 1	F VALUE: 0.1065	
		DENOMINATOR: 364.088	DF: 220	PROB >F : 0.7445	
TEST: TESTHI2		NUMERATOR: 665.978	DF: 1	F VALUE: 1.8292	
		DENOMINATOR: 364.088	DF: 220	PROB >F : 0.1776	

The equations for the 2 lines are:

$$\text{pred}(t2, g1) = g1 + t1 * gs1$$

$$\text{pred}(t2, g2) = g2 + t1 * gs2.$$

From Table 2, it is apparent that group 1 appears to have a less likely violation of additivity than group 2 which has a very low slope

value. For group 1 then, we may be willing to assume a constant level of change across group membership while group 2 members have varying levels of change depending on initial values.

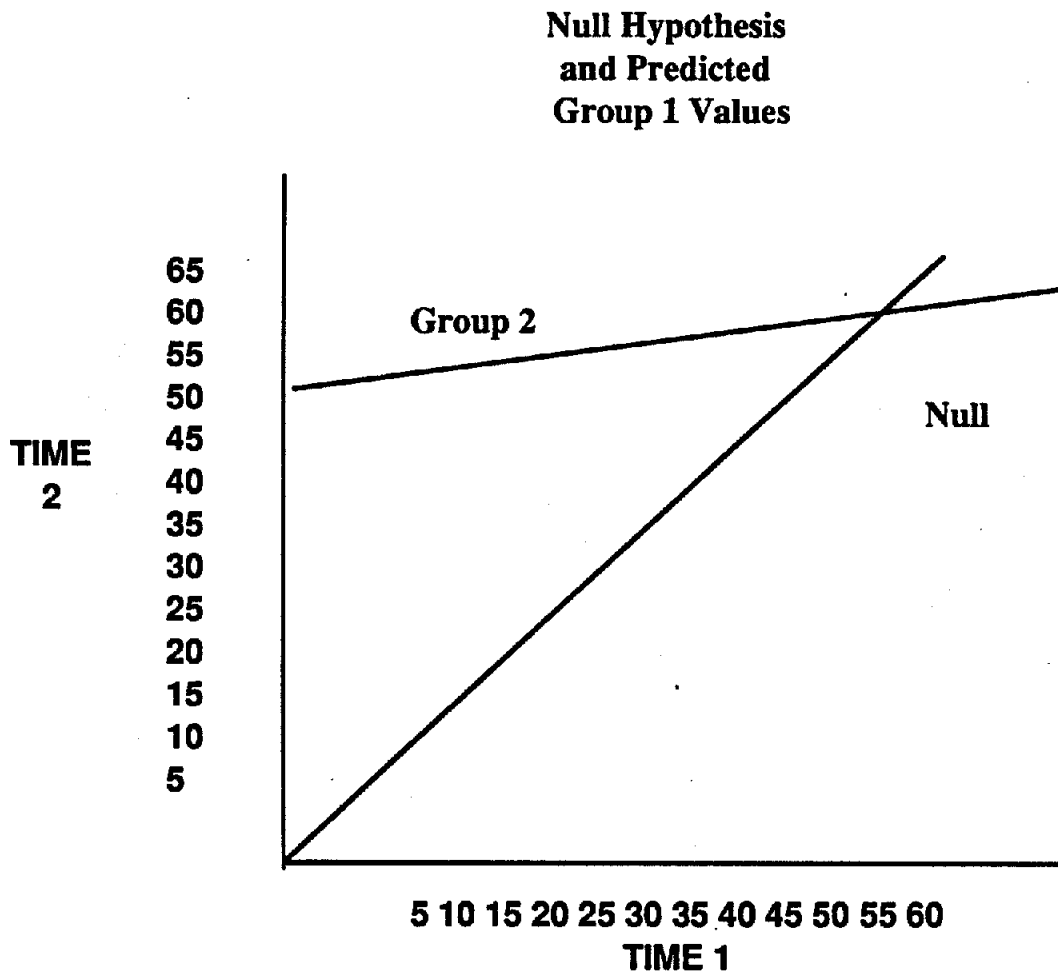
Hypothesis Testing for Time Changes Based on Initial Scores

For the remainder of this paper we will focus on examining the levels of differential change within group 2. As can be seen from Figure 1, those who initially scored low have increased dramatically while those with initially high scores have remained almost constant. The question of interest now becomes, "Below which level of the pretest do significant changes in time exist?" Answering this question is best served by an examination of Figure 2. Here the line of best fit for group 2 has been plotted against what can be called the null hypothesis line. This line depicts no change across all individuals in group 2 and crosses the best fit line around the value $t1=50$. This point represents the $t1$ value at which no change in typing performance is expected. Below this value scores are likely to have increased across time, while above the value scores have decreased. (It is possible that decreasing scores for experts may be expected depending on the degree of relearning necessary or prior habituation to overcome.)

While decisions may be made on the basis of this no change value, it may be of interest to the points at which statistically significant changes have occurred between the time points. Once again PROC REG through use of the TEST statement allows a direct examination of such hypotheses. Loosely stated, we wish to test a series of hypotheses that the predicted value for an observation in group 2 with a pretest score of some value is equal to that time 1 value. A number of test statements were presented in the PROC REG statement above. These TEST statements examined equality of time 1 and predicted time 2 values for time 1 values of 5, 25 and 50.

From the output given in Table 2, it can be seen that individuals having a score at time 1 of 5 show a significant change to time 2 (TESTLOW2 DF:1,220 F:161.8 PROB >F: 0.0001); likewise with those who scored 25 initially (TESTMED2 DF:1,220 F:181.02 PROB >F:0.0001). For those initial typist who scored 50 wpm, we see no significant change at time 2 indicating that the intervention has an impact primarily on those with low time 1 scores. Such information is relevant in relation to selection of individuals who are likely to benefit from an

Figure 2



intervention. Note that no significant differences were found for individuals in group 1 at any of the time points.

DISCUSSION

This paper stresses the additional information to be gained from the examination of differential time effects within and across groups. While the data presented are artificial, numerous components resemble aspects found in many repeated measures situations. Of central importance is the estimation of a 'no change point' on the x (time 1) axis, and possibly the point below which significant differences may be reported. In this analysis, such a point serves to suggest that poor initial performers have more to

benefit from than those already performing proficiently. It should be mentioned that such a point is contingent partially on sample size which suggests that the 'no change point' may provide less sample dependent information than the significance testing which is partly sample dependent.

Richard W. Cole
Systems Analyst
Computation Center
University of Texas @ Austin
Austin, TX 78701-1110

INTERNET: rcole@emx.cc.utexas.edu
BITNET: rwc@utxvm