Testing Individual vs Group Mean Differences in Social Science Research

Randall E. Schumacker **Lauren F. Holmes**
University of Alabama

A true experimental design requires random selection and random assignment of subjects to control and experimental groups. A hypothesized statistically significant mean difference in the dependent variable between these two groups is typically specified. This methodology is also referred to as a randomized clinical trial when testing for group differences. Individual differences are generally not considered, rather the focus is on the average control group and experimental group difference. This article offers another approach that illustrates testing for individual differences over time.

The true experimental design conducts a test of control group versus experimental group average dependent variable difference using analysis of variance statistical tests (Maxwell & Delaney, 2004). This methodology is also referred to as a randomized clinical trial when testing for group mean differences (Machin & Fayers, 2010). Oftentimes a true experimental design is not possible, so the researcher uses a quasi-experimental design. A quasi-experimental design uses a comparison group rather than a control group. The typical quasi-experimental design considers a pre-test measure, followed by treatment, and then a similar post-test measure for the subjects in the comparison group and the experimental group. In the statistical analysis, individual post-test measure differences are adjusted for individual pre-test measure differences to control for bias. This adjustment is referred to as analysis of covariance and expressed in the general linear model as:

$$Y_{Post} = b_0 + b_1 X_{\text{Pre}} + e \ ;$$

where: Y_{Post} = post-test measures
X_{Pre} = pre-test measures
b_1 = estimated sample regression weight, and
e = residual error.

The computation of the adjusted post-test group means are shown as (Hinkle, Wiersma & Jurs, 2003, p. 504):

$$\bar{Y}_k^{\,i} = \bar{Y}_k - b_w(\bar{X}_k - \bar{X})$$

where: $\bar{Y}_k^{\,i}$ = adjusted group mean on the dependent variable (post-test measure)
$\bar{Y}_k$ = original group mean on dependent variable (post-test measure)
b_w = pooled within group regression coefficient
$\bar{X}_k$ = group mean on the covariate (pre-test measure)
$\bar{X}$ = grand mean on the covariate (pre-test measure)

The analysis of covariance methodology however has been criticized for several reasons. One important reason is that the assumption of a linear relation between the covariate variable (pre-test measure) and the dependent variable (post-test measure) is rarely met. Another reason, is that the research question no longer relates to the original post-test group mean differences, rather interprets the adjusted post-test mean differences (Tracz, Nelson, Newman & Beltran, 2005). The authors contend that the analysis of covariance approach changes the hypothesis, Type I error, and interpretation.

There are further drawbacks in only testing for comparison group and experimental group mean differences, whether adjusted or not. One reason is the selection of comparison group subjects, e.g. matching, blocking. Today, a new approach termed *propensity score matching* (Polkinghorne, McDonald, Atkins, & Kerr, 2004; Holmes, 2014) has proven useful in choosing cohort subjects for the comparison group. Another important reason is that individual differences can mask the averages obtained in each group. Some subjects may improve given treatment as expected, others might stay the same, or some subjects may decline. These individual subject outcomes would not be identified when testing for only group mean differences.

Another approach by researchers is to test for mean gain differences between groups. A paired *t*-test would test whether the mean gain difference was statistically significant. Using heuristic created sample data, Tables 1 and 2 show the mean gain, standard error, *t*-values and *p*-values for the control group and

Table 1. Control Group Mean Gain

Time	Mean Gain	*SE*	*t*	*p*
Pre-Treatment (t_1-t_3)	5	3.16	1.58	.19
Post-Treatment (t_4-t_6)	7	2.00	3.50	.02

Table 2. Experimental Group Mean Gain

Time	Mean Gain	*SE*	*t*	*p*
Pre-Treatment (t_1-t_3)	4	5.97	0.67	.54
Post-Treatment (t_4-t_6)	-14	2.75	-5.09	.0006

experimental group's pre-treatment and post-treatment mean gain differences, respectfully, over a six-month time-period.

The null and alternative hypotheses for mean gain difference can be expressed as:

$$H_O: (t_1\text{-}t_3) = (t_4\text{-}t_6)$$
$$H_A: (t_1\text{-}t_3) \neq (t_4\text{-}t_6)$$

For control group: $H_A: (t_1\text{-}t_3) \neq (t_4\text{-}t_6) = (137 - 142) \neq (136 - 143) = 5 \neq 7$;
where paired *t*-test can be expressed as:

$$paired.t = \frac{MeanGain_{Post} - MeanGain_{Pre}}{(SE_{Post} + SE_{Pre})/2}$$

$$paired.t = \frac{7-5}{(2.00+3.16)/2} = \frac{2}{2.58} = .78$$

For experimental group: $H_A: (t_1\text{-}t_3) \neq (t_4\text{-}t_6) = (166 - 170) \neq (127 - 113) = 4 \neq -14$;
where paired *t*-test can be expressed as:

$$paired.t = \frac{MeanGain_{Post} - MeanGain_{Pre}}{(SE_{Post} + SE_{Pre})/2}$$

$$paired.t = \frac{-14-4}{(5.97+2.75)/2} = \frac{-18}{4.36} = -4.13$$

The mean gain difference results indicated that the experimental group post-treatment mean gain was statistically significantly different from their pre-treatment mean gain ($t = -4.13$, $p < .01$), in contrast to the control group post-treatment mean gain difference from their pre-treatment mean gain ($t = 0.78$, $p > 0.05$). In this type of analysis, we are able to determine the statistical significance of the control group and experimental group mean gain differences. The results are what we would expect if the treatment given to the experimental group subjects was effective. However, we do not know whether all subjects in each group stayed the same, increased, or decreased between the pre-test and post-test six-month time period.

Another approach commonly used by researchers involves comparing slope differences between the control group and experimental group using two separate regression equations. The null hypothesis is specified as: Ho: $b_{contro} = b_{experimental}$, where *b* is the slope regression coefficient of each group. This permits an *F*-test to determine the statistical significance between the two regression analysis model R^2 values. If the *F*-test is statistically significant then the regression slopes of the two regression lines would be significantly different. The *F*-test for difference in R^2 values is given by:

$$F = \frac{(R_1^2 - R_2^2)/df_1}{(1 - R_1^2)/df_2}$$

Although this analysis is relevant in testing whether change (slopes) between the two groups (control and experimental) are statistically significantly different, it still does not indicate individual subjects change over time. Our interest is not in testing the group mean difference, mean gain difference or slope difference, rather determining individual subject change over time. We therefore approach the problem differently by computing a separate regression equation for each subject.

The general linear model equation yields a R^2 value which is interpreted as a variance accounted for effect size (Schumacker, 2015). This is appropriate when testing for a linear trend over time. If a subject's dependent variable increases over time, then decreases over time after treatment, a curvilinear trend occurs as expected. In this instance, the eta-squared (η^2) or partial eta-squared ($\eta^2_{Y1.12}$) is appropriate for a non-linear trend over time (Pedhazur, 1973). In many instances, $\eta^2 = R^2$ when linear. The variance explained in the treatment outcome from a general linear model equation is tested for statistical significance using the *F*-test (Hinkle, Wiersma & Jurs, 2013). The *F*-test is given as:

$$F = \frac{R^2 / k}{(1 - R^2)/(n - k - 1)}$$

where: R^2 = multiple correlation squared
k = number of predictor variables
n = sample size

The *F*-test is *simultaneously* testing whether all the regression coefficients (*b* values) are statistically significant in the equation. Individual subject slope values are not interpreted. We therefore propose the testing of whether individual subject *b* values are significant, that is statistically different from zero while controlling for the effects of the other predictor variables (Pedhazur, 1973).
The general linear regression model can calculate individual regression coefficients to indicate change due to treatment. The general linear model *b* coefficient computation can be expressed in matrix form as:

$$b = (X'X)^{-1} X'y$$

where: b = column vector a (intercept) plus b_k regression coefficients
X = n by 1 + k matrix with unit vector and k column vector of scores
X' = transpose of matrix X
$(X'X)^{-1}$ = inverse of $(X'X)$
y = n by one column of dependent variable scores

A test of the statistical significance of the individual subjects *b* regression coefficients is obtained by:

$$t = \frac{b}{SE_b}$$

The general linear model has been used for testing slope differences or rate of change in repeated measure designs (Schumacker, 2015). Therefore, testing only mean differences (intercepts) has been expanded using the general linear model to include testing for slope differences, i.e. differences in the rate of change. Newman & Schumacker (2012) demonstrated the use of regression-discontinuity techniques to test for slope differences, intercept differences, and to examine change in individuals. In past developments, researchers have applied discrete-time survival analysis techniques to investigate the duration and timing of event occurrence (Singer & Willett, 1993). The modeling of change and event occurrence has become more popular over the years in the statistical analysis of data (Singer & Willett, 2003).

The interpretation of an individual's change over time should have a more meaningful application in statistical analysis. The key design issue in testing for individual change is that it requires three measures over time, hence the basic pre-test and post-test design does not yield important subject change interpretations. A graph could be produced for each subject over measured time periods. A positive *b* coefficient would indicate an increase, a negative *b* coefficient would indicate a decrease, and a zero *b* coefficient would indicate no change.

Methods and Procedures

Data Source

The experimental design would randomly assign subjects to a control group and an experimental group. A heuristic data set was created with the first five subjects in the control group (1 to 5) and the last five subjects in the experimental group (6 to 10). The heuristic data set of 10 subjects has six (6) smoking measures per subject as shown in Appendix A. The data set shows the 6 measurements for each group with 3 measures before treatment and 3 measures after treatment. Treatment consisted of counseling to quit smoking for subjects in the treatment group and distribution of stop smoking pamphlets in the control group.

Regression Analysis

The statistical analysis of data involved computing the intercept and slope of each subject using separate general linear model equations. The regression program was written using *R* (*R* Core Team (2020), which is easily written and reported in Appendix B. The regression analysis computed a subject's regression coefficient over the six smoking measures. The statistical analysis can therefore report individual subject coefficients, standard errors, test of statistical significance, and level of statistical significance (*p*-value).

The general linear model equation can be expressed as:

$$Y_i = b_0 + b_i X_i + e_i ;$$

where: Y_i = individual subject smoking measures
X_i = individual subject treatment time (1 to 6 months)
b_0's = estimated individual subject intercept coefficients
b_i's = estimated individual subject slope coefficients, and
e_i = residual error

Our interest was in computing individual subject intercept and slope regression coefficients over time. We hypothesized that the control group subjects who received only a stop smoking pamphlet would indicate no increase or a modest change (not statistically significant), while the experimental group subjects would have a statistically significant decrease in smoking after counseling, hence negative slope coefficients.

Results

The individual subject regression analysis provided individual subject intercept and slope values. The results in Table 3 indicated that 4 out of 5 subjects in the control group did not change significantly, while each subject in the experimental group did show a decreased change in smoking.

It is possible to now determine if any subjects in either group increased or decreased their smoking behavior. Our example showed that the slope values decreased significantly for subjects in the treatment group. One subject (ID = 4) in the control group decreased smoking behavior significantly after receiving a quit smoking pamphlet. Individual statistical analysis would clearly show better results for interpretation than simply testing group mean differences. The regression coefficients listed in Table 3 can be displayed using a simple EXCEL scatter plot (Figure 1) that visually displays their slope values, where b-values above and below a 0 value indicate change.

The issue is how can we analyze our research data to test for subject change due to treatment rather than interpret only group mean differences. This can be accomplished by computing individual subject regression equations. The general linear model can compute the intercept and slope for each subject. The intercept value for each subject can be interpreted as a baseline measure or starting point. The slope value for each subject can be interpreted as a rate of change. The computed individual subject slope value is divided by its standard error to compute a *t*-value with an accompanying *p*-value for statistical significance. The practical interpretation is readily available since a positive regression coefficient (*b*-value) indicates an increase in the measured outcome variable; a negative regression coefficient indicates a decrease; and a zero regression coefficient indicates no change. The regression coefficient interpretation can also be made in the context of change from an intercept value (baseline measure).

Our results indicated a subject in the control group had a significant negative regression coefficient, thus changed. In contrast, all subjects in the experimental group had significant negative regression coefficients. We can therefore examine each individual subject to know whether one, a few, or all benefited in a study; thus, whether individual subjects change in either the control or experimental group can easily be tested and interpreted. This is more advantageous than simply having significant mean differences between groups where individual results are not readily interpreted.

Table 3. Subject Intercept and Slope Regression Coefficients

ID	Group	Intercept	Slope	*SE*	*t*	*p*
1	Control	83	-0.14	1.39	-0.10	.92
2	Control	81	-0.14	1.56	-0.09	.93
3	Control	86	-0.28	1.45	-0.19	.85
4	Control	90	-6.85	0.38	-17.73	.0001
5	Control	78	1.42	1.18	1.21	.29
6	Experimental	55	-7.57	0.50	-15.09	.0001
7	Experimental	71	-10.71	0.41	-25.98	.00001
8	Experimental	49	-7.42	0.50	-14.81	.0001
9	Experimental	70	-10.00	0.0005	-19.23	.00001
10	Experimental	81	-10.71	0.42	-25.98	.00001

Summary

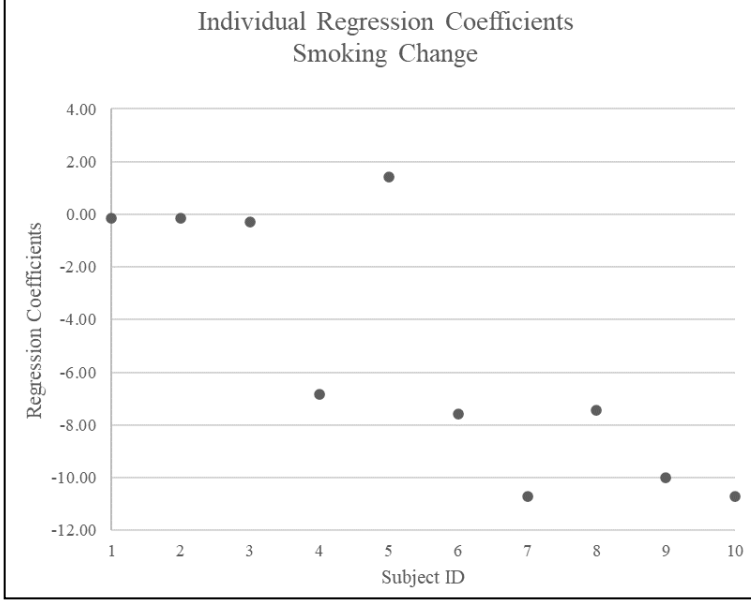
There have been a few statistical approaches used to analyze mean differences between groups, the most popular being analysis of variance; analysis of covariance; and mean gain score analysis using paired *t*-test. There have been substantiated published articles over the past several decades criticizing the use of analysis of covariance and gain score analysis. The use of analysis of variance in true experimental designs has been the gold-standard of research designs. Unfortunately, researchers have not always been able to conduct true experimental designs and struggled with creating comparison groups in quasi-experimental designs. The use of an experimental design or a quasi-experimental design using propensity score matching for the cohort of subjects in a comparison group however does not preclude conducting a regression statistical analysis to reveal individual subject treatment effectiveness.

Another emerging analysis method, regression-discontinuity, was developed for program evaluation (Trochim, 1984). The methodology was eventually extended to randomized clinical trials to test slope differences between control and experimental groups (Trochim, 1992). This proved to be a useful method to determine whether a program or treatment was effective. The test of slope differences between groups via regression-discontinuity however does not provide the individual subject treatment results. Individual change can still show individual gains, losses, or no change in treatment. The ability to examine individual subject treatment status can be deduced from a regression-discontinuity design if proper dummy coding of subject vectors is employed. This is generally not done in applied social research studies, so the emphasis is still not on the individual subject change, rather only group prediction. For example, was a certain stop smoking program effective?

Multilevel modeling (HLM; Bickel, 2007; Schumacker & Lomax, 2016) has also demonstrated the use of regression and structural equation modeling to generate individual intercept and slope values across time. It is recommended that the intra-class correlation coefficient be interpreted (ICC; Shrout & Fleiss, 1979) to determine whether a common regression line or individual regression lines should be interpreted. In addition, the *design effect* adjustment is used in some studies where cluster randomized control trials are conducted (Bland, 2004). The design effect uses the intra-cluster correlation ($D_{eff} = 1 + (m - 1)p$ to assess the effect of clustering, where m is the sample size in a cluster and p (ICC) is the intra-cluster correlation. Clustering may result in *p*-values and confidence intervals which are biased if cluster size is large, the number of clusters is small, or the intra-cluster correlation coefficient is large.

A recent SAS approach, varying time estimation method (VTEM), estimates regression coefficients in a time-varying effect model (TVEM SAS Macro, 2017). It provides an end user SAS macro (%VTEM) to make longitudinal analysis using regression equations easier to execute (Li, Dziak, Tan, Huang, Wagner, & Yang, 2017). This SAS macro permits fixed or time varying variables in the equations and highlights the pivotal work by Singer & Willet (1993; 2003) who earlier demonstrated SAS code for time varying variables in longitudinal data analysis. Additionally, the centering approach in VTEM, which is used in the graphical display of *b*-value deviations around zero (0) supports the earlier work by Aiken & West (1991). Basically, if the graph shows *b*-values with confidence intervals *not* capturing the zero point, then the fixed or time varying variable effect is statistically significant. The model selection fit function criteria is based on Akaike information criterion (AIC; non-parametric) and Bayesian information criterion (BIC - parametric), although VTEM is considered a non-parametric approach. These two fit functions are commonly used for choosing the best predictor subset models in regression where lower values suggest a model closer to a true model. A new

Figure 1. Plot of Smoking Change (Regression Coefficients)



parametricness index (PI) has been introduced to assess whether the best regression model selected should be judged by AIC or BIC fit criteria in estimating the regression function (Liu & Wang, 2011).

All of the aforementioned approaches are important techniques used to assess change over time. The importance of analyzing individual subject change over time needs a more prominent place in our social science research. It is the most important practical question in longitudinal modeling. We feel that individual treatment effectiveness is more important than group mean or group slope differences. Several methods test group mean differences (*t*-test, analysis of variance, analysis of covariance), however, more suitable methods should be used that provide individual intercept and slope values (multilevel modeling, survival analysis, longitudinal analysis). The basic *R* program makes it easy to compute individual subject treatment results over time. Any statistical analysis that masks individual contribution by only computing and interpreting group mean and/or group slope differences are not recommended.

References

Aiken, L.S. & West, S.G. (1991). *Multiple Regression: Testing and Interpreting Interactions*. Thousand Oaks: CA.

Bickel, R. (2007). *Multilevel Analysis for Applied Research: It's Just Regression*. Guilford Press: New York, NY.

Li, R., Dziak, J. J., Tan, X., Huang, L., Wagner, A. T., & Yang, J. (2017). *TVEM (time-varying effect model) SAS macro users' guide* (Version 3.1.1). University Park: The Methodology Center, Penn State. Retrieved from http://methodology.psu.edu

Hinkle, D.E., Wiersma, W. & Jurs, S.G. (2013). *Applied Statistics for the Behavioral Sciences*, Houghton-Mifflin Co.: New York, NY.

Holmes, W.M. (2014). *Using Propensity Scores in Quasi-Experimental Designs.* Sage Publications, Thousand Oaks: CA.

Liu, W. & Wang, Y. (2011). Parametric or Nonparametric? A Parametricness Index for Model Selection? *The Annals of Statistics*, *39*(4), 2074-2102.

Machin, D. & Fayers, P.M. (2010). *Randomized Clinical Trials: Design, Practice and Reporting*, Wiley–Blackwell: Hoboken, New Jersey.

Maxwell, S.E. & Delaney, H.D. (2004). Designing Experiments and Analyzing Data: A Model Comparison Perspective (2nd Edition). Lawrence Erlbaum Associates, Mahwah: NJ.

Newman, I. & Schumacker, R. E. (2012). Regression-Discontinuity Designs in Medical Settings: An Alternative to Quasi-Experimental Designs in Testing Treatment Effectiveness. *Multiple Linear Regression Viewpoints, 38*(*1*), 16-25.

Pedhazur, E.J. (1973). *Multiple Regression in Behavioral Research: Explanation and Prediction* (3rd Edition). Holt, Rinehart & Winston, NY.

Polkinghorne, K.R., McDonald, S.P., Atkins, R.C. & Kerr, P.G. (2004). Vascular Access and All-Cause Mortality: A Propensity Score Analysis, *Journal of the American Society of Nephrology*, *15*, 477-486..

R Core Team (2020*). R: A language and environment for statistical computing*. *R* Foundation for Statistical Computing, Vienna: Austria. URL http://www.*R*-project.org/

Shrout, P.E. & Fleiss, J.L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 20-428.

Schumacker, R.E. (2015). *Learning Statistics Using R.* SAGE Publications: Thousand Oaks, CA.

Schumacker, R.E. & Lomax, R.G. (2016). *A Beginner's Guide to Structural Equation Modeling (4th Edition)*. Routledge (A Taylor and Francis Group): New York, NY.

Tracz, S., Nelson, S., Newman, I. & Beltran, A. (2005). The Misuse of ANCOVA: The Academic and Political Implications of Type VI Errors In Studies of Achievement and Socioeconomic Status. *Multiple Linear Regression Viewpoints*, *31*(1), 16-21.

TVEM SAS Macro (Version 3.1.1) [Software]. (2017). University Park: The Methodology Center, Penn State. Retrieved from http://methodology.psu.edu

Singer, J.D. & Willett, J.B. (1993). It's About Time: Using Discrete-Time Survival Analysis to Study Duration and the Timing of Events. *Journal of Educational Statistics*, *18*(*2*), 155-195.

Singer, J.D. & Willett, J.B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford University Press: New York, NY.

Trochim, W.M.K. & Cappelleri, J.C. (1992). Cutoff assignment strategies for enhancing randomized clinical trials. *Controlled Clinical Trials, 13*, 190–212.

Trochim, W.M.K. (1984) *Research Design for Program Evaluation: The Regression-Discontinuity Approach.* Sage, Beverly: CA.

Send correspondence to: Randall E. Schumacker
University of Alabama
Email: rschumacker@ua.edu

Appendix A: Smoking Data per Month (Heuristic Data)

ID	Y	Time	ID	Y	Time
1	80	1	6	50	1
1	85	2	6	40	2
1	90	3	6	30	3
1	75	4	6	25	4
1	80	5	6	20	5
1	85	6	6	10	6
2	75	1	7	60	1
2	80	2	7	50	2
2	90	3	7	40	3
2	85	4	7	30	4
2	80	5	7	20	5
2	75	6	7	5	6
3	90	1	8	40	1
3	85	2	8	35	2
3	85	3	8	30	3
3	75	4	8	20	4
3	85	5	8	10	5
3	90	6	8	5	6
4	85	1	9	60	1
4	75	2	9	50	2
4	70	3	9	40	3
4	65	4	9	30	4
4	55	5	9	20	5
4	50	6	9	10	6
5	85	1	10	70	1
5	75	2	10	60	2
5	80	3	10	50	3
5	85	4	10	40	4
5	90	5	10	30	5
5	85	6	10	15	6

Appendix B: Regression Program via R

```
# Smoking data
# Compute intercept and slope of each subject
# ID = subject id;   Y = number of cigarettes smoked ; Time = months
of treatment ( 1 to 6)
# CTRL = control ;TRT = treatment; CT = 1,2,3; TT = 4,5,6; P1 to P10
are dummy coded
# Input data

mydata=read.table("c:/regression.csv",header=TRUE,sep=",")
mydata

# Compute individual intercept and slope values
# All subjects  ( i = 10 subjects)
K = 1
L = 6
for (i in 1:10) {
j = lm(Y[K:L] ~ Time[K:L],data = mydata)
z = summary(j)
print(z)
K = K + 6
L = L + 6
}
# Individual regression results copied into Table 3
```